INFORMATICA, 2008, Vol. 19, No. 2, 227–254
© 2008 *Institute of Mathematics and Informatics, Vilnius*

Acoustic Modelling for Croatian Speech Recognition and Synthesis

Sanda MARTINČIĆ–IPŠIĆ
Department of Informatics, Faculty of Philosophy, University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
e-mail: smarti@ffri.hr

Slobodan RIBARIĆ
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
e-mail: slobodan.ribaric@fer.hr

Ivo IPŠIĆ
Faculty of Engineering, University of Rijeka
Vukovarska 58, 51000 Rijeka, Croatia
e-mail: ipsic@riteh.hr

Received: May 2007

Abstract. This paper presents the Croatian context-dependent acoustic modelling used in speech recognition and in speech synthesis. The proposed acoustic model is based on context-dependent triphone hidden Markov models and Croatian phonetic rules. For speech recognition and speech synthesis system modelling and testing the Croatian speech corpus VEPRAD was used. The experiments have shown that Croatian speech corpus, Croatian phonetic rules and hidden Markov models as the modelling formalism can be used to develop speech recognition and speech synthesis systems in parallel for a highly flective and free order language like Croatian. We propose an evaluation procedure for speech synthesis, which combines an objective and a subjective evaluation approach and we present the achieved evaluation results. The proposed procedures for Croatian acoustic modelling were developed as parts of speech interfaces in a spoken dialog system for a limited weather forecast domain.

Keywords: acoustic modelling, context-dependent acoustic units, large vocabulary speech recognition, hidden Markov models based speech synthesis.

1. Introduction and Related Work

The paper describes procedures for context-dependent acoustic modelling of Croatian speech in the speech recognition and in the speech synthesis module of 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 (TTS) synthesis. In this paper ASR and TTS modules based on data-driven statistical and rule-based knowledge approach are 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 and 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 and Telksnys, 2006; Vaičiūnas and Raškinis, 2005) and Estonian (Alumäe and Võhandu, 2004) with applications for dialog systems (Žibert *et al.*, 2003), dictation (Psutka *et al.*, 2003) or automatic transcriptions (Skripkauskas and Telksnys, 2006) have been reported lately.

In (Lihan *et al.*, 2005) the Czech and Slovak telephone (*SpeechDat-E* corpus) monolingual and cross lingual speech recognition systems are presented. The 9675 word-internal triphone acoustic model of the Czech speech recognition system is trained out of 42 Czech phonemes and allophones and 3 phonemes in foreign names. In the Slovak acoustic model 9121 word-internal triphone models were trained for 43 phonemes and 8 allophones, including diphthongs. Authors considered 33 out of 51 modelled phonemes identical. The Czech recognition system was trained from 36449 utterances containing 19313 different words. The Slovak recognition system was trained on 32855 utterances containing 14907 different words. The speech feature vectors were 39-dimensional Mel-Frequency Cepstral Coefficients (MFCC), including the first and the second order deltas. The number of continuous Gaussian-mixture components was increased to 32. The number of triphone HMM states was tied up to 13.4% in the Slovak and up to 14.7% in the Czech speech recognition system. For both languages the six recognition scenarios were evaluated: from recognizing simple yes/no answers with 0.43% Slovak word error rate (WER) and 0.93% Czech WER, to recognizing 2586 words with 10.46% Slovak WER and 7.18 Czech WER.

In (Skripkauskas and Telksnys, 2006) automatic phonetic transcription of Lithuanian text is presented. The phoneme-to-grapheme transformation is based on predefined set of rules and a dictionary with 22000 entries. Proposed automatic transcription will be used in the large vocabulary Lithuanian speech recognition system. The system automatically transforms text into 225 phonemes (68 basic and 157 diphthongs) with 94–98% correctness depending on the text genre.

In (Vaičiūnas and Raškinis, 2005) the language modelling of Lithuanian for LVASR is presented. The trigram, class, cache and morphological LM are compared according to the WER of ASR system. The ASR system was trained on 13 hours of one speaker speech (4461 different words, 1096 utterances). The 39 dimensional acoustic feature vectors with 13 MFCC coefficients and their deltas and accelerations were computed every 10 ms over a 25 ms window. The test was performed on the 1000 acoustically most probable word sequences. The best WER of 43.56% was achieved with the class language model, and 46.76% WER was achieved with morphological LM.

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 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 standard phoneme set includes 30 phonemes, where the set of vowels is extended with the vibrant vowel /r/.

The use of anaphora is very frequent in Croatian and the referenced term can be found in previous sentences but also in the previous paragraphs or sections which aggravates the automatic extraction of meaning and understanding of the Croatian speech. In the development of speech applications for Croatian large number of out of vocabulary words (OOV) is expected, mainly due to the high flectivity. The number of 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). The TONGUES project continued with this research in direction towards large Croatian vocabulary recognition system and unit selection speech synthesis (Black *et al.*, 2002). The Serbian, Bosnian and Croatian dictation system trained on broadcasted news is discussed in (Scheytt *et al.*, 1998). The speech was recorded from the local television and radio programs and from local speakers, with no exact differentiation between Croatian, Bosnian and Serbian speech. High OOV rates for Croatian are reported in (Geutner *et al.*, 1998) as well as fast speaking rates in the field recordings.

The HMM based trainable synthesis systems for Japanese (Tokuda *et al.*, 2000), English (Acero, 1999; Donovan and Woodland, 1999; Tokuda *et al.*, 2002b) and for a few other languages (Barros *et al.*, 2005; Vesnicer, 2003) were developed. For Croatian speech synthesis, so far only experiments using unit selection speech synthesis has been reported (Black *et al.*, 2002).

In (Tokuda *et al.*, 2002b) an HMM-based speech synthesis system originally developed for Japanese was applied to English. The same research group developed the HTS Toolkit (HTS, 2004) as an extension to the HTK Toolkit (Young *et al.*, 2002). The system was trained on speech from 524 sentences from the CMU Communicator database. The speech signal is sampled at 16 KHz, windowed by a 25 ms Blackman window with a 5 ms shift. The speech feature vectors consist of spectral part (25 mel-cepstral coefficients with delta and delta-delta coefficients) and excitation part (log fundamental frequency with delta and delta-delta coefficients). The monophone multispace probability distribution HMMs (MSD-HMM) were trained for the English phoneme set. The clustering of the context dependent MSD-HMM's states considered 40 different contextual factors like: phoneme predecessor, successor, position in the syllable, number of phonemes in syllable, number of syllables in word, stress of syllable, accent of syllable, etc. The separated decision trees were constructed for spectrum, fundamental frequency and duration, because of the different contextual factors influence. The speech signal is generated using the mel log spectral approximation (MLSA) filter from stream of mel-cepstral coefficients and fundamental frequency generated from MSD-HMMs according to the sequence of phonemes in the input text and their duration.

In (Vesnicer, 2003) the Slovene HMM-based speech synthesis system was trained on 39 minutes of Slovene speech. The monophone MSD-HMMs were trained for 38 Slovene phonemes (differentiation between short and long vowels). The triphone context-dependent models were constructed from monophones taking only the immediate preceding or succeeding phoneme as contextual information. The 57 Slovene phonetic rules were used for state clustering. The decision trees were constructed for spectral and fundamental part: for each state and each phoneme separately. Other parameters of the HMM-TTS systems were the same as in (Tokuda *et al.*, 2002b). The synthesized speech was evaluated subjectively by the preference test. The 14 evaluator compared different speech signals generated from monophone or triphone HMM-TTS (with or without dynamic part of the features) with an academic diphone concatenation synthesis system. The generated speech was evaluated objectively as well. The Euclid distance computed with the dynamic time warping procedure was calculated for different kinds of generated speech feature vectors and natural speech features vectors.

In (Barros *et al.*, 2005) the HMM based speech synthesis system is applied to Portuguese. The system was trained on 21 minutes of speech in 104 utterances. The monophone MSD-HMMs were trained for the Portuguese phoneme set. The 46 different contextual factors were considered in context dependent modelling. For tree based context-clustering 37 Portuguese phonetic categories were considered. Other MSD-HMM parameters are the same as in the (Tokuda *et al.*, 2002b). The generated Portuguese speech was evaluated by the comparative mean opinion score (MOS) test in which HMM-TTS was positioned between the best commercial TTS system and an academic TD-PSOLA system. The subjective test was performed with 27 listeners.

The common formal (HMM) and procedural foundations of developed speech system enabled the transfer of linguistic knowledge, acoustic modelling knowledge and technological knowledge from the speech recognition to the speech synthesis systems (Ostendorf and Bulyko, 2002). The use of speech recognition technology in speech synthesis, especially for the limited domain of interest (for instance weather forecast domain) has emerging interest in the research community (Ostendorf and Bulyko, 2002). With this motivation we have based this work on fundamental assumptions that systems are based on the same:

- acoustic units, used for speech recognition and speech synthesis acoustic modelling,
- acoustic modelling procedure, used for speech recognition and synthesis training,
- speech data, used for acoustic model training,
- automatic segmentation procedure and
- phonetic rules, used for state tying in context dependent acoustic modelling.

In this work we have shown that the use of the common approach for acoustic modelling in speech recognition and synthesis systems 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 VEPRAD 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 and the speech synthesis system. The proposed combination of subjective and objective evaluation of synthesized speech enables verification of subjective evaluation scores and better insights of user opinions and needs.

The Croatian speech recognition and speech synthesis 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), while for speech synthesis training the HTS tool (HTS, 2004), which is as an extension of the HTK, was used.

Since the main resource in a spoken dialog system design is the collection of speech material, the Croatian domain related speech corpus is presented in Section 2. Orthographic-to-phonetic rules used in the phonetic dictionary preparation are shown as well. Further the acoustic modelling 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 at the end of the Section 3. Section 4 explains the Croatian HMM based speech synthesis, which is based on the same context-dependent acoustic model as the one used in the speech recognition experiments. The evaluation procedure for speech synthesis and achieved results are discussed. We conclude with discussion on advantages of the same acoustical modelling approach for Croatian speech recognition and speech synthesis and description of current activities and future research plans.

2. The Croatian Speech Corpus

The Croatian speech corpus VEPRAD includes news, weather forecasts and reports spoken within broadcasted shows of the national radio (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 VEPRAD corpus is a multi-speaker speech database and contains 13 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 corpus statistics is shown in Table 1.

The first part, VEPRAD radio 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. VEPRAD 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 VEPRAD speech database is strictly domain oriented. From the VEPRAD radio part one male and two female speakers were selected for the broadcasted news recording and the voice of the selected male speaker was used for speech synthesis training.

The second part, VEPRAD telephone corpus, 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 VEPRAD 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 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.

Table 1

Croatian speech corpus statistics

VEPRAD	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
Overall VEPRAD	**1464**	**12707**	**16**	**21**	**235430**	**10898**	**1135**

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). The VEPRAD speech database is used for training of acoustic and language models of the speech recognition as well as for the speech synthesis system.

2.1. *Phonetic Dictionary*

We have 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
- word with suffixes ***naest*** were phonetically transcribed as [*n a j s t*].

The phonetic dictionary comprises all words in transcription texts. All word formats (different cases, genders and numbers of the same basic word format) are considered as a new word in the dictionary. The current phonetic dictionary contains 10898 different words. The fact that Croatian language is highly flected 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. All accented vowels were marked with a colon. For the words that can be pronounced in more correct ways the position of the really accented vowel was marked. The accented vowels were used in the acoustic modelling procedure for the speech synthesis system.

3. Speech Recognition

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. 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 kth mixture in state s_j,
Σ_{jk} is covariance matrix of the kth mixture in state s_j,
M is the number of mixture components and
c_{jk} is the weight for the kth mixture in state s_j satisfying the condition:

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

For the estimation of continuous HMM parameters iterative Baum–Welch procedure is used. The 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 (Duda *et al.*, 2001; Huang *et al.*, 2000). 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 (Huang *et al*., 2000; Rabiner, 1989). Viterbi algorithm is a dynamic programming algorithm that decodes the state sequence according to the observed output sequence.

Fig. 1 presents main steps performed in the Croatian speech recognition system de-

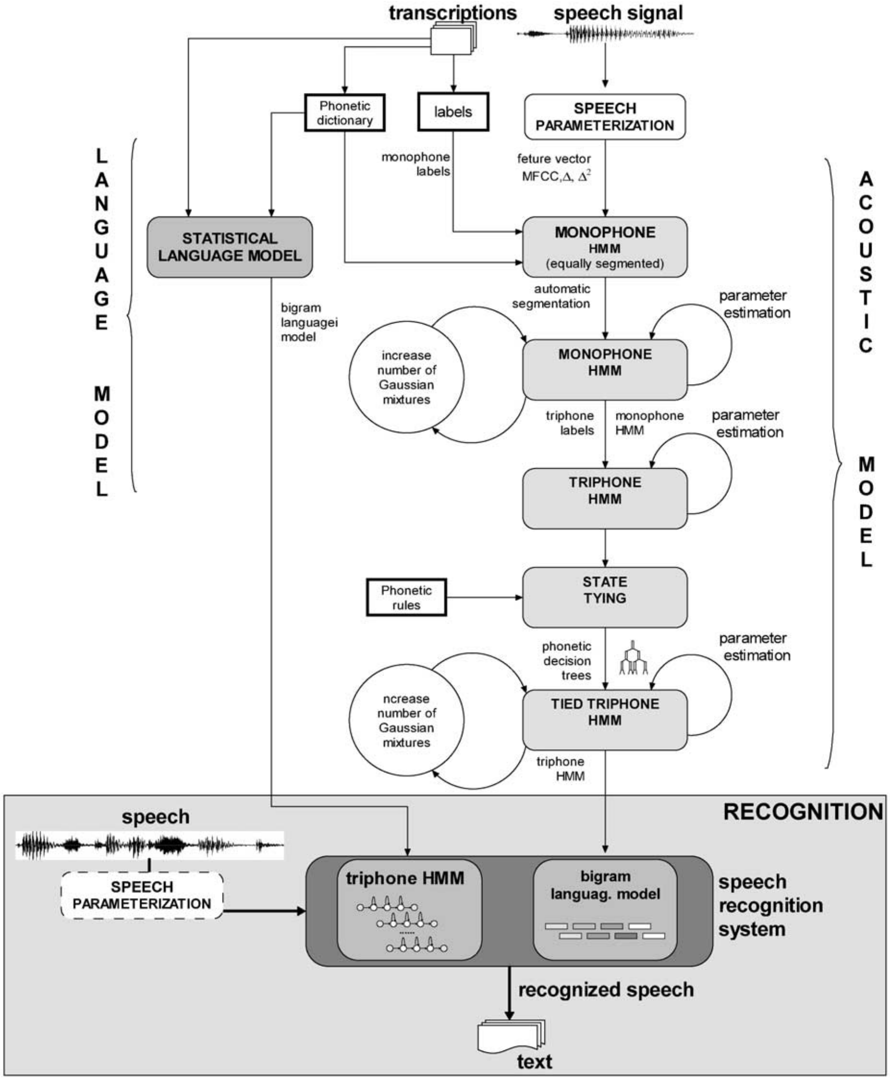

Fig. 1. Development of the Croatian speech recognition system.

velopment, 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 HMM 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.

3.1. *Acoustic Modelling*

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 modelling chosen acoustic unit HMMs. 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 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 and Lamel, 2003; Lee *et al.*, 1990) or sub phonemic units like senones (Hwang *et al.*, 1993). Some speech recognition systems are modelling syllables (Shafran and 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).

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-word 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
The number of monophone and triphone models and seen triphones percentage per parts of the VEPRAD corpus

VEPRAD	**No. monophones**	**No. triphones**			**% seen**
		possible	**all**	**seen**	
radio weather	29+4	$33^3 = 35937$	31585	4042	12.80%
radio news	30+4	$34^3 = 39304$	36684	7931	21,62%
telephone	29+4	$33^3 = 35937$	31585	4618	14.62%

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. The lack of speech data is overcome by a phonetically driven state tying procedure.

Speech feature vector

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

Context independent acoustic model

The training of speech recognition and synthesis 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 in the used monophone recognition system.

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 83 phonetic rules (166 Croatian phonetic questions about left and right context (Martinčić–Ipšić and 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.

State tying

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

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

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.

In the speech synthesis system the same 166 Croatian phonetic questions (83 rules) from the speech recognition system were used. The clustering trees for spectral parameters, fundamental frequency F0 and duration were built separately as shown at Fig. 2

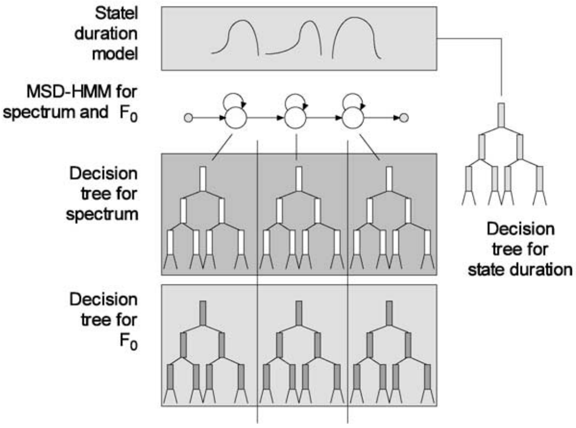
Statel duration model
MSD-HMM for spectrum and F_0
Decision tree for spectrum
Decision tree for F_0
Decision tree for state duration

Fig. 2. Decision trees for spectrum, pitch and duration in the Croatian speech synthesis system.

(Yoshimura *et al*., 1999). The clustering trees were built separately because different context clustering factors are relevant for spectral part clustering, pitch clustering and duration clustering but the same Croatian phonetic rules were used.

Table 3 contains the most frequently used Croatian phonetic questions in the phonetic decision trees in the speech recognition and speech synthesis 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).

Table 3

The most frequently used Croatian phonetic questions in speech recognition and synthesis

Recognition				**Synthesis**	
VEPRAD radio		**VEPRAD telephone**		**VEPRAD radio**	
Phonetic question	**No.**	**Phonetic question**	**No.**	**Phonetic question**	**No.**
R_Front	811	R_Front	522	L_Front	65
L_Front	797	L_Front	498	L_Central	49
L_Vowel-Open	635	L_Central	348	R_Central	46
L_Central	594	R_Vowel-Open	336	R_Vowel-Unacccented	36
R_Vowel-Open	561	R_Central	312	L_Nasal	36
L_Consonant-Voiceless	432	L_Vowel-Open	312	R_Vowel-Accented	34
R_Vowel	384	L_Consonant-Voiceless	222	R_e:(Accented)	33
R_Consonant-Voiceless	357	R_Vowel	221	R_n	33
D_Central	355	D_Consonant-Voiceless	216	D_Front	31
L_Nasal	338	L_Consonant-Closed	201	L_Vowel-Accented	31

As expected and reported for other languages (Gauvain and 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 result are presented for left and right coarticulation context and for the stable part of the phoneme summarized, 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.

In the speech synthesis phonetic clustering task beside the most common Croatian phonetic rules are phonetic rules investigating the position of the single phoneme in the coarticulated context. For example checking for phoneme /n/ in the right context was performed 33 times. In the speech synthesis phonetic clustering left-right context pairs are not so obvious.

3.2. *Language Modelling*

Language model is an important part of the speech recognition system. 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 $W = w_1, w_2, .., w_n$

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

where $P(w_i|w_1, w_2, .., w_{i-1})$ 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

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

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}.

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 VEPRAD radio bigram language model is 11.17 for weather domain and 17.16 for the news domain and perplexity of the VEPRAD telephone is 17.97.

3.3. *Speech Recognition Experiments*

In VEPRAD radio speech recognition system 4135 (71%) utterances form 8 male and 8 female speakers were used for training and 1710 (29%) utterances from 3 male and 6 female speakers were used for testing. In VEPRAD telephone speech recognition system 2036 utterances (62%) where used for acoustic modelling and parameter estimation of context dependent phone models and 1241 utterances (38%) were used for recognition. Speech from 3 female and 3 male meteorologists was used for training and speech from 2 male and 4 female meteorologists was used for testing.

Speech recognition results

Speech recognition results for context-dependent and speaker independent recognition of the "clean" radio and noisy telephone speech are presented respectively in the Table 4. The number of different words is in the first raw. Word error rate (WER) results are given for 20 Gaussian mixtures. WER is computed according to

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

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.

Table 4

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

	VEPRAD		
	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	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**

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.

4. Croatian HMM based Speech Synthesis

In HMM speech synthesis the modified Viterbi algorithm is used to find the most probable path through HMMs states that can generate the speech signal feature vectors. The speech signal can be then synthesised from so generated feature vectors using the source-filter model. For phones duration period the feature vectors consisting of mel-cepstrum parameters and excitation parameters are generated.

Fig. 3 presents the HMM speech synthesis system. The context dependent triphone HMMs are trained on Croatian speech the same way as for the speech recognition, except that the speech feature vector is extended with pitch information. The cepstral, pitch and duration parameters of the trained HMM are used to generate the speech signal for the input text. For the speech generation the source-filter model is used.

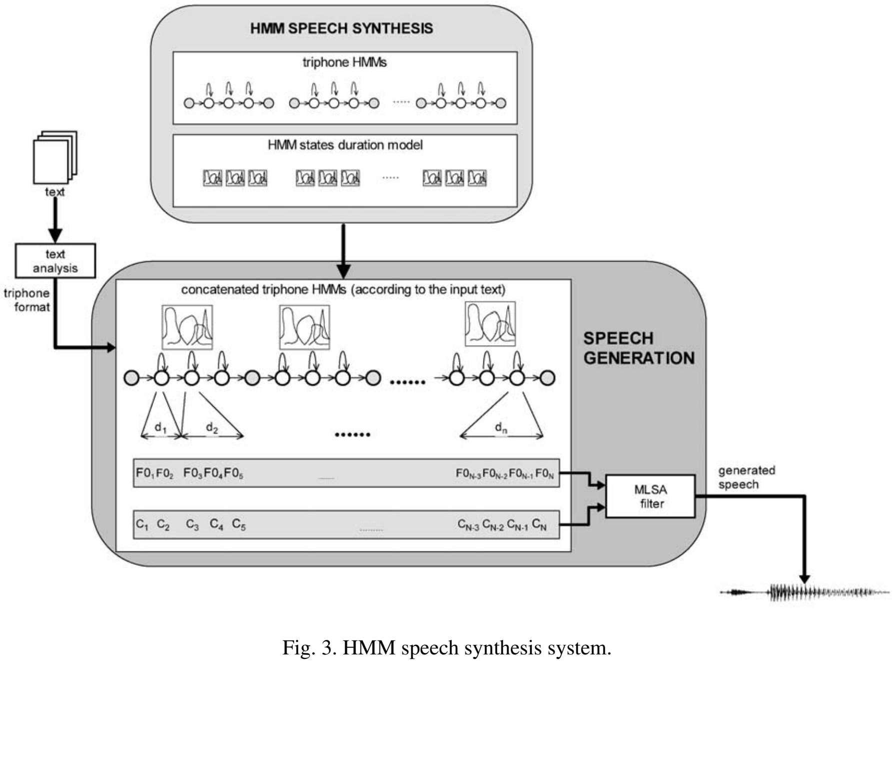
HMM SPEECH SYNTHESIS
triphone HMMs
HMM states duration model
text
text analysis
triphone format
concatenated triphone HMMs (according to the input text)
SPEECH GENERATION
d_1
d_2
d_n
$F0_1 F0_2$ $F0_3 F0_4 F0_5$ $F0_{N-3} F0_{N-2} F0_{N-1} F0_N$
C_1 C_2 C_3 C_4 C_5 C_{N-3} C_{N-2} C_{N-1} C_N
MLSA filter
generated speech

Fig. 3. HMM speech synthesis system.

4.1. *The Source-Filter Speech Model*

The source-filter model of human speech production is the basis of many speech synthesis approaches. Speech can be viewed as the output of a linear filter excited by a sound source. Typically the sound source has a voiced sound component and an unvoiced sound component. The filter simulates the frequency response of the vocal tract and shapes the spectrum of the signal generated by the source. In a source filter model an impulse train for voiced sounds or white noise for unvoiced sounds is used as input to a time varying filter. For voiced fricatives mixed model is used by combining impulse train and white noise at the same time.

The z-transform of the speech signal S(z) can be modeled using

$$S(z) = U(z)H(z), \tag{10}$$

where $U(z)$ is the excitation model and $H(z)$ the transfer function of the filter model representing the vocal tract response. The vocal tract response $H(z)$ can be constructed from cepstral coefficients $c(m)$

$$H(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, \tag{11}$$

where M is the number of mel cepstral coefficients, $\tilde{z}$ is frequency transformation function

$$\tilde{z} = \frac{1-\alpha z^{-1}}{z^{-1}-\alpha}, \quad |\alpha| < 1, \tag{12}$$

and α is frequency compression parameter, which is used to compress mel-scale in order to approximate the human sensitivity to the frequencies of the speech signal.

The transfer function $H(z)$ can be realized by the exponential function in the MLSA filter (Imai, 1983). The speech signal can be reconstructed from MFCC parameters approximating the transfer function of the source-filter model using the MLSA filter (Fukada *et al.*, 1992).

$$H(z) = \exp\left(\sum_{m=0}^{M} b(m)\Theta_m(z)\right), \tag{13}$$

where

$$\Theta_m(z) = \begin{cases} 1, & m = 1, \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)}, & m \geqslant 1, \end{cases}$$

$b(m)$ coefficients are linear transformation of cepstral coefficients $c(m)$

$$b(m) = \begin{cases} c(M), & m = M \\ c(m) - \alpha b(m+1), & 0 \leqslant m \leqslant M-1. \end{cases} \tag{14}$$

4.2. *The HMM Speech Synthesis*

The context independent and context dependent (triphone) HMMs were embedded-trained on the features vectors consisting of spectrum, pitch and their dynamic features simultaneously in a unified framework of multi-space probability distribution HMMs (Tokuda *et al.*, 2000). Since the observation sequence of fundamental frequency is composed of one dimensional continuous function for voiced voices and a constant for the unvoiced speech segments, multispace probability distributions were used (Tokuda *et al.*, 2002a). Although the HMMs were trained simultaneously the context-dependent clustering of Gaussian distributions was performed independently for spectrum, pitch and duration because of the different clustering factor influence (Yoshimura *et al.*, 1999). In state tying the phonetic knowledge captured in the same 83 Croatian phonetic rules as in the speech recognition system was used. State duration densities for the speech feature vectors generation are estimated by probabilities obtained in the last iteration of embedded re-estimation. State duration densities were modelled by multivariate Gaussian distributions.

Speech feature vector

The speech signals where windowed using a 25 ms Blackman window and 5 ms frame shift. The feature vector consists of spectral and excitation (pitch) parameters as shown in Fig. 4. The spectral feature vector consists of 25 mel-cepstral coefficients including the zeroth coefficient and its delta and acceleration coefficients. The pitch feature vector consists of logF0 and its dynamic parameters (delta and acceleration).

Speech signal generation

In the synthesis part, from the set of concatenated context-dependent HMMs according to the symbols in the entry text, the speech parameters are generated. The generated excitation parameters and mel-cepstrum parameters are used to generate the speech signal using the source-filter model. Since the last step in the training procedure was HMM parameters generation for unseen triphones, according to their classification in the phonetic decision trees, the unseen triphones can be synthesized as well.

Let us have the HMM Φ that contains concatenated elementary triphone or monophone HMMs that correspond to the symbols in the word w, which has to be synthesized.

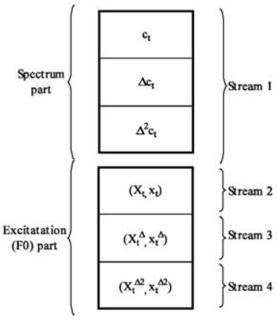

Fig. 4. The HMM state output feature vector.

The aim of the speech synthesis is to find the most probable sequence of states features vectors x from the HMM Φ. x_{s_i} is the M-dimensional generated feature vector at the state s_i of the model Φ

$$x_{s_i} = \left(x_1^{(s_i)}, x_2^{(s_i)}, .., x_M^{(s_i)}\right)^T. \tag{15}$$

From model Φ we want to generate a sequence of features vectors $X' = x_{s_1}x_{s_2}..x_{s_L}$ of length L maximizing the overall likelihood $P(X|\Phi)$ of a HMM

$$X' = \arg\max_X \left\{P(X|\Phi)\right\} = \arg\max_X \left\{\sum_S P(X|S,\Phi)P(S|\Phi)\right\}, \tag{16}$$

where the $S = s_1, s_2, .., s_L$ is the path through the states of the model Φ. The overall likelihood of the model $P(X|\Phi)$ is computed by adding the product of joint output probability $P(X|S,\Phi)$ and state sequence probability $P(S|\Phi)$ over all possible paths S.

Practically the Viterbi approximation is used because theoretically all possible paths through the model have to be searched for, which is too time consuming

$$X' \approx \arg\max_X \left\{P(X|S,\Phi)P(S|\Phi)\right\}. \tag{17}$$

The state sequence S' of the model Φ can be maximized independently of X

$$S' = \arg\max_S \left\{P(S|\Phi)\right\} \tag{18}$$

Let's assume that the output probability distribution of each state s_i is one Gaussian density function with a mean vector μ_i and covariance matrix Σ_i. The HMM model Φ is set of all means and covariance matrices for all N states:

$$\Phi = (\mu_1, \Sigma_1, \mu_2, \Sigma_2.., \mu_N, \Sigma_N). \tag{19}$$

Then the log-likelihood of (17) is given by

$$\begin{aligned}\ln P(X|S,\Phi) &= -\frac{LM}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{L}\ln|\Sigma_{s_t}| \\ &\quad - \frac{1}{2}\sum_{t=1}^{T}(x_t - \mu_{s_t})^T\Sigma_{s_t}^{-1}(x_t - \mu_{s_t}).\end{aligned} \tag{20}$$

Maximizing X in (17) leads to the trivial solution$X' = (\mu_{s_1}\mu_{s_2}..\mu_{s_L})$, where the sequence is equal to the means of the corresponding states. Such a solution does not represent well the speech because of the discontinuities at the state boundary. This can be solved by extending the feature vectors with first and second differentials

$$X_{s_i} = \left((x_{s_i})^T, (\Delta x_{s_i})^T, (\Delta^2 x_{s_i})^T\right)^T. \tag{21}$$

In (Tokuda *et al.*, 1995b) a fast algorithm is given for the solution of Eq. 20.

According to the phoneme sequence in text labels the context-dependent HMMs were concatenated. State durations of the sentence are determined by maximizing the likelihood of state duration densities. According to the obtained state sequence and state duration the sequence of mel-cepstral coefficients and F0 values including voiced/unvoiced decisions are determined by maximizing the output probability of HMMs. Finally the speech is synthesized from generated mel-cepstral feature vectors and pitch values using the MLSA filter (Tokuda *et al.*, 1995a).

4.3. *Speech Synthesis Training*

The hidden Markov model based trainable speech synthesis use the Croatian speech for the training of context-independent and context-dependent acoustic models and uses HMM as a generative model for speech production.

The Croatian HMM based speech synthesis was trained on 2.5 hours of selected male speaker's speech from the VEPRAD corpus (Martinčić–Ipšić and Ipšić, 2006b). In the synthesis part 2332 utterances were used for speech synthesis training. For 6222 different words phonetic dictionary contains accented words and phonetic transcriptions. The speech synthesis system, in contrary to the speech recognition system, differentiates between accented and non accented vowels. Accented vowels are marked by a: including the occurrence of r as a vowel.

The HMMs were embedded-trained on the features vectors consisting of spectrum, pitch and their dynamic features simultaneously in a unified framework of multi-space probability distribution HMMs and multi-dimensional Gaussian distributions. Since the observation sequence of fundamental frequency is composed of one dimensional continuous function for voiced voices and a constant for the unvoiced speech segments, multi-space probability distribution is used (Tokuda *et al.*, 2000). The HMM state output feature vector consists of spectrum and excitation in a multispace probability distribution part as shown in Fig. 4. State duration densities for the speech feature vectors generation are estimated by probabilities obtained in the last iteration of embedded re-estimation.

4.4. *Speech Synthesis Evaluation*

The synthesized speech was evaluated subjectively by the mean opinion score of the 21 evaluator, and objectively by the speech recognition system. The combined approach compared objective and subjective results. Fig. 5 presents evaluation of synthesized speech using: the subjective (a), the objective (b) and the combined (c) test method. For the test purposes 30 s (76 words in 9 sentences) of synthesized Croatian speech concerning weather forecast, temperature, sea and wind conditions, visibility at the see, the river water level was generated.

Subjective evaluation

The generated speech was played to the evaluators which were asked to fill in the questionnaire. The questionnaire consists of three parts: the first part is investigating the

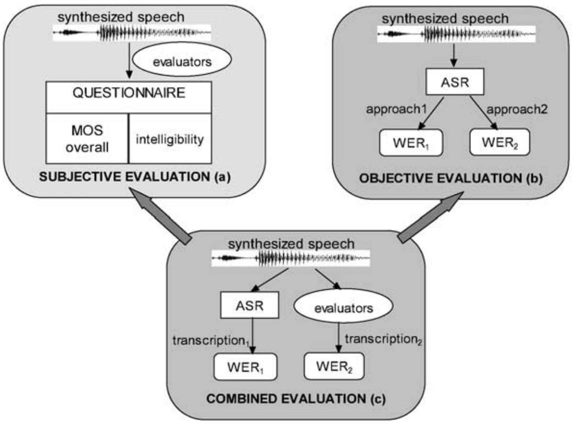

Fig. 5. Evaluation of synthesized speech using: the subjective (a), the objective (b) and the combined (c) test method.

evaluator's opinion about the overall quality of generated speech, intelligibility, naturalness, pronunciation and usability in the telephony applications. For each attribute, except usability, five possible answers were offered in the questionnaire, and three possible answers were offered for usability. Later, to each qualitative answer score form 1 to 5, or 1 to 3 was assigned. The second part of the questionnaire checked the intelligibility with 4 questions. The third part of the questionnaire allowed evaluators to write down their remarks and opinion.

The female (12) and male (9) evaluators were mainly students of informatics, students of linguistics, some professionals with experience in speech technologies, linguists, and one person with vision impairment who uses a screen reader on daily basis. The results were collected according to the mean opinion score (MOS) scale. Fig. 6 presents the MOS results. The overall score is 3.02. The highest score is achieved in intelligibility and the lowest at overall quality.

The second part of the questionnaire investigates the intelligibility of synthesized speech by 4 questions about weather condition at the Adriatic coast, about the air temperature, about visibility at the sea and about the river water level. The results are presented

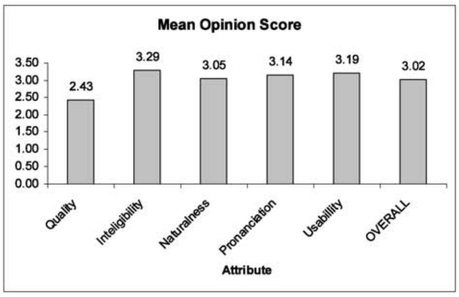

Fig. 6. The MOS results for the synthesized speech.

in Fig. 7. Intelligibility test showed high percentage of correct answers (88.69%), which is consistent wit the MOS rate.

Objective evaluation

The developed speech recognition system for weather was used for the objective evaluation of the synthesized speech. Achieved correctness and accuracy are presented in Table 5. In the first approach (1) all incorrectly recognized silences and inspirations are counted as errors. In the second approach (2) all, correctly or incorrectly, recognized silences and inspirations were ignored.

Combined evaluation

The 5 evaluators were additionally asked to write down the transcription of the synthesized speech. Two of them had experience with speech transcription, one with the weather domain and two of them had no previous contact with synthesized speech at all. Their transcription correctness and accuracy was calculated in the HTK environment which makes it comparable with the correctness and accuracy achieved in the objective test. In the objective test the same synthesized speech was recognized by the speech recognition system. In the evaluation incorrectly recognized silences and inspirations were ignored. Fig. 8 presents the comparison of results achieved by speech recognition system and by evaluators. Better results are achieved by evaluators. The 31% WER achieved by the recognizer is consistent with the MOS score of 3.02. Fig. 9 compares achieved ASR results with overall evaluator's results. The results are compared using the second approach.

The evaluators in their remarks noticed that the intelligibility of the synthesized speech is lowered because of the fast speaking rate, "sharp" pronunciation and because of the too short pauses between sentences. Two of them noted that certain amount of effort

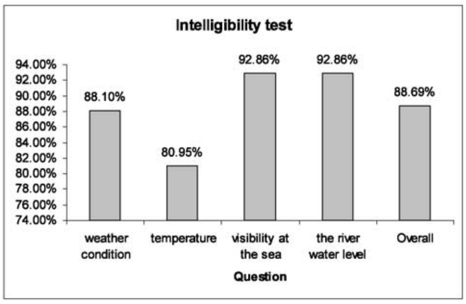

Fig. 7. Intelligibility test results.

Table 5

ASR correctness and accuracy of synthesized speech

Synthesized speech	No. words	Corr. %	Accurac. %
approach (1)	92	75.00	51.09
approach (2)	76	84.21	68.42

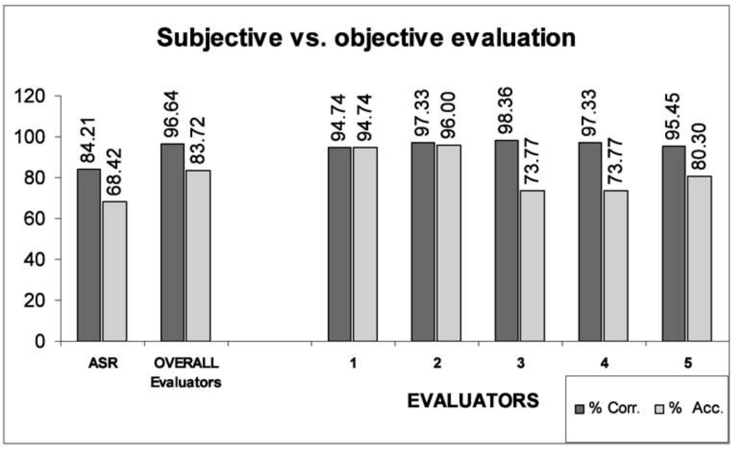

Fig. 8. The comparison of subjective and objective synthesis test correctness and accuracy.

is needed for complete understanding. The remark about higher attention was expected since majority of evaluators (15) had no previous contact with artificial speech of any kind.

The combination of objective and subjective evaluation results put different perspectives on the quality of the synthesized speech. Although the overall score is only with 3.02, the human evaluators performed better on the comparison test. Subjective evaluation enabled better understanding of user opinions and needs and therefore should not be omitted in the synthesis evaluation, while the objective evaluation enabled different perspective on the generated speech.

5. Conclusion

In the paper we described the context-dependent acoustic modelling of Croatian speech in the speech recognition and speech synthesis systems. The same Croatian speech corpus and Croatian phonetic rule were used for context-dependent hidden Markov models based speech recognition and speech synthesis. Presented speech recognition system for radio and telephone data and HMM based speech synthesis are 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 synthesized speech was qualified as intelligible with an overall rate 3.02 according to MOS by evaluators. The developed weather ASR system was used for objective evaluation of synthesized speech and 48.91% WER and 31.58% WER where achieved.

The combined approach, which compares objective ASR results and subjective overall evaluator's transcription results, confirmed that the objective results are in the 15% range with the evaluator's opinion. The achieved results for telephone speech recognition and speech synthesis are promising for further actions in development of the dialog system.

In this work we have shown that the common approach for speech recognition and synthesis 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 VEPRAD 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 speech recognition and speech synthesis system. The proposed combination of subjective and objective evaluation of synthesized speech enables better insights of user opinions and needs.

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 synthesis quality must be considered, mainly some improvement in the prosody modeling. Additionally work on including more speech from the same male speaker in the corpus is in progress. Further research activities are also planed towards development of the speech understanding module in the dialog system.

Acknowledgment. The research reported here have been performed within the projects “Speech Technologies" and “Theory, Modelling and Application of Autonomy-Oriented Computing Structures" within the research program “Autonomy Oriented Computing Structures" supported by the Croatian Ministry of Science, Education and Sports.

References

Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Proc. EUROSPEECH'99*. Budapest, Hungary. pp. 1047–1050.

Alumäe, T., and L. Võhandu (2004). Limited-vocabulary Estonian continuous speech recognition systems using hidden Markov models. *Informatica*, **15**(3), 303–314.

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

Barros, M.J., R. Maia, K. Tokuda, F. Resende and D. Freitas (2005). HMM-based European Portuguese TTS system. In *Proc. INTERSPEECH'05–EUROSPEECH*. Lisbon, Portugal. pp. 2581–2584.

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

Donovan, R.E., and P.C. Woodland (1999). A hidden Markov model-based trainable speech synthesizer. *Computer Speech and Language*, **23**, 223–242.

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

Dusan, S., and L.R. Rabiner (2005). On integrating insights from human speech perception into automatic speech recognition. In *Proc. INTERSPEECH'05–EUROSPEECH*. Lisbon, Portugal. pp. 1233–1236.

Frederking, R., A. Rudnicky and C. Hogan (1997). Interactive speech translation in the DIPLOMAT project. In *Proc. Spoken Language Translation Workshop*. Madrid. pp. 61–66.

Fukada, T., K. Tokuda, T. Kobayashi and S. Imai (1992). An adaptive algorithm for melcepstral analysis of speech. In *Proc. IEEE ICASSP'92*. San Francisco, California, vol. 1. pp. 137–140.

Furui, S. (2005). 50 years of progress in speech and speaker recognition. In *Proc. SPCOM'05*. Patras, Grece. pp. 1–9.

Furui, S., M. Nakamura and K. Iwano (2006). Why is automatic recognition of spontaneous speech so difficult? In *Proc*. *Large-Scale Knowledge Resources*. Tokyo, Japan. pp. 83–90.

Gauvain, J.L., and L. Lamel (2003). Large vocabulary speech recognition based on statistical methods. In W. Chou and B.W. Juang (Eds.), *Pattern Recognition in Speech and Language Processing*, ch. 5. CRC Press LLC, Florida, USA.

Geutner, P., M. Finke, P. Scheytt, A. Waibel and H. Watclar (1998). Transcribing multilingual broadcast news using hypothesis driven lexical adaptation. In *Proc. BNTUW-98 of the 1998 DARPA Broadcast News Trancription and Understanding Workshop*. Landsdown, Virginia.

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

Department of Computer Science, Nagoya Institute of Technology, Japan (2004). *HTS HMM Based Speech Synthesis System 1.0*. Department of Computer Science, Nagoya Institute of Technology, Japan (2004). http://hts.ics.nitech.ac.jp/

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. In *Proc. IEEE ICASSP'93*, vol. 2. pp. 311–314.

Imai, S. (1983). Cepstral analysis synthesis on the mel-frequency scale. In *Proc. IEEE ICASSP'83*, vol. 1. pp. 93–96.

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

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. In *ACL HLT Conference*. New York, USA. pp. 487–494.

Lee, K., H. Hon and R. Reddy (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **38**(1), 35–45.

Lihan, S., J. Juhar and A. Čižmar (2005). Crosslingual and bilingual speech recognition with Slovak and Czech SpeechDat-E databases. In *Proc. INTERSPEECH'05–EUROSPEECH*. Lisbon, Portugal. pp. 225–228.

Martinčić–Ipšić, S., and I. Ipšić (2004). Recognition of Croatian broadcast speech. In *Proc. XXVII. MIPRO 2004*. Opatija, Croatia, vol. CTS + CIS. pp. 111–114.

Martinčić–Ipšić, S., and I. Ipšić (2006a). Croatian telephone speech recognition. In *Proc. XXIX. MIPRO 2008*. Opatija, Croatia, vol. CTS + CIS. pp. 182–186.

Martinčić–Ipšić, S., and I. Ipšić (2006b). Croatian HMM based speech synthesis. *Journal of Computing and Information Technology-CIT*, **14**(4), 299–305.

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

Ostendorf, M., and I. Bulyko (2002). The impact of the speech recognition on speech synthesis. In *IEEE TTS Workshop*. Santa Monica. California, USA. pp. 99–106.

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. In *Proc. EUROSPEECH´03*. Geneva, Switzerland. pp. 1821–1824.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. IEEE*, **77**(2). pp. 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. In *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*, **17**, 311–328.

O'Shaughnessy, D. (2003). Interacting with computers by voice: automatic speech recognition and synthesis. In *Proc. of IEEE*, 91(9). pp. 1271–1305.

Skripkauskas, M., and L. Telksnys (2006). Automatic transcription of Lithuanian text using dictionary. *Informatica*, **17**(4), 587–600.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura (2000). Speech parameter generation algorithm for HMM-based speech synthesis. In *Proc. IEEE ICASSP'00*. Istanbul, Turkey, vol. 3. pp. 1315–1318.

Tokuda, K., T. Masuko, T. Yamada, T. Kobayashi and T. Kitamura (1995a). Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP'95*. Detroit. pp. 660–663.

Tokuda, K., T. Masuko, T. Yamada, T. Kobayashi and S. Imai (1995b). An algorithm for speech parametar generation from continous mixture HMM with dynamic features. In *Proc. EUROSPEECH´95*. pp. 757–760.

Tokuda, K., and T. Masuko, N. Miyazaki and T. Kobayashi (2002a). Multi-space probability distribution HMM. *IEICE Trans. Inf. & System*, **E85-D**(3), 455–464.

Tokuda, K., H. Zen and A. Black (2002b). An HMM-based speech synthesis system applied to English. In *IEEE TTS Workshop'02*. Santa Monica. California, USA.

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

Vesnicer, B. (2003). *Umetno tvorjenje govora z uporabo prikitih Markovovih modelov*. Magistarsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani. Ljubljana, Slovenija (in Slovene).

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH'99*. Budapest, Hungary, vol. 5. pp. 2347–2350.

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. In *ARPA HLT Workshop*. Plainsboro, NJ. Morgan Kaufman Publishers. pp. 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. In *Proc. EUROSPEECH´03*. Geneva, Switzerland, vol. 1. pp. 1917–1920.

S. Martinčić–Ipšić received the BSc degree in computer science from the Faculty of Computer Science and Informatics, University of Ljubijana in 1995, the MSc degree in informatics from the Faculty of Economy, University of Ljubljana, Slovenia in 1999 and the PhD degree in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia in 2007. Currently she works as an assistant at the Department of Informatics, University of Rijeka, Croatia. Her research interests are in speech recognition, speech synthesis, speech corpora development and spoken dialog systems, with special focus on Croatian language.

S. Ribarić received the BS degree in electronics, the MS degree in automatics, and the PhD degree in electrical engineering from the Faculty of Electrical Engineering, Ljubljana, Slovenia, in 1974, 1976, and 1982, respectively. He is currently a full professor at the Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His research interests include pattern recognition, artificial intelligence, biometrics, computer architecture and robot vision. He has published more than one hundred and fifty papers on these topics, and he is author of four books (*Microprocessor Architecture*, *The Fifth Computer Generation Architecture*, *Advanced Microprocessor Architectures*, *CISC and RISC Computer Architecture*) and co-author of one (*An Introduction to Pattern Recognition*). Dr. Ribarić is a member of the IEEE, ISAI and IAPR.

I. Ipšić received the BSc, MSc and PhD degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Slovenia in 1988, 1991 and 1996, respectively. From 1988–1998 he was a staff member of the Laboratory for Artificial Perception, at the Faculty of Electrical Engineering, University of Ljubljana. Since 1998 I. Ipšić is a professor of computer science at the University of Rijeka, Croatia. His current research interests belong to the field of multilingual speech recognition, spoken dialog systems and artificial intelligence.

Akustinis modeliavimas kroatų kalbos atpažinimui ir sintezei

Sanda MARTINČIĆ–IPŠIĆ, Slobodan RIBARIĆ, Ivo IPŠIĆ

Šiame straipsnyje pateikiamas kroatų kalbos priklausomas nuo konteksto akustinis modeliavimas, naudojamas kalbos atpažinime ir sintezėje. Pasiūlytas akustinis modelis remiasi nuo konteksto priklausomais trigarsių paslėptais Markovo modeliais ir kroatų kalbos fonetinėmis taisyklėmis. Kalbos atpažinimo ir sintezės sistemų modeliavimui ir testavimui buvo naudojamas kroatų kalbos garsynas VEPRAD. Eksperimentai parodė, kad kroatų kalbos garsynas, kroatų kalbos taisyklės ir paslėpti Markovo modeliai kaip modeliavimo formalizmas gali būti lygiagrečiai naudojami kalbos atpažinimo ir sintezės sistemų kūrimui tokiai smarkiai kaitomai ir turinčiai laisvą žodžių tvarką kalbai kaip kroatų. Mes pasiūlėme įvertinimo procedūrą kalbos sintezei, kuri derina objektyvų ir subjektyvų vertinimą ir pateikiame gautus įvertinimo rezultatus. Pasiūlytos procedūros kroatų kalbos akustiniam modeliavimui buvo sukurtos kaip kalbos sąsajos dalys kalbos dialogo sistemoje, skirtoje ribotai oro prognozės sričiai.