Using Morphological Information for Robust Language Modeling in Czech ASR System

Pavel Ircing, Josef V. Psutka, and Josef Psutka

***Abstract*—Automatic speech recognition, or more precisely language modeling, of the Czech language has to face challenges that are not present in the language modeling of English. Those include mainly the rapid vocabulary growth and closely connected unreliable estimates of the language model parameters. These phenomena are caused mostly by the highly inflectional nature of the Czech language. On the other hand, the rich morphology together with the well-developed automatic systems for morphological tagging can be exploited to reinforce the language model probability estimates. This paper shows that using rich morphological tags within the concept of class-based n-gram language model with many-to-many word-to-class mapping and combination of this model with the standard word-based n-gram can improve the recognition accuracy over the word-based baseline on the task of automatic transcription of unconstrained spontaneous Czech interviews.**

***Index Terms*—Language models, speech recognition and synthesis.**

I. Introduction

In the recent decade, the automatic processing of languages other than English has been gradually receiving more attention as both the availability of computation resources and the relative success of English automatic speech recognition and natural language processing systems made this field of research rather attractive. However, when researchers began to develop systems for languages that belong to different language groups, it turned out that some of the methods that worked well for English do not yield satisfactory results. This statement is true also for the Czech language.

Czech, as well as other Slavic languages (such as Russian and Polish, to name the most known representatives), is a richly inflected language. The declension of Czech nouns, adjectives, pronouns, and numerals has seven cases. Case, number (singular or plural), and gender (masculine, feminine, or neuter) are usually distinguished by an inflectional ending; however, sometimes the inflection affects the word stem as well. The declension follow 16 regular paradigms but there are some additional irregularities.

Manuscript received August 01, 2008; revised December 03, 2008. Current version published April 01, 2009. This work was supported in part by the Ministry of Education of the Czech Republic under Projects MSMT LC536 and MSMT 2C06020 and in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) program under EC Grant IST-FP6-034434. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tanja Schultz.

The authors are with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, 306 14 Plzeň, Czech Republic (e-mail: ircing@kky.zcu.cz; psutka_j@kky.zcu.cz; psutka@kky.zcu.cz).

Digital Object Identifier 10.1109/TASL.2009.2014217

The conjugation of Czech verbs distinguishes first, second, and third person in both singular and plural. The third person in the past tense is marked by gender. The conjugation is directed by 14 regular paradigms but many verbs are irregular in the sense that they follow different paradigms for different tenses.

Word order is grammatically free with no particular fixed order for constituents marking subject, object, possessor, etc. However, the standard order is subject-verb-object. Pragmatic information and considerations of topic and focus also play an important role in determining word order. Usually, topic precedes focus in Czech sentences.

In order to make a language with such free word order understandable, the extensive use of agreement is necessary. The strongest agreement is between a noun and its adjectival or pronominal attribute: they must agree in gender, number, and case. There is also agreement between a subject (expressed by a noun, pronoun, or even an adjective) and its predicate verb in gender and number, and for pronouns, also in person. Verbal attributes must agree in number and gender with its related noun, as well as with its predicate verb (double agreement). Possessive pronouns exhibit the most complicated type of agreement—in addition to the above-mentioned triple attributive agreement with the possessed thing they must also agree in gender and number with the possessor. Objects do not have to agree with their governing predicate verb but the verb determines their case and/or preposition. Similarly, prepositions determine the case of the noun phrase following them [1].

An interesting phenomenon occurring in the Czech language is a considerable difference between the written form of the language (Standard or Literary Czech) and the spoken form (Common Czech). This difference occurs not only on the lexical level (usage of Germanisms and Anglicisms), but also on the phonetic and morphological level. Some of the differences can even be formalized (see for example [2]).

II. Robust Language Modeling for Czech Large Vocabulary Continuous Speech Recognition

A. Challenges

The properties of the Czech language described above pose a challenge for large vocabulary continuous speech recognition (especially for language modeling), mainly because of the following problems [3].

1) A rapid vocabulary growth and a high out-of-vocabulary (OOV) rate.
   The vocabulary size grows very rapidly with the size of the training corpus. This problem is caused by the aforementioned high degree of inflection (potentially up to

300/20/200 word forms for a single verb/noun/adjective, but with frequent cases of systematic homography) and also by a high degree of derivation (use of prefixes and suffixes). Since, from the ASR point of view, a word is defined by its spelling, two differently spelled inflections or derivations of the same basic word form are considered different words.

Consequently, the vocabularies extracted from (necessarily limited) training data have an inevitably higher OOV rate than comparable lexicons of English.

2) A high perplexity of word-based n-gram language models. This fact is usually attributed to the free word order. However, experiments with a trigram model with permutations [4] showed that the free word ordering does not really pose such a serious problem, especially in the short-term dependencies represented by the n-gram language models.

   We suspect that the high perplexity is again closely connected with the highly inflectional nature of the Czech language. The idea is as follows—even though available Czech text corpora already reached the size that would be sufficient for training a decent language model for English, the parameter estimates for Czech still remain unreliable due to the higher number of distinct words (and consequently the language model parameters).
3) Lack of language model training data for spontaneous speech.

   This problem stems from the substantial difference between the spoken and the written form of the language. Since most text corpora consist of written text (newspaper articles, books, broadcast news transcripts), it is usually hard to find appropriate training data for estimating the language model that could be used in a system for transcribing spontaneous speech.

Our paper describes an attempt to address mainly the second problems listed above, that is, the poor language model probability estimation, by exploiting available morphological information. We have, however, listed all the challenges as we perceive them since we feel that especially the first two problems (rapid vocabulary growth and high perplexity) are closely connected and that the third one (discrepancy between speech transcripts and written text) is actually responsible for a rather modest benefit of the "background" language model described later in the paper.

B. Available Morphological Information

Czech computational morphology is a field that has been extensively studied during the last 20 years. An elaborate tagset and a set of well-developed tools for automatic morphological processing of the Czech texts are therefore available [5]. Every tag in this positional system is represented as a string of 15 symbols. Each position in the string corresponds to one morphological category in the following order—part of speech, detailed part of speech, gender, number, case, possessor's gender, possessor's number, person, tense, degree of comparison, negation, and voice. Positions 13 and 14 are currently unused and finally position 15 is used for various special purposes (such as marking colloquial and archaic words or abbreviations). Nonapplicable values are denoted by a single hyphen (-). A more detailed description of the tag system is beyond the scope of this paper and can be found in [5].

For example, the tag `VB-S---3P-AA---` denotes the verb (`V`) in either the present or the future tense (`B`), singular (`S`), in the third person (`3`), in the present tense (`P`), affirmative (`A`), and in the active voice (`A`).

The usage of such morphological tags is likely to be beneficial for the Czech language modeling because of the following factors.

- The number of distinct tags occurring even in the large text corpus is very small in comparison with the number of distinct words in the same corpus (there approximately 1500 unique tags versus several tens of thousands distinct words) and therefore estimates of the tag-based language model parameters are going to be very reliable.
- Using this type of morphological information seems to be suitable for the language modeling of the Czech language because of the extensive use of agreement already described in the Introduction. Since all morphological categories involved in the agreement rules (gender, number, case, etc.) are included in the morphological tags, there should exist some dependencies between adjacent tags that can be captured even by an n-gram language model.

Note that, due to the frequent homography, the word-to-tag mapping is highly ambiguous, since the ambiguity does not have to be a result of a word having several possible parts of speech—the difference can appear in any position of the tag.

The morphological tagging can be performed automatically using a serial combination of the morphological analyzer and tagger (see for example [6]). The current accuracy of automatic tagging is over 95% [7].

C. Incorporating Tags Into the Language Model and Into the Decoder

Let us state again that we would like to exploit the dependencies that (as we believe) exist between the adjacent morphological tags and at the same time we of course still want to have surface word forms at the output of the ASR decoder. Using a class-based language model with many-to-many word-to-class mapping (remember the homography of the words described above) is thus a natural choice. The n-gram probability of the word w_i given the history h_i is in such a model defined by the formula

$$P(w_i|h_i) = \sum_{c_{i-n+1},\cdots,c_{i-1},c_i} P(w_i|c_i) P\left(c_i|c_{i-n+1}^{i-1}\right) \quad (1)$$

where $P(w_i|c_i)$ denotes the probability of the word w_i given the class c_i and $P(c_i|c_{i-n+1}^{i-1})$ represents the n-gram probability that the class c_i will follow the previous $(n-1)$ classes $c_{i-n+1}, \ldots, c_{i-1}$.

In the following paragraphs, we will present a way of an efficient incorporation of the language model (1) into the ASR decoder developed by AT&T Labs-Research. This decoder is built on the basis of weighted finite-state transducers (FST) [8], [9] and we feel that a brief overview of the FST terminology and basic operations is necessary for the comprehensibility of our approach. Please note that a manner of representing a class-

based model as a finite-state automaton has already been introduced for example in [10] but it dealt with many-to-one word-to-class mapping only. We will show that many-to-many word-to-class mapping can be also consistently represented within the finite-state machine paradigm.

1) Finite-State Transducers Essentials and Their Use in Speech Recognition: The definition of weighted finite-state machines depends on the algebraic structure called semiring, $(\mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1})$ [11]. A semiring is a set $\mathbb{K}$ with two binary operations, collection $(\oplus)$ and extension $(\otimes)$, such that $\oplus$ is associative and commutative with identity $\bar{0}$, $\otimes$ is associative with identity $\bar{1}$ and $\otimes$ distributes over $\oplus$.

For example, $(\mathbb{R}, +, ., 0, 1)$ is a semiring. Since in speech recognition we use negative natural log probabilities and the Viterbi approximation, the proper semiring is defined by $(\mathbb{R}_+ \cup \{\infty\}, \min, +, \infty, 0)$. This structure is called the tropical semiring.

The most general finite-state machine, a weighted finite-state transducer [9] over a given semiring is an 8-tuple

$$T = (Q, Q_0, F, \Sigma, \Delta, E, \gamma, \rho) \tag{2}$$

where Q is the finite set of states, $Q_0 \subseteq Q$ is the set of initial states, $F \subseteq Q$ is the set of final states, Σ is the input alphabet, Δ is the output alphabet, $E \subseteq Q \times \Sigma \times \Delta \times \mathbb{K} \times Q$ is the finite set of transitions, γ is the initial weight function mapping $Q_0 \to \mathbb{K}$ and ρ is the final weight function mapping $F \to \mathbb{K}$.

A transition $t = (q, a, b, c, r) \in E$ can be viewed as an arc from the source state q to the destination state r, labeled with the input symbol a, the output symbol b and the weight c.

A path π in T is a set of consecutive transitions from q to q', that is

$$\pi = ((q_1, a_1, b_1, c_1, r_1), \ldots, (q_n, a_n, b_n, c_n, r_n)) \tag{3}$$

where $q_1 = q$, $r_n = q'$ and $r_i = q_{i+1}$ for $i = 1, \ldots, n-1$. A successful path π' is a path from an initial state to a final state. The input label ι of the path π' is the concatenation of the labels of its constituent transitions, i.e.,

$$\iota(\pi') = a_1 a_2 \ldots a_n \tag{4}$$

and analogically the output label o of the path π' is defined as

$$o(\pi') = b_1 b_2 \ldots b_n. \tag{5}$$

The weight ω associated to π' is the $\otimes$-product of the initial weight function value for a given initial state q_1, the weights of its constituent transitions and the final weight function value for a given final state r_n, that is

$$\omega(\pi') = \gamma(q_1) \otimes c_1 \otimes c_2 \ldots \otimes c_n \otimes \rho(r_n) \tag{6}$$

A string $x \in \Sigma^*$ (the asterisk denotes the Kleene closure) is accepted by T if there exists a successful path π' labeled with the input string x (i.e., $\iota(\pi') = x$). The weight associated by T to the sequence x is then the $\oplus$-sum of the weights of the successful paths π' labeled with the input label x and an output label $y \in \Delta^*$.

Such mapping from $\Sigma^* \times \Delta^*$ to $\mathbb{K}$ is called a weighted transduction of a given automaton T and is defined as

$$L_T(x, y) = \bigoplus_{\pi' \in q \overset{x:y}{\rightsquigarrow} q'} \omega(\pi') \tag{7}$$

where $\bigoplus$ represents the summation using the collection operator $\oplus$ and $\pi \in q \overset{x:y}{\rightsquigarrow} q'$ denotes the set of paths from q to q' labeled with the input string x and the output string y.

The AT&T FSM Library offers software tools for operations with finite-state automata, such as, for example, union, concatenation, and Kleene closure and also tools for automata determinization and minimization. Let us present the exact definitions of the two operations that are essential for the application of the finite-state transducers to speech recognition—projection and composition [12].

Each transduction $L_T : \Sigma^* \times \Delta^* \to \mathbb{K}$ has two associated weighted languages—the first (input) projection $\xi_1(L_T) : \Sigma^* \to \mathbb{K}$ and the second (output) projection $\xi_2(L_T) : \Delta^* \to \mathbb{K}$ defined by

$$\xi_1(L_T)(x) = \bigoplus_{y \in \Delta^*} L_T(x, y) \tag{8}$$

$$\xi_2(L_T)(y) = \bigoplus_{x \in \Sigma^*} L_T(x, y). \tag{9}$$

A composition of two transductions $L_T : \Sigma^* \times \Delta^* \to \mathbb{K}$ and $L_S : \Delta^* \times \Gamma^* \to \mathbb{K}$ is defined by

$$L_R(x, z) = L_T(x, y) \circ L_S(y, z) = \bigoplus_{y \in \Delta^*} L_T(x, y) \otimes L_S(y, z) \tag{10}$$

where $x \in \Sigma^*$, $y \in \Delta^*$, and $z \in \Gamma^*$. The transducer R then represents a composition of the automata $T \circ S$ and provides a mapping $\Sigma^* \times \Gamma^* \to \mathbb{K}$. It is clear that the composition is useful for combining different information sources or different levels of representation.

Using the composition, the speech recognizer can be represented by the so-called recognition cascade $H \circ C \circ L \circ G$, where each component is a weighted finite-state transducer over the tropical semiring—H represents an acoustic model, C transduces context-dependent phones to context-independent ones, L represents a pronunciation lexicon and finally G is a word-based language model (see Fig. 1). The decoder task of finding the best word sequence $\hat{W}$ can be then expressed in terms of FST operations as

$$\begin{aligned} \hat{\pi}' &= \arg\min_{\pi'} \xi_2(O \circ H \circ C \circ L \circ G) \\ \hat{W} &= \iota(\hat{\pi}') \end{aligned} \tag{11}$$

where O is the input sequence of acoustic features which can of course be transformed to a trivial finite-state machine as well. The $C \circ L \circ G$ part of the cascade is constructed beforehand whereas the composition with O and H is performed during the decoder run.

Let us now go back to the class-based model defined by (1). The probability of the entire word sequence

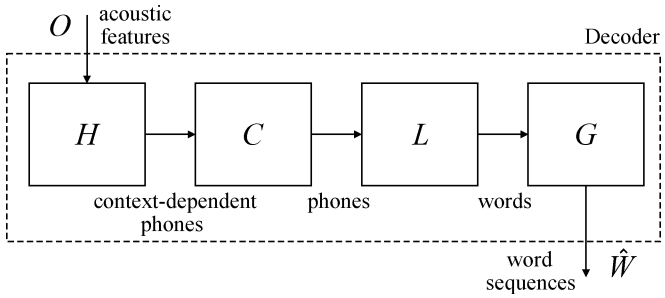

Fig. 1. Recognition cascade.

$W = w_1, w_2, \ldots, w_K$ can be expressed using this model as

$$P(W) = \prod_{i=1}^{K} \sum_{c_{i-n+1},\cdots,c_{i-1},c_i} P(w_i|c_i) P\left(c_i|c_{i-n+1}^{i-1}\right) \quad (12)$$

which can be rewritten as

$$P(W) = \sum_{C} \prod_{i=1}^{K} P(w_i|c_i) \prod_{i=1}^{K} P\left(c_i|c_{i-n+1}^{i-1}\right) \quad (13)$$

where C denotes all possible class sequences $c_1, \ldots, c_K$. Now if we replace the arithmetic sum and product with the general $\oplus$ and $\otimes$ operations, we see that the two product components of (13) constitute a weighted transductions [see (6) and (7)] and the entire right-hand side of (13) can be rewritten as

$$\bigoplus_{c \in \Psi^*} L_T(w, c) \otimes L_V(c, c) \quad (14)$$

where Ψ denotes the vocabulary of class symbols, Θ denotes the word vocabulary and consequently $c \in \Psi^*$, and $w \in \Theta^*$ represent the string of class symbols and words, respectively. It is evident that (14) corresponds to the transducer composition formula (10) and hence the class-based language model (1) can be represented by a composition of two finite-state transducers $T \circ V$, where T realizes a mapping from word-class pairs (w_i, c_i) to $-\ln P(w_i|c_i)$[1] and V is a transducer that maps class sentences to $-\ln P(c_i|c_{i-n+1}^{i-1})$ and thus constitutes the exact analogy of the word-based n-gram G, just with classes instead of words.

So it seems that if we want to use the class-based n-gram model instead of the word-based one, we can simply replace the transducer G with $T \circ V$ in the recognition cascade, but in that case we would obtain the best class sequence instead of the best word sequence in the output of the decoder. However, the AT&T decoder is built so that it produces not only the best sequence but also a lattice. The lattice is an acyclic finite-state transducer containing the most probable paths through the recognition cascade for a given utterance. It has context-dependent phones on the input side and output labels from the right-most transducer in the cascade on the output side.

[1]The reader should bear in mind that we are working with natural log probabilities within the tropical semiring.

Therefore, we can retrieve the best word sequence $\hat{W}$ even from the class-based lattice Z using the following operations:

$$\hat{\pi}' = \arg\min_{\pi'} \xi_2 \left(\xi_1(Z) \circ C_{\bar{1}} \circ L_{\bar{1}}\right)$$
$$\hat{W} = \iota(\hat{\pi}') \quad (15)$$

where $C_{\bar{1}}$ and $L_{\bar{1}}$ are special variants of C and L with all weights set to $\bar{1}$.

However, the transducer $C \circ L \circ T \circ V$ often becomes too large due to a high degree of ambiguity caused by many-to-many word-to-class mapping. Thus, the first recognition run is usually performed with a simple word-based n-gram G. Then, the language model score is stripped from the word-based output lattices (X) and the resulting lattices are rescored with $T \circ V$. In terms of FSM operations, the best word sequence $\hat{W}$ is determined by

$$\hat{\pi}' = \arg\min_{\pi'} \xi_1 \left(\xi_2(X \circ G_-) \circ T \circ V\right)$$
$$\hat{W} = \iota(\hat{\pi}') \quad (16)$$

where G_- denotes the original language model with negative weights.

It is generally known (and our preliminary experiments proved it) that class-based language models yield more robust probability estimates than word-based models but at the same time they have worse discrimination ability ("sense of detail"). Thus, word-based and class-based language models are usually combined in some manner. The FST framework offers a natural way of model combination—we can simply retain the word language model score in the output lattices and then compose the lattices with $T \circ V$. We have found out empirically that better results are achieved when the transducer T just maps words to classes but does not associate any probability with this mapping. Such model combination can be expressed formally as

$$\hat{\pi}' = \arg\min_{\pi'} \xi_1 \left(\xi_2(X) \circ T_{\bar{1}} \circ V\right)$$
$$\hat{W} = \iota(\hat{\pi}') \quad (17)$$

where again $T_{\bar{1}}$ is the transducer T with all weights set to $\bar{1}$.

The ratio between the word-based n-gram G and the class-based n-gram V contributions must be carefully set up to ensure good recognition results (which is not surprising as the so-called scaling factors must be optimized also when using a word-based model alone). Within the FST framework, the scaling of the language model component is performed by multiplying each transition weight in a transducer by a scaling factor. The scaling factors for both G and V are usually optimized on a development data set.

Let us now summarize that finding the best word sequence according to (17) with scaled G and V components corresponds to the usage of a language model

$$\hat{P}(w_i|h_i) = \left(P\left(w_i|w_{i-l+1}^{i-1}\right)\right)^{\theta_w} \left\{ \max_{c_{i-n+1},\cdots,c_i} P\left(c_i|c_{i-n+1}^{i-1}\right) \right\}^{\theta_c} \quad (18)$$

where θ_w and θ_c are the word-based and class-based model scaling factors, respectively. Note that although this model is deficient (in the sense that the expression $\sum_{w_i} \hat{P}(w_i|h_i)$ does

TABLE I
TRANSCRIBED SPEECH DATA

	Training (336)		Test (10)	
	Male	Female	Male	Female
Speakers	145	191	5	5
Hours transcribed	36.25	47.75	13.15	9.7

not sum up to 1) it consistently outperforms the word-based n-grams used alone (see the evaluation section).

III. EXPERIMENTAL EVALUATION

A. Training and Test Corpora

The presented language modeling approach have been tested using the Czech part of the ASR training and test data prepared within the MALACH project [13]. The ultimate goal of this project (which ended September 2007) was to use advanced ASR and IR techniques to facilitate access to the large multilingual spoken archives created by the Visual History Foundation. These archives consist of testimonies given by the survivors and witnesses of the Holocaust. The entire collection contains almost 52 000 interviews in 32 languages, a total of 116 000 hours of audio and video.

The Czech portion of the archives consists of 346 interviews that we divided into 336 speakers used for the ASR training and ten test speakers. A 15-min segment was transcribed from each of the training speakers, yielding a total of 84 h of annotated speech. The testimonies of the test speakers were transcribed completely, yielding approximately 23 h of transcribed speech. The ratio between males and females in terms of the number of speakers and the amount of transcribed speech is shown in Table I.

The test set was further divided into the development data (randomly selected 500 sentences from the test set) that were used for tuning of the language model scaling factors and the evaluation data (the rest of the test set—6368 sentences).

The training part of this speech corpus was of course used for acoustic model parameter estimation but also served as a basic corpus for the language modeling purposes (see Section III-C for details)

B. Front-End and Acoustic Models

The acoustic models were trained using 84 h of transcribed speech (see Section III-A). The data was parameterized as 17-dimensional PLP cepstral features including their delta and delta-delta derivatives (resulting into 51-dimensional feature vector). These features were computed at a rate of 100 frames per second. Cepstral mean subtraction was applied on a per-utterance basis. The resulting cross-word-triphone-based models were trained using the HTK toolkit [14] and had approximately 6 k states and 107 k Gaussians. The performance of the acoustic models could be of course further enhanced by using state-of-the-art speaker adaptation techniques but the improvements of those techniques were shown to be additive and we therefore considered such tuning of the acoustic models to be beyond the scope of this paper.

TABLE II
BASIC PROPERTIES OF DIFFERENT LEXICONS

	Corpus Size	Lexicon Size	OOV Rate
Words	606,242	41,249	5.07%
Tags	606,242	1,180	0.11%

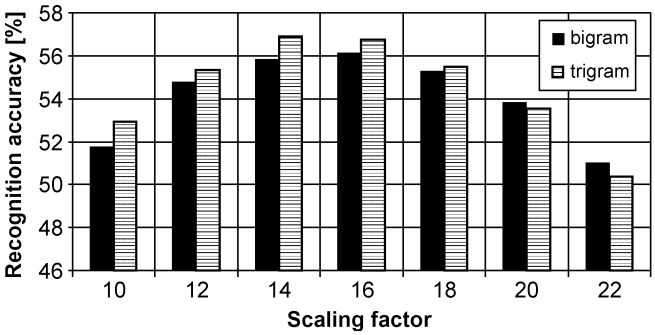

Fig. 2. Recognition accuracy on the development data.

C. Language Models

The previous experiments with the MALACH data [15] hinted that the transcripts of the interviews used for acoustic model training, although relatively small from the language modeling point of view (approximately 600 k tokens), constitute quite a good language modeling corpus to start with. Therefore, only those transcripts were used in the first round of experiments. The text was processed using a serial combination of the morphological analyzer and tagger [6] in order to obtain data for the proposed class-based language models. Basic properties of both word and tag versions of the corpus are summarized in Table II (the OOV rate is measured on the development set).

Bigram and trigram language models were estimated from both the word (transducer G) and the tag corpus (transducer V). All models employed Katz's backing-off scheme and were estimated using the SRILM toolkit [16]. We have also constructed the transducer mapping words to tags ($T_{\overline{1}}$) from the "parallel" word and tag corpora. Then we successively put the bigram and trigram word-based models G into the recognition cascade and tuned their scaling factors using the development data. The behavior of the recognition accuracy of those word-based models depending on the scaling factors is depicted on Fig. 2. Word lattices for all the tested settings were also generated during the decoder runs.

The scaling factor values that yielded the best results (16 for the bigram, 14 for the trigram) were then used to recognize also the evaluation data. Resulting development and evaluation data accuracies and perplexities (PPL) are shown in Table III.

As can be seen from both the development and the evaluation data results, the trigram model does not outperform the bigram. This might be a little surprising for readers accustomed to English ASR but the very limited contribution of the trigram model to the recognition accuracy is a phenomenon that has been consistently observed in all Czech ASR system across different do-

TABLE III
BASELINE RESULTS WITH TUNED WORD-BASED MODELS

	Devel. data		Eval. data	
Language model	PPL	Acc [%]	PPL	Acc [%]
bigram ($\theta_w = 16$)	291.38	56.17	374.02	52.54
trigram ($\theta_w = 14$)	277.82	56.89	360.36	52.45

TABLE IV
EVALUATION RESULTS WITH COMBINATION OF MODELS

Language model	Acc [%]
word bigram $\circ$ class bigram	54.67
word bigram $\circ$ class trigram	55.27

mains. Our hypothesis is that it is again caused by a rich morphology that leads to excessive number of unique trigrams and therefore makes the trigram estimates unreliable.

The development data lattices obtained using the bigram word model with various scaling factors (the trigram lattices were not used as their baseline accuracy did not promise a substantially better results) were then rescored with the class-based bigram and trigram models, again scaled with a set of different factors. The rescoring was performed by composition with the mapping transducer $T_{\bar{1}}$ and the tag model V according to (17)—that is, word-based and class-based model probabilities were combined in a way described by (18). The development data tests showed that there is quite a wide range of scaling factor values where the combination of word-based and tag-based model outperforms the system with only word-based models (the plot is not shown here as the 3-D graph is not very transparent)—the very best performing scaling factor combinations were again used to process the evaluation data and the results are presented in Table IV.

Now the results on both the development and the evaluation data indicate that the combination of word-based and tag-based models consistently outperforms the usage of word-based models alone. We have performed Wilcoxon signed-rank test to assess whether the difference is statistically significant—the outcome corroborates our hypothesis as the p-value is virtually zero (at the level of 10^{-45}).

It could be argued that such a significant improvement of the recognition accuracy is actually due to the weak estimates of the baseline word-based model. In order to test this hypothesis, we have estimated the "background" language model using the LN text corpus. This corpus consists of texts from Czech daily newspapers Lidové Noviny. The data were collected by the Institute of Czech National Corpus, cleaned at the Institute of Formal and Applied Linguistics and now they constitute a part of the Prague Dependency Treebank 1.0 (http://ufal.mff.cuni.cz/pdt/). The texts contained in the corpus were published during the period 1991 through 1995 and constitute approximately 33 million tokens in about 2.3 million sentences.

The complete lexicon of the LN corpus has approximately 650 k distinct words; we have, however, estimated a bigram word-based language model on this corpus using just the 41 k lexicon derived from the MALACH speech transcripts. We have used this fixed lexicon in order to filter out the potential effect of lower OOV rate of a larger lexicon from the LN corpus. This way, any benefit of the background language model should come purely from better n-gram estimates.

TABLE V
TUNING OF THE INTERPOLATION WEIGHT

λ	PPL	Acc [%]	λ	PPL	Acc [%]
1.0	291.38	56.17	0.4	320.36	56.64
0.9	264.28	57.39	0.3	352.81	55.91
0.8	264.34	57.50	0.2	402.26	55.02
0.7	270.56	57.50	0.1	490.17	53.10
0.6	281.56	57.14	0.0	898.76	47.60
0.5	297.67	56.88			

TABLE VI
EVALUATION RESULTS WITH INTERPOLATED MODEL

Language model	Acc [%]
word bigram Tr	52.54
word bigram Int	54.15
word bigram Int $\circ$ class bigram	55.76
word bigram Int $\circ$ class trigram	56.32

The background language model (LN) was then linearly interpolated with the bigram model estimated from the transcripts which was used in the previous set of experiments (Tr) in the following way:

$$P_{Int}(w_i|h_i) = \lambda P_{Tr}(w_i|h_i) + (1-\lambda)P_{LN}(w_i|h_i). \quad (19)$$

The value of the interpolation weight λ was optimized using the development data. Such an interpolation was already successfully used in the experiments with both English and Czech MALACH data [13]. The language model scaling factor was kept fixed to the value that was found to be optimal for the Tr model ($\theta_w = 16$). The behavior of the development set perplexity and recognition accuracy is shown in Table V.

The best performing interpolated model ($\lambda = 0.8$) was then used to process the evaluation data. Two version of this model were actually employed—one with scaling factor $\theta_w = 16$ for the evaluation of the word-based bigram accuracy and one with $\theta_w = 14$ for generating lattices for the subsequent rescoring with the class-based model. The rescoring was again performed according to (18) and both the mapping transducer $T_{\bar{1}}$ and the tag n-gram model V remained the same as in the previous set of experiments. The evaluation data results with the interpolated word-based language model (Int) are summarized in Table VI, together with the baseline result for the model estimated from acoustic data transcripts only (Tr).

The results in Table VI suggest that the improvements in the recognition performance caused by the proposed class-based models are additive and that the tag-based model is able to improve also the accuracy of the recognition systems with more robust baseline language models. This hypothesis was also validated by the experiments performed by one of the authors of this

TABLE VII
COMPARISON OF MANY-TO-MANY AND MANY-TO-ONE MAPPING

	Many-to-many		Many-to-one	
test set - LM	Acc [%]	Time [mins]	Acc [%]	Time [mins]
devel. - bigram	59.17	9	58.50	4
devel. - trigram	59.50	70	58.42	15
eval. - bigram	55.76	-	55.14	-
eval. - trigram	56.32	-	55.45	-

paper on a broadcast news task where the accuracy improved from 70.20% (word bigram) to 72.73% (rescoring with class trigram) [17]. Note that the more prominent contribution of the tag-based model in the broadcast news task is to be expected as the sentences in the news are naturally mostly grammatically correct and thus also the tag model itself is more robustly estimated.

Finally, we would like to present one more set of comparative experiments to justify the use of many-to-many word-to-class mapping as opposed to many-to-one mapping. In the case of many-to-one word-to-class mapping, the probability of the word w_i given the history h_i is given by

$$P(w_i|h_i) = P(w_i|c_i)P\left(c_i|c_{i-n+1}^{i-1}\right) \tag{20}$$

and the absence of the summation (in comparison with (1)) results into the model that is less computationally expensive.

In order to test the model (20), we have modified the mapping transducer T_{I} in such a way that it maps each word to a single tag, namely the one that is most frequently associated with such word in the training data. The tag n-gram model V remains unchanged. We have performed a set of rescoring experiments using the best available word lattices, i.e., the ones generated with the interpolated bigram model. The results comparing the performance of models (1) and (20) on both the development and the evaluation set are given in Table VII, together with the running times of both models on the development data.

The results in the table essentially conform with our intuition that allowing each word to have only one part-of-speech tag is an oversimplification of the real language behavior, as the models with many-to-many word-to-class mapping consistently outperform the models with many-to-one mapping. On the other hand, the expected greater computational load of the former model also turned out to be true.

IV. CONCLUSION

The aim of our paper was to demonstrate that the rich morphology of the Czech language that poses challenges for the language modeling can be, on the other hand, employed to reinforce the language model probability estimates. We have shown that the well-known concept of the class-based n-gram language model with many-to-many word-to-class mapping can be efficiently represented within the finite-state-transducer framework and that the class-based models that use morphological tags for determining class membership can (in combination with the standard word-based n-grams) significantly improve the recognition accuracy. Note that the improvement was achieved even on a challenging task of automatic transcription of unconstrained, spontaneous interview where the syntactic rules are often not followed very closely.

It can be expected that the proposed technique would yield good results also for other languages from the Slavic family as their morphology demonstrates essentially the same patterns as the Czech language.

ACKNOWLEDGMENT

The access to the METACentrum computing facilities provided under the research intent MSM6383917201 is highly appreciated because otherwise the large experiments reported in this paper could not have been finished in time.

REFERENCES

[1] M. Grepl, Z. Hladká, M. Jelínek, P. Karlík, M. Krčmová, M. Nekula, Z. Rusínová, and D. Šlosar, *Příruční mluvnice češtiny*. Praha, Czech Republic: NLN, 1996.
[2] , P. Sgall, J. Hronek, A. Stich, and J. Horecký, Eds., *Variation in Language: Code Switching in Czech as a Challenge for Sociolinguistics*. Amsterdam, The Netherlands: John Benjamins, 1992.
[3] F. Jelinek, "Proposal for speech recognition of a Slavic language: Czech," Johns Hopkins Univ., Baltimore, MD, 1997, Tech. Rep..
[4] E. Whittaker, "Statistical Language Modelling for Automatic Speech Recognition of Russian and English," Ph.D. dissertation, Univ. of Cambridge, Cambridge, U.K., 2000.
[5] J. Hajič, *Disambiguation of Rich Inflection. (Computational Morphology of Czech)*. Prague, Czech Republic: Karolinum, 2004.
[6] J. Hajič and B. Hladká, "Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset," in *Proc. COLING-ACL Conf.*, Montreal, QC, Canada, 1998, pp. 483–490.
[7] D. J. Spoustová, J. Hajič, J. Votrubec, P. Krbec, and P. Květoň, "The best of two worlds: Cooperation of statistical and rule-based taggers for Czech," in *Proc. Balto-Slavonic Natural Lang. Process. Workshop, ACL*, Prague, Czech Republic, 2007, pp. 67–74.
[8] M. Mohri, "Finite-state transducers in language and speech processing," *Comput. Linguist.*, vol. 23, no. 2, pp. 269–311, 1997.
[9] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," in *Proc. ASR2000, Int. Workshop Autom. Speech Recognition: Challenges for the Next Millennium*, Paris, France, 2000, pp. 97–106.
[10] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic automata for language modeling," *Comput. Speech Lang.*, vol. 10, pp. 265–293, 1996.
[11] W. Kuich and A. Salomaa, *Semirings, Automata, Languages*. Berlin, Germany: Springer-Verlag, 1986.
[12] F. Pereira and M. Riley, "Speech recognition by composition of weighted finite automata," in *Finite-State Lang. Process.*, E. Roche and Y. Schabes, Eds. Cambridge, MA: MIT Press, 1997.
[13] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajič, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 420–435, Jul. 2004.
[14] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, U.K.: Entropic, 2000.
[15] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovský, and S. Gustman, "Large vocabulary ASR for spontaneous Czech in the MALACH project," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 1821–1824.
[16] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, CO, 2002, pp. 901–904.
[17] P. Ircing, "Large vocabulary continuous speech recognition of highly inflectional language (Czech)," Ph.D. dissertation, Univ. of West Bohemia-Západočeská Univerzita v Plzni, Plzeň, Czech Republic, 2003.

Pavel Ircing received the M.Sc. degree equivalent in cybernetics in 1999 and the Ph.D. degree in cybernetics in 2004, both from the University of West Bohemia, Plzeň, Czech Republic.

He was a Research Assistant at the Department of Cybernetics, University of West Bohemia, from 1999. He is currently an Assistant Professor at the same department. He also worked as the Visiting Scholar at the Johns Hopkins University, Baltimore, MD, in 1999 and 2000. His research interests include language modeling, large-vocabulary ASR, and speech retrieval.

Josef V. Psutka received the M.Sc. degree equivalents in cybernetics in 2001 and in mathematics in 2005, and the Ph.D. degree in cybernetics in 2007, all from the University of West Bohemia, Plzeň, Czech Republic.

He was a Research Assistant in the Department of Cybernetics, University of West Bohemia, from 2001. He is currently an Assistant Professor at the same department. His research interests include mainly speech signal parameterization and acoustic modeling methods for automatic speech recognition.

Josef Psutka received the M.Sc. degree equivalent in electrical engineering and the Ph.D. degree in cybernetics from the Czech Technical University, Prague, Czech Republic, in 1974 and 1980, respectively.

He worked as an Assistant Professor in the Technical Institute, Plzeň, Czech Republic, from 1978 to 1991. In 1991, he joined the Department of Cybernetics, University of West Bohemia, Plzeň, as an Associate Professor, and became a Full Professor in 1997. His research interests include speech signal processing, acoustic modeling, large-vocabulary ASR, speech synthesis, and pattern recognition.