ISSN 0005-1144
ATKAFF 52(2), 147–157(2011)

Sanda Martinčić-Ipšić, Miran Pobar, Ivo Ipšić

Croatian Large Vocabulary Automatic Speech Recognition

UDK 004.934'1:811.163.42
IFAC 5.9.3

Original scientific paper

This paper presents procedures used for development of a Croatian large vocabulary automatic speech recognition system (LVASR). The proposed acoustic model is based on context-dependent triphone hidden Markov models and Croatian phonetic rules. Different acoustic and language models, developed using a large collection of Croatian speech, are discussed and compared. The paper proposes the best feature vectors and acoustic modeling procedures using which lowest word error rates for Croatian speech are achieved. In addition, Croatian language modeling procedures are evaluated and adopted for speaker independent spontaneous speech recognition. Presented experiments and results show that the proposed approach for automatic speech recognition using context-dependent acoustic modeling based on Croatian phonetic rules and a parameter tying procedure can be used for efficient Croatian large vocabulary speech recognition with word error rates below 5%.

Key words: Acoustic modeling, Automatic speech recognition, Context-dependent acoustic units, Language modeling

Automatsko raspoznavanje hrvatskoga govora velikoga vokabulara. Članak prikazuje postupke akustičkog i jezičnog modeliranja sustava za automatsko raspoznavanje hrvatskoga govora velikoga vokabulara. Predloženi akustički modeli su zasnovani na kontekstno-ovisnim skrivenim Markovljevim modelima trifona i hrvatskim fonetskim pravilima. Na hrvatskome govoru prikupljenom u korpusu su ocjenjeni i uspoređeni različiti akustički i jezični modeli. U članku su uspoređeni i predloženi postupci za izračun vektora značajki za akustičko modeliranje kao i sam pristup akustičkome modeliranju hrvatskoga govora s kojim je postignuta najmanja mjera pogrešno raspoznatih riječi. Predstavljeni su rezultati raspoznavanja spontanog hrvatskog govora neovisni o govorniku. Postignuti rezultati eksperimenata s mjerom pogreške ispod 5% ukazuju na primjerenost predloženih postupaka za automatsko raspoznavanje hrvatskoga govora velikoga vokabulara pomoću vezanih kontekstno-ovisnih akustičkih modela na osnovu hrvatskih fonetskih pravila.

Ključne riječi: akustičko modeliranje, automatsko raspoznavanje govora, kontekstno-ovisne jedinice, jezično modeliranje

1 INTRODUCTION AND RELATED WORK

Speech is the most natural and effective way people communicate. Thus, one of the challenges facing researchers is building natural conversational systems, which enable people to communicate with machines in a natural way, without a keyboard, mouse or monitor. One of the goals of such conversational systems is automatic speech recognition which has to recognize spoken words represented by a stream of input feature vectors calculated from the acoustic signal and transform the spoken words into text.

Speech recognition systems are mainly based on data-driven statistical approaches [1,2]. Statistical pattern recognition and segmentation algorithms and methods for stochastic modeling of large speech quantities are used. The statistical approach uses hidden Markov models (HMM) as state of the art formalism for speech recognition. Many large vocabulary automatic speech recognition (LVASR) systems use mel-cepstral speech analysis, hidden Markov modeling of acoustic subword units, n-gram language models (LM) and n-best search of word hypothesis [1,3,4,5]. Automatic speech recognition research in languages like English, German and Japanese [6] puts its focus on recognition of spontaneous and broadcast speech. For highly flective Slavic languages the research focus is more narrowed down, mainly due to the lack of speech resources like corpuses. Large or limited vocabulary speech recognition for Slovene [7], Czech [8,9], Slovak [8,10] with applications in spoken dialog systems [7] or dictation [9] have been reported. Some interest in the research and development of speech applications for Croatian can be noticed over the past decades. The speech translation system DIPLOMAT between Serbian and Croatian on the one side and English on the other is reported in [11]. The TONGUES project continued with this research in a direction towards large Croatian vocabulary recognition system [12]. The Serbian, Bosnian and Croatian dictation system trained on broadcast news is discussed in [13]. A work on the issues of acoustic modeling of Croatian speech has recently been reported in [14].

Acoustic and language models for the Croatian LVASR are proposed in this paper. Linear prediction and cepstral-based feature vectors for Croatian LVASR are compared in terms of speech recognition results and the best speech feature vectors are proposed for Croatian acoustic modeling. The acoustic model is based on context-dependent subunits (triphones) and Croatian acoustic rules used for better parameter estimation. The developed Croatian speech corpus includes speech recorded from different sources and from different channels. The presented results of speech recognition experiments show that the collected speech is sufficient to estimate reliable parameters of the HMM-based acoustic model which can be used for speaker independent speech recognition applications. Furthermore, different Croatian language models are presented. The experimental results encourage the use of the proposed LVASR system for spontaneous Croatian speech applications like spoken dialog systems.

As the main resource in speech technology development is the collection of speech material, the Croatian speech corpus is presented in Section 2. Furthermore, the acoustic modeling procedures of the speech recognition system including phonetically driven state tying procedures as well as language modeling methods are provided in Section 3. Conducted speech recognition experiments and large vocabulary speech recognition results are presented in the Section 4. The paper concludes with a discussion on the advantages of presented work for activities in Croatian speech technology's future research.

2 THE CROATIAN SPEECH CORPUS

For the purpose of this work, the Croatian speech corpus VEPRAD [16] which includes news, weather forecasts and reports spoken within the national radio broadcast shows was extended with read tales and spontaneous speech. The used speech material is divided into several groups: weather forecasts read by professional speakers within national radio news, daily news read by professional speakers, stories read and partially acted out by a professional speaker and finally spontaneous speech uttered by non-professional speakers.

The used speech corpus is a multi-speaker speech database and contains 15 hours of transcribed speech spoken in the studio acoustical environment and 1 hour of spontaneous speech. Each spoken utterance has its word level textual transcription. The corpus statistics are displayed in Table 1.

The first two parts of the corpus consist of transcribed weather forecasts and news recorded from the national radio programmes. This is a multi-speaker database, which contains speech utterances of 11 male and 14 female professional speakers with 9431 utterances and duration of 13 hours. The transcribed sentences contain 183000 words, where 10227 words are different. A relatively small number of 1462 different words in the weather forecast domain show that this part of speech database is strictly domain-oriented.

The third part of the corpus contains 10 Grimm's fairy tales read and partially acted out by the same male speaker whose speech was collected in the weather and news part. The tales part contains 2532 utterances lasting over 2 hours and containing 18984 spoken words, where 5268 words are different.

The last part of the corpus contains spontaneous speech form 17 male and 17 female speakers, mostly students from the Department of Informatics. Each speaker uttered 45 simple questions and answers simulating the weather domain-related dialog.

This corpus has been developed over the past 8 years as the research infrastructure for Croatian speech technologies: speech recognition [28], text-to-speech synthesis [17,18] and spoken dialog systems [19]. The presented corpus has been intentionally developed for acoustic modeling of Croatian speech with the following characteristics: 16 hours of speech, over 208K uttered words, 15K of different words in the dictionary and almost 60 male and female speakers. The phonetic dictionary comprises all words in transcription texts and their phonetic transcription based on the phonological-morphological principle which enables automation of Croatian phonetic transcription [29].

Table 1. Croatian speech corpus statistics

	Number		**Speakers**		**Words**		**Dur.**
	Rec.	***Utt.***	***M***	***F***	***All***	***Diff.***	***min***
Radio weather for.	1057	5456	11	14	77322	1462	482
Radio news	237	3975	1	2	105678	9923	294
Read tales	10	2532	1		18984	5268	121
Weather dialogs	34	1530	17	17	6664	78	56
OVERALL	**1338**	**13493**	**28**	**31**	**208648**	**14551**	**953**

3 AUTOMATIC SPEECH RECOGNITION

The goal of automatic speech recognition procedures is to recognize the spoken words represented by a stream of input feature vectors calculated from the acoustic signal. The major problems in continuous speech recognition arise due to the nature of spoken language: there are no clear boundaries between words, the phonetic beginning and ending are influenced by neighbouring words, there is a great variability in different speakers' speech: male or female, fast or slow speaking rate, loud or whispered speech, read or spontaneous, emotional or formal and the speech signal can be affected with noise. Further, spontaneous speech is often ill-structured and ungrammatical and usually produced in degraded acoustic environments.

To avoid these difficulties the data-driven statistical approach based on large quantities of spoken data is used [6]. Statistical pattern recognition and segmentation algorithms and methods for stochastic modeling of time varying speech signals are used. The data-driven statistical approach uses hidden Markov models (HMM) as state of the art formalism for automatic speech recognition [2]. Hidden Markov models are stochastic finite-state automata consisting of a finite set of states and state transitions where HMMs can simulate the human speech production system. The state sequence is hidden, but in each state according to the output probability function an output observation can be produced. Output observations are presented by probability distributions of speech feature vectors which are modeled with mixtures of Gaussian distributions.

For the estimation of continuous HMM parameters, the iterative Baum-Welch procedure is used. The Baum-Welch algorithm, also known as the Forward-Backward algorithm iteratively refines the HMM parameters by maximizing the likelihood of a speech signal feature sequence given an HMM. The algorithm is based on the optimization technique used in the Expectation-Maximization (EM) algorithm for estimation of Gaussian mixture densities parameters [1,20]. For the search of an optimal path through the HMM network of acoustic models the Viterbi algorithm is used [1,2]. The Viterbi algorithm is a dynamic programming algorithm that decodes the state sequence according to the observed output sequence. Due to the large number of states which have to be considered when computing the Viterbi alignment, a state pruning technique has to be used to reduce the size of the search space. The Viterbi beam-search technique which expands the search only to states whose probability falls within a specified beam was used. The probability of reaching a state in the search procedure cannot fall short of the maximum probability by more than a predefined ratio. During the forward search in the HMM n-best word sequences are generated using acoustic models and a bigram language model.

3.1 Acoustic Modeling

An acoustic model should represent all possible variations in speech. Speech variations can be caused by a speaker's characteristics, co-articulation, surrounding acoustic conditions, channel etc. Therefore selection of an appropriate acoustic unit, which can capture all speech variations, is crucial for acoustic modeling. Enough acoustic material should be available for acoustic unit HMMs training. At the same time the chosen acoustic unit should enable construction of more complex units, like words [21]. In continuous speech recognition systems the set of acoustic units is modeled by a set of HMMs. Since the number of units is limited (by the available speech data) usually subword acoustic units are modeled. The subword units are as follows: monophones, biphones, triphones, quinphones [22,23] or subphonetic units like senones [24]. All these units enable the construction of more complex units and recognition of unseen units which are not included in the training procedure.

3.2 Context-ndependent Acoustic Model

The training of speech recognition acoustic models started with defining the Croatian phoneme set according to SAMPA. For each of the 30 Croatian phonemes a context-independent left-to-right (L-R) monophone hidden Markov model was defined. Each monophone model consists of 5 states, where the first and last state have no output functions as shown in Fig. 1. The transition probabilities from state$s_i$to state$s_j$are noted asa_{ij}, and output vector probabilities are $b_j(x_t)$. A monophone model is presented for the phoneme /h/.

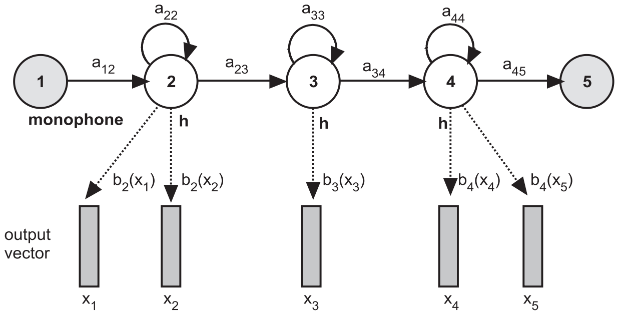


Fig. 1. The monophone L-R HMM with 5 states

Initially, the monophone models with continuous Gaussian output probability functions described with diagonal covariance matrices were trained. The initial training of the Baum-Welch algorithm on HMM monophone models resulted in a monophone recognition system, which was used for the automatic segmentation of the speech signals. The automatic segmentation of the speech signal from the word to the phone level is performed using the forced alignment [16] of the spoken utterance and the corresponding word level transcriptions. The results of automatic segmentation are estimated time intervals for each phone. Furthermore, the monophone models were trained by 10 passes of the

Baum-Welch algorithm and the resulted monophone models were used for the initialization of context-dependent triphone hidden Markov models. The number of mixtures of output Gaussian probability density functions per state was increased up to 20 in the used monophone recognition system.

3.3 Context-dependent Acoustic Model

The triphone context-dependent acoustic units were chosen due to the quantity of available speech and possibility for modeling both the left and right coarticulation context of each phoneme. Context-dependent cross-word triphone models with continuous density output functions (up to 20 mixture Gaussian density functions), described with diagonal covariance matrices were trained for Croatian LVASR. The triphone HMMs also consist of 5 states, where the first and last states have no output functions.

The state tying procedure proposed in [25] allows classification of unseen triphones in the test data into phonetic classes by tying the parameters of each phonetic class. The phonetic rules describe the classes of the phonemes according to their linguistic, articulatory and acoustic characteristics and were used for the construction of a phonetic decision tree [29]. A phonetic tree is a binary tree where the phoneme's left or right phonetic context is investigated in each node. The phonemes are thus clustered into phonetic classes according to the phonetic rules which examine the phoneme's left and right context. State tying enables clustering of the states that are acoustically similar, which allows all the data associated with one state to be used for more robust estimation of the model parameters (mean and variance). This enables more accurate estimation of output probabilities of the Gaussian mixtures and consequently better handling of the unseen triphones.

3.4 Language Modeling

The language model is an important part of the speech recognition system. A bigram statistical language model was used in this work [5]. As the weather domain corpus contains a limited amount of sentences a bigram language model is used to approximate the probability of the word sequence $P(W)$. The probability $P(w_i/w_{i-1})$ of the word$w_i$after wordw_{i-1} in a bigram language model is calculated by the fraction of the frequency of word pairs (w_{i-1}, w_i) with the frequency of the word w_{i-1}.

Discounting and smoothing techniques are well-known methods which cope with low and zero n-gram frequencies. The smoothing techniques try to adjust the maximum likelihood estimate of probabilities to produce more accurate probabilities. These techniques adjust low probabilities such as zero probabilities upward, and high probabilities downward. Not only do smoothing methods generally prevent zero probabilities, but they also attempt to improve the accuracy of the model as a whole [26]. In discounting techniques, when there is not enough data for estimating an n-gram probability, a better estimate can be provided by the corresponding back-off (n-1)-gram. A simple method of combining the information from lower-order n-gram models in estimating higher-order probabilities is linear interpolation, and a general class of interpolated models is described by Good-Turing discounting [26].

In practise Good-Turing smoothing is combined with some other methods which combine higher order n-grams with lower order n-grams. Katz proposed additional back-off smoothing scheme [5] where an n-gram estimate is combined with (n-1)-gram estimates in cases when n-gram is not present or not observed frequently enough in the training data.

4 CROATIAN LVASR EXPERIMENTS AND RESULTS

Different acoustic and language models for Croatian LVASR systems have been evaluated using the speech corpus presented in the Section 2 and modeling procedures described in the Section 3. In the first part of this Section different speech features vectors are evaluated on the weather part of the speech corpus in terms of word error rates of developed ASR. The central part considers different aspects of acoustic modeling in Croatian ASR, the third part presents the use of developed ASR for spontaneous speech in spoken dialog applications and the final part compares language models for Croatian LVASR.

The training of the ASR system was performed using the HTK toolkit [15]. The HTK toolkit, developed at the Cambridge University Engineering Department, is a general-purpose toolkit for modeling time series. It is widely used for building HMM based acoustic speech models and statistical language models for automatic speech recognition.

In the Croatian LVASR development the following steps were performed: context-independent (monophone) and context-dependent (cross-words triphone) acoustic models were trained and bigram language models were built. Initially, the speech signal is parameterized with feature vectors and their first and second order differences. Speech transcriptions and speech signal feature vectors are used to train parameters of the monophone HMMs. The automatic segmentation of the word-level to the phone-level for each utterance in corpus was performed using the monophone HMMs. The results of automatic segmentation were used for training (estimating) the parameters of monophone HMMs by repeating the Baum-Welch re-estimation procedure; the transition and output probability distribution parameters at states of HMM models. The training procedure is repeated for each increase of the Gaussian mixture component.

In the next step cross-words context-dependent triphones are constructed from monophones in a way that each triphone has in the left and in the right context the preceding and the succeeding phone even across neighbouring words. The triphone HMMs are constructed from monophone HMMs and the parameters are estimated with the Baum-Welch procedure. The triphone states with estimated parameter values are tied according to the proposed Croatian phonetic rules. The state tying procedure ensures enough acoustic material to train all context-dependent HMMs and enables acoustic modeling of unseen acoustic units that are not present in the training data. The parameters of tied triphone HMMs are estimated by repeating the Baum-Welch re-estimation procedure and by increasing the number of Gaussian mixtures. The prepared textual transcriptions of speech utterances and phonetic dictionary are used to build a back-off bigram language model. The triphone HMMs and bigram language model are used for large vocabulary Croatian speech recognition.

4.1 Speech Feature Vectors

Acoustic models for three different kinds of speech parameterisation which are most commonly used in ASR systems were compared: Linear Prediction Coefficients (LPC), Mel-Cepstral Coefficients (MFCC) and Perceptual Linear Prediction Coefficients (PLP) [1,27].

Linear prediction (LPC) model of speech signal uses a linear combination of previous samples of a signal to predict the current sample. For the purposes of the ASR system training, speech was modeled using an LPC filter of order 12.

The Mel-frequency Cepstrum Coefficients (MFCC) are derived from speech signal by firstly obtaining the amplitude spectra of a windowed short-time signal and then transformed using non-linear mel frequency band pass filters which approximate the characteristics of the human hearing. In tested ASR system the first 12 mel-frequency cepstral coefficients and the zeroth coefficient, and their first and second order differences (delta and delta-delta) were used. The feature coefficients were computed using a Hamming window every 10 ms for a speech signal frame length of 20 ms.

The perceptual linear prediction (PLP) coefficient representation is obtained using the same basic algorithm as in LPC, only the coefficients are not computed in the time domain, but on a perceptually motivated mel frequency scale. Feature vectors used in the ASR system comprise first 12 PLP coefficients, their first and second order differences.

For the speech feature comparison experiments only the radio weather part of the speech from the corpus was used. The 30 standard monophone models and 4 additional models for special acoustic events like pause, breathing etc. were trained using 4135 speech utterances from 8 male and 8 female speakers (71%) and the rest of 1710 utterances from 3 male and 6 female speakers (29%) were used for testing. The influence of LPC, PLP and MFCC speech features vectors in terms of Word Error Rates (WER) for different number (1, 10 and 20) of mixtures of Gaussian output probability density function is given in Table 2. WER is computed according to:

$$WER = 100\% \left(\frac{W_S + W_D + W_I}{N} \right) \tag{1}$$

where W_S, W_D and W_I are substituted, deleted and inserted words, while N is the total number of words. W_S, W_D and W_I are computed using the Levenshtein distance between the transcribed and recognized sentences.

Table 2. WER for the monophone ASR of the weather task with LPC, PLP and MFCC features vectors

WER	LPC	PLP	MFCC
1mix	59,99	18,8	18,7
10mix	33,95	11,6	11,57
20mix	29,23	10,61	10,54

The cepstral based features, MFCC and PLP, are expectedly better due to the better following of auditory scale. Similar results are reported for other languages as well [4]. According to the slightly better achievement of the MFCC over PLP features for acoustic modeling in Croatian LVASR the use of MFCC speech feature vectors is proposed. So, speech feature vectors in all subsequent experiments consist of mel-frequency cepstral coefficients with the zeroth coefficient (energy) and their first and second order differences.

4.2 Croatian Acoustic Models Experiments and Results

The main issue of this work was investigation of speech quantity and state tying procedures influence on system performance. In addition, the test results of speaker independent ASR in real speech dialogs between users and system were investigated as well.

For Croatian acoustic modeling experiments four different ASR models were built using four parts of the speech corpus: the weather, news, tales and dialogs. The speech was cumulatively used for estimating parameters for each acoustic and language model: for weather 8 hours, for news 13 hours and for tales 15 hours of speech, always from the same 8 male and 8 female speakers. The test set in all experiments remained the same 1710 weather related

Table 3. Monophone ASR results for weather, news, and tales: WER computed for different number of Gaussian mixtures

WER	**ASR**			**weat2tales improv.**
	weather	*news*	*tales*	
1mix	18,7	18,49	15,36	**3,34**
10mix	11,57	11,36	9,46	**2,11**
20mix	10,54	10,58	8,73	**1,81**
1-20 mix improv.	**8,16**	**7,91**	**6,63**	

utterances from remaining 3 male and 6 female speakers. For each experiment the acoustic and the language model were trained according to the procedures described in Section 3.

Table 3 presents the context-independent monophone recognition results in terms of WER for weather, news and tales recognition tasks. In each monophone acoustic model the number of Gaussian mixtures have been increased from 1 to 20. In the last column of the Table 3 the absolute reduction in WER between weather and news task was calculated. The average improvement of 2,5% in WER points to the importance of having enough speech material for acoustic model training. But, still it is worth noticing that standard modeling technique with increasing number of Gaussians mixtures for the output probability distributions of HMMs contributes additional WER reduction of 7,6% in average.

The monophone models were used for initial setting of the triphone models according to the procedures presented in the previous Section. Table 4 shows the number of theoretically possible cross-word triphones, the number of all possible triphones where special acoustic events have no left and right context and finally the number of actually seen triphones in the training data. The percentage of seen triphones in the corpus has risen from 12 to 25% percent, which was one of the aims of this corpus. The increase of seen triphones resulted in reduction of the WER of tested models by 3,5%. But still, there is not enough acoustic material for modeling all possible triphone models. The severe under training of the model can be a real problem in the speech recognition system performance. The lack of speech data is overcome by a phonetically driven state tying procedure already presented in Section 3. In Croatian LVASR 83 proposed phonetic rules (166 Croatian phonetic questions about left and right context [29]) are used to build phonetic decision trees for HMM state clustering of acoustic models.

Table 5 presents the context-dependent recognition results in terms of WER. The results are presented from left to right for weather, news and tales ASR task. For each acoustic model the number of Gaussian mixtures was increased from 1 to 20. In the last column, the absolute reduction in WER between weather and news task was calculated with an average WER improvement of 2.6%. Increased number of Gaussians mixtures contributed with 6% WER reduction in average.

Table 4. The number of possible and seen triphones per parts of the corpus

	No. triphones			**%**
	possible	***all***	***seen***	***seen***
weather	35937	31585	4042	12,80
+news	39304	34684	7931	22,87
+tales	39304	34684	8700	25,01
+dialogs	39304	34684	8757	25,25

Table 5. Triphone ASR results for weather, news and tales: WER computed for different number of Gaussian mixtures

WER	**ASR**			**weat2tales improv.**
	weather	*news*	*tales*	
1mix	17,27	16,63	13,79	**3,48**
10mix	11,28	11,03	9,1	**2,18**
20mix	10,61	10,5	8,55	**2,06**
1-20mix improv.	**6,66**	**6,13**	**5,24**	

Monophone vs. triphone ASR results for weather, news and tales recognition tasks are compared in Fig. 2. It is likely to be noticed that the acoustic model which has 12 or more Gaussian mixtures for monophone output distributions (weather speech) performs almost as well as a triphone based ASR, which reflects the fact that there is not enough speech material. The ratio between monophone and triphone recognition WER trained on news part keep the constant difference about 2%. The comparison for the tales ASR shows an average of approximately 0,5% triphone outperforming monophone model, while at higher mixes the monophone and triphone results equate. This behaviour was caused by the fact that the number of words used for training is almost 15K which resulted in a higher percentage of triphones seen exactly once and the tales speech was spoken by only one speaker. So it is possible to argue that monophone models at higher mixes are as good as triphones so why context-dependent modeling at all. The main cause for these results stems from the number of different speakers used for training. Since models were trained on a limited number of speakers' speech (8 male and 8 female) there was not enough variability in speech captured in triphone models. However, at lower mixes the triphone models outperform monophones models. For a more in-depth explanation of such behaviour, a deeper look into more detailed context-dependent state tying performance is required in the following subsection. We should finally point out that for the same recognition task (the same 1710 weather sentences) the WER is reduced by more than 10%: from 18,7% in weather monophone 1mix lowered to 8,55% in tales triphone 20mix model, which is a firm argument for the use of context-dependent subword acoustic unit modeling.

4.3 The Influence of State Tying on Recognition Accuracy

The phonetically driven state tying presented in Subsection 3.3 is a standard technique of coping with sparse speech data. The influence of speech tying on the performance of ASR system is presented in Fig. 3. In all tied models the same set of 83 proposed Croatian phonetic rules were used.

The results are prepared for a different percentage of tied states depending on the threshold parameters set for tying and for 1, 5 and 10 mixtures. The results in terms of WER are presented for radio ASR, news ASR and tales ASR. In each ASR the states are tied into more compact form. So for the radio ASR the number of state was tied from 66% down to 3%, for news ASR from 36% to 2% and for tales from 59% to 2%.

All three ASR systems are displaying the same behaviour, tying the states improves the overall WER. The best result is achieved at 2% or 3 % of all states which is quite close to the monophone model. But this also implicates that there is not enough variability of speech used for training: 16 speakers reading and all sharing the same channel, although the number of seen triphones had increased from 12 to 25%.

Reported context-dependent ASR result leaves no obvious research space for further improvement in the acoustic modeling by increasing the quantity of good quality broadcasted and read Croatian speech. Therefore further research was motivated by acoustic modeling of spontaneous speech in different acoustic conditions. This means modeling the speech produced over different channels (GSM, telephone etc.) with different background noises (car noise, speech or music in background, etc.) and different speaker's emotional states.

4.4 Spontaneous Speech Recognition Experiments and Results

The dialog ASR was trained on 15,5 hours of speech: the whole radio, news, tales part of the corpus and additional with the spontaneous dialogs from 12 male and 12 female speakers (70% of the dialogs). The dialog ASR was tested on speech of the remaining 5 male and 5 female speakers (30% of the dialogs).

The results in Table 6 show the WER of speaker independent automatic dialog speech recognition. For presented speaker independent recognition results the triphone context-dependent acoustic models on available speech from 23 male and 26 female speakers containing large vocabulary of 14551 words were trained. Table 6 presents WER for 5 male (dm) and 5 female (dz) speakers each uttering 45 sentences for 1, 10 and 20 Gaussian mixtures. For 10 Gaussian mixtures triphones overall WER is 4.27%. This result encourages the use of large vocabulary Croatian automatic speech recognition in the real spoken dialog systems. In real world application however telephone access to the system is expected and according to the previous research [28] expected telephone speech ASR results have higher (almost double) word error rates.

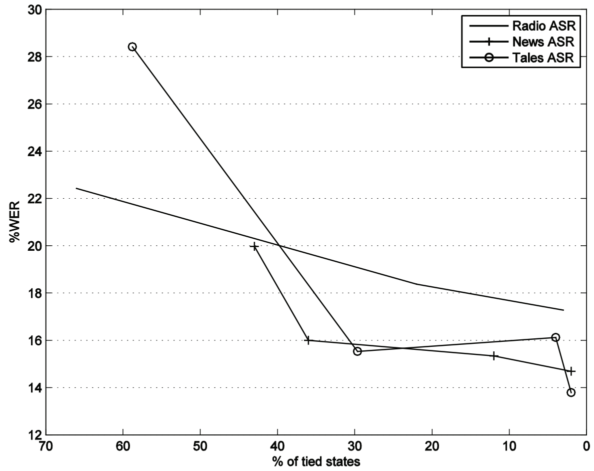


Fig. 3. WER for different % of tied states in radio, news and tales ASR

Table 6. WER of triphone dialog ASR system per speaker and 1, 10 and 20 Gaussian mixtures

	WER		
Speaker	***1mix***	***10mix***	***20mix***
dm013	13,64	1,75	3,15
dm014	10,49	8,39	6,29
dm015	12,59	4,55	5,59
dm016	3,85	3,85	3,5
dm017	13,64	1,05	0,7
dz013	13,99	1,75	1,4
dz014	21,33	11,54	11,89
dz015	6,29	1,75	3,5
dz016	18,18	3,15	4,55
dz017	10,84	4,9	5,24
OVERALL	**12,48**	**4,27**	**4,58**

Figure 4 compares monophone vs. triphone recognition tasks in dialog ASR for 1 to 20 Gaussian mixtures. The triphone results are better for 2,7% in average. This proves the dominance of used context-dependent over context-independent modeling approach in Croatian spontaneous LVASR.

Figure 5 summarizes the state tying results in the dialog ASR. The results are given for a different number of tied states: 59%, 30%, 22%, 14%, 4%, and 3% of states.

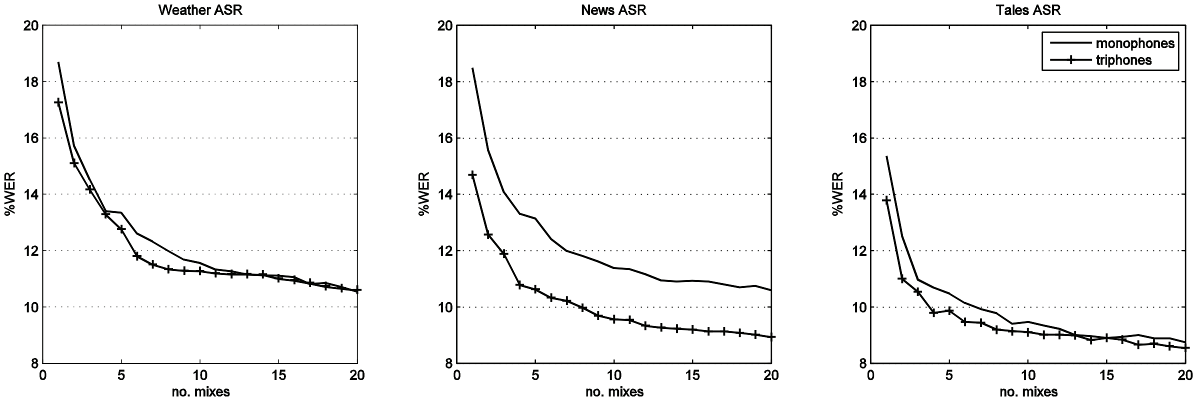


Fig. 2. Comparison of monophone vs. triphone ASR for the weather, news and tales tasks

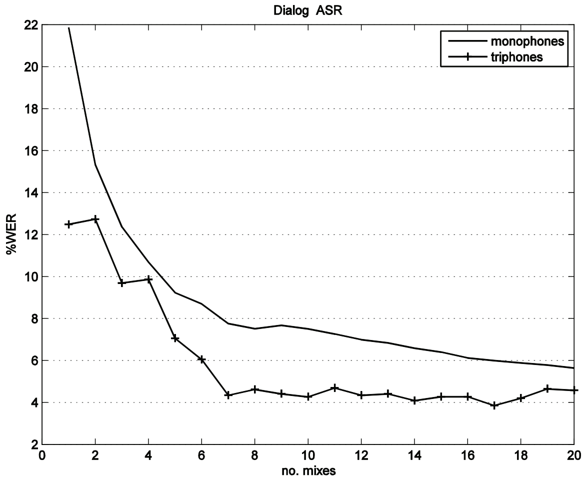


Fig. 4. Comparison of monophone vs. triphone ASR for spontaneous dialogs task

The behaviour of the tying procedure shows different trend then the tying procedures results shown in Fig. 3. In the dialog ASR the models with 20-25% of tied states are performing better, even with higher number of Gaussian mixtures. This confirms the context-dependent triphone acoustic model is outperforming the context-independent monophone model due to the good variability (different speakers and seen triphones) in available speech for acoustic modeling in Croatian LVASR.

Further improvement of these results can be achieved with language modeling. Thus, the next part of the paper presents the influence of different language models on automatic speech recognition results.

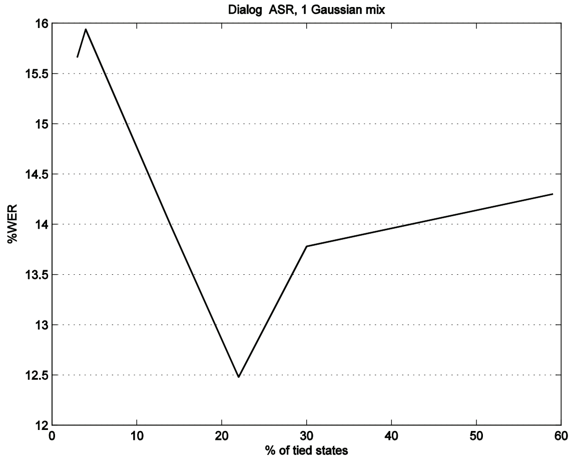


Fig. 5. WER for different % of tied states in dialog ASR

4.5 Language Model

The Croatian LVASR language model consists of 14551 different words, which determine the size of a bigram language model to14551^2. However, like speech, the textual data is sparse. Only a subset of all possible word pairs is present in collected texts, so the majority of possible bigrams have zero frequencies. In the actual texts from corpus out of the theoretically possible number of bigrams 211731601 only 46397 (0,02%) were present. Hence in each ASR in this work bigram language model with smoothing and discounting procedures from Subsection 3.3 were used. Perplexities and entropies for each language model constructed from the vocabulary and text transcription of each part of the speech corpus are presented in Table 7.

For Croatian language modeling experiments, previously trained context-dependent dialog ASR speaker inde-

Table 7. The perplexity, entropy and bigram counts of language models

	perplexity	entropy	# bigram
weather	11,17	3,48	7999
news	17,16	4,01	31475
tales	27,59	4,79	46323
dialogs	2,71	1,44	146

pendent acoustic model from Croatian LVASR was used. The basic idea in language modeling experiments was to compare if clustering of the large vocabulary to the limited domain vocabulary plays a significant role in the ASR system performance.

The results in Table 8 are prepared for the same 450 utterances (5 male and 5 female speakers) that have been used for speaker independent testing of the dialog ASR. The results are in terms of WER for 1, 10 and 20 Gaussian mixtures. The LM1 in the first column from the left was prepared from the whole vocabulary containing 14551 different words and from the text of all 17603 utterances used for the acoustic model training with perplexity of 16,94. The LM2 in the middle column was prepared from the weather vocabulary containing only 1494 different words and from the text of all 17603 utterances with perplexity of 15,47. The last LM3 was prepared from the weather vocabulary containing only 1494 different words and from the text of 10137 weather related utterances with perplexity of 14,77. As expected the most limited vocabulary slightly lowered the perplexity and overall WER. This result suggests checking on different LM for the same recognition task, as some improvement in WER can be achieved for the domain-oriented tasks. The importance of used text for language modeling shows dependence on the number of present bigrams in text, so for further development additional texts with sufficient number of bigrams should be collected.

The presented results of language modeling for Croatian ASR are in their preliminary stage. Furthermore, higher order Croatian n-gram, model should be considered with different smoothing and discounting methods. Additionally, as Croatian is a highly flective language some class based models including Part-of-Speech tags as the class of each word in the dictionary will also be considered.

5 CONCLUSION

The large vocabulary Croatian ASR is presented in the paper. The acoustic and language modeling aspects of Croatian ASR are discussed through word error rates of four different systems: weather, news, tales and dialog ASR trained from the speech of the Croatian speech corpus. The main issue was investigation of context-independent and context-dependent subword acoustic units and state tying procedures influence on the tested system performance. In addition, speaker independent ASR simulating real dialog between the user and system was investigated. The language modeling experiments compared the size of vocabulary with the limited domain texts once more through performance of the dialog ASR system.

Table 8. WER for the dialog ASR with 3 different language models

	LM1	LM2	LM 3
No. words	14551	1494	1494
No. utterances	17603	17603	10137
No. bigrams	46397	11969	9992
Perplexity	16,94	15,47	14,77
	WER		
1mix	12,48	12,62	11,92
10mix	4,27	4,3	4,48
20mix	4,58	5,07	4,72

For the purposes of this work four different ASR systems were built using four parts of speech from the following corpus: weather, news, tales and dialogs, so they are named: weather ASR, news ASR, tales ASR and dialog ASR system. The speech was cumulatively used for the training of each new acoustic model: for weather ASR 8 hours of speech, for news ASR training 13 hours of speech and for tales ASR 15 hours of speech, always from the same 8 male and 8 female speakers. The test set in all experiments remained the same 1710 weather related utterances from 3 male and 6 female speakers. The dialog ASR was trained on 15,5 hours of speech: the whole radio, news, tales part of the corpus and additionaly with the spontaneous dialogs from 12 male and 12 female speakers (70% dialogs). The dialog ASR was tested on speech of remaining 5 male and 5 female speakers (30% dialogs). The dialog ASR is large vocabulary (over 14500 different words) speaker independent (59 speakers) automatic speech recognition system which achieved WER below 5%.

The paper compares the linear prediction and the cepstral-based feature vectors for acoustic modeling in Croatian LVASR as well. The cepstral-based features, MFCC and PLP, are better than linear prediction LPC features due to its better following the human auditory scale. According to the slightly better performance of the MFCC over PLP features for Croatian LVASR the MFCC speech feature vectors were used.

The presented work shows that the common approach for automatic speech recognition using context-dependent acoustic modeling based on Croatian phonetic rules state tying procedures performed on sufficient quantity of speech can be used for Croatian large vocabulary ASR with error rates below 5%. Additionally the paper proves the importance of state tying in context-dependent acoustic modeling and finally points toward certain aspects of language model smoothing and discounting in LVASR.

Further research activities are planned towards applications in the field of Croatian speech technologies especially on development of the Croatian spoken dialog system. One of the problems for spoken dialog systems is speech/non-speech detection [30] which should be addressed in the future.

ACKNOWLEDGMENT

This work is supported by the Ministry of Science, Education and Sports of the Republic of Croatia filed under projects 318-0361935-0852 and 009-0361935-0852.

REFERENCES

[1] X. D. Huang, A. Acero, H. W. Hon, *Spoken Language Processing: A Guide to theory, Algorithm and System Development*, Prentice Hall, New Jersey, USA, 2000.

[2] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, vol. 77(2), pp. 257-286, 1989.

[3] S. Furui, 50 Years of Progress in Speech and Speaker Recognition, *SPCOM'05*, Grece, 2005, pp. 1-9.

[4] D. O'Shaughnessy, Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis, *Proc. of IEEE*, vol. 91(9), pp. 1271-1305, 2003.

[5] F. Jelinek, *Statistical Methods for Speech Recognition*, The MIT Press, USA. 1999.

[6] S. Furui, M. Nakamura, K. Iwano, Why is automatic recognition of spontaneous speech so difficult, *LKR 2006*, pp. 83-90. 2006.

[7] J. Žibert, S. Martinčić-Ipšić, M. Hajdinjak, I. Ipšić, F. Mihelič, Development of a Bilingual Spoken Dialog System for Weather Information Retrieval, *EUROSPEECH´03*, 2003, vol. 1, pp. 1917-1920.

[8] S. Lihan, J. Juhar, A. Čižmar, Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases, *INTERSPEECH'05-EUROSPEECH*, 2005, pp. 225-228.

[9] J. Psutka, P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovsky, S. Gustman, Large Vocabulary ASR for Spontaneous Czech in the MALACH Project, *EUROSPEECH´03*, 2003, pp. 1821-1824.

[10] J. Kacur and G. Rozinaj, ch. Practical Issues of Building Robust HMM Models Using HTK and SPHINX Systems, in *Speech recognition: technologies and applications*, F. Mihelič and J. Žibert, Eds., I-Tech Education and Publishing, 2008, pp. 171-191.

[11] R. Frederking, A. Rudnicky, C. Hogan, Interactive Speech Translation in the DIPLOMAT Project, *Spoken Language Translation Workshop*, 1997, pp. 61-66.

[12] A. Black, R. Brown, R. Frederking, R. Singh, J. Moody, E. Steinbrecher, TONGUES: Rapid development of a speech–to-speech translation system, *HLT'02*, 2002, pp. 2051-2054.

[13] P. Scheytt, P. Geutner, A. Waibel, Serbo-Croatian LVCS on the dictation and broadcast news domain, *ICASSP'98*, 1998.

[14] B. Dropuljić and D. Petrinović, Development of Acoustic Model for Croatian Language Using HTK, *Automatika*, vol. 51(1), pp. 79-88, 2010.

[15] S. Young et al., *The HTK Book*, Cambridge University Engineering Department, Cambridge, 2002.

[16] S. Martinčić-Ipšić, M. Matešić, I. Ipšić, Korpus hrvatskog govora, *Govor: časopis za fonetiku*, vol. 21(2), pp. 135-150, 2004.

[17] S. Martinčić-Ipšić, I. Ipšić, Croatian HMM Based Speech Synthesis, *Journal of Computing and Information Technology,* vol. 14(4), pp. 299-305, 2006.

[18] M. Pobar, S. Martinčić-Ipšić, I. Ipšić, Text-to-Speech Synthesis: A Prototype System for Croatian Language, *Engineering Review*, vol. 28(2), pp. 31-44, 2008.

[19] A. Meštrović, L. Bernić, M. Pobar, S. Martinčić-Ipšić, I. Ipšić, Overview of a Croatian Weather Domain Spoken Dialog System Prototype*, ITI2010*, 2010, pp. 103-108.

[20] R. Duda, P. Hart, D. Stork, *Pattern Classification*, John Wiley, Canada, 2001.

[21] J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. dissertation, Queen's College, University of Cambridge, Cambridge, UK, 1995.

[22] J. L. Gauvain, L. Lamel, Large Vocabulary Speech Recognition Based on Statistical Methods, ch. 5 in *Pattern Recognition in Speech and Language Processing*, (ed.) Chou, W., (ed.) Juang, B. W., CRC Press LLC, Florida, USA, 2003.

[23] K. Lee, H. Hon, R. Reddy, An Overview of the SPHINX Speech Recognition System, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38(1), pp. 35-45, 1990.

[24] M. Y. Hwang, X. Huang, F. Alleva, Predicting unseen triphones with senones, *IEEE ICASSP'93*, 1993, vol. 2, pp. 311-314.

[25] S. Young, J. Odell, P. Woodland, Tree-Based State Tying for High Accuracy Acoustic Modeling, *ARPA HLT Workshop*, 1994, pp. 307-312.

[26] S. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, *Computer Speech and Language*, vol. 13(4), pp. 359-394, 1999.

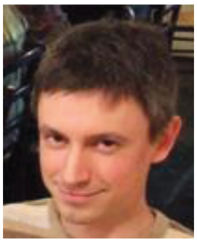
[27] J. Picone, Signal Modeling Techniques in Speech Recognition, *Proc. IEEE*, vol. 81(9), pp. 1215-1247, 1993.

[28] S. Martinčić-Ipšić and I. Ipšić, Croatian Telephone Speech Recognition, *MIPRO06*, vol. CTS + CIS, 2006, pp. 182-186.

[29] S. Martinčić-Ipšić, S. Ribarić, I. Ipšić, Acoustic Modeling for Croatian Speech Recognition and Synthesis, *Informatica*, vol. 19(2), pp. 227-254, 2008.

[30] J. Žibert, N. Pavešić, F. Mihelič, Speech/non-speech segmentation based on phoneme recognition features, *EURASIP journal on applied signal processing*, vol. 2006, p. 1-13, 2006.

Sanda Martinčić-Ipšić obtained her B.Sc. degree in computer science in 1994, from the University of Ljubljana, Faculty of Computer Science and Informatics and her M.Sc. degree in informatics from the University of Ljubljana, Faculty of Economy in 1999. In 2007 she obtained the Ph.D. degree in computer science from the University of Zagreb, Faculty of Electrical Engineering and Computing. She currently works as an Assistant Professor at the University of Rijeka, Department of Informatics. Her research interests include automatic speech recognition, speech synthesis, speech corpora development and spoken dialog systems, with a special focus on the Croatian language.



Miran Pobar was born in 1983 in Rijeka, Croatia. In 2007 he obtained his B.Sc. degree in electrical engineering from the University of Rijeka, Faculty of Engineering. He currently works at the University of Rijeka, Department of Informatics, as an Assistant. In 2008 he commenced a Ph.D. postgraduate study programme at the University of Zagreb, Faculty of Electrical Engineering and Computing. His research interests include speech synthesis, speech recognition and speech technologies for Croatian language.

Ivo Ipšić obtained his B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from the University of Ljubljana, Faculty of Electrical Engineering, in 1988, 1991 and 1996, respectively. From 1988–1998 he was a staff member of the Laboratory for Artificial Perception, at the University of Ljubljana, Faculty of Electrical Engineering. Since 1998 Ivo Ipšić has been a professor of computer science at the University of Rijeka, teaching computer science courses. His current research interests lie within the field of speech and language technologies.

AUTHORS' ADDRESSES

Asst. Prof. Sanda Martinčić-Ipšić, Ph.D.
Miran Pobar, B.Sc.
University of Rijeka,
Department of Informatics,
Omladinska 14, 51000, Rijeka, Croatia
email: {smarti, mpobar}@uniri.hr

Prof. Ivo Ipšić, Ph.D.
University of Rijeka,
Faculty of Engineering,
Vukovarska 58, 51000, Rijeka, Croatia
email: ipsic@riteh.hr

Received: 2010-11-09
Accepted: 2011-06-08