Real-time Large Vocabulary Spontaneous Speech Recognition for Spoken Dialog Systems

Jan Švec
Center of Applied Cybernetics
Department of Cybernetics
University of West Bohemia
Pilsen, Czech Republic

Luboš Šmídl
Center of Applied Cybernetics
Department of Cybernetics
University of West Bohemia
Pilsen, Czech Republic

***Abstract*—This paper describes the method for modifying the baseline speech recognition system to be suitable for a use in spoken dialog system with mixed initiative and natural user's input. We present three approaches for extending the recognition vocabulary to ensure the spoken dialog system is able to recognize all entities in the given domain. The colloquial text normalization method is proposed. The experiments performed on spontaneous speech corpus suggested that the proposed method is very important for languages where the formal written language and a common colloquial speech are very different. The overall word error rate was reduced by 16.7%.**

***Index Terms*—speech recognition, language modeling, spoken dialog system**

I. Introduction

The advances in speech recognition and understanding research allows us to construct a spoken dialog system which is not constrained with the rules describing the expected behavior of the user. Nevertheless a large portion of spoken dialog systems is built using a knowledge based resources such as (probabilistic) context free grammars or predefined lists of commands and menus. In [1] Young presented a statistical framework for building a spoken dialog system without the need of an expert knowledge. It uses statistical methods to estimate a robust model of a given domain. This paper describes the speech recognition system which is built up for use in such a dialog system.

The described methods allows the adaptation of a widely used speech recognition systems based on Hidden Markov models (HMM) and n-gram language models [2] for the use in a dialog system. The adaptation is necessary because training the language model only from transcribed data is not sufficient – there could be entities that are not observed in the training data and therefore the dialog goals, which can be satisfied by recognizing these entities, are unreachable. Therefore in the first part of this paper we explore three methods for extending the recognition vocabulary and language model to contain such entities.

The second part deals with a novel approach for recognizing colloquial speech which is very common in spontaneous dialog and most closely mirrors dialog. The target language of our dialog system is Czech and colloquial Czech substantially differs from standard Czech due the existence of a phenomenon called *diglossia* [3] . Standard Czech is defined by orthographic, morphological, lexical and syntactic rules governed by the Czech language normative bodies and is used in most of Czech written materials as well as in official public speeches, such as TV news, in schools etc. Colloquial Czech is used in unofficial communication, particularly in spontaneous speech. Main problems of colloquial Czech are: pronunciation variants (as found in English and many other languages); changes in morphology; the length of vowels can be shortened or prolonged, depending on the particular word (a distinctive phenomenon in Czech); endings and prefixes are often changed; differences in syntax (less common). The described statistical method converts a sequence of colloquial orthographic words into a sequence of normalized words together with the increasing of the recognition performance. Please note that a form used in colloquial Czech in some context (e.g., a particular case, gender and number) may be equal to some standard form in a different context. This makes automatic mapping from standard forms to colloquial and vice versa in general difficult.

The described methods were experimentally evaluated on Czech data. The impact on recognition performance is described and discussed. Although the methods were developed for Czech, they are not language specific and can be used for any language where colloquial speech is different from formal or where is the need for extending the recognition vocabulary.

II. Speech recognition

A. Acoustic modeling

As a basic speech unit of the recognition system a triphone is used. Each individual triphone is represented by 3 state left-to-right HMM with a continuous output probability density function assigned to each state. Each density is expressed as a mixture of multivariate Gaussians, where each Gaussian has a diagonal covariance matrix. Since a variety of noise sounds, e.g. loud breath, click on the microphone and noise of a telephone channel can appear in an utterance, a set of noise HMM was introduced and trained in order to capture these non-speech events.

The speech data was parameterized as 12-dimensional PLP cepstral features [4] including their delta and delta-delta derivatives (resulting into 36-dimensional feature vectors). These features were computed at a rate of 100 frames per second. Cepstral mean subtraction was applied per

speaker. The resulting triphone-based model was trained using HTK Toolkit [5]. The number of clustered states and number of Gaussians mixtures per state was optimized to get high accuracy and real-time response and had 1500 states and 16 mixtures per state. The state-of-the-art speaker discriminative training algorithms were employed to further improve the quality of the acoustic models [6], [7].

B. Decoding

We used a real-time large vocabulary continuous speech recognizer (LVCSR) to achieve a very high degree of interactivity. The LVCSR system [8] is based on Hidden Markov Models, lexical trees and Viterbi search using n-gram language models.

C. Language modelling

The most often used type of stochastic language models are the n-gram language models which model the probability of i-th word w_i given the last $(n-1)$ words [9]. The last $(n-1)$ words are called the history and we will denote it by $h_i = \{w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}\}$. In this paper we will also use the term back-off history which is defined as $\bar{h}_i = \{w_{i-n+2}, w_{i-n+3}, \ldots, w_{i-1}\}$.

The n-gram language model approximates the probability $P(W)$ with the product of probabilities with a limited history of length $(n-1)$:

$$P(W) \approx \prod_{i=1}^{N} P(w_i|h_i) = \prod_{i=1}^{N} P(w_i|w_{i-n+1}, \ldots, w_{i-1}) \quad (1)$$

The maximum likelihood estimation of $P_{MLE}(w_i|h_i)$ is given by the ratio:

$$P_{MLE}(w_i|h_i) = \frac{C(h_i, w_i)}{C(h_i)} \quad (2)$$

where $C(h_i, w_i)$ is the number of n-gram (h_i, w_i) occurring in the training data and $C(h_i)$ is the number of occurrences of $(n-1)$-gram h_i. In practice the size of the training data is limited and therefore the counts $C(h_i, w_i)$ are zero for a large number of n-grams. Therefore the smoothing techniques are involved. In our experiments we used a back-off smoothing where the n-gram conditional probability is given by a recursive equation:

$$P(w_i|h_i) = \begin{cases} \alpha(h_i, w_i) P_{MLE}(w_i|h_i) & C(h_i, w_i) \neq 0 \\ \beta(h_i, w_i) P(w_i|\bar{h}_i) & C(h_i, w_i) = 0 \end{cases} \quad (3)$$

In other words the conditional probability of an unseen n-gram (h_i, w_i) is approximated with a back-off conditional probability of an $(n-1)$-gram $(\bar{h}_i, w_i)$ and the functions $\alpha(h_i, w_i)$ and $\beta(h_i, w_i)$ ensures $P(w_i|h_i)$ to be a valid probability distribution.

In this work we used the 3-gram back-off language model and the functions $\alpha(h_i, w_i)$, $\beta(h_i, w_i)$ were determined using the Witten-Bell discounting scheme described in more detail in Sec. V-A.

TABLE I
DETAILS OF THE HUMAN-HUMAN TRAIN TIMETABLE CORPUS

	# of dialogs	5443
	# of turns	76k
Normalized	*User*	
	# of tokens	270k
	vocabulary size	5830
	Operator	
	# of tokens	259k
	vocabulary size	3747
Orthographic	*User*	
	# of tokens	279k
	vocabulary size	7737
	Operator	
	# of tokens	262k
	vocabulary size	4696

III. EXPERIMENTAL DATA

We used a unique Czech human-human train timetable (HHTT) corpus containing recordings of phone calls to a train information service [10]. The phone calls consisted of inquiries regarding train connections such as departure and arrival times, ticket prices, details on restrictions on hand luggage etc. The dialog took place between two humans - an operator and a user. The speech was spontaneous and unformal. The corpus is comprised of a recorded speech of both the operator and the user in single channel recordings. In total the corpus consists of 5443 dialogs (74k turns). Each turn starts with a speaker change. The corpus statistics are summarized in Tab. I.

The dialogs were manually processed and the orthographic transcription and normalized transcription were assigned. The orthographic transcription also contains non-speech events such as *inhale*, *hesitation* and *noise*. In addition a semantic annotation process was performed and each turn assigned an abstract semantic tree representing the meaning of the sentence.

Only the user's turns were used for the language modelling and speech recognition. The corpus was divided into train (72% dialogs, 209k normalized tokens, 28k turns), development (8%, 3742 normalized tokens, 775 turns, 23 minutes of speech) and test data (20%, 9629 normalized tokens, 2073 turns, 56 minutes of speech). Overlapping turns in the development and test data were tagged and left out from further speech recognition experiments. Speaker overlaps in the training data are not taken into account since the training data are used only for language modelling. The recognizer parameters (word insertion penalty and language model weight) were optimized on the development data.

IV. BASELINE EXPERIMENT

Two baseline systems consist of an acoustic model described in Sec. II-A and a 3-gram back-off language model with Witten-Bell discounting trained from normalized respective orthographic transcriptions and evaluated on the corresponding test data.

We used standard *correctness* ($Corr$) and *word error rate* (WER) measures [5] defined as:

TABLE II
BASELINE ASR RESULTS

Transcription	$\|V\|$	PP	OOV [%]	$Corr$ [%]	WER [%]
Normalized	5127	44.3	1.19	67.74	37.79
Orthographic	6768	57.2	1.62	69.94	34.63

$$Corr = \frac{H}{N} \cdot 100\% \tag{4}$$

$$WER = \frac{S + D + I}{N} \cdot 100\% \tag{5}$$

where N is the total number of words in reference transcript, H is the number of correctly recognized words, S, D, I denotes the number of substitution, deletion and insertion errors. In addition we evaluated *perplexity* (PP) and *out-of-vocabulary rate* (OOV) [11] of the language model evaluated on the development data:

$$PP = 2^{-\frac{1}{N}\log_2 P(W)} \tag{6}$$

$$OOV = \frac{O}{N} \cdot 100\% \tag{7}$$

where $P(W)$ is a language probability assigned by a language model and O is the number of tokens in the development data which are not included in the recognition vocabulary V [11]. We also present the size of the recognition vocabulary ($|V|$) since this parameter also influences the recognition performance.

Table II shows the results of a baseline system trained from orthographic and normalized transcriptions. The performance of the normalized language model evaluated on the normalized data is worse than the performance of an orthographic language model evaluated on orthographic transcriptions. We suppose that the orthographic vocabulary better models the colloquial speech which can be found in the test data.

In addition the rather small vocabulary of the baseline system does not include all the words necessary to successfully understand the sentence and correctly satisfy the goal of the dialog. Therefore we have explored three methods for adding new words into a recognition vocabulary (Sec. V). In addition our analysis of the results of an orthographic-based recognition system shows that there are many recognition errors caused by the substitution of one orthographic word with a similar word and both of them can be mapped to one normalized word (eg. *prosi* and *prosim* and corresponding normalized word *prosím*, lit. *please*). To remove these ambiguities and improve an overall performance we designed a method for colloquial text normalization (Sec. VI).

V. EXTENDED RECOGNITION VOCABULARY

The intended used of the described ASR system is a spoken dialog system. The very common need in this application of speech recognition is to extend a language model with new words which represents entities generated from some list or from a database. The domain of the described ASR system is an information service providing the times of departure and arrival for trains in the Czech Republic. Therefore there is a need to recognize all the names of railway stations in the country. In the Czech Republic there are about 2800 railway stations with the name mainly in the form of the name of the city (eg. *Klatovy*) or composed by the name of a city and a district (eg. *Plzeň-Koterov*). Sometimes the name of the station is composed of the name of the city and the name of the station (eg. *Praha-Masarykovo nádraží*). The name of the station can be shortened and often occurs in a colloquial form (eg. *Masarykovo nádraží*, *Masarykáč*). In addition the station name can be in four different grammatical cases: *nominative*, *genitive*, *accusative*, *locative*. Note that Czech has seven cases and in general the case is not determined by the word form but also by its context (e.g. by the preposition).

Because the training data for the language model was collected only in one regional call center, the distribution of station names is different from the general distribution of these names in a language. There are many out-of-vocabulary words because many destinations are not mentioned in a finite number of dialogs. To cover the whole domain of the dialog we developed a set of rules which takes the station name in nominative and generates shortened and colloquial forms in all four cases. Although the names of railway stations are unique within a whole country, the shortened and colloquial forms generally are not. The generated lists of station names in a given grammatical case were used to enrich the recognition vocabulary and the language model. We have explored three methods:

- Using a discounting method for assigning a non-zero probability to the new words.
- Estimating an open vocabulary language model (LM which models the probability of unseen word) and distributing the probability mass of the unseen word between the new words.
- Training a class-based language model and estimating the class member probabilities.

While the first method only modifies the recognition vocabulary and a unigram conditional probability, the last two methods also take a history of an n-gram into account and modifies both the trigram and the bigram conditional probabilities.

A. Discounted language model

The basic and straightforward method for adding new words into the vocabulary of the language model is the use of smoothing method for assigning a non-zero probability to an unseen event – in this case to a new word. We use the *Witten-Bell discounting scheme* [12]. This smoothing method employs the number $c(h)$ which equals to the number of different words following the history h. The smoothed probability of the word w_i given its history h_i is then given by:

$$P(w_i|h_i) = \begin{cases} \frac{C(h_i,w_i)}{C(h_i)+c(h_i)} & C(h_i,w_i) > 0 \\ \frac{c(h_i)}{C(h_i)+c(h_i)} \cdot \frac{P(w_i|\bar{h}_i)}{\sum_{w \in S_i} P(w|\bar{h}_i)} & C(h_i,w_i) = 0 \end{cases} \tag{8}$$

where the set $S_i = \{w : C(h_i, w) = 0\}$ is the set of words with the history h_i not occurring in the training data and the generalized distribution $P(w_i|\bar{h}_i)$ is a probability of word w_i given the shortened history $\bar{h}_i$.

The principle of this method is to smooth the unigram probability of the word $P(w_i)$ so that the probability of a new word is non-zero. Define the set of new words V' and the original recognition vocabulary V. We suppose that $V \cap V' = \emptyset$. Therefore $P(w'_i) = 0$ for every $w'_i \in V'$. Using the general Witten-Bell equation (Eq. 8) for the unigram probability $P(w_i)$ (the history $h_i = \emptyset$) we can derive the smoothed unigram probability $P'(w_i)$:

$$P'(w_i) = \begin{cases} \frac{C(w_i)}{|T|+|V|} & w_i \in V \\ \frac{1}{|V'|} \cdot \frac{|V|}{|T|+|V|} & w_i \in V' \end{cases} \quad (9)$$

where $|T|$ is the number of tokens (running words) in the training data.

B. Open vocabulary language model

This method for adding new words to the recognition vocabulary uses an open vocabulary language model. This type of model uses a special symbol for modelling the unknown previously unseen words. In this paper we will use the unk symbol representing an unknown word [11]. The main issue of this method is the definition of unknown words to ensure their occurrence in the training data. This can be achieved by using a vocabulary which is not a superset of training data. One of the possible approaches uses a vocabulary estimated from an independent data set and then uses it during the language modelling from the training data set and marks all words outside this vocabulary as unknown words. Another approach limits the vocabulary estimated from training data so that only words occurring more than k-times ($k > 1$) are included in the vocabulary. Words with unigram counts $C(w) \leq k$ are marked as unknown words.

With respect to the size of development data we decided to use the second approach with $k = 1$. The procedure for adding new words then consists of marking all singleton words (words occurring just once in the training data) as unknown words and replacing them with the unk symbol. Then the singleton words are added into a set of new words V' (again $V \cap V' = \emptyset$). The probability $P(w_i|h_i)$ for $w_i \neq unk, w_i \in V$ stays unchanged and the probability mass of the unk symbol is uniformly distributed between the words in V' given the $P(w'_i|unk), w'_i \in V'$:

$$P(w'_i|h_i) = \frac{1}{|V'|} \cdot P(unk|h_i), w'_i \in V' \quad (10)$$

This definition of an extended vocabulary and language model also models the probability of a new word given its history h_i.

C. Class-based language model

This method is based on an expert knowledge – the known structure of a given entity in the language. In our experiments these knowledge is represented by the list of known station names and their variants.

TABLE III
DEFINITION OF CLASSES

Class	$C_N(c_k)$	$C_O(c_k)$	$\|c_k\|$	*Prep.*	*Gr. case*
c_the	2106	2064	4348	-	nominative
c_from	2988	2929	12132	z, ze	genitive
c_to	3816	3704	6066	do	genitive
c_toward	761	749	4946	na	accusative
c_in	1302	1232	12390	v, ve	locative

$C_N(c_k)$ and $C_O(c_k)$ denotes the number of occurrences of class c_k in the normalized transcription data respectively in the orthographic transcription data. The number of members of class c_k is represented by $|c_k|$. The column Prep. *contains the assigned preposition and* Gr. case *contains the grammar case of the given members.*

Since the training data consists of orthographic or normalized transcriptions, we have to detect the class occurrences in the data. This is ensured by an algorithm which takes the list of members of the given class and replaces it with a class identifier. To avoid class definition with overlapping members we also included the preposition associated with a given grammar case. The definition of classes is summarized in Tab. III. The class members can consist of more than one word therefore the replacement rules are applied with the priority determined by the number of words the rule replaces. For example the rule *do Sušice* → *c_to* is applied before the rule *Sušice* → *c_the* because the former replaces two words and the latter only one word.

After replacing the occurrences of class members with the class identifiers the standard 3-gram back-off language model is trained. Then during the recognition the class members are used instead of the class identifier and the probability of the i-th member m_{ik} of class c_k is given by the distribution $P(m_{ik}|c_k)$ [13]:

$$P(m_{ik}|h_i) = P(m_{ik}|c_k)P(c_k|h_i) \quad (11)$$

where $P(c_k|h_i)$ is a n-gram conditional probability of a class occurring in a context h_i. The probability distribution $P(m_{ik}|c_k)$ cannot be estimated from data because only small portion of class members occurs in the training data (see columns $C_N(c_k)$ and $C_O(c_k)$ in comparison with $|c_k|$ in Tab. III). Since the class members are related to railway station names and these names are related to the names of towns we used the number of citizens of the nearest town to weight the members of the same class – the stations in cities with higher number of citizens gains higher probability than the stations in smaller towns. The probability distribution $P(m_{ik}|c_k)$ was heuristically determined to minimize the perplexity of the language model on the development data. In our experiments we use:

$$P(m_{ik}|c_k) = \frac{\sqrt{z(i)}}{\sum_j \sqrt{z(j)}} \quad (12)$$

where $z(i)$ denotes the number of citizens of the town which is the nearest to the railway station represented by words m_{ik}.

TABLE IV
EXTENDED VOCABULARY ASR RESULTS

Method	$\|V\|$	PP	OOV	$Corr$	WER
Normalized transcriptions					
Discounted LM	11595	45.6	1.01	66.63	38.01
Open vocabulary LM	11595	46.9	1.01	67.11	37.89
Class-based LM	11595	54.0	1.01	65.49	39.81
Open vocabulary, $V' = \emptyset$	2595	41.8	1.81	67.01	37.95
Orthographic transcriptions					
Discounted LM	13211	56.5	1.43	68.88	35.02
Open vocabulary LM	13211	58.8	1.43	70.59	34.40
Class-based LM	13211	65.7	1.43	68.61	36.09
Open vocabulary, $V' = \emptyset$	3197	50.7	2.41	68.33	35.49

D. Results

The results of the three described methods are shown in Tab. IV. The ASR systems with modified language model were evaluated on both the normalized and orthographic transcriptions. In general the results of normalized language model are worse then the results of orthographic language model although the sizes of recognition vocabulary, the out-of-vocabulary rates and perplexities are larger. This conclusion follows the results of the baseline experiment described in Sec. IV. The language model with the open vocabulary has a slightly better performance than the language model based only on discounting, and the open vocabulary language model's performance is comparable or rather better than the baseline despite the much larger vocabulary. The enriched vocabulary is approximately more than twice as large as the baseline vocabulary.

Note that the language model based on classes has the worse performance of the presented methods. Its performance is also significantly lower than the baseline. This is caused by the effect described in Sec. V – the corpus used in evaluation (both the training and test data) was collected in one regional call center. Since the first two methods only add new words into the recognition vocabulary, the probability of known words remaining unchanged is high allowing the corresponding language model to better predict words during recognition.

The class-based language model changes also the probabilities of known words (class members) according the Eq. 12. It is also shown in the column PP of Tab. IV. The perplexity of class-based language model is significantly higher than the perplexity of the other two methods. The effect of class-based language model could be evaluated on a data with a "global" distribution of station names. This data will be collected during the operation of the spoken dialog system. The adaptation to a new class member distribution is possible only in the class-based language model.

VI. COLLOQUIAL TEXT NORMALIZATION

The result from the baseline experiment described in Sec. IV is that the performance of an ASR system with a language model trained from normalized transcriptions is worse than the performance of language model trained from orthographic transcriptions. The difference is about 3% absolute. The normalized output of an ASR system is important for natural language understanding because the vocabulary is smaller and the speech understanding is more robust. Therefore we explored a novel method for post-processing the orthographic ASR result and generate a normalized output which simplifies the understanding module and also increases the performance of an ASR system.

Let's have an output of an ASR system trained from orthographic transcriptions. Define the sequence of words as $W_O = \{w_{O,1}, w_{O,2}, \dots w_{On}\}, w_{O,i} \in V_O$. The goal of colloquial text normalization is to generate the sequence of normalized words $W_N = \{w_{N,1}, w_{N,2}, \dots w_{N,m}\}, w_{N,i} \in V_N$ where $n \geq m$. The normalized sequence of words is shorter because the colloquial form contains a large number of word fragments, repetitions and non-speech events which cannot be mapped to any of the normalized words. The normalization is context dependent. For example the Czech colloquial word *sem* can have two different normalized forms and two different meanings – the first one is *jsem* (lit. *I am*) and the second one is *sem* (lit. *here*). Therefore we used a noisy channel model described by the following equation:

$$P(W_N|W_O) = \frac{P(W_O|W_N) \cdot P(W_N)}{P(W_O)} \tag{13}$$

where $P(W_N)$ is a normalized language model, $P(W_O)$ is a orthographic language model which normalizes the product $P(W_O|W_N) \cdot P(W_N)$. The conditional probability $P(W_O|W_N)$ describes the probability of observing orthographic words W_O given some normalized words W_N. For a given W_O we can compute the normalized sequence of words W_N^* using a MAP criterion:

$$W_N^* = \arg\max_{W_N} P(W_O|W_N) \cdot P(W_N) \tag{14}$$

The normalized language model $P(W_N)$ is the standard trigram back-off language model used in a baseline experiment. To model the conditional distribution $P(W_O|W_N)$ we first introduce the sequence $W'_N = \{w'_{N,1}, \dots, w'_{N,n}\}$ which has the same number of elements as W_O and the elements are from the set $V'_N = V_N \cup \{\epsilon\}$. In addition W'_N satisfies:

$$W_N = \{w'_i : w'_i \in W'_N, i = 1, 2, \dots n; w'_i \neq \epsilon\} \tag{15}$$

The orthographic words which do not have counterparts in a normalized word sequence are mapped to the ϵ symbol. We used the following assumptions: the language probabilities of W'_N and W_N are equal: $P(W'_N) = P(W_N)$ and the probability of observing $w_{O,i}$ is conditioned only by $w_{N,i}$. Then we can use the following model:

$$W_N^* = \arg\max_{W_N} P(W_O|W'_N) \cdot P(W_N) \tag{16}$$

$$P(W_O|W'_N) \approx \prod_{i=1}^{n} P(w_{O,i}|w'_{N,i}) \tag{17}$$

TABLE V
ASR RESULTS WITH COLLOQUIAL TEXT NORMALIZATION

Method	$Corr$	WER
Baseline results		
Normalized LM	67.74	**37.79**
Orthographic LM	69.94	34.63
Colloquial text normalization		
Orthographic LM, normalized	74.15	31.59
Discounted LM, normalized	73.95	**31.47**
Open vocabulary LM, normalized	74.06	31.84
Class-based LM, normalized	72.58	32.98

The conditional probability $P(w_O|w'_N)$ is estimated from training data. First of all the dynamic programming based on Levenshtein distance [14] is used to align the orthographic and normalized transcriptions and the list of confusions $R = \{(w_{O,i}, w'_{N,i}) : i = 1, \dots |T|\}$ is generated. If an orthographic word $w_{O,i}$ cannot be matched with a normalized word $w'_{N,i}$ then we assign $w'_{N,i} = \epsilon$. The conditional probability is then given by MLE:

$$P(w_O)|w'_N) = \frac{C(w_O, w'_N)}{\sum_{w_O} C(w_O, w'_N)} \tag{18}$$

where $C(w_O, w'_N)$ is the number of occurrences of the tuple (w_O, w'_N) in the confusion list R.

A. Results

The method for colloquial text normalization described above was applied on ASR results in orthographic form both during the evaluation on the development data and during the final test data evaluation. In total 89% orthographic words can be directly mapped to a normalized word (mostly the forms are the same). Each of the remaining 11% of orthographic words were assigned on average 2.88 normalized words. We applied the colloquial text normalization on both the baseline results (Tab. II) and the results with extended recognition vocabulary (Tab IV).

The table shows that the colloquial text normalization significantly improves the recognition performance. The correctness increased by 6.4% and the word error rate decreased by 6.3%. The relative decrease in word error rate is 16.7%.

VII. CONCLUSION

We have presented two ways of modifying the baseline speech recognition system built up to use in a spoken dialog system with mixed initiative and natural user's input [15]. The first one describes three approaches for extending the recognition vocabulary to ensure the spoken dialog system is able to recognize all entities in the given domain. The second one presents the colloquial text normalization method. The normalization is very important for speech recognition in languages where the formal written language and a common colloquial speech are very different. The output of the colloquial text normalization is suitable for further processing in a spoken language understanding module. The use of normalized text improves the robustness of the understanding process and simplifies the semantic interpretation because the normalized vocabulary is smaller than the orthographic vocabulary. The combination of the presented methods enriches the recognition vocabulary to be more than twice large and at the same time it reduces the overall word error rate by 16.7% measured on normalized transcriptions.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education of the Czech Republic under project No. 1M0567 (CAK) and by the grant of the University of West Bohemia, project No. SGS-2010-054. The access to the MetaCentrum computing facilities provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is greatly appreciated.

REFERENCES

[1] S. Young, "Talking to machines (statistically speaking)," in *Seventh International Conference on Spoken Language Processing*. ISCA, 2002, pp. 9–16.

[2] F. Jelinek, *Statistical methods for speech recognition*, ser. Language, speech, and communication. MIT Press, 1997.

[3] J. Psutka, P. Ircing, J. Hajič, V. Radová, J. Psutka, W. Byrne, and S. Gustman, "Issues in annotation of the Czech spontaneous speech corpus in the MALACH project," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004.

[4] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, APR 1990.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Entropic, 2000.

[6] D. Povey, "Discriminative training for large vocabulary speech recognition." Ph.D. dissertation, University of Cambridge, Cambridge, United Kingdom, 2003.

[7] J. Vaněk, J. Psutka, J. Zelinka, A. Pražák, and J. Psutka, "Discriminative training of gender-dependent acoustic models," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, V. Matoušek and P. Mautner, Eds., vol. 5729. Springer Berlin / Heidelberg, 2009, pp. 331–338.

[8] A. Pražák, P. Ircing, J. Švec, and J. Psutka, "Efficient combination of n-gram language models and recognition grammars in real-time lvcsr decoder," in *9th International Conference on Signal Processing, 2008. ICSP 2008.*, oct. 2008, pp. 587 –591.

[9] D. Jurafsky and J. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, ser. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2009.

[10] F. Jurčíček, J. Zahradil, and L. Jelinek, "A Human-Human Train Timetable Dialogue Corpus," in *Proceedings of EUROSPEECH*, Lisboa, Portugal, 2005.

[11] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Proceedings of the international conference on spoken language processing*, vol. 2, 2002, pp. 901–904.

[12] I. Witten and T. Bell, "The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1085 –1094, jul 1991.

[13] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001, foreword By-Reddy, Raj.

[14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Cybernetics and Control Theory*, vol. 10, no. 8, pp. 707–710, 1966, original in *Doklady Akademii Nauk SSSR* 163(4): 845–848 (1965).

[15] J. Švec and L. Šmídl, "Prototype of czech spoken dialog system with mixed initiative for railway information service," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer Berlin / Heidelberg, 2010, vol. 6231, pp. 568–575.