Speech Recognition of Slovenian and Croatian Weather Forecasts

Sanda Martinčić-Ipšić[*], Janez Žibert[†], Ivo Ipšić[*],
France Mihelič[†]

[*]Faculty of Philosophy
University of Rijeka, Croatia
{smarti, ivoi}@pefri.hr

[†] Faculty of Electrical Engineering
University of Ljubljana, Slovenia
{janez.zibert, mihelicf}@fe.uni-lj.si

Abstract

In the paper we present some results of a joint project in speech data collection and speech recognition of Slovenian and Croatian weather forecasts. In the paper we describe the procedures we have performed in order to obtain a domain specific speech database from broadcast programmes. Additionally the speech recognition experiments are described and some speech recognition results for the Croatian and Slovenian speech are presented.

1. Introduction

The paper describes the development of a Slovenian and a Croatian speech database of weather forecasts. The work is done within a joint bilateral Slovenian Croatian project, which aim is the creation of a bilingual speech database. The speech databases are used in continuous speech recognition experiments as well as for comparison of speech recognition results of the Slovenian and Croatian language. The bilingual speech database can be further used for the development of a bilingual spoken dialog system for weather forecasts. In such a system a user could ask questions about weather forecasts in two different languages. The dialog system would provide information about weather in different regions and for different time periods in the two languages. The paper describes the process of speech data acquisition, segmentation and transcription of the Croatian and Slovenian speech databases of weather forecasts. Additionally the speech recognition experiments for the Croatian and Slovenian language are described and some preliminary results are presented.

2. Data acquisition and transcription

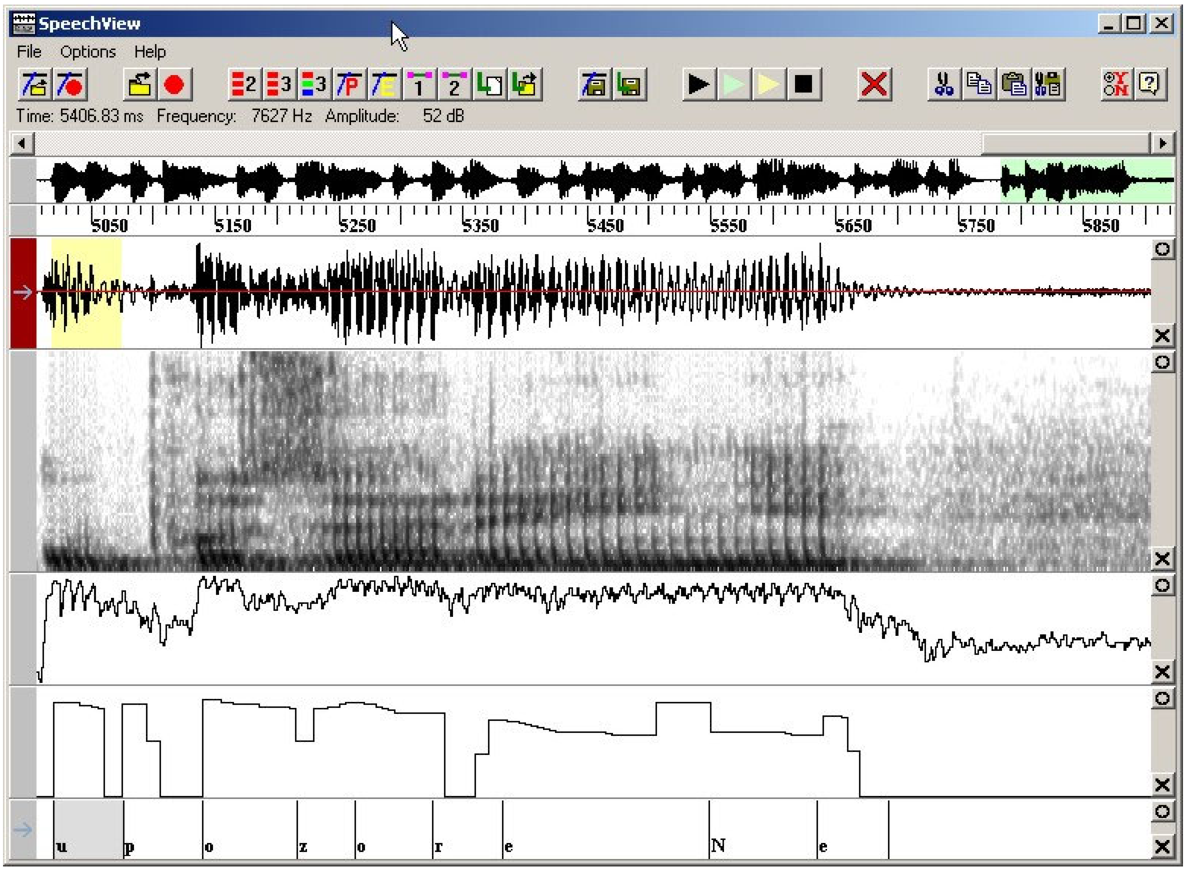
The first step in spoken dialog system development is the collection of speech material and creation of speech databases. To enable a comparison between speech recognition results for two different languages we decided to collect similar data. Thus we are collecting weather forecast spoken within news broadcast of national TV and radio programmes.

The Croatian speech database VEPRAD consists of weather forecasts read by professional speakers within the news broadcasts of the national radio. Some of the recording contain spontaneous speech, where professional meteorologists speak about weather. The recordings are performed since March 2002 and are still going on. The weather forecasts are recorded four times a day using a PC with an additional Hauppauge WINTV/radio card. The speech signals are sampled with 16 kHz and stored in a 16-bit PCM encoded waveform format. At he same time texts of weather forecasts for each day are collected from the web site of the Croatian meteorological institute. The texts are used for speech transcription and will be also used for training of a language model for the weather forecast speech recognition experiments.

The Slovenian speech database is divided into two parts: television weather forecasts VNTV and radio related weather forecasts, named VNRAD. The recordings were performed from October 1999 till February 2000. The VNTV speech database consists of recordings of weather reports captured three times a day on the national TV programme TVSLO1. The speech material for the VNRAD database was collected from the national radio programme once a day, every time after the morning news bulletin.

Television weather forecasts were recorded using the PC with ATI All in Wonder graphical card with a built-in TV receiver. Radio data was collected by a sound blaster with a FM stereo tuner. Recording conditions for both databases were always the same and consequently the quality of the recorded data is even. The recordings were sampled at a 22050 Hz sample rate and stored as 16-bit PCM encoded mono waveform sample files. The files were formatted according to Windows WAVE headers and are arranged on CD-ROMs for publication. All weather forecasts are also divided into individual sentences and stored separately.

In the first phase only read speech of weather forecast is being transcribed. The transcribing process involves listening to speech parts of weather forecasts until a natural break is found. The sentences or parts of speech signals are cut out and a transcription file is generated. The speech file and the transcription file have the same name with different extensions. The transcription is performed on the word level with some additional marks (silence, breath, paper noise, coughs and restarts). Since the sentences and speech parts are cut out of the broadcast programmes there was no need to transcribe non speech parts using standard approaches in broadcast programmes annotation (Graff, 2002). In the process of generating speech files and their transcription we



Time: 5406.83 ms Frequency: 7627 Hz Amplitude: 52 dB

Figure 1: CSLU SpeechView tool. From top to bottom window: recorded speech signal, a portion of the signal, its spectrogram, energy contour, pitch contour and a label file.

used the SpeechView tool from the CSLU Speech Toolkit (Sutton, 1998). SpeechView enables inspection (visual and audio) of speech signals and their spectrograms. Figure 1 shows a recorded speech signal, a portion of the signal, its spectrogram, energy contour, pitch contour and a label file in underlying windows respectively.

In Slovenian weather forecasts speech data we distinguished between two kinds of speaking modes: planned and spontaneous. Speech in television forecasts (VNTV) have been prepared in advance and have been read, they were identified as planned. Speech in radio (VNRAD) forecasts was unscripted and marked as spontaneous. This kind of speech typically contained more disfluencies and/or hesitations as planned speech.

Word transcriptions of television weather reports are done in two stages. In the first stage we received texts of TV weather forecasts by the Environmental Agency of the Republic of Slovenia (EARS) who prepared broadcasts for the TV. The texts were not the exact transcriptions and we had to correct them, but they were a good start.

For each waveform file there is one document (transcript) file containing the word transcriptions of each sentence, the number of sentence and start and end frame of sentence boundaries. At the beginning of transcriptions they are attributes to identify the file name, the speaker and the date of the recorded data. We also used special symbols enclosed in < . > brackets representing disfluencies or hesitations in speech signal. All transcriptions were made with the Transcriber tool (Transcriber, 2000).

3. Structure of the database

The Croatian speech database of weather forecast VEPRAD contains speech files of 11 male and 11 female speakers. The database contains 298 weather forecast lasting 4 hours and 57 minutes. The average duration of a weather forecast is approximately 4 minutes. In the first phase of transcription we have transcribed 1077 sentences and speech parts, which were cut out of 104 weather forecasts. The transcribed part of the speech database is approx. 75 minutes. Table 1 shows some statistics of the Croatian speech database. The transcribed sentences contain 11837 words, where 612 are different. The small number of different words shows that the speech database is strictly domain oriented. The transcribed data is used for initial segmentation and first word recognition experiments for the Croatian language.

The Slovenian speech database is divided into two parts: VNTV and VNRAD.

There are 5 speakers (1 female + 4 male) in the VNTV database and 9 male speakers in the VNRAD database (4 are the same as in the VNTV).

The data in the VNTV database were collected from one to two minutes long television weather forecasts. We collected 178 forecasts of approximately 252 minutes of speech material. The data in the VNRAD database represents 62 weather forecasts of approximately 87 minutes of

months	# forecasts		duration
	male	female	
March	9	12	52 min
April	48	56	75 min
May	52	60	120 min
June	31	30	50 min

Table 1: *Statistics of the VEPRAD speech database.*

speech.

The sentence corpus of the VNTV speech data consists of 3882 (different) sentences. The vocabulary includes 2857 words extracted from corpus of 41277 words (all words in the database).

We also divided the database into a test and training part. The test set includes 1389 sentences (93 minutes of speech data) which represents 36% of the data.

The basic characteristics of collected material per speaker are shown in the table 2.

speaker	# forecasts	# sentences	duration
01f	36	789	51 min
01m	43	1078	70 min
02m	32	578	39 min
03m	39	965	59 min
04m	28	472	32 min
Overall	178	3882	252 min

Table 2: *Statistics of the VNTV speech database.*

4. Speech recognition experiments

Using the Slovenian and Croatian weather forecasts speech databases we have performed several experiments. The training of acoustic and language models was performed using the HTK toolkit (Young, 2000). The speech recognition system is based on continuous hidden Markov models of monophones and triphones (Rabiner, 1989).

The speech signal feature vectors consist of log energy, 12 mel–cepstrum features and their derivatives and acceleration coefficients. The feature coefficients are computed every 8 ms for a speech signal window of 20 ms.

To built a phonetic dictionary we have proposed a set of phonetic symbols to transcribe the words from the Croatian speech database (Mihelič, 2002). The selected phonemes are derived according to (Turk, 1992). SAMPA symbols used for the transcription of the Slovenian speech database have been proposed in (Dobrišek, 1998).

Using the phonetic transcription symbols a Croatian and Slovenian dictionary have been developed. The dictionary comprises all words which occur in the recordings with their phonetic transcription.

In the first step we trained monophone models with continuous density output function (three mixture Gaussian density functions), described with diagonal covariance matrices. Since the transcription of the speech files is on the word level we performed first training procedures for context independent monophone models. The initial training of HMM monophone models resulted in a monophone recogniser, which is used for segmentation of the speech signals. Table 3 shows speaker independent word recognition results of the Croatian speech material. Word models are constructed from monophone models. Additional models for silence, breath noise, paper noise and restarts are used. In all experiments for the Croatian speech a bigram language model of perplexity 8,15 is used.

Word accuracy WA is computed from:

$$WA = 100\% \cdot (1 - \frac{W_S + W_D + W_I}{N}), \qquad (1)$$

where W_S, W_D and W_I are substituted, deleted and inserted words, while N is the number of words.

Word correctness is computed from:

$$WC = 100\% \cdot (1 - \frac{W_S + W_D}{N}), \qquad (2)$$

speaker	% correct.	% accuracy	% sentence corr.
m10	92.62	90.71	43.53
m11	92.63	90.09	51.92
z10	91.01	88.17	32.69
z11	86.25	85.00	28.12

Table 3: *Croatian speech recognition results of the VEPRAD speech data in terms of correctness and accuracy.*

An increase of word and sentence accuracy for the Croatian speech was achieved using Gaussian density functions with 6 mixtures to represent the HMM output probability functions. Table 4 shows these results for speaker independent recognition.

speaker	% correct.	% accuracy	% sentence corr.
m10	94.49	92.72	49.41
m11	94.41	91.19	51.92
z10	93.98	89.91	28.85
z11	92.00	90.75	40.62

Table 4: *Croatian speech recognition results using monophone models.*

Further we have performed experiments using context dependent phone models. Tables 5 and 6 show Croatian word recognition results using triphone models with three and six mixture Gaussian density functions. The increased number of mixtures for the output function does not increase the accuracy since the speech material is not big enough and a great number of triphones are not present in the training data.

In the table 7 results of Slovenian word recognition scores in terms of correctness and accuracy for different speakers are depicted. The recognition results are obtained using triphone models with 3 mixture Gaussian density functions per state. The Slovenian recognizer uses a bigram language model with perplexity 18.65.

speaker	% correct.	% accuracy	% sentence corr.
m10	95.98	94.75	64.71
m11	95.67	92.78	63.64
z10	96.95	95.04	65.38
z11	94.24	90.00	50.00

Table 5: *Croatian speech recognition results using triphone models with 3 mixture density functions.*

speaker	% correct.	% accuracy	% sentence corr.
m10	95.98	94.31	65.53
m11	95.67	95.67	63.46
z10	96.37	96.37	57.69
z11	93.33	93.33	40.62

Table 6: *Croatian speech recognition results using triphone models with 6 mixture density functions.*

In order to compare Slovenian and Croatian recognition results a further experiment was performed on the Slovenian speech material. Speech data were downsampled to 16 kHz. We performed two speech recognition experiments where different HMM models were trained. In the first case context independent monophone HMM models with mixture of 5 density functions per state were built and applied to the test part of the VNTV speech database. Table 8 shows the results of the VNTV test part.

In the second experiment triphone HMMs with mixture of 3 density functions per state were trained. Table 9 shows the word and sentence recognition results of the test part of the VNTV speech database.

The experiments have shown similar recognition results for the Croatian and Slovenian speech data (figure 2). The reason for this is the same way in data acquisition and selection of ”clear” read speech from the weather forecasts for acoustic training. The developed acoustic models for the Croatian and Slovenian speech can be used for further training and recognition of more spontaneous data from broadcast programmes.

speaker	% correct.	% accuracy	% sentence corr.
01f	96.32	93.51	75.24
01m	93.23	89.41	60.15
02m	93.53	89.67	57.52
03m	92.92	90.21	57.07
04m	93.81	90.10	60.23

Table 7: *Slovenian speech recognition results in terms of correctness and accuracy. HMMs were build using mixture of 3 density functions per state.*

speaker	% correct.	% accuracy	% sentence corr.
01f	93.94	93.27	62.26
01m	91.04	90.06	46.78
02m	90.62	89.77	43.62
03m	87.57	86.68	38.40
04m	93.12	92.69	58.20

Table 8: *Slovenian speech recognition results in terms of correctness and accuracy. Monophone HMMs with mixture of 5 density functions per state were used.*

speaker	% correct.	% accuracy	% sentence corr.
01f	94.04	92.21	55.66
01m	93.82	91.53	53.73
02m	95.48	94.11	67.79
03m	93.11	91.56	55.67
04m	93.61	90.66	49.18

Table 9: *Slovenian speech recognition results in terms of correctness and accuracy. Triphone HMMs with mixture of 3 density functions per state were used.*

5. Conclusion

In the paper we have presented the work we have done within a joint project in speech material collection. The paper gives some statistical data about the so far collected speech data, their transcription procedures and describes a speech recognition experiment for the Croatian and Slovenian language. The recognition results for the two languages are very similar, since the speech databases cover the same topic and have been collected in the same way. Further work will be done in speech data transcription, thus more speech material will be prepared for training of the speech recognition systems for the Slovenian and Croatian language. Existing applications of the Slovenian part of database for subtitling (Žibert, 2000) and speech–synthesis (Vesnicer, 2001) will be extended for the Croatian language. Bilingual speech recognition system simulation together with language identification experiments are also planed for the future work.

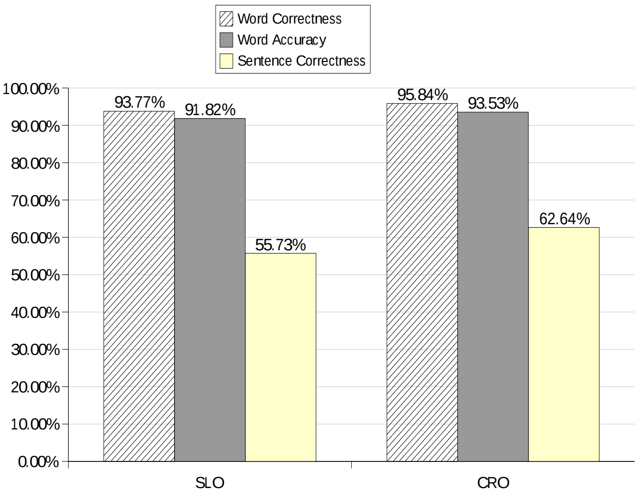


Figure 2: Slovenian (SLO) and Croatian (CRO) speech recognition results.

6. Acknowledgment

This work was partially funded within a bilateral project SLO–HRV 23 by the Croatian and Slovenian Ministry of Science and Technology.

The authors of the Slovenian part of the speech database wish to thank the Environmental Agency of the Republic of Slovenia for their contribution in sending already prepared texts of the TV weather forecasts used in the transcription process.

7. References

D. Graff. 2002. An overview of Broadcast News corpora. *Speech Communication,* Vol. 37, Issues 1–2, May 2002, pp. 15–26.

S. Dobrišek, J. Gros, F. Mihelič, and N. Pavešić. 1998. Recording and labeling of the GOPOLIS Slovenian speech database. *Proc. 1st Int.Conf. on Language Resources & Evaluation*, vol. 2, ESCA, pp. 1089–1096.

F. Mihelič, I. Ipšić, J. Žibert, S.Martinčić-Ipšić. 2002. Development of a SLO-CRO Bilingual Speech Database. *Proc. SoftCom 2002*

M. Turk. 1992. *Fonologija hrvatskog jezika*. Biblioteka Dometi.

F. Kubala, et. al. 1996. Toward automatic recognition of broadcast news. *In Proc. DARPA Speech Recognition Workshop*.

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

S. Young, J. Odell, D. Ollason, V. Vatchev, and P. Woodland. 1985. *The HTK Book.* Cambridge University Engineering Department, Cambridge, Great Britain.

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 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, January 2000.

S. Sutton, R. A. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro, and M. Cohen. 1998. Universal Speech Tools: The CSLU Toolkit. *Proc. of the International Conference on Spoken Language Processing 1998 (ICSLP98)* , vol. 7, pp. 3221-3224.

J. Žibert, F. Mihelič, S. Dobrišek. 2000. Automatic subtitling of TV weather forecasts. *Proceedings of 9th Electrotechnical and Computer Science Conference ERK 2000,* vol. B, pp. 165–168.

B. Vesnicer, N. Pavešić, and F. Mihelič. 2001. Corpus based speech synthesis. *Proceedings of 10th Electrotechnical and Computer Science Conference ERK 2001*, vol. B, pp. 253–255.