8

Croatian Speech Recognition

Ivo Ipšić and Sanda Martinčić–Ipšić
University of Rijeka
Croatia

1. Introduction

In the chapter we describe procedures for Croatian speech recognition which are used in a limited domain spoken dialog system for Croatian speech. The dialog system would provide information about weather in different regions of Croatia for different time periods (Žibert et al., 2003). The spoken dialog system includes modules for automatic speech recognition (ASR), spoken language understanding and text-to-speech synthesis. In this work ASR module based on data-driven statistical and rule-based knowledge approach is discussed. Data driven statistical approach is based on large quantities of spoken data collected in the speech corpus. Rule based approach is based on Croatian linguistic and phonetic knowledge. Both approaches must be combined in a spoken dialog system because there is not enough speech data to statistically model the human speech and there is not enough knowledge about the processes in human mind during speaking and understanding (Dusan & Rabiner, 2005). Speech recognition today, as in the past decades, is mainly based on data driven statistical approaches (Huang et al. 2000; Rabiner, 1989). Statistical pattern recognition and segmentation algorithms and methods for stochastic modelling of large speech quantities are used. The data driven statistical approach uses hidden Markov models (HMM) as the state of the art formalism for speech recognition. Many large vocabulary automatic speech recognition systems (LVASR) use mel-cepstral speech analysis, hidden Markov modelling of acoustic sub word units, n-gram language models (LM) and n-best search of word hypothesis (Furui, 2005; O'Shaugnessy, 2003; Huang et al., 2000; Jelinek, 1999). Speech recognition research in languages like English, German and Japanese (Furui et al., 2006) has focus in recognition of spontaneous and broadcasted speech. For highly flective Slavic and agglutinative (Kurimo et al., 2006) languages the research focus is still more narrowed mainly due to the lack of speech resources like corpuses. Large or limited vocabulary speech recognition for Slovene (Žibert et al., 2003), Czech (Lihan et al., 2005; Psutka et al., 2003), Slovak (Lihan et al., 2005), Lithuanian (Skripkauskas & Telksnys, 2006; Vaičiūnas & Raškinis, 2005) and Estonian (Alumäe & Võhandu, 2004) with applications for dialog systems (Žibert et al., 2003), dictation (Psutka et al., 2003) or automatic transcriptions (Skripkauskas & Telksnys, 2006) have been reported lately.

Croatian is a highly flective Slavic language and words can have 7 different cases for singular and 7 for plural, genders and numbers. The Croatian word order is mostly free, especially in spontaneous speech. The unstressed word system is complex because the possible transition of the accent from a stressed word to the unstressed one is conditioned by the position of the word in a sentence, which is mostly free. Standard Croatian

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

pronunciation rules sometimes allow more different word accents. Mostly free word order, a complex system of unstressed words and nondeterministic pronunciation rules make the development of pronunciation dictionary and prosodic rules difficult. On the other hand Croatian orthographic rules based on phonological-morphological principle are quite simple which simplifies the definition of orthographic to phonetic rules and process of phonetic transcription.

The number of Croatian native speakers is less then 6 millions. Still some interest in the research and development of speech applications for Croatian can be noticed. The speech translation system DIPLOMAT between Serbian and Croatian on one side and English on the other is reported in (Frederking, et al., 1997; Scheytt, et al., 1998; Black, et al., 2002). The TONGUES project continued with this research in direction towards large Croatian vocabulary recognition system.

Croatian orthographic-to-phonetic rules are proposed for phonetic dictionary building. The developed Croatian multi-speaker speech corpus was successfully used for the development of speech applications. Proposed Croatian phonetic rules captured adequate Croatian phonetic, linguistic and articulatory knowledge for state tying in acoustical models of the speech recognition system.

The Croatian speech recognition system is based on continuous hidden Markov models of context independent (monophones) and context dependent (triphones) acoustic models. The training of speech recognition system was performed using the HTK toolkit (Young et al., 2002; HTK, 2002).

Since the main resource in a spoken dialog system design is the collection of speech material, the Croatian speech corpus is presented in Section 2. Orthographic-to-phonetic rules used in the phonetic dictionary preparation are shown as well. Further the acoustic modeling procedures of the speech recognition system including phonetically driven state tying procedures are given in Section 3. Conducted speech recognition experiments and speech recognition results are presented in section 4. We conclude with discussion on advantages of the proposed acoustical modelling approach for Croatian speech recognition and description of current activities and future research plans.

2. The Croatian speech corpus

The Croatian speech corpus includes news, weather forecasts and reports spoken within broadcasted shows of the national radio and television news broadcasted at the national TV (Martinčić-Ipšić and Ipšić, 2004). The collected speech material is divided into several groups: weather forecasts read by professional speakers within national radio news, weather reports spontaneously spoken by professional meteorologists over the telephone, other meteorological information spoken by different reporters and daily news read by professional speakers.

The speech corpus is a multi-speaker speech database which contains 16,5 hours of transcribed speech spoken in the studio acoustical environment and 6 hours of telephone speech. The spoken utterance has its word level transcription.

The first part of the speech corpus consists 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. The radio part consists of 9431 utterances and lasts 13 hours. The transcribed sentences contain 183000 words, where 10227 words are different. Relatively small number of 1462 different words in

the weather forecast domain shows that this part of the speech database is strictly domain oriented.
The second part contains weather reports given by 7 female and 5 male professional meteorologists over the telephone. The 170 transcribed weather reports are lasting 6 hours and contain 1788 different words in 3276 utterances. Most of the speech captured in the telephone part can be categorized as semi-spontaneous. This data is very rich in background noises such as door slamming, car noise, telephone ringing and background speaking and contains noise produced by channel distortions and reverberations. All this special events and speech disfluencies and hesitations are annotated in transcriptions by < >.
The third part of the speech database consists of TV News broadcasted at the national TV – HTV. The news data is not domain oriented. Diversity of subjects and topics is noticeable in the number of all words compared to the number of different words. Further the number of speakers is also significantly bigger then in the weather part of the database. The news data is also very rich in different background noises, including music, it also contains commercials, reports in foreign languages and so on. All of this features where captured and annotated during the transcription. The transcribed part of TV News consists of 18632 words where 9326 are different. The transcribed part of TV News is 3 hours and 28 minutes. long. Statistics of TV News is also shown in the bottom part of Table 1.

	Number		Speakers		Words		Dur.
	Reports	Utts.	Male	Fem.	All	Diff.	[min]
Radio weather forecasts	1057	5456	11	14	77322	1462	482
Radio news	237	3975	1	2	105678	9923	294
Overall RADIO	1294	9431	11	14	183000	10227	775
Teleph. weather reports	170	3276	5	7	52430	1788	360
BCN	6	280	217		18632	9326	208
Overall	1470	12987	253		254062	15998	1343

Table 1. Croatian speech corpus statistics.

2.1 Data acquisition and transcription

The broadcasted radio news with weather forecasts and telephone weather reports were recorded four times a day using a PC with an additional Haupage TV/Radio card. The speech signals are sampled with 16 kHz and stored in a 16-bit PCM encoded waveform format. At the same time texts of weather forecasts for each day were collected from the web site of the Croatian Meteorological Institute. The texts were used for speech transcription and for training of a bigram language model for the weather forecast speech recognition system. For the telephone weather reports and daily news no adequate text existed so the whole transcription process was manual. The transcribing process involved listening to speech until a natural break was found. The utterances or parts of speech signals were cut out and a word level transcription file was generated. The speech file and the transcription file have the same name with different extensions.
During the transcription some basic rules were followed: all numbers and dates were textually written, all acronyms and foreign names were written as pronounced and not as

spelled and all other words were written according to the Croatian writing rules (Anić and Silić, 2001). Word transcriptions of TV news have been done in two stages. In the first stage we collected texts from TV NEWS at the internet site of the national TV (HTV). The texts were not the exact transcriptions and we had to correct them, but they were a good start.
All final transcriptions of Croatian BCN (Broadcast News) were made with the Transcriber tool (Barras, et al., 2000). Transcriber is a tool for assisting in the creation of speech corpora enabling manual segmentation and transcription as well as annotation of speech turns, topic and acoustic condition. The data format follows the XML standard with Unicode support for multilingual transcriptions (Graff, 2000).

2.2 Phonetic dictionary

For the word segmentation and recognition task we have developed a phonetic dictionary, where we proposed a set of phonetic symbols to transcribe the words from the Croatian speech database. The selected symbols are derived according to the Speech Assessment Methods Phonetic Alphabet (SAMPA) (SAMPA, 1997). The standard phoneme set includes 30 phonemes, where the set of vowels is extended with the vibrant vowel /r/.
Croatian orthographic rules are based on the phonological-morphological principle which enables automatisation of phonetic transcription. Standard definition of orthographic to phonetic rules, one grapheme to one phonetic symbol was extended with additional rules for example:

- words with group ds were phonetically transcribed as [c] and
- words with suffixes naest were phonetically transcribed as [n a j s t].

The phonetic dictionary comprises all words in transcription texts. All word forms (different cases, genders and numbers of the same basic word form) are considered as a new word in the dictionary. The current phonetic dictionary contains 15998 different words. The fact that Croatian language is highly flective reflects to the size of the phonetic dictionary. The dictionary can contain few entries for the same basic word format. For example the word bura, which denotes the northern wind type, is represented by 4 different word forms: bura, bure, burom, buru. Since all foreign names were written as pronounced there was no need for writing the orthographic to phonetic rules for languages like English, German, Italian, Chinese, Arab, etc.
The accent position is embedded in the dictionary with differentiation between accented and non–accented vowels. For the words that can be pronounced in more correct ways the position of the really accented vowel was marked.

2.3 Segmentation

Since the transcription of the speech files is on the word level for the training procedures the utterances have to be segmented on the phone level. The initial segmentation is performed using automatic alignment of speech signals and word transcriptions, which is based on hidden Markov monophone models. The automatic segmentation is performed using the monophone speech recognizer described in section 3.
Typical segmentation errors detected during manual inspections of automatically determined speech segments can be roughly classified as transcription errors and real segmentation errors. Similar automatic segmentation error taxonomy for English is presented in (Kominek, et al., 2003).
Transcription errors are errors in the speech transcription stage of speech corpora development. Some words or special acoustic events were incorrect or inaccurate typed or

were not typed at all. For example if breathing noise (inspiration) was not marked in the textual transcription in a utterance, the whole inspiration was segmented as a really long phoneme.
Real segmentation errors occurred when transcriptions were correct but the segment interval was not determined correctly. Typical segmentation errors occurred:

- at infrequent phones like /lj/ or /dž/,
- at two following vowels which are seldom in Croatian words like /ea/ and
- at too tightly segmented phones combinations where one of the phones was not pronounced like /ije/.

Automatically segmented speech utterances were manually inspected and segmentation errors were corrected in the speech database.

3. Acoustic and language modelling

The goal of speech recognition system 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. To avoid these difficulties the data driven statistical approach based on large quantities of spoken data is used (Furui et al., 2006). Statistical pattern recognition and segmentation algorithms and methods for stochastic modelling of time varying speech signals are used (Rabiner et al., 1989; Huang et al., 2000; Duda et al., 2001). Additionally statistical language models are used in order to improve the recognition accuracy (Jelinek et al., 1999).
The data driven statistical approach uses hidden Markov models (HMM) as the state of the art formalism for speech recognition. Hidden Markov models are stochastic finite-state automata consisting of finite set of states and state's transitions. The state sequence is hidden, but in each state according to the output probability function an output observation can be produced.
The HMM Φ is defined by a triplet $\Phi=(A,B,\Pi)$ where A is state transition probability matrix, B is speech signal feature output probability matrix and Π is the initial state probability matrix. The output probability density function is represented by a mixture of Gaussian probability density function $b_j(x)=N(x,\mu_{jk},\Sigma_{jk})$ (Huang et al., 2000)

$$b_j(x) = \sum_{k=1}^{M} c_{jk} N(x, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^{M} c_{jk} b_{jk}(x) \quad \text{for } j = 1..N \text{ and } t = 1..T, \tag{1}$$

where

x	is the speech signal feature vector,
$b_j(x)$	is a Gaussian probability density function associated with state s_j,
μ_{jk}	is mean vector of the k^{th} mixture in state s_j,
Σ_{jk}	is covariance matrix of the k^{th} mixture in state s_j
M	is the number of mixture components and
c_{jk}	is the weight for the k^{th} mixture in state s_j satisfying the condition:

$$\sum_{k=1}^{M} c_{jk} = 1, \text{ and } c_{jk} \geq 0, \quad 1 \leq j \leq N, \ 1 \leq k \leq M. \tag{2}$$

For the estimation of continuous HMM parameters iterative Baum-Welch procedure is used. The in Baum-Welch also known as the Forward-Backward algorithm iteratively refines the HMM parameters by maximizing the likelihood of a speech signal feature sequence X given a HMM Φ, $P(X|\Phi)$. The algorithm is based on the optimisation technique used in the EM algorithm for the estimation of Gaussian mixture densities parameters. The Baum-Welch algorithm uses iteratively forward and backward probabilities which define the probability of the partial observation sequence X_t at time t in state i, given the HMM Φ (Duda et al., 2001; Huang et al., 2000).

For the search of an optimal path in the HMM network of acoustic models the Viterbi algorithm is used (Rabiner, 1989). Viterbi algorithm is a dynamic programming algorithm that decodes the state sequence according to the observed output sequence.

For speech modelling and recognition the speech signal feature vectors consist of 12 mel-cepstrum coefficients (MFCC), frame energy and their derivatives and acceleration coefficients. The feature coefficients were computed every 10 ms for a speech signal frame length of 20 ms.

Figure 1 presents main steps performed in the Croatian speech recognition system development, where acoustic and language models are trained. The speech signal is parameterized with MFCC feature vectors and their dynamic components, where the spectral resolution of the human ear is modelled. Speech transcriptions and speech signal feature vectors are used to train parameters of the monophone HMMs. The automatic segmentation is performed using monophone HMMs. The results of automatic segmentation are time intervals for each spoken phone. The automatically segmented phones are used for training (estimating) the parameters of monophone HMMs by repeating the Baum-Welch re-estimation procedure. The training procedure is repeated for each increase of the Gaussian mixture component. The 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. The triphone HMMs are constructed from monophone HMMs and the parameters are estimated with the Baum-Welch procedure.

The triphone states with estimated parameters value are tied according to the proposed Croatian phonetic rules. The state tying procedure insures enough acoustic material to train all context dependent HMMs and enables acoustic modelling 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 bigram language model. The triphone HMMs and bigram language model are used for Croatian speech recognition.

The acoustic model should represent all possible variations in speech. Variations in speech can be caused by speaker characteristics, coarticulation, surrounding acoustical conditions, channel etc. Therefore selection of an appropriate acoustic unit, which can capture all speech variations, is crucial for acoustic modelling. Enough acoustic material should be available for HMMs modelling of chosen acoustic unit. At the same time the chosen acoustic unit should enable construction of more complex units, like words (Odell, 1995). In continuous speech recognition systems the set of acoustic units is modelled by a set of HMMs. Since the

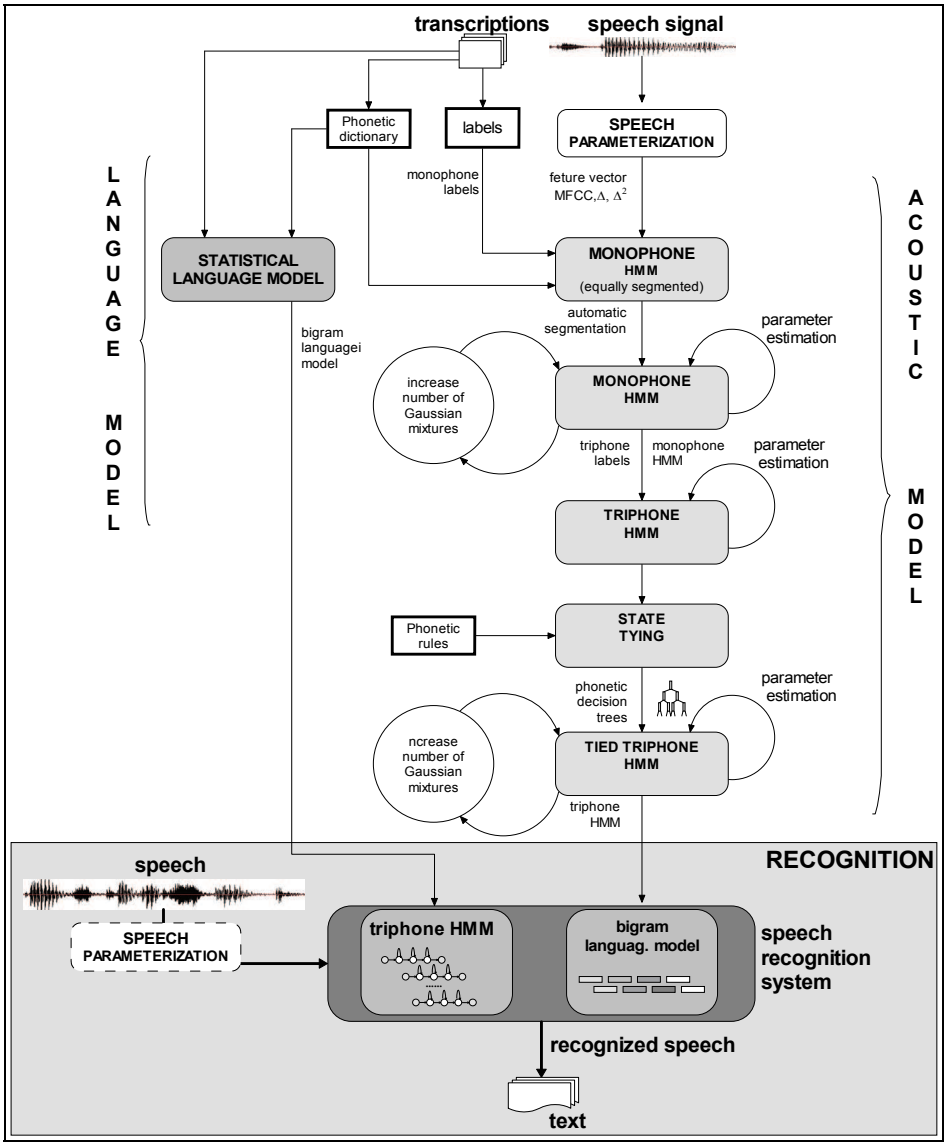

Fig. 1. Development of the Croatian speech recognition system.

number of units is limited (by the available speech data) usually the subword acoustic units are modelled. The subword units are: monophones, biphones, triphones, quinphones (Gauvain & Lamel, 2003; Lee et al., 1990) or sub phonemic units like senones (Hwang et al., 1993). Some speech recognition systems are modelling syllables (Shafran & Ostendorf, 2003) or polyphones (Schukat-Talamazzini, 1995). All these units are enabling construction of the more complex units and recognition of the units not included in the training procedure (unseen units).

3.1 Context independent acoustic model

The training of speech recognition acoustic models started with defining the Croatian phoneme set according to SAMPA (SAMPA, 1997). For each of 30 Croatian phonemes a context independent monophone hidden Markov model was defined. Initially the monophone models with continuous Gaussian output probability functions described with diagonal covariance matrices were trained. Each monophone model consists of 5 states, where the first and last states have no output functions. 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 to the phone level is performed using the forced alignment (Young et al., 2002) of the spoken utterance and the corresponding word level transcriptions. The results of automatic segmentation are exact time intervals for each phone. Further, 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.

3.2 Context dependent acoustic model

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

Table 2 shows the number of cross-word seen triphones in the training data used for radio speech recognition training. Evidently there was not enough acoustical material for modelling all possible triphone models. The severe under training of the model can be a real problem in the speech recognition system performance (Hwang et al., 1993). The lack of speech data is overcome by a phonetically driven state tying procedure.

	No.		No. triphones		%
	monophones	possible	all	seen	seen
radio weather	29+4	35937	31585	4042	12.80%
radio news	30+4	39304	36684	7931	21,62%
telephone	29+4	35937	31585	4618	14.62%

Table 2. The number of monophones and triphones and seen triphones percentage per parts of the speech corpus.

3.3 Croatian phonetic rules and decision trees

The state tying procedure proposed in (Young et al., 1994) allows classification of unseen triphones in the test data into phonetic classes and tying of the parameters for each phonetic class. In our system 108 phonetic rules (216 Croatian phonetic questions about left and right context (Martinčić-Ipšić & Ipšić, 2006a)) are used to build phonetic decision trees for HMM state clustering of acoustic models. The phonetic rules are describing the classes of the phonemes according to their linguistic, articulatory and acoustic characteristics. A phonetic decision tree is a binary tree, where in each node the phoneme's left or right phonetic

context is investigated. The phonemes are classified into phonetic classes depending on the phonetic rules which examine the phoneme's left and right context. Some Croatian phonetic rules used for the training of phonetic classes are shown in Table 3.

Vowel	a, e, i. o, u
High Vowel	i, u
Medium Vowel	o, e
Back	k, g, h, o, u
Affricate	c, C, cc, dz, DZ
Velar	k, g, h
Glide	j, v
Apical	t, d, z, s, n, r, c, l
Strident	v, f, s, S, z, Z, c, C, DZ
Constant Consonant	v, l, L, j, s, S, z, Z, f, h
Unvoiced Fricative	f, s, S, h
Compact Consonant	N, L, j, S, Z, C, cc, dz, DZ, k, g, h

Table 3. Examples of Croatian phonetic classes.

Figure 2 presents an example of phonetic decision tree for Croatian phoneme /h/. It classifies triphones with the phoneme /h/ in the middle in eight possible classes. At each node the binary question (from the set of 108 phonetic rules) about left and right context is asked and YES/NO answers are possible. The triphones in the same class are sharing the same parameters (state transition probabilities and output probability density functions of HMMs).

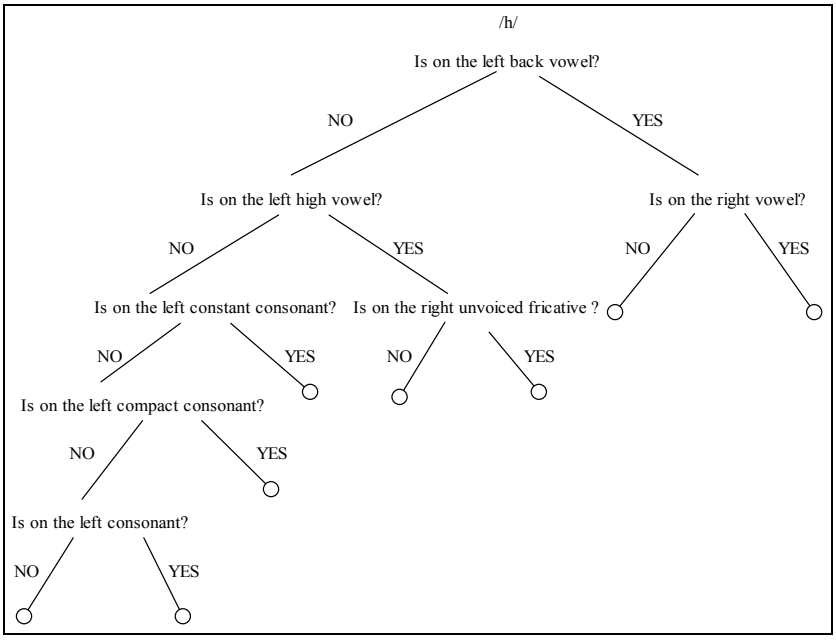

Fig. 2. The decision tree of phonetic questions for the left and right context of phoneme /h/.

For the construction of the phonetic decision tree from phonetic rules and from parameters of triphone HMM states a state tying procedure proposed in (Young et al., 1994) is used. 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. This enables more accurate estimating mixtures of Gaussian output probabilities and consequently better handling of the unseen triphones.
For each phoneme a decision tree is built using a top-down sequential optimization procedure (Odell, 1995). Initially all states are placed in the root node. So, all states are initially tied together and log likelihood is calculated for this node. The tying procedure iteratively applies phonetic rules to the states of the triphone models and partitions the states into subsets according to the maximum increase in log likelihood. When the threshold is exceeded the tied states are no further partitioned.
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 Gaussian mixtures output probabilities and consequently better handling of the unseen triphones.
For the speech recognition task the state clustering procedure uses a separate decision tree for initial, middle and final states of each triphone HMM which is built using a top-down sequential sub-optimal procedure (Odell, 1995). Initially all relevant states are placed in the root node. So, all states are initially tied together and log likelihood is calculated for this node. The tying procedure iteratively applies phonetic rules to the states of the triphone models and partitions the states into subsets according to the maximum increase in log likelihood. When the threshold is exceeded the tied states are no further partitioned.
For a set S of HMM states and a set F of training vectors x the log likelihood L(S) is calculated according to (Young et al., 1994) by

$$L(S) = \sum_{f=1}^{F} \sum_{s=1}^{S} \log(P(x_f, \mu(S), \Sigma(S))) \xi_s(x_f), \tag{3}$$

where $P(x_f,\mu(S),\Sigma(S))$ is the probability of observed vector x_f in state s under the assumption that all tied states in the set S share a common mean vector $\mu(S)$ and variance $\Sigma(S)$. $\xi_s(x_f)$ is the posterior probability of the observed feature vector x_f in state s and is computed in the last pass of the Baum-Welch re-estimation procedure (Young et al., 2002).
The node with states from S is partitioned into two subset S_y and S_n using phonetic question Q which maximizes the ΔL:

$$\Delta L = L(S_y) + L(S_n) - L(S), \tag{4}$$

where S_y is set of states which are satisfying the investigated phonetic question Q and in the S_n set are the rest of the states. Further the node is split according to the phonetic question which gives the maximum increase in log likelihood. The procedure is then repeated until it exceeds the threshold. The terminal nodes share the same distribution so the parameters of the final nodes can be estimated accurately, since the tying procedure provides enough training data for each final state.
The state tying procedure is presented in figure 3. From the top first is shown a monophone HMM for phoneme /h/. At the second level are HMMs for triphones o-h+r, e-h+a and a-h+m. Then the triphone states where tied and states sharing the same parameters are clustered using the phonetic decision trees. And at the bottom are the same tied states with

increased number of mixtures of Gaussians probability functions evaluated by the Baum-Welch parameter reestimation procedure.

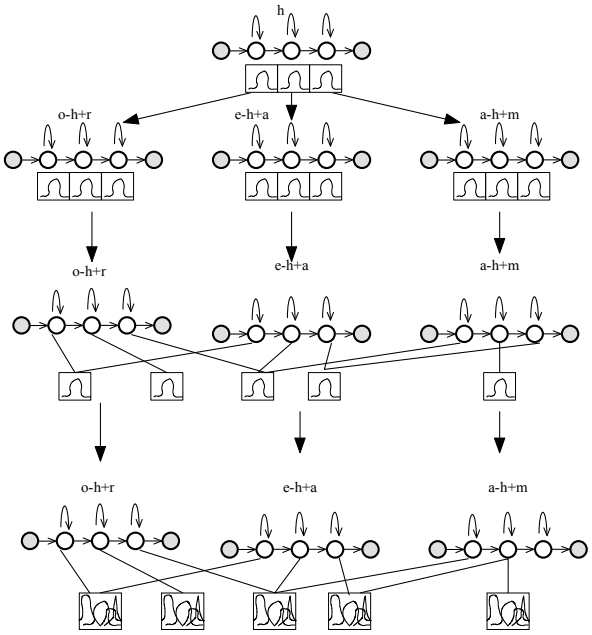

Fig. 3. The state tying procedure for the triphones with /h/ in the middle.

Table 4 contains the most frequently used Croatian phonetic questions in the phonetic decision trees in the speech recognition systems. Phonetic questions in the table are abbreviated. For instance the R-Front is the abbreviated phonetic question: Is the phoneme in the right context from the articulatory class front? Phonetic questions are ranked according to the appearance frequency in the decision trees. For the speech recognition part the frequency is calculated over 3 different sets of phonetic trees with different number of tied states (clusters).

Radio speech		Telephone speech	
Phonetic question	No.	Phonetic question	No.
R_Front	811	R_Front	522
L_Front	797	L_Front	498
L_Vowel-Open	635	L_Central	348
L_Central	594	R_Vowel-Open	336
R_Vowel-Open	561	R_Central	312
L_Consonant-Voiceless	432	L_Vowel-Open	312
R_Vowel	384	L_Consonant-Voiceless	222
R_Consonant-Voiceless	357	R_Vowel	221
D_Central	355	D_Consonant-Voiceless	216
L_Nasal	338	L_Consonant-Closed	201

Table 4. The most frequently used Croatian phonetic questions in radio and telephone speech recognition.

As expected and reported for other languages (Gauvain & Lamel, 2003) the most common Croatian phonetic rules (front, central, vowel) are the most frequently used for phonetic clustering in the speech recognition system. Since the results are presented for left and right coarticulation context and for the stable part of the phoneme, the phonetic rules are in left-question, right-question pairs. Phonetic questions investigating the presence of the single phoneme in the coarticulated context are the less frequent one, and used only in phonetic trees with higher number of tied states.

3.4 Language modelling

Language model is an important part of the speech recognition system. The language model estimates the probabilities of word sequences which are derived from manual transcriptions of the speech database and from normalized text corpora. In this work statistical language model was used (Jelinek, 1999). N-gram statistical language models are modelling the probability P(W) for the sequence of words $W=w_1,w_2,..,w_n$

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, .., w_{i-1}) \tag{5}$$

where $P(w_i \mid w_1,w_2,..,w_{i-1})$ is probability that word w_i follows the word sequence $w_1,w_2,..,w_{i-1}$. Since the weather domain corpus contains a limited amount of sentences a bigram language model is used to approximate P(W). The probability of the word w_i after word w_{i-1} in a bigram language model is calculated by

$$P(w_i \mid w_{i-1}) = \frac{N(w_{i-1}, w_i)}{N(w_{i-1})} \tag{6}$$

where:

$N(w_{i-1},w_i)$ is the frequency of the word pair (w_{i-1},w_i),

$N(w_{i-1})$ is the frequency of the word w_{i-1}.

One major problem with standard N-gram models is that they are estimated from some corpus, and because any particular training corpus is finite, some perfectly acceptable N-grams are bound to be missing from it (Jurafsky & Martin, 2000). To give an example from the domain of speech recognition, if the correct transcription of an utterance contains a bigram $w_{i-1}w_i$ that has never occurred in the training data, we will have $p(w_i \mid w_{i-1})=0$ which will preclude the recognition procedure from selecting the correct word sequence, regardless of how unambiguous the acoustic signal is.

Smoothing is used to address this problem. The term smoothing describes techniques for adjusting 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 recognition.

Perplexity of the language model represents the branching factor of the number of possible words branching from a previous word. Perplexity PP is defined as:

$$PP = 2^{H(L)} \tag{7}$$

where H(L) represents the entropy of the language and is approximated by:

$$H(L) = -\frac{1}{n}\log_2 P(w_1, w_2, \cdots, w_n) \tag{8}$$

where $P(w_1,w_2,..,w_n)$ is probability of the word sequence $w_1,w_2,..,w_n$, and n is the number of words in a sequence.
In all experiments bigram language model was used. Estimated perplexity of the radio part of the speech database bigram language model is 11.17 for weather domain and 17.16 for the news domain and perplexity of the telephone part of speech database is 17.97.

4. Experiments and results

The word recognition procedure computes the word sequence probability using the Viterbi search in the network of word hidden Markov models and a bigram language model. Word models are constructed from triphone models as shown in figure 4. Additional models for silence, breath noise, paper noise and restarts are used.
All word models are concatenated in parallel and form a single Hidden Markov Model, which is represented by a huge network of nodes. The analysis of an unknown observation sequence is performed by the Viterbi algorithm, producing the maximum a posteriori state sequence of the model with respect to the observed input vectors. Knowing the state sequence of the HMM we can decode the input sequence and transform it into a string of words. Because of 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. We use the Viterbi beam--search technique which expands the search only to states which probability falls within a specified beam. 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.

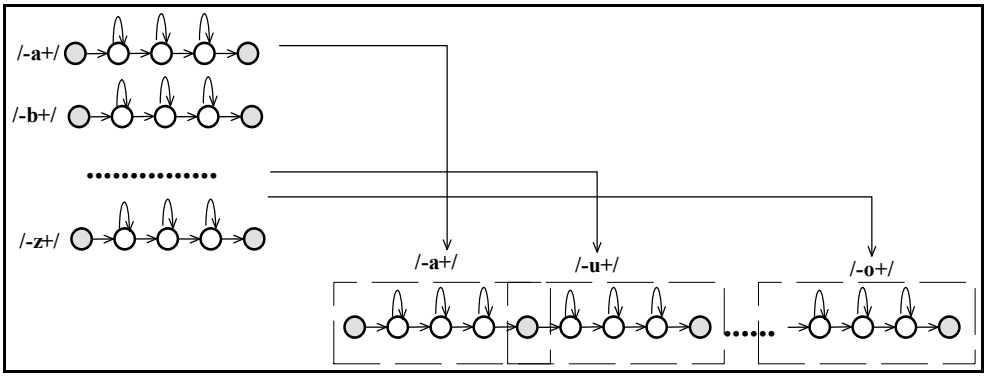

Fig. 4. Word models construction from triphone models.

So far we have performed speech recognition experiments using the radio speech database. The speech database contains weather forecast and news recordings. One part of the database (71%) was used for acoustic modelling and parameter estimation of context dependent phone models, while a smaller part (29%) of the database was used for recognition. All results are given for speaker independent recognition (2 male and 4 female speakers).

Speech recognition results for context-independent and context-dependent speaker independent recognition of the "clean" radio and noisy telephone speech are presented in tables 5 and 6 respectively. Word error rate (WER) results are given for 20 Gaussian mixtures. WER is computed according to:

$$WER = 100\% \left(\frac{W_S + W_D + W_I}{N} \right), \tag{9}$$

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.

The increase of the acoustic material in Croatian radio speech recognition resulted with 1.68% decrease of WER. Since the access to the weather information spoken dialog system is planned by telephone, the WER for the telephone data is quite promising. The word error rate for telephone data must be bellow 20% which will be achieved by incorporating more telephone speech in the acoustical model training procedure. And finally both recognition systems performed better when the number of tied states was reduced (using the same phonetic rules) and the number of Gaussian mixtures increased which indicates that more speech should be incorporated in the training of both recognizers for the use in the spoken dialog system.

	RADIO		**TELEPHO.**
	weath. forec.	**news**	**weath. repor.**
Duration [h]	8	13	6
No. words trained	1462	10230	1788
No. words recognized	1462	1462	1788
perplexity	11.17	17.16	17.97
No. Gauss. mix	**% WER**	**%WER**	**%WER**
1	18.7	18.49	30.41
5	13.35	13.13	25.21
10	11.57	11.36	23.18
15	11.11	10.91	22,52
20	**10.54**	**10.58**	**21.76**

Table 5. Croatian speech recognition results: WER computed using monophone HMMs with different number of Gaussian mixtures.

	RADIO		**TELEPHO.**
	weath. forec.	**news**	**weath. repor.**
No. Gauss. mix	**% WER**	**%WER**	**%WER**
1	17.27	14.69	27.16
5	12.76	10.63	21.82
10	11.28	9.56	20.83
15	11.02	9.20	20.49
20	**10.61**	**8.93**	**20.06**

Table 6. Croatian speech recognition results: WER computed using triphone HMMs with different number of Gaussian mixtures.

Graphs in figures 5 and 6 show the word accuracy for monophone and triphone Croatian speech recognition for radio and telephone speech for different numbers of Gaussian mixtures. Word accuracy WA is computed according to:

$$WA = 100\% \left(1 - \frac{W_S + W_D + W_I}{N} \right), \tag{10}$$

The presented recognition results are obtained using 553 tied states for 'clean' radio speech and 377 tied states for telephone speech. Further increase of Gaussian mixture did not increase the accuracy since the speech material is not big enough and a great number of triphones are not present in the training data.

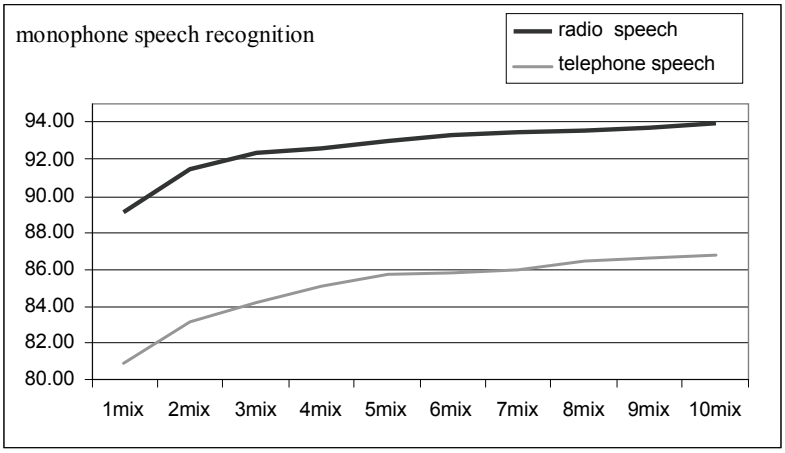

Fig. 5. Word accuracy using monophones for radio and telephone speech.

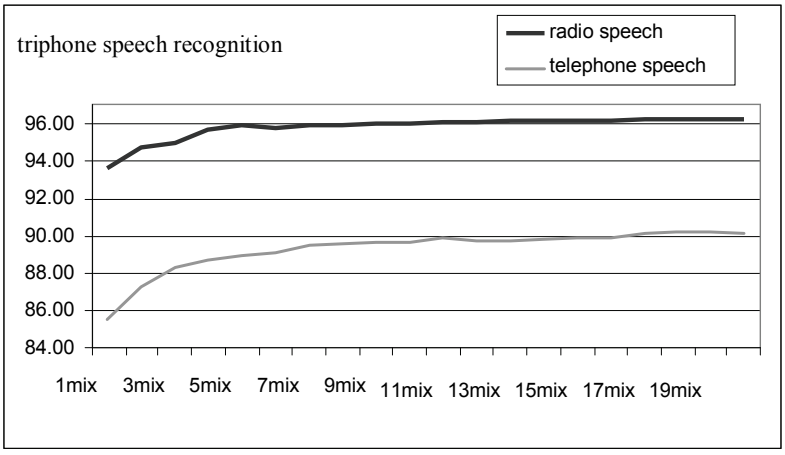

Fig. 6. Word accuracy using triphones for radio and telephone speech.

5. Conclusion

In the paper we described the context-dependent acoustic modelling of Croatian speech in the speech recognition system. An application specific Croatian speech corpus and Croatian phonetic rule were used for context-dependent hidden Markov models based speech recognition. Presented speech recognition system for radio and telephone data is planed for use in the Croatian weather information spoken dialog system.

Speech recognition experiments using context-independent and context-dependent acoustic models were prepared for "clean" radio and for noisy telephone speech. The WER for the radio weather domain is reduced to 10.61% by increasing the number of Gaussian mixtures. The radio speech WER was further reduced to 8.93% by adding the news related speech into acoustical modelling. For the telephone speech 20.06% WER was achieved. The achieved results for telephone speech recognition are promising for further actions in development of the dialog system.

In this work we have shown that the approach for speech recognition using context-dependent acoustical modelling is appropriate for rapid development of limited domain speech applications for low-resourced languages like Croatian. Croatian orthographic-to-phonetic rules are proposed for phonetic dictionary building. The developed Croatian multi-speaker speech corpus was successfully used for development of speech applications. Proposed Croatian phonetic rules captured adequate Croatian phonetic, linguistic and articulatory knowledge for state tying in acoustical models for the speech recognition system. Main advantage of the used approach lies in the fact that speech applications can be efficiently and rapidly ported to other domains of interest under the condition that an adequate speech and language corpus is available.

Since the telephone access to the spoken dialog system is planed, further improvements in speech recognition must be considered. Additionally work on including more speech especially spontaneous speech from different speakers in the corpus is in progress. Further research activities are also planed towards development of the speech understanding module in the dialog system and the speech synthesis module.

6. References

Alumäe, T. and L. Võhandu (2004). Limited-Vocabulary Estonian Continuous Speech Recognition Systems using Hidden Markov Models, *Informatica*, Vol.15(3), 303-314.

Anić, V. and J. Silić (2001). *Pravopis hrvatskoga jezika*, Novi liber. Zagreb. (in Croatian)

Barras, C., Geoffrois, E., Wu, Z. and M. Liberman (2000) Transcriber: use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools.* Vol. 33, No. 1-2.

Black, A., R. Brown, R. Frederking, R. Singh, J. Moody and E. Steinbrecher (2002). TONGUES: Rapid development of a speech–to-speech translation system, Proc. *HLT Workshop*, San Diego, California, pp. 2051-2054.

Duda, R., P. Hart and D. Stork (2001). *Pattern Classification*, John Wiley, Canada, 2001.

Dusan, S. and L. R. Rabiner (2005). On Integrating Insights from Human Speech Perception into Automatic Speech Recognition, *Proc. INTERSPEECH'05-EUROSPEECH,* Lisbon, Portugal, pp. 1233-1236.

Frederking, R., A. Rudnicky and C. Hogan (1997). Interactive Speech Translation in the DIPLOMAT Project, *Proc. Spoken Language Translation Workshop*, Madrid, 61-66.

Furui, S. (2005). 50 Years of Progress in Speech and Speaker Recognition, *Proc. SPCOM'05*, Patras, Grece, 1-9.

Furui, S., M. Nakamura and K. Iwano (2006). Why is Automatic Recognition of Spontaneous Speech So Difficult? *Proc. Large-Scale Knowledge Resources*, Tokyo, Japan, 83-90.

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

Graff, D.(2002) An overview of Broadcast News Corpora. *Speech Communication*, Vol. 37, Issues 1--2, pp. 15-26.

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

Hwang, M. Y., X. Huang and F. Alleva (1993). Predicting unseen triphones with senones, Proc. *IEEE ICASSP'*93, 1993, vol. 2, 311-314.

Jelinek, F. (1999). *Statistical Methods for Speech Recognition*, The MIT Press, USA.

Jurafsky, D., and J. Martin (2000). *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kominek, J., Bennett, C. and A. W. Black (2003). Evaluation and correcting phoneme segmentation for unit selection synthesis, *EUROSPEECH ´03*. ISCA. pp. 313-316. Geneva, Switzerland.

Kurimo, M., A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe and M. Saraclar (2006). Unlimited vocabulary speech recognition for agglutinative languages, *ACL HLT Conference,* 487-494. NewYork, USA.

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

Lihan, S., J. Juhar and A. Čižmar (2005). Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases, *Proc. INTERSPEECH'05-EUROSPEECH*, Lisbon, Portugal, 225-228.

Martinčić-Ipšić, S. and I. Ipšić (2004). Recognition of Croatian Broadcast Speech, *Proc. XXVII. MIPRO 2004,* Opatija, Croatia vol. CTS + CIS, p. 111-114.

Martinčić-Ipšić, S. and I. Ipšić (2006a). Croatian Telephone Speech Recognition, *Proc. XXIX. MIPRO 2008,* Opatija, Croatia, vol. CTS + CIS, 182-186.

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

Psutka, J., P. Ircing, J. V. Psutka, V. Radová, W. Byrne, J. Hajič, J. Mírovsky and S. Gustman (2003). Large Vocabulary ASR for Spontaneous Czech in the MALACH Project, *Proc. EUROSPEECH´*03, Geneva, Switzerland, 1821-1824.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, vol. 77, no. 2, 257-286.

SAMPA, ESPRIT project1541 Speech Assesment Method, created 1997 on initiative of Bakran and Horga, Phonetics and Linguistics University College London. (accessed May, 2002) `http://www.phon.ucl.ac-uk/hone/sampa/croatian.htm`

Scheytt, P., P. Geutner, A. Waibel (1998). Serbo-Croatian LVCS on the dictation and broadcast news domain, *Proc. IEEE ICASSP'*98, Seattle, Washington.

Schukat-Talamazzini, E. G. (1995). Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algoritmen, Vieweg Verlag, Braunschweig.

Shafran, I. and M. Ostendorf (2003). Acoustic model clustering based on syllable structure, *Computer Speech and Language*, vol. 17, 311-328,

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

Skripkauskas, M. and L. Telksnys (2006). Automatic Transcription of Lithuanian Text Using Dictionary, *Informatica*, 17(4), 587-600.

Vaičiūnas A. and G. Raškinis (2005). Review of statistical modeling of highly inflected Lithuanian using very large vocabulary, *Proc. INTERSPEECH'05-EUROSPEECH*, Lisbon, Portugal, 1321-1324.

Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland (2002). *The HTK Book*, (for HTK Version 3.2). Cambridge University Engineering Department, Cambridge, UK.

Young, S., J. Odell and P. Woodland (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling, *ARPA HLT Workshop*, Plainsboro, NJ, Morgan Kaufman Publishers, 307-312.

Žibert, J., S. Martinčić-Ipšić, M. Hajdinjak, I. Ipšić and F. Mihelič (2003). Development of a Bilingual Spoken Dialog System for Weather Information Retrieval, *Proc.* EUROSPEECH´03, Geneva, Switzerland, vol. 1, 1917-1920.

Hidden Markov Model Toolkit, Version 3.2, Cambridge University Engineering Department, Cambridge, UK, 2002. http://htk.eng.cam.uk/

Advances in Speech Recognition

Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ivo Ipsic and Sanda Martincic-Ipsic (2010). Croatian Speech Recognition, Advances in Speech Recognition, Noam Shabtai (Ed.), ISBN: 978-953-307-097-1, InTech, Available from: http://www.intechopen.com/books/advances-in-speech-recognition/croatian-speech-recognition

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821