Modeling of language distinctive features for Ukrainian real-time speech recognition system

Mykola Sazhok, Valentyna Robeiko

The presented research is focused on features that are specific for most Slavonic languages and for Ukrainian particularly. Given arguments confirm the necessity to distinguish stressed and unstressed vowels in the phoneme alphabet. Lexical stress irregularity implies expert involvement for stress assignment. To automate this procedure we propose a data-driven stress prediction algorithm that represents words as sequences of substrings (morphemes). The formulated criterion that validates a substring sequence is based on a set of words with manually pointed stresses and a large text corpus. The described search algorithm finds N-best symbol sequences with a hypothetical stress. As a Slavonic language, Ukrainian is highly inflective and tolerates relatively free word order. These features motivate transition from word- to class-based statistical language model. Spontaneous speech recognition experiments confirmed efficiency of the stressed phoneme introduction and performance comparability of both class and word n-gram language models. We also describe several tools developed to visualize HMMs, to predict word stress, and to manage equivalence class-based language modeling.

spontaneous speech recognition • real-time • stress prediction • word equivalence classes • language models

---

Представленное исследование сосредоточено на специфических особенностях большинства славянских языков вообще и украинского в частности. Представленные аргументы подтверждают необходимость различать ударные и безударные гласные в алфавите фонем. Нерегулярность лексического ударения подразумевает участие эксперта при разметке ударений. Для автоматизации этой процедуры предложено выводить из данных алгоритм предсказания ударения, представляющий слова в виде последовательности подстрок (морфем). Сформулированный критерий допустимости последовательности подстрок основан на множестве слов с расставленными вручную ударениями с использованием большого текстового корпуса. Описанный алгоритм поиска находит N-лучших последовательностей символов с гипотетическим ударением. В украинском, как и в любом славянском языке, наблюдается обилие словоформ и относительно свободный порядок слов. Эти особенности мотивировали переход от словарной лингвистической статистической модели к классовой. Экспериментальные исследования распознавания спонтанной речи показали эффективность введения ударных гласных и сравнимость производительности лингвистических моделей на основе как классов, так и слов. Также описывается инструментарий, разработанный для визуализации HMM, предсказания словесного ударения и моделирования языка на основе классов эквивалентности.

распознавание спонтанной речи • реальное время • предсказание ударений • классы эквивалентности слов • лингвистические модели

INTRODUCTION

Specific features of Slavonic languages are observable on different levels of speech pattern hierarchy. Lexical stress is among features neglected in speech recognition works for Western languages. However, for several Slavonic languages a stressed vowel acts as if a specific phoneme that is separate from its unstressed counterpart. Lexical stress position irregularity is the reason we must develop robust predictive algorithms.

More generic features of Slavonic languages are high inflectiveness and relatively free word order, which leads to rapid growth of the recognition vocabulary (8-10 times larger than in English for the same domain) 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 computational platforms 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 ability for the user to dictate in conditions of home and office inside and perhaps outside.

Another application of speech-to-text technology is transcribing of meetings, e.g. parliamentary speech. Though in such case the immediate user interaction is not required, near real time signal processing is desirable for the entire system efficiency.

In previous work [3] we described a speech-to-text system that operated in real time with a 100 000 vocabulary tightly covering common and news domain (politics, economics, culture, education, sports, and weather). Nevertheless, we must move towards a vocabulary for million words to reach the desired OOV for the arbitrary speech signal.

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.

LEXICAL STRESS ANALYSIS

The phenomenon of lexical stress plays significant role for many languages. Prosodic features like duration, pitch, and loudness are used to describe phonetic distinctions for stressed segments of the word. So every text-to-speech system

must implement lexical stress prediction. Letter-to-sound rules practically always work for vowels under lexical stress even for highly spontaneous pronunciation manner. And this property might be useful for spontaneous speech recognition tasks. So do we need to introduce both stressed and unstressed vowels to the phoneme alphabet?

Answering positively to this question we rely on phonetic, lexical and acoustical facts for Ukrainian. Stressed vowels normally acts as phonemes changing word grammatical function and meaning that we observe in about 10% of words in arbitrary texts.

To explore the acoustical side of the problem we trained stressed and unstressed vowels as if they are different phonemes and inspected dissimilarities particularly by means of the HMM visualization tool [4]. Following Fig. 2 we can see the difference between models for unstressed and stressed context-independent phonemes of “a” and “i” trained on 40 hour multi-speaker AKUEM subset [5].

The presented central state contains 32 GMMs estimated in MFCC feature space accomplished with energy coefficient and mean subtraction that makes total 13 coefficients. The dotted line corresponds to a zero value. Visually, a stressed model looks like a subset for most coefficients. Overlaps rather than inclusions with respective coefficients in the stressed model are proper in cases like the 5th coefficient for “a” and the first coefficient for “i”. More HMMs are available from the tool’s web-page.

Analyzing transition matrices, we can see that diagonal values corresponding to emitting states are 1.5–2 times greater for stressed models, which confirms the essential difference in duration. In Fig. 2 transition matrix diagonal values are equivalent to petals.

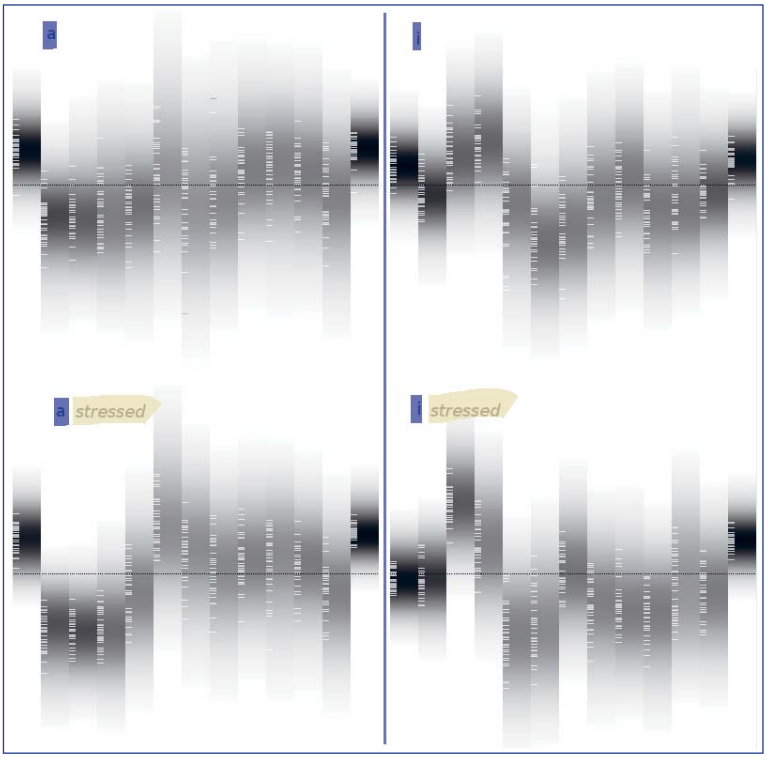

Figure 1. HMM visualization for unstressed (above) and stressed monophones (below)

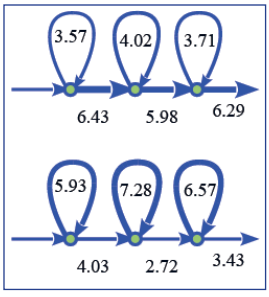

*Figure 2. State transition probabilities for acoustic models of unstressed (above) and stressed (below) monophone **a***

Lexical stress position is irregular and it can be changed even within forms of the same word. Referring to lexical stress irregularity example for English *phóto/photography* [6] we can see that derivations from the same stem in Ukrainian cause 5 different stress positions: *fóto, fotóhraf, fotohrafíja, fotohrafíchnyj, fotohrafuváty*. Therefore, lexical stress irregularity is essentially more proper for Slavonic languages. Anyway, it is not acceptable to point stresses manually for the entire lexicon. Hence, we propose a word stress prediction procedure based on the known vocabulary and a large text corpus [7].

Grapheme-to-phoneme conversion methods like [8] could be directly used also for lexical stress modeling; however they have no provisions to account for the structural properties of stress. In our work, rather than modifying an existing algorithm, we prefer to construct a model concentrated on stress properties and then convert the stressed text to phoneme sequences by means that allow for counting to specific pronunciation properties provided by the technique that requires about 30 find-replace-and-step rules for Ukrainian [9].

Let us consider all possible segmentations S for a word with unknown stress. The i-th segmentation of S

$$S_i = (q_{i,1}, q_{i,2}, \dots, q_{i,j}, \dots, q_{i,L_i}) \quad (1)$$

has length of L_i. Here $q_{i,j}$ is a j-th symbol (a character or a phoneme) within the i-th segment of S. Now we introduce a vector Θ_L that indicates the stress level (e.g. 0, 1, 2) for each of L items. Thus, we can estimate a probability of stress position given the segment S_i:

$$P(\Theta_{L_i} \mid S_i) \approx \frac{c(S_i, \Theta_{L_i})}{c(S_i)} \quad (2)$$

where $c(S_i, \Theta_{L_i})$ is count of segments S_i with stress position defined by a stress indication vector Θ_{L_i} and $c(S_i)$ is the number of S_i total occurrence. All counts are taken from the text corpus but the words are not included in stress vocabulary.

Finally, we search through all valid segmentations S and stress positions Θ^S that satisfy the expression:

$$\underset{S,\theta^S}{\operatorname{argmax}} \prod_{S_i,\theta_{L_i}} P(\Theta_{L_i} \mid S_i). \quad (3)$$

We constructed a dynamic programming graph where finding the shortest trajectory is equivalent to the search (3). Memorizing N prospective arrows in nodes of the graph we can extract N-best word stress positions supplemented with the probability estimation.

We estimated stress prediction model parameters on 250 M text corpus. Special word boundary symbol "|" is included. More than 60 000 character segments detected for length one to four. In Figure 3 an example of one-best stress prediction is

shown for a proper name “Obama” missing from the basic Ukrainian vocabulary. The respective symbol sequence (|, *o*, *b*, *a*, *m*, *a*, |) is represented as a concatenation of all valid symbol segments where the largest segment length is limited to four. Each input symbol introduces a set of valid segments. Potentially optimal arcs are either shown or coded with the name of a previous node. Partial criteria are log probability based. The optimal path, respective nodes and criteria are bold.

Here we account for the structural properties of stress that is avoiding of running stressed syllables. Therefore, the segment “mA|” in column 7 (cirled) accepts potentially optimal path from node “a” rather than “obA”.

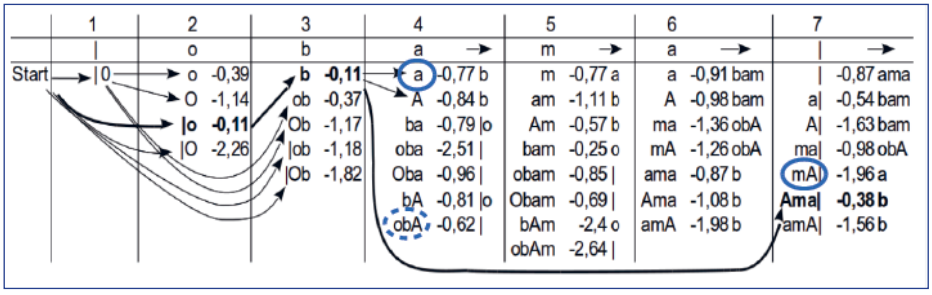

Figure 3. Stress prediction for an out-of-vocabulary word “Obama”

Stress error rate estimation in not as obvious procedure, since in specific cases it is unclear what is a mistake, e.g. the stress is predicted in misspelled words but if the prediction looks wrong why we should count it erroneous? Anyway, the proposed lexical stress prediction algorithm has been evaluated in frames of the entire speech recognition system. Therefore, we describe the data used and results attained in the generic experimental chapter, which follows the analysis of language modeling distinctive features.

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 a comparable language model for Ukrainian, theoretically, a 6 time larger vocabulary is required. Moreover, relatively free word order toleration leads to perplexity and data sparsity growth. Analysis of these features motivates a

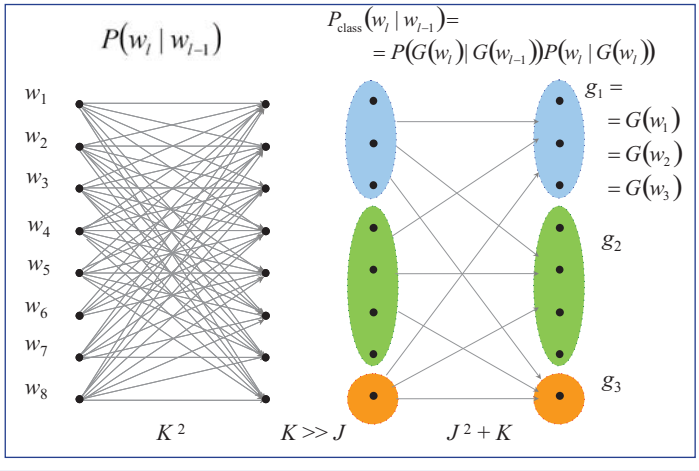

Figure 4. Class 2-gram model introduction

transition from word- to class-based statistical language model that operates with transition probability and membership probability [10].

In Fig. 4 we compare complexity for both word 2-gram model and class 2-gram model. Here a vocabulary consists of $K = 8$ words and $J = 3$ equivalence classes are introduced. A classification function, $G(w_i)$, maps first three words from the vocabulary to g_1 class so $P_{\text{class}}(x \mid w_1) = P_{\text{class}}(x \mid w_2) = P_{\text{class}}(x \mid w_3)$ for any word x from the vocabulary; 4th to 7th words are mapped to g_2 and, finally, g_3 class has the only member, w_8. Introducing a class 2-gram model we change the amount of model parameters from K^2 to $J^2 + K$, which means significant reduction since $K >> J$.

Word clustering procedure tries to optimize the perplexity improvement criterion

$$F_G = \sum_{g,h \in G} C(g, h) \log C(g, h) - 2 \sum_{g \in G} C(g) \log C(g) \qquad (4)$$

where (g, h) means a class g follows a class h from the set of equivalence classes G and function $C(\bullet)$ counts its argument occurrence in the training corpus. An exchange algorithm described in [10] implies iterations in which each word is tested for a better class and consequently moved there (Fig. 5).

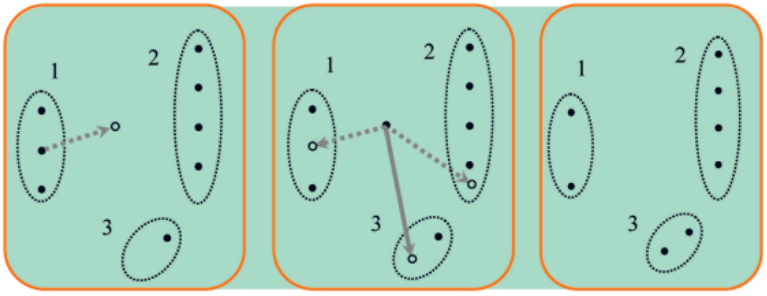

Figure 5. Word exchange algorithm illustration for three classes. A word from Class 1 has been moved to Class 3 providing a better criterion (4)

While implementing the algorithm we came to an alternative formulation of criteria computation refinement.

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

Assuming that a preceding single classification function $G^{-}(\bullet)$ 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 (4), having the most complicated computations, $O(G^2)$, can be expressed as

$$\sum_{i,j} \log C_{ij} = \sum_{\substack{i,j \\ \{i\,j\} \cap \{u,v\} = \varnothing}} \log C_{ij} + \sum_{\substack{i=u,v \\ j}} \log C_{ij} + \sum_{\substack{j=u,v \\ i \neq u,v}} \log C_{ij}. \qquad (5)$$

Thus, the analyzed sum is decomposed in three components (Fig. 6) where the most expensive for computations component, still $O(G^2)$, might be expressed as a recursion relatively to the predecessor:

$$\sum_{\substack{i,j \\ \{i,j\} \cap \{u,v\} = \varnothing}} \log C_{ij} = \sum_{\substack{i,j \\ \{i\,j\} \cap \{u,v\} = \varnothing}} \log C^{-}_{ij} = \sum_{i,j} \log C^{-}_{ij} - \left(\sum_{\substack{i=u,v \\ j}} \log C^{-}_{ij} + \sum_{\substack{j=u,v \\ i \neq u,v}} \log C^{-}_{ij} \right) \qquad (6)$$

arriving to the computation time complexity of $O(G)$.

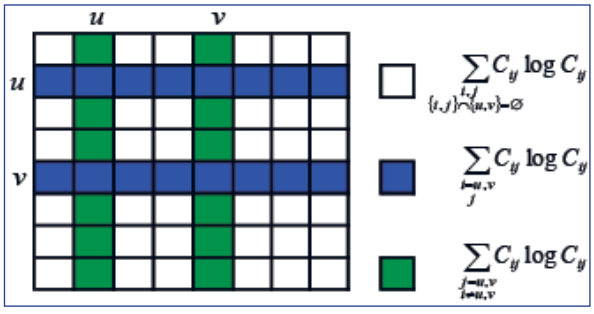

Figure 6. Decomposition of the most expensive for computations criterion component (5)

Proceeding from (4)–(6) we have developed an efficient tool for word clustering and assigning a given word, accomplished with bigram counts, to one of existing classes [11].

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

Most word classes have an obvious syntactic interpretation, such as nouns in genitive form, or plural adjectives. Table 1 shows several word classes that have been obtained by bigram clustering on large text 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. 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 on phonetic context. For instance, the conjunction word corresponding to *and* has three forms normally used relatively to context: between consonants, between vowels and in other cases. All these forms were automatically assigned to different classes.

Table 1

Bigram clustering homogeneous 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

Another version of the table. 1

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

EXPERIMENTAL RESEARCH

A general structure for the basic speech-to-text conversion system is shown in Fig. 7. The real-time component implements *Recognizer* itself that refers to *Data and Knowledge Base* developed off-line by means beside the illustrated components. To create the speech recognition experimental 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* (*VAD*) suggests beginnings of speech segments for *Pre-processor* that extracts acoustic features. 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 [12]. The output, presented as a lattice or a confusion network, is passed to *Decision Maker* that forms a *Recognition response* considering the history and performing necessary mappings to symbols and actions.

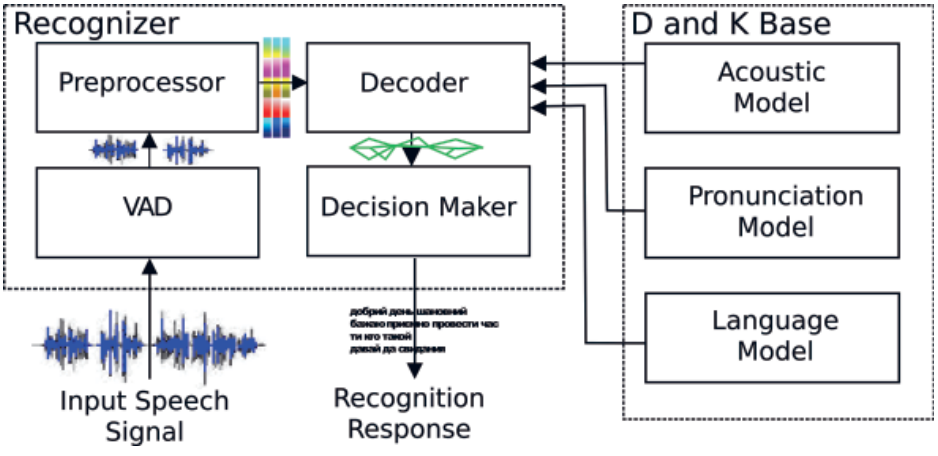

Figure 7: A speech-to-text system general structure

To estimate *Data and Knowledge Base* parameters the system was complemented with text data and automated means for lexical stress prediction and for word clustering. The text data consists of basic dictionary and text corpus.

The basic dictionary is extracted from the electronic lexicography system [13] subset containing 151 962 lemmas, including over 10 thousand names, that makes 1.90 million word forms. Due to shared spelling the actual word form vocabulary consists of 1.83 million words that have different either spelling or primary 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. A text filter, used for text corpus processing, provides conversion of numbers and symbolic characters to relevant letters, by means referred below in *Pronunciation Model* description, as well as removing improper text segments and paragraph repetitions. Hereafter, we refer to the basic text corpus as 275M corpus. In accordance to the corpus summary shown in Table 2, we observe 6.64 word forms per lemma in average, whereas this relation is twice greater, 12.3, within the dictionary [13]. Adding 200 000 most frequent words to the vocabulary we reduce OOV to less than 0.5%. Words that have 2 or more valid stress positions, referred as homographs, take over 6% of the entire text.

Table 2

Basic text corpus 275M summary

Running 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 a realignment stage.

215 000 segments were used to predict lexical stress for words in 275M text corpus by the developed means that implement (1)–(3). Incorrect stress position was detected for less than 1% known words. Stress detection for 5 000 OOV words was incorrect for 21.1% words or 5.3% syllables. However, over a half of incorrectly stressed words have strong foreign origins.

Perhaps, the most interesting is the case of stress moving with some morphological derivations. Checking Ukrainian derivations from phóto (*fóto, fotóhraf, fotohráfija, fotohrafíchnyj, fotohrafuváty* and their forms) we found that incorrect stress has been assigned only in one case (*fotohraf́*).

Acoustic model parameters are estimated with and without account to lexical stress on a subset of the AKUEM speech corpus [5] by means of [14],[15]. The basic phoneme alphabet consists of 56 phonemes including stressed and unstressed versions for 6 vowels. Context-dependent phoneme models are covered with about 14000 GMMs. Several stress positions are assumed as valid for homographs and most frequent two- and three-syllable words. On realignment stage the most appropriate pronunciation is assigned.

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 [9]. For grapheme-to-phoneme mapping an expert formulated about 30 find-replace-and-step rules with exceptions. The rules only partially model the individual speaker peculiarities as well as co-articulation and reduction of sounds in speech flow. Finally, each word produces about 1.05 transcriptions on average. The same algorithm with other 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 275M corpus and supplementary vocabularies covering speech corpus, social and local dialects, proper names, abbreviations etc. A recognition vocabulary is formed taking a specified amount of top-frequent words from the system vocabulary.

Language model parameters were estimated proceeding from the recognition vocabulary and a text corpus subset consisting of sentences containing below the specified portion of OOV words. For the recognition vocabulary of 100 000 words, 88.5 million distinct 3-grams are detected in the subset of the 275M 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 250M corpus. Consequently, we got OOV words occupy 2.5% 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 as well as emotion and attitude expressions (laugh, applauds etc.).

Applying language modeling tool [15] to 250M corpus we have received a text file in ARPA format that occupies 5 GB reduced to 1.3 GB by a module of the decoder tool [12], which is a baseline 3-gram LM. To estimate class *n*-gram models, firstly, we mapped recognition vocabulary words to 1000 equivalence classes by the developed means that implement (4)–(6). Then we converted 250M corpus to the sequence of classes. Finally, we estimated transition probabilities for classes and membership probabilities for words (Fig. 4). This way we prepared both class 3-gram (290 MB) and 4-gram (1.2 GB) models to be used in the decoder.

The real-time modules are used to build a basic speech-to-text conversion system for experimental research and for 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].

We present two experimental setups combining acoustic, pronunciation and linguistic models. For Setup 1, the AKUEM speech corpus was randomly divided into non-overlapping test and training sets in relation 1/8. Acoustic model parameters were estimated with and without account to the lexical stress. For class *n*-gram language models, 100 000 words were automatically assigned to 1000 classes and rest of words were ascribed to the unknown category. Introducing multiple new words to the word *n*-gram LM in same manner caused essential accuracy degradation, so we do not consider this case. Sentences for Setup 2 are taken from manually selected recordings in order to reduce share of the forensic domain from 70% to 32%. Dissimilarity of OOV word distribution for both setups (Table 3) proves that Setup 2 is lexically richer.

Table 3

Out-of-vocabulary word share (%)

Vocabulary size (words)	Setup 1	Setup 2
100 000	5,27	5,62
200 000	2,38	4,15

It is notable that 275M corpus used to build LMs contains no explicit text transcriptions of spontaneous speech. Therefore, the constructed language models are in general irrelevant to the test data domain and style. Essential part of records is the highly spontaneous speech since more than half of the corpus is court show records where participants widely express their attitude and emotions, lots of words are interrupted or misspelled.

Both setups are adjusted so that the decoder performs in real-time (less than 0.75RT on *i7* processor). This means that the accuracy keeps on increasing, meanwhile, the recognition process time may exceed the signal duration due to larger amount of parameters in the acoustic model. Noteworthy is a fact that decoding takes longer time for the models where lexical stress is ignored.

The results of experiments are shown in Table 4. Word error rate is estimated in terms of accuracy and correctness (%). The latter ignores incorrect word insertions. Thus, roughly every 4th error is caused by insertion. Trying to approximate an OOV word in most cases the system produces numeral insertions. Anyway, for dictation purposes, a speaker interacts with the recognition system in cooperative manner and errors are expected to reduce in two or even more times.

Table 4

Experimental results for spontaneous speech recognition

Language Model type, order, vocabulary size	Setup 1, no stress		Setup 1		Setup 2	
	Accuracy	Correctness	Accuracy	Correctness	Accuracy	Correctness
Word-based, n = 3, 100 000	62,03	71,77	64,01	72,94	64,62	72,49
Class-based, n = 3, 200 000	58,96	69,51	61,24	71,18	66,07	74,17
Class-based, n = 4, 200 000	59,16	69,53	61,55	71,22	66,29	74,22

CONCLUSIONS

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.

The proposed stress prediction procedure allows for assigning most hypothetically possible one or more lexical stresses in unknown words. However, stress disambiguation even for known words is necessary for further introduction of the semantic level.

Distance to closest alien classes should give a clue to predicting homographs and consequent semantic word decomposition that may lead to more homogeneous classes. From the other hand, words that are confused while recognizing must be assigned to different classes. 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 the membership probabilities.

Stressed phoneme introduction to the basic alphabet is undoubtedly beneficial for recognition systems; however, wider context-dependency on acoustic level may neglect the accuracy gain. Slight error difference between word and class *n*-gram models requires more detailed analysis and explanation.

To improve the overall system performance more conventional techniques should be used like speaker/feature adaptation, discriminative training and wider context application. Data for more languages is available and we plan to present Russian, English and Tatar in further research.

REFERENCES

1. *T. Vintsiuk, M. 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. *V. Robeiko, M. Sazhok.* Real-time spontaneous Ukrainian speech recognition system based on word acoustic composite models. In Proc. UkrObraz'2012, Kyiv, 2012, pp. 77-81.
4. www.cybermova.com/speech/visual-hmm.htm
5. *Н. Васильева, В. Пилипенко, А. Радуцкий, В. Робейко, Н. Сажок.* Корпус украинской эфирной речи // Речевые технологии, №2, 2012, pp. 12–21.
6. *Black A., Lenzo K., Pagel V.* Issues in Building General Letter to Sound Rules. 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, (1998) 77-80.
7. *M. Sazhok, V. Robeiko.* Lexical Stress-Based Morphological Decomposition and Its Application for Ukrainian Speech Recognition. In TSD 2013, LNAI 8082, pp. 327–334, 2013. Springer-Verlag Berlin Heidelberg 2013.
8. *M. Bisani and H. Ney.* "Joint-Sequence Models for Grapheme-to-Phoneme Conversion." Speech Communication, Volume 50, Issue 5, May 2008, pp. 434–451.
9. *В. Робейко, Н. Сажок.* Преобразование между орфографическим и фонемным текстами для моделирования спонтанного произношения // Речевые технологии, №2, 2012, pp. 33–42.
10. *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
11. *M. Sazhok, V. Robeiko.* Language Model Comparison for Ukrainian Real-Time Speech Recognition System. In SPECOM 2013, LNAI 8113, pp. 211–218, 2013. Springer International Publishing Switzerland 2013.
12. *A. Lee, T. Kawahara.* Recent Development of Open-Source Speech Recognition Engine Julius. APSIPA ASC, 2009, pp. 131-137.
13. *Широков В.А., Манако В.В.* Організація ресурсів національної словникової бази // Мовознавство. – №5. – 2001 р. – С. 3–13.
14. *Young S.J. et al.,* The HTK Book Version 3.4, Cambridge University, 2006.
15. *Bo-June (Paul) Hsu and James Glass.* Iterative Language Model Estimation: Efficient Data Structure & Algorithms. In Proc. Interspeech, 2008.

Information about the authors:

Mykola Sazhok

Speech Science and Technology Department, IRTC, Kyiv, Ukraine,
mykola@cybermova.com

Valentyna Robeiko

CyberMova, Kyiv, Ukraine, valya.robeiko@gmail.com