Lexical Stress-based Morphological Decomposition and Its Application for Ukrainian Speech Recognition

Mykola Sazhok[1 2] and Valentyna Robeiko[2]

[1] Hlushkov Institute of Cybernetics, Kyiv, Ukraine
mykola@cybermova.com
[2] International Research/Training Center for Information Technology and Systems, Kyiv, Ukraine
{sazhok,valia.robeiko}@gmail.com

Abstract. This paper presents an approach to word morphological decomposition based on lexical stress modeling. Word segmentation quality is estimated by a hidden variable that assigns the lexical stress. The formulated segmentation criterion is based on a training set of words with manually pointed stresses and a large text corpus. The described search algorithm finds one or more segmentations with the best likelihood. Given arguments confirm the necessity to distinguish stressed and unstressed vowels in the phoneme alphabet for Ukrainian speech recognition systems. The developed tool allows to assign primary lexical stress in unknown words. Experimental research is described and results are discussed.

Keywords: lexical stress, morphological decomposition, speech recogntition, Ukrainian

1 Introduction

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 word segments. So any 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.

In [1] authors assume that morphological decomposition is required for lexical stress prediction particularly for cases where a local context is insufficient. Moreover, presenting words as a sequence of reasonable segments or morphemes is a key to model the word formation and to strive beyond vocabulary limitations particularly in speech understanding systems. Known methods of morphological decomposition relies solely on orthography [2], [3]. In our research, lexical stress prediction and morphological decomposition are considered as a result of the same process through which phonetic, syntactic and semantic hidden features can be discovered from word spelling.

In the next section we consider motives, prerequisites and possible applications, in Section 3 we describe segmentation procedure formalization and implementation, Section 4 is devoted to data description, in Section 5 we report and discuss experimental results followed by Conclusion.

2 Motivation

In Ukrainian, stress position is irregular and it can be changed even within forms of the same word. Fortunately, the available electronic lexicography system contains more than 1.8 million words with manually assigned lexical stress covering all valid word forms [4]. The web-based basic text corpus contains 275 million unchecked word samples, which makes a vocabulary for about two million words. Relatively to the lexicography system, half of which is detected in the basic corpus, OOV words make 2.5%. At least 200 thousand more words allow for OOV reduction to less than 0.5%. So stress prediction is a way to assign lexical stress for the large amount of both new and known words.

The reason to introduce stressed vowels to the text-to-speech system is obvious due to necessity to meet a human perception of duration, pitch, and loudness. In speech recognition, feature extraction models are mostly invariant to prosodic features. However, we believe that introduction of both stressed and unstressed vowels to the phoneme alphabet, at least for Ukrainian, is essential due to phonetic, lexical, and acoustical facts. Stressed vowels normally act as distinctive phonemes changing word grammatical function and meaning that we observe for more than 5% of words in the basic text corpus. In English language, stress shift, normally, causes changes in phonetic content of the pronunciation (compare: récord and recórd). Therefore, such argumentation might be rather inapplicable for a specific language.

Grapheme-to-phoneme conversion methods like [5] could be directly used also for modeling stress, however they have no provisions to account for the structural properties of stress. Here, rather than modifying an existing algorithm we prefere to construct a model concentrated on stress properties and then convert the stressed text to phoneme sequences by means that allow for counting specific pronunciation properties provided by the method that needs about 30 find-replace-and-step rules for Ukrainian [6].

2.1 Application to speech recognition

To explore the acoustical side of lexical stress we estimated phoneme model parameters considering stressed and unstressed vowels as different phonemes and inspected dissimilarities particularly by means of the HMM visualization tool [7]. Following Fig. 1 we can see the difference between models for unstressed and stressed phonemes of “a” and “i” trained on 40 hour multi-speaker speech corpus for Ukrainian [8].

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 corresponded to emitting states are 1.52 times greater for stressed models that confirms the essential difference in duration.

Introducing both stressed and unstressed vowels is relatively small overhead for Ukrainian language since only 6 vowel phonemes are distinguished: a, e, y, i, o, and u.

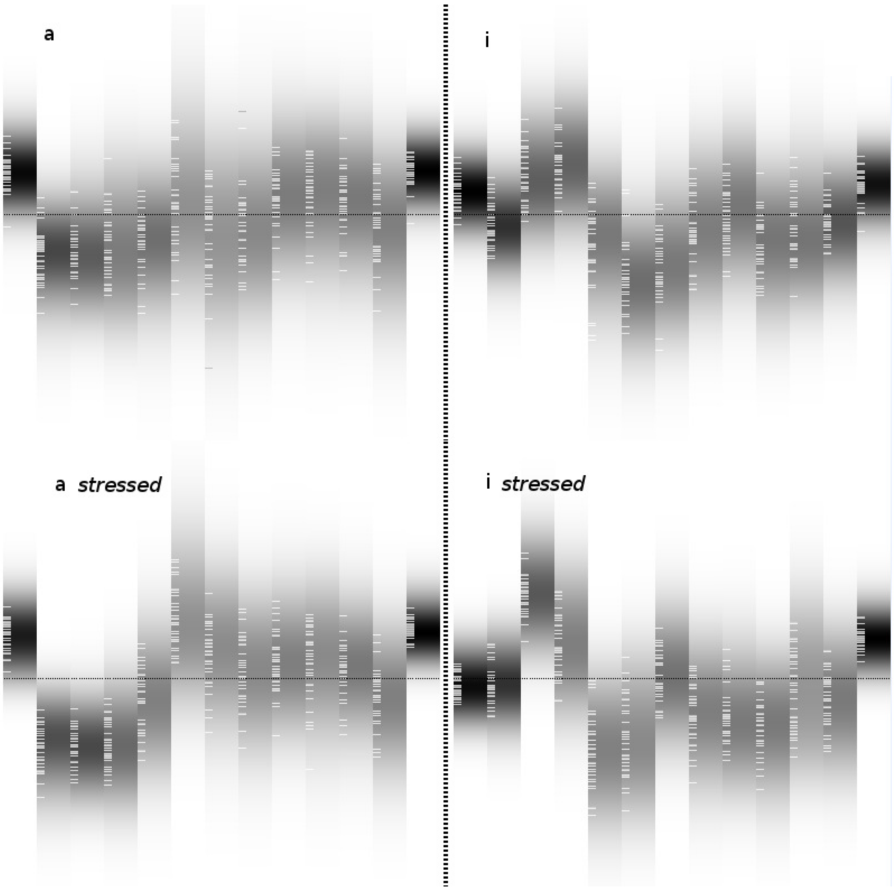

Fig. 1. HMM visualization for Ukrainian unstressed and stressed monophones a and i

However, such extension is more expensive for phoneme alphabets containing significantly more vowels (due to diphthongs, nasalization, etc.).

Perhaps, the most convincing argument for stressed vowel models became the analysis of Ukrainian speech recognition preliminary experiments. A multi-speaker and a single speaker > 40 hour training sets were used to learn models for both versions of the phoneme alphabet (49 and 55 monophones). For word recognition we applied a bi- and a trigram language models and a free order grammar was used for syllable and phoneme recognition. To make the results comparable we ignored the stress information from recognized word and phoneme sequences. In all cases we observed a $12 - 23\%$ improvement relatively to the word/phoneme error rate. Therefore, our concern about possible confusion of stressed and unstressed phonemes was rather exaggerated.

The benefit of morphological decomposition is a possibility to present the entire lexicon with a limited set of morpheme level segments irrespective to their correspondence to morphemes in a sense of classical linguistics.

3 Methodology

3.1 Lexical stress-based word segmentation model

A training vocabulary W contains words with assigned attributes like a lexical stress. Each word w from the vocabulary W can be decomposed into a sequence of symbols $q^{(w)} = (q_1, q_2, \ldots, q_{K_w})$ taken from an alphabet of letters or phonemes Q.

We consider subsequences of $q^{(w)}$ as segments of a certain segmentation $s^{(w)}$ among all valid segmentations $S^{(w)}$ for the word w. The i-th segment of segmentation $s^{(w)}$

$$s_i^{(w)} = (s_{i1}, s_{i2}, \ldots, s_{iL_i^{(w)}}) \tag{1}$$

together with other segments of $s^{(w)}$ cover the entire $q^{(w)}$ without overlaps that is for any $w \in W$

$$\sum_i L_i^{(w)} = K_w,\, 1 \leqslant L_i^{(w)} \leqslant \min\{L_{\max}, K_w\}, \tag{2}$$

$$I(s_{11}^{(w)}) = 1 \text{ and } I(s_{i1}^{(w)}) = I\left(s^{(w)}_{(i-1)L^{(w)}_{i-1}}\right) + 1,\, i > 1, \tag{3}$$

where $I(\cdot)$ returns a segment item index within $q^{(w)}$. The constrain for the largest segment length, $L_{\max}$, determines a model order. More segmentation constrains might be introduced, e.g. restriction on running primarily stressed syllables.

Uniting all segments of valid segmentations for all words belonging to the vocabulary W we form a training set of segments

$$S = \bigcup_{w \in W,\, s^{(w)},\, i} s_i^{(w)} \tag{4}$$

and consider each segment in this set, s_i, with no relation to words.

A stress level, $\theta_k^{(w)} = \{0, 1, 2\}$, assigned to each symbol forms a corresponding attribute sequence $\theta^{(w)} = (\theta_1, \theta_1, \ldots, \theta_k, \ldots, \theta_{K_w})$. We assume the stress level other than zero can be assigned to symbols that introduce a syllable, at least potentially. Normally, these are vowels that may be accomplished with specific consonants like “r” in Slovenian [3]. For other symbols the stress level is not applicable and is always equal to zero. Stress level values may be limited only to 0 and 1 that means only a primary lexical stress is considered. On the contrary, we may introduce more values corresponding to symbol attributes that might be hidden in spelling like reduction, lengthening and modifiers, as well as symbol attribute combinations. Thus, we generally refer to $\theta^{(w)}$ as an attribute sequence for corresponding symbols in w.

Obviously, the returned index in (3) is the same within $\theta^{(w)}$, which subsequences are assigned to $s_i^{(w)}$. Attribute sequences assigned to each segment of $s^{(w)}$ segmentation, in turn, form a set $\Theta^{(w)}$.

We can estimate a probability of the attribute sequence θ given a segment s_i observable in the training set:

$$P\left(\theta | s_i\right) \approx \frac{c(s_i, \theta)}{c(s_i)} \tag{5}$$

where $c(s_i, \theta)$ is a count of segments s_i with stress assignment defined by a stress indication vector θ and $c(s_i)$ is a number of s_i occurrence. All counts are taken from the text corpus for words included in the stress vocabulary. For the segments with low occurrence a smoothing technique should be applied.

Finally, we search through all valid segmentations $s^{(w)}$ and attribute sequences θ that satisfy the expression:

$$\left(\hat{s}^{(w)}, \hat{\Theta}\right) = \underset{s^{(w)},\, \Theta^{(w)}}{\arg\max} \prod_{i,\theta} P\left(\theta | s_i^{(w)}\right) \tag{6}$$

For known words, θ is determined for each segment $s_i^{(w)}$ uniquely, otherwise, all valid attribute sequences are being searched.

Thus, to carry through morphological decomposition we introduced a segmentation model based on features, hidden in word spelling, like lexical stress. However, not all obtained segments are valid morphemes due to potentially more strict morpheme constrains like presence of at least one vowel. Therefore, unifying such a segment with adjacent one is a way to compose a valid morpheme.

4 Segmentation graph analysis

We constructed a dynamic programming graph where finding the shortest trajectory is equivalent to the search (5). Each input symbol introduces a set of valid pairs (segment, attributes), which are nodes on the graph where the partial criterion is accumulated. Connections between nodes correspond to valid segmentations. Memorizing N prospective arrows entering to nodes we can extract N-best word segmentations.

In Fig. 2 an example of one-best stress prediction search (6) is shown for a proper name Obama missing from the basic Ukrainian vocabulary. The word is represented as concatenation of all valid character segments where the largest segment length is limited to five. Input items are down-cased letters accomplished with the word boundary symbol "|". Each input item introduces a set of valid segments with attributes. For compact presentation we show only the result of the attribute application. Thus, a notation "obAm" in column 5 means the segment (o, b, a, m) with the attribute vector (0, 0, 1, 0). Potentially optimal arcs are either shown or coded with the name of previous node. Indicated partial criteria are log probability based. The optimal trajectory, respective nodes and criteria are bold.

In the example we illustrate two running stressed syllables forbidding: in column 7 the segment "mA" follows the segment "a" rather than "obA". As far no constrains for the segment content are introduced, segments may contain a single consonant like "b" in column 3. This is the way to guarantee a successful search (6) for any word. The system may decide that such a segment and an adjacent segment may belong to the same morpheme depending on constrains given by the expert. In accordance to morphology knowledge, for Ukrainian language, only the first and the last morphemes may consist of consonants solely. To compose a formally valid morpheme we may append the segment "b" to the preceding segment preferring more frequent morphemes and coming to Ob-áma. We can see that a foreign word has been approximated with native morphemes. The model updated with rather correctly stressed new word samples learns new morphemes, which make linguistically more justified decomposition of the considered word and its forms possible: Obám-a, Obám-y, etc.

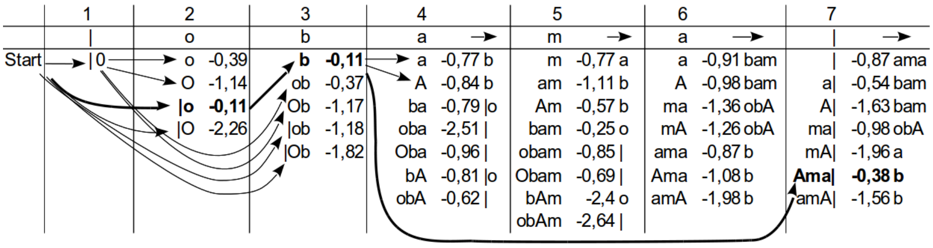
1 2 3 4 5 6 7
| o b a m a |
Start |0 o -0,39 b -0,11 a -0,77 b m -0,77 a a -0,91 bam | -0,87 ama
O -1,14 ob -0,37 A -0,84 b am -1,11 b A -0,98 bam a| -0,54 bam
|o -0,11 Ob -1,17 ba -0,79 |o Am -0,57 b ma -1,36 obA A| -1,63 bam
|O -2,26 |ob -1,18 oba -2,51 | bam -0,25 o mA -1,26 obA ma| -0,98 obA
|Ob -1,82 Oba -0,96 | obam -0,85 | ama -0,87 b mA| -1,96 a
bA -0,81 |o Obam -0,69 | Ama -1,08 b Ama| -0,38 b
obA -0,62 | bAm -2,4 o amA -1,98 b amA| -1,56 b
obAm -2,64 |

Fig. 2. Stress prediction for an out-of-vocabulary word “obama”

4.1 Implementation

To implement the described word segmentation algorithm a set of basic three tools has been developed. Currently, these tools operate only with primary stress information.

The first tool, *putstress*, prepares the data necessary for estimation of probabilities eq:stressAttrBySegmentProb by input data and knowledge base, word frequency vocabulary and, optionally, corrected homograph frequency proportions. For each word the tool tries to retrieve records about stress position and stores found words accomplished with stress-mark and frequency. Found homographs are saved with frequencies updated in accordance to their corrected proportions. An expert may correct more proportions and run the tool again.

The second tool, *guessstress*, implements the search procedure (6) extracting $N-$best sequences of segments with corresponding attributes. A frequency vocabulary of words with assigned stresses is the source to estimate probabilities for hypothetical symbol subsequences.

Finally, the *prep_stressvcb* tool forms a stress vocabulary by the extracted segments. Several additional tools allow for extracting various information from input data, estimated models, and segmentations. All modules are written in Perl language.

5 Data

The stress 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 [4]. Due to the shared spelling an actual word form the 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 the spontaneous speech. A 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 [4]. Adding to the vocabulary 200 000 most frequent words we reduce OOV to less than 0.5%.

Table 1. 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 two or more valid stress positions, referred as homographs, take 6% of the average text. However, homographs may occur with quite different frequency that may considerably affect occurrence for certain segments. Therefore, an expert may correct occurrence proportions in the homograph dictionary containing over 14 000 spellings.

6 Experiments

Known and OOV words were evaluated separately. The evaluation of known words was made in order to learn how a large part of the vocabulary can be coded without explicit lexical stress information. The largest model order, $L_{\max}$, was set to five, and 4-best segmentations were analyzed to form a stress vocabulary. The expert has corrected occurrence proportions in the homograph dictionary for 500 most frequent spellings.

The system detected about one million pairs (segment, stress). Frequencies for different segment lengths are shown in Table 2.

Table 2. Detected segment counts

Segment length, L	1	2	3	4	5
Item count	46	1781	35280	233816	721575
Occurance (million)	2115.652	1848.766	1581.879	1314.993	1070.579

215 000 segments were used to predict a lexical stress for words in 275M text corpus. Stress position for less than 1% known words was detected incorrectly. Stress de-

tection for 5 000 OOV words was incorrect in 21.1% words or 5.3% syllables. However, over 50% of incorrectly stressed words have strong foreign origins.

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

7 Conclusion

The proposed morpheme level segmentation model allows for simultaneous extraction the features generally ignored in word spelling. However, these features are hypothetical and a word context must be used to choose either right stress position or letter modifiers or their absence that is typical for homographs. The segment level context introduction is a way to further model improvement.

The model refinement is possible in expectation-maximization manner, especially with a slight supervising for stress correction between iterations. Future research should also concern a model order, letter modifiers and phonemic input, as well as more languages should be considered.

References

1. Black, A., Lenzo, K., Pagel, V.: Issues in Building General Letter to Sound Rules. 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, (1998), pp. 77–80.
2. Creutz, Mathias; Lagus, Krista 2004. Induction of a simple morphology for highly-inflecting languages. In: Proc. 7th Meeting of the ACL Special Interest Group in Com putational Phonology (SIGPHON), Barcelona, pp. 43–51.
3. Gams, Matjaž, et al. Automatic lexical stress assignment of unknown words for highly inflected Slovenian language. In: Text, Speech and Dialogue. Springer Berlin Heidelberg, 2006, pp. 165–172.
4. http://lcorp.ulif.org.ua/dictua/
5. M. Bisani and H. Ney. "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". Speech Communication, Volume 50, Issue 5, May 2008, pp. 434–451.
6. V. Robeiko, M. Sazhok. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian // In Proc. UkrObraz'2012, Kyiv, 2012, pp. 43–46.
7. www.cybermova.com/speech/visual-hmm.htm
8. Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, Nina Vasylieva, Oleksandr Radoutsky. Ukrainian Broadcast Speech Corpus Development // In Proc. Specom'2011, Kazan, RF, pp. 244–247.