Context-Dependent Acoustic Modelling of Croatian Speech

Sanda Martinčić – Ipšić[*)] and Ivo Ipšić[+)]

[*)] Department of Informatics,
Faculty of Philosophy
University of Rijeka,
Omladinska 14, 51000 Rijeka, Croatia
Phone: (385) 51-345 046 Fax: (385) 51-345 207
E-mail: smarti@ffri.hr

[+)] Technical Faculty,
University of Rijeka,
Vukovarska 58, 51000 Rijeka, Croatia,
Phone: (385) 51-651 421
E-mail: ipsic@riteh.hr

Abstract

This paper presents experiments of Croatian speech modelling used in speech recognition as well as 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 experiments a common Croatian speech corpora is used. The experiments have shown that Croatian speech corpora, Croatian phonetic rules and hidden Markov models as the modelling formalism can be used to develop speech recognition and speech synthesis systems in parallel. The proposed procedures for Croatian acoustic modelling were developed as speech interfaces in a spoken dialog system.

Kontekstno odvisno akustično modeliranje hrvaškega govora

V članku so opisani postopki akustičnega modeliranja, ki so bili uporabljeni pri razpoznavanju in sintezi hrvaškega govora. Predlagani akustični model temelji na kontekstno odvisnih trifonskih modelih in na fonetičnih pravilih hrvaškega govora. Za učenje akustičnih modelov je izbran formalizem prikritih Markovovih modelov in korpus hrvaškega govora. Razviti postopki za razpoznavanje in sinetezo hrvaškega govora so del enotnega sistema za govorni dialog.

1. Introduction

The paper describes procedures for acoustic modelling of Croatian speech. The proposed context-dependent acoustic model is used in the speech recognition as well as in the speech synthesis module of a spoken dialog system for Croatian speech in the domain of weather forecasts. Using such a system a user can ask questions about weather conditions and forecasts. The dialog system would provide information about weather in different regions of Croatia and for different time periods, collecting the information from the available web sites over the Internet (Žibert et al., 2003). The spoken dialog system includes modules for speech recognition, spoken language understanding and speech synthesis. The speech recognition and synthesis module relay on data-driven statistical and rule-based knowledge approach. Data driven statistical approach is based on large quantities of spoken data collected in speech corpora. 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 processes in human mind during speaking and understanding.

Since the main resource in a spoken dialog system design is the collection of speech material the Croatian domain related speech corpora is presented. Further the acoustic modelling procedures of the speech recognition system including phonetically driven state tying procedures are given. Conducted speech recognition experiments and speech recognition results are presented in the third section. The fourth part explains the Croatian trainable speech synthesis, which is based on the same context-dependent acoustic model as the one used in the speech recognition experiments. Some advantages of the same acoustical modelling approach for Croatian speech recognition and speech synthesis are discussed. We conclude with the description of current activities and future plans in Croatian speech technologies

2. The Croatian speech database

The Croatian speech corpora VEPRAD includes weather forecasts and reports spoken within broadcast news of national radio (Martinčić-Ipšić et al., 2004). The collected speech material is divided in several groups: weather forecasts read by professional speakers within national radio news, weather reports spontaneously spoken by professional meteorologists, other meteorological information spoken by different reporters and radio News.

The VEPRAD corpus is a multi-speaker speech database and contains 13 hours of the transcribed speech spoken in the studio acoustical environment and the telephone speech. The spoken utterance has its word level transcription. The corpora statistics is shown in Table 1

The first part, VEPRAD radio database, of the collected speech material consists of transcribed weather forecast. This is a multi-speaker database, which contains speech utterances of 11 male and 14 female professional

speakers. VEPRAD radio part consists of 3566 utterances and lasts 6 hours and 17 minutes. The transcribed sentences contain 57896 words, where 1354 are different. Relatively small number of different words shows that the VEPRAD speech database is strictly domain oriented.

From the VEPRAD radio database one male speaker was selected for speech synthesis voice. For the selected speaker additional 85 minutes of radio news speech was recorded and transcribed. The synthesis part database includes 1111 utterances with 3840 different words.

The third part, VEPRAD telephone database, contains weather reports given by 7 female and 5 male professional meteorologists daily over the telephone. The 158 transcribed weather reports are lasting 5 hours and 39 minutes and contain 1803 different words in 3223 utterances. Most of the speech captured in the VEPRAD telephone database 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 transcribing process involved listening to speech parts until a natural break is 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. In the process of generating speech files and their transcription we used the Speech Viewer from the CSLU Speech Toolkit (Sutton et. al. 1998) and Transcriber (Barras et al. 2000). Manual correction of automatically segmented phones was performed using the Wavesurfer tool (KTH, 2004). An utterance example z07060702102 is shown in Table 2.

```
z07060702102
postupna naoblaka <uzdah> mjestimice s
pljuskovima i grmljavinom vjetar slab <sil>
a na jadranu povremeno umjeren
jugozapadnjak i jugo <uzdah>
```

Table 2. Example of one transcribed utterance.

VEPRAD	Dur. [min]	No.			Speakers	
		Sentences	Words	Diff. words	Male	Female
RADIO						
Radio weather forecasts	377	3566	57896	1354	11	14
Radio news-(synthesis)	85	1111	12265	3840	1	
Overall RADIO	**462**	**4677**	**70161**	**4504**	**11**	**14**
TELEPHONE						
Telephone weather reports	339	3223	51187	1803	5	7
Overall VEPRAD	**801**	**7900**	**121348**	**5344**	**16**	**21**

Table 1. Croatian speech database statistics.

3. Context-dependent modeling

The Croatian speech recognition and speech synthesis system is based on continuous hidden Markov models of monophones and triphones. The training of speech recognition system was performed using the HTK toolkit (Young et al., 2002), while for speech synthesis training the HTS tool (HTS, 2004), which is as an extension of the HTK, was used. Croatian weather forecasts speech database VEPRAD was used for training of all acoustic models.

3.1. Monophones

The training of speech recognition and synthesis acoustic models started with defining the Croatian phoneme set according to SAMPA (Bakran and Horga, 1996). For each Croatian phoneme 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 models 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 recognizer, 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 of the spoken utterance and the corresponding word level transcriptions. The number of mixtures of output Gaussian probability density functions per state was increased to 20 in the used monophone recognizer.

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.

Additional models for silence, breath, restarts, hesitations, cough, telephone ringing, modem noise, car beeping and driving, paper turning, door slamming, background speaking and mispronounced words where made. The resulting monophone set has 5 additional models for handling unexpected events and noise in radio speech and 7 additional models for the telephone speech. Since telephone speech is rich in channel distortions and additive environmental noise for telephone data 2 additional models were trained. Additive noise can be

stationary (like computer ventilators, cars or air conditioners) and has a power spectral density that does not change over period of time. Nonstationary noise, in contrary, changes over time and is produced by door slamming, background speaking, telephone ringing, coughing, breathing etc. The channel distortions can be caused by reverberation in telephone network or distortions in broadcasted radio signal. This unexpected events handling approach is known as the explicit noise modelling approach and enables explicit handling of noise errors and acoustic events (Ward, 1989).

3.2. Triphones

In the next step we trained context-dependent cross-words triphone models with continuous density output functions (one to six mixture Gaussian density functions), described with diagonal covariance matrices. The triphone models also consist of 5 states, where the first and last states have no output functions.

The number of cross-word seen triphones in the training data used for radio speech recognition training is 6054 (about 16% of all possible triphones), for telephone speech recognition training is 6054 (about 12% of all possible) and the number in speech synthesis is 8290 (about 13% of the number of all possible triphones) (Martinčić-Ipšić and Ipšić, 2006a).

Therefore there is evidently not enough acoustical material for modelling all possible triphone models. The severe undertraining of the model can be a real problem in the recognizer performance. The lack of speech data is overcome by a phonetically driven state tying procedure.

3.2.1. 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 216 Croatian phonetic rules are used to build phonetic decision trees for HMM state clustering of acoustic models. The phonetic rules are describing the class of the phonemes according to their articulatory and acoustic characteristics. Some defined Croatian phonetic rules used for the training of phonetic classes are shown in Table 3.

Vowel	a, e, i, o, u, a:, e:, i:, o:, u:, r:
Accented Vowel	a:, e:, i:, o:, u:, r:
High Vowel	i, u, i:, u:
Medium Vowel	o, e, 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 rules.

En example of a phonetic decision tree for the Croatian phoneme /h/ is presented in Figure 1. It classifies triphones with the phoneme /h/ in the middle in eight possible classes. At each node the binary question about left and right context is asked and YES/NO answers are possible. The triphones in the same class are sharing the same parameters of state output probability density functions of HMMs.

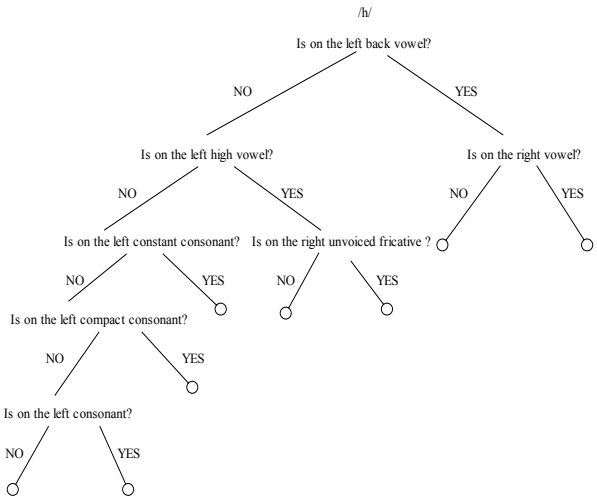

Figure 1: The decision tree of phonetic questions for the left and right context for phoneme /h/.

3.2.2. 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 (Gaussian distribution mean and variance). This enables more accurate estimating mixtures of Gaussian output probabilities and consequently better handling of the unseen triphones.

In the speech recognition state clustering procedure a separate decision tree for initial, middle and final states of each triphone HMM is built using a top-down sequential optimization 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.

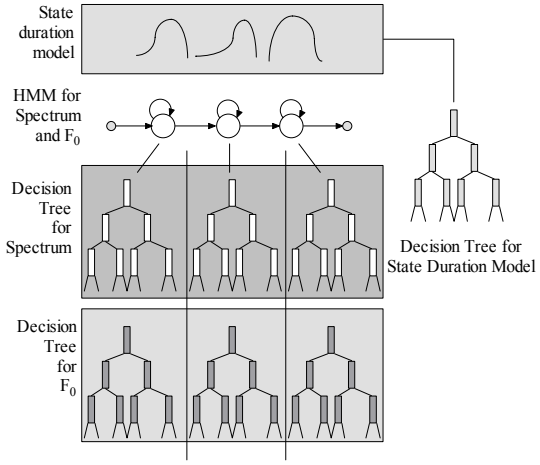

Figure 2. Decision trees for spectrum, pitch and duration in the Croatian speech synthesis system.

In the speech synthesis system the same 216 Croatian phonetic rules used in the speech recognition system were used. The clustering trees for spectral parameters, fundamental frequency F0 and duration were built separately as shown at Figure 2 (Yoshimura, 2000). 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.

4. Speech recognition experiments

So far we have performed speech recognition experiments using the VEPRAD radio (Martinčić-Ipšić and Ipšić, 2004) and VEPRAD telephone speech database (Martinčić-Ipšić and Ipšić, 2006a).

In VEPRAD radio speech recognition system 4135 (71%) utterances form 8 male and 8 female speakers were used for training and 1712 (29%) utterances from 3 male and 6 female speakers were used for testing.

In VEPRAD telephone speech recognition system 1982 utterances (61%) where used for acoustic modelling and parameter estimation of context dependent phone models and 1241 utterances (39%) 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.

In all experiments bigram language model was used. Estimated perplexity of the VEPRAD radio bigram language model is 13.23 and perplexity of the VEPRAD telephone is 18.09.

Table 4 compares data used for separate training and testing of radio and telephone speech recognition system and for speech synthesis experiments. Bottom part of the table compares the number of monophone models used in each subsystem as well as the number of seen triphones compared to the number of all possible triphones. Number of monophone models trained for speech synthesis is expanded for accented vowels including the occurrence of r as a vowel in Croatian language, and additional models for silence, breathing noises, mispronounced words and noise.

	VEPRAD		
	RADIO	TELEPHONE	SYNTHESIS
# diff. words	4504	1803	3840
perplexity	13.23	18.09	23.6
training			
# utterances	4135	1982	1111
# speakers	8m+8f	3m+3f	1m
testing			
# utterances	1712	1241	41
# speakers	3m+6f	2m+4f	1m
#monophones	30+5	29+7	36+5
triphones			
# all	36756	37597	60521
# seen	6054	4610	8290
% seen	16.47%	12.26%	13.70%

Table 4. The comparison of the data used for training and testing of radio and telephone speech recognition and for speech synthesis.

4.1. Speech feature vector

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

4.2. 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 5. The number of different words is in the first raw. Results are given in terms of correctness and accuracy per different number of tied states and different number of Gaussian mixtures, always using the same proposed 216 Croatian phonetic rules set.

	VEPRAD			
	RADIO		TELEPHONE	
#words	**4504**		**1803**	
	% Corr.	**% Acc.**	**% Corr.**	**%Acc.**
# state	**1011**		**646**	
mix 1	78.58	75.74	86.18	72.14
mix 3	79.82	77.52	88.47	76.13
mix 6	**80.60**	**78.86**	**89.30**	**77.54**
# state	**1235**		**980**	
mix 1	78.39	75.41	86.62	72.31
mix 3	79.75	77.51	88.86	74.94
mix 6	80.34	78.51	89.16	76.92
# state	**1977**		**1872**	
mix 1	78.23	74.70	86.73	71.39
mix 3	79.50	76.75	88.15	74.02
mix 6	80.09	77.68	87.99	73.64
# state	**3099**		**2273**	
mix 1	78.03	74.34	86.49	70.45
mix 3	79.34	76.37	87.99	73.25
mix 6	79.61	76.93	87.08	72.02

Table 5. Speech recognition results for radio and telephone data in terms of correctness and accuracy.

The results for radio and telephone speech recognition are in the same error range. At the firs glance this is surprising, but this was actually expected since the number of different word is more then double (4504) as the number of different word in the radio data (1803). This indicates that use of trigram language models should be considered for the radio speech recognizer. Further, since the access to the weather information spoken dialog system is planned by telephone, the speech recognition accuracy for the telephone data is quite promising. The word error rate for telephone data, for the same reason, 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.

5. Speech synthesis experiments

The hidden Markov model based trainable speech synthesis use speech corpora for the training of context-

dependent acoustic model, and uses HMM as a generative model for speech production. Similar speech synthesis systems were already developed for Japanese and English (Tokuda, 2002), Slovene (Vesnicer, 2004) and Portuguese (Baross, 2005).

The Croatian speech synthesis system (Martinčić-Ipšić and Ipšić, 2006b) was trained on selected male speaker speech, as presented in the third column of the Table 4. In the synthesis part 1111 utterances of selected male speaker from the VEPRAD radio were used for speech synthesis training and 41 for testing. For 3840 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.

5.1. 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. 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).

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, multispace 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 Figure 3. State duration densities for the speech feature vectors generation are estimated by probabilities obtained in the last iteration of embedded reestimation.

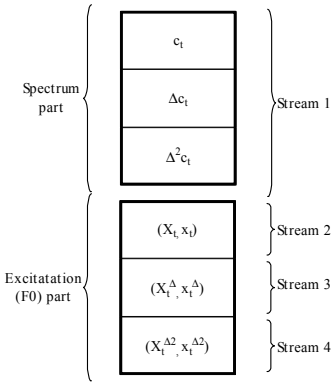

Figure 3. The HMM state output feature vector.

5.2. Speech signal generation

The speech synthesis part used prepared context-dependent HMMs, and state duration and pitch trees for generating the sequence of feature vectors for the test text. 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.

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 the sequence of mel-cepstral coefficients and F0 values including voiced/unvoiced decisions are determined by maximizing the output probability of HMM. State duration densities were modeled by multivariate Gaussian distribution. The dimensionality of state duration density is equal to the number of states of corresponding HMM. Finally the speech is synthesized from generated mel-cepstral feature vectors and pitch values using the MLSA filter (Tokuda et al. 1995).

5.3. Speech synthesis results

The text-to-speech test included 41 Croatian sentences. The text labels were transformed into triphone format. For each sentence the speech in raw format, pitch and duration were generated. Figure 4 presents the result of generated speech for the sentences:

"*Vjetar u unutrašnjosti većinom slab, na Jadranu umjerena i jaka bura. <uzdah> Najviša dnevna temperatura od minus jedan do plus tri stupnja na Jadranu od deset do petnaest.*"

From the top the pitch, spectrogram and raw signal are shown.

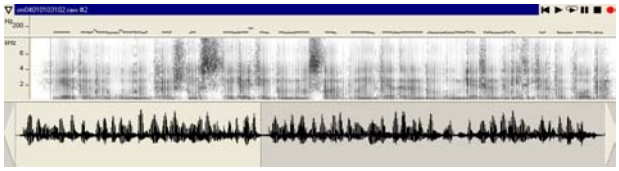

Figure 4. Pitch and spectrogram of generated speech signal for utterance sm04010103102.

Figure 5 shows the pitch and spectrogram of the corresponding part of original signal.

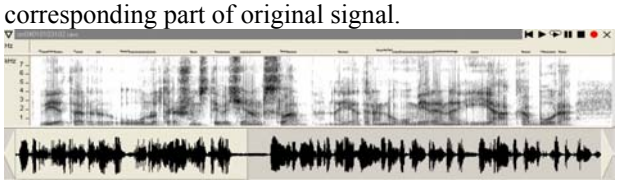

Figure 5. Pitch and spectrogram of original signal for utterance sm04010103102.

This approach where HMMs are used as generative model for speech production is very effective in rapid development of the TTS system for new language and for a new domain of interest within the same language. Although the quality of generated speech is "vocoded" buzzy speech it can be understood.

Since this speech synthesis will be incorporated in Croatian weather information spoken dialog system some improvement in the TTS quality should be considered.

In order to improve the context-dependent phone models used for synthesis more Croatian speech material

for selected speaker will be recorded and annotated. Further quality improvements will be done by manually inspecting the time boundaries of automatically segmented phones, since the overall automatically segmented phones correctness is 78.62%. And finally intelligibility and naturalness of synthetic speech can be improved also by using different Croatian speaker's speech in the speech synthesis system acoustic model training (Latore et al., 2006).

6. 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 corpora 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 fact that recognition accuracy for telephone speech is in expected range are very promising for further actions in development of the dialog system.

Since the telephone access to the spoken dialog system is planed, further improvements in speech synthesis quality must be considered. When the quality of the speech synthesis is satisfactory further work on evaluation of intelligibility, naturalness and functionality of synthetic speech will be done. The human experts and users will evaluate the system. The rate for intelligibility, overall quality, naturalness and functionality will be collected.

In spoken dialog system development the actions toward linguistic and semantic analysis are in progress. The Wizard-of–OZZ experiments for collecting the possible dialog scenarios are planned as the first stage of the dialog manager development.

This work showed a common approach for speech recognition and speech synthesis context-dependent acoustical modeling. Main advantage of the used approach is in fact that can be efficiently and rapidly ported to the other domains of interest under condition that adequate Croatian speech and language corpora is included.

7. References

Bakran, J., Horga, D.(1996). SAMPA for Croatian. *Govor*. XIII, Vol.1-2, p. 99-104.

Barras C, Geoffrois E, Wu Z and Liberman M. (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,

Barros, M. J., et al. (2005). HMM-Based European Portuguese TTS System*, INTERSPEECH '05*, Lisbon, Portugal, p.p. 2581-2584.

Department of Computer Science, Nagoya Institute of Technology, HTS HMM Based Speech Synthesis System 1.0. http://hts.ics.nitech.ac.jp/, Japan, 2004. [09.2005.]

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for melcepstral analysis of speech*, Proc. of ICASSP,* vol.1, pp.137–140, 1992.

Latorre, J., Iwano, K., Furui, S. New approach to pologlot synthesis: how to speak any language with anyone's voice. XXXX.2006.

Martinčić-Ipšić, S.; Matešić, M.; Ipšić, I. (2004). Croatian Speech Corpora. Govor: časopis za fonetiku. XXI (2); 135-150. (in Croatian).

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

Martinčić-Ipšić, S., Ipšić, I. (2006a). Croatian Telephone Speech Recognition. *XXIX. MIPRO 2006,* Opatija, Vol. CTS + CIS, p. 182-186.

Martinčić-Ipšić, S., Ipšić, I. (2006b). Croatian HMM Based Speech Synthesis. 28th International Conference on Information Technology Interfaces, ITI 2006, Cavtat, Croatia. 2006.

Odell, J. The Use of Context in Large Vocabulary Speech Recognition, PhD Thesis, Queen's College, University of Cambridge, Cambridge, 1995.

Sutton S, et. al. (1998). Universal Speech Tools: The CSLU Toolkit. Proc. of the International Conference on Spoken Language Processing 1998 (ICSLP98), vol. 7, p. 3221-3224.

Tokuda, K. et al. (1995).An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. *EUROSPEECH 95*, 1:p. 757-760.,

Tokuda, K., Zen, H., Black, (2002) A. An HMM-Based Speech Synthesis System Applied to English. IEEE *TTS Workshop 2002*. Santa Monica. California, USA.

Tokuda, K. et al. (2000). Speech Parameter Generation Algorithm for HMM-Based Speech Synthesis. HMM, *Proc. ICASSP*., Vol. 3. p. 1314-1318.

Vesnicer, B., Mihelič, F. (2004). Sinteza slovenskega govora z uporabo prikritih Markovovih modelov. *Elektrotehniški vestnik*., vol. 71, no. 4, str. 223-228.

Ward, W., (1989). Modelling Non-Verbal Sounds for Speech Recognition, Proc Speech and Natural Language Workshop, Cape Cod, Morgan Kauffman, pp.311-318.

WaveSurfer, ver. 1.7.5., Centre for Speech Technology (CTT),KTH, Stocholm, Sweden, 2004. http://www.speech.kth.se/wavesurfer/

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1999) Simoultaneous Modelling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. Eurospeech 99, Budapest, pp. 2347-2350.

Young S, et. al. (2002). The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department, Cambridge, Great Britain.

Young, S., Odell, J., Woodland, P.C. (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling, ARPA *HLT Workshop*, Plainsboro, NJ, Morgan Kaufaman Publishers, p307-312.

Žibert, J., Martinčić-Ipšić, S., Melita, H., Ipšić, I., Mihelič, F.. (2003). Development of a Bilingual Spoken Dialog System for Weather Information Retrieval. *EUROSPEECH '03*. Geneva, Switzerland. Vol. 1, p. 1917-1920.