INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 221–232, 2003

Spoken Language Resources at LUKS of the University of Ljubljana

FRANCE MIHELIČ, JERNEJA GROS, SIMON DOBRIŠEK, JANEZ ŽIBERT AND NIKOLA PAVEŠIĆ
Faculty of Electrical Engineering, University of Ljubljana, Laboratory of Artificial Perception, Systems and Cybernetics, Tržaška 25, 1000 Ljubljana, Slovenia
mihelicf@fe.uni-lj.si

Abstract. This paper presents the Slovene-language spoken resources that were acquired at the Laboratory of Artificial Perception, Systems and Cybernetics (LUKS) at the Faculty of Electrical Engineering, University of Ljubljana over the past ten years. The resources consist of:

- isolated-spoken-word corpora designed for phonetic research of the Slovene spoken language;
- read-speech corpora from dialogues relating to air flight information;
- isolated-word corpora, designed for studying the Slovene spoken diphthongs;
- Slovene diphone corpora used for text-to-speech synthesis systems;
- a weather forecast speech database, as an attempt to capture radio and television broadcast news in the Slovene language; and
- read- and spontaneous-speech corpora used to study the effects of the psycho physical conditions of the speakers on their speech characteristics.

All the resources are accompanied by relevant text transcriptions, lexicons and various segmentation labels. The read-speech corpora relating to the air flight information domain also are annotated prosodically and semantically. The words in the orthographic transcription were automatically tagged for their lemma and morphosyntactic description. Many of the mentioned speech resources are freely available for basic research purposes in speech technology and linguistics. In this paper we describe all the resources in more detail and give a brief description of their use in the spoken language technology products developed at LUKS.

Keywords: speech corpora, Slovene language, read speech, spontaneous speech

1. Introduction

The domain of spoken language technologies ranges from systems for speech input and output to complex understanding and generation systems, including multi-modal systems of widely differing complexity (such as automatic dictation machines) and multilingual systems (for example, automatic dialogue and translation systems). The definition of standards and evaluation methodologies for such systems involves the specification and development of a highly specific spoken language corpus and lexicon resources, as well as measurement and evaluation tools.

In the beginning, standards for these areas were derived by consensus within the spoken language community that has become established in a number of European and national projects and with reference to important initiatives in the US and Japan. Foremost among these have been the SAM projects (centered on component technology assessment and corpus creation), SQALE (for large vocabulary systems assessment) and both SUNDIAL and SUNSTAR (for multimodal systems).

Past and present projects with significant outputs in the domain of assessment and resources include ARS, RELATOR, ONOMASTICA and SPEECHDAT,

as well as major national projects and programs of research such as VERBMOBIL in Germany. This has led to an initial documentation of existing practice, which was relatively comprehensive but, in many respects, heterogeneous and widely dispersed. The lack of generic technologies and resources and the wide diversity of formats and specifications had hindered the effective reutilization of existing resources. In 1993, the EAGLES (Expert Advisory Group on Language Engineering Standards) initiative was launched within the framework of the CEU's DGXIII Linguistic Research and Engineering (LRE) Programme, to accelerate the provision of standards for developing, exploiting and evaluating large-scale language resources. A special working group has been set up for this purpose, named the Spoken Language Working Group (SLWG). The project resulted in the publication of comprehensive guidelines documenting existing working practices in Europe and guidelines for spoken language resource creation and description (Gibbon et al., 1997).

1.1. Slovene Speech Corpora

The motivation behind collecting special Slovene speech corpora lies in the fact that the Slovene spoken language differs widely from other spoken Slavic languages. Most importantly, it has a most distinctive tonemic accent, the so-called acute accent. Without proper knowledge on how it is manifested, no Slovene TTS system can be built. Further, the Slovene language is very rich with inflectional forms, and it features dual along with singular and plural number categories. Also there is the rather free accent position within a word and numerous other specifics that make it necessary to build, study and analyze Slovene speech corpora.

For the Slovene language, several attempts to collect speech data have been made in the past, resulting in variety of speech corpora:

- SNABI (Kačič and Horvat, 1998),
- LUKS diphones (Gros et al., 1996),
- GOPOLIS (Dobrišek et al., 1998),
- SPEECHDAT-Slovene (Kaiser and Kačič, 1998; Kačič and Horvat, 1998), distributed by European Language Resources Association (ELRA), and
- Telephone Speech Corpus (Kačič et al., 2000).

The collected speech data represented mainly the domain of intended applications and in the past was not available for distribution. An exception is the Slovene SpeechDat(II) FDB-1000 corpus, which contains phonetically rich sentences. The corpus consists of read and spontaneous speech and was recorded through an ISDN card (1000 speakers). A phonetic lexicon with canonical transcriptions in SAMPA is also provided. The availability of this corpus for research purposes is rather limited, unfortunately due to its relatively high cost. The aim of this paper is to present spoken resources of the Slovene language, which were acquired at the Laboratory of Artificial Perception, Systems and Cybernetics (LUKS) at the Faculty of Electrical Engineering, University of Ljubljana over the past ten years, along with a brief description of their uses in spoken language technology products developed at LUKS. Most of the resources are, or will be, freely available for research purposes in speech technology and linguistics, thus enabling the comparison of results and methods in the same or similar applications for Slovene and other (Slavic) languages.

2. Isolated-Spoken-Word Corpus K211d

K211d[1] is a multi-speaker isolated-word corpus designed for phonetic research studies of the Slovene spoken language. The main aim of this database was to collect speech samples incorporating significant numbers of all the Slovene allophones, as well as Slovene diphthongs.

2.1. Corpus Description

The K211d lexicon consists of 251 carefully selected words derived from various Slovene texts ranging from literature, newspapers and dictionaries. Some of the words in the lexicon are also Slovene first and last names. The procedure for the word selection was done automatically, taking into account statistical analysis and automatic grapheme-to-phoneme conversion of the selected texts to provide a representative sample of all Slovene allophones, with special attention to 26 diphthongs. The criteria for selection were the achievement of minimal frequency for rare Slovene allophones (at least 50 pronunciations) and all Slovene diphthongs (at least 20 pronunciations).

Ten speakers (five female and five male) were selected to participate for the recording. We were looking for middle-aged speakers with no strong accent. Two of the male speakers were linguists, and one was a professional actor. With this selection of speakers we wanted to provide some standard of word pronunciation.

Table 1. Allophones in K211d, using Slovene MRPA (Dobrišek et al., 1996) notations.

Allophones groups	Allophones in narrow phonetic transcriptions
Vowels	i: e: E: @: a: O: o: u: I E @ a O u
Sonorants	j I l l' r v w W U m F n n' N
Nonsonorants	p b f t d s z ts dz S Z tS dZ k g x

2.1.1. Recording Conditions. The speech was recorded using a close-talk microphone at a 16-kHz sampling frequency and a 16-bit linear quantization. All the speakers uttered all 251 words stored. The recordings were stored in 2510 files. The average duration of each record (including a silent part before and after each uttered word) was 1.35 seconds.

2.2. Annotation and Labeling

The corpus consists of 16,947 phones derived from 44 different allophones (Table 1), and a frequency distribution as given in Figs. 1 and 2.

All the recordings were phonetically transcribed and labeled by a human expert using our special annotation software (Fig. 3).

2.3. Applications

The K211d database was used primarily in speech analysis studies in order to determine the appropriate selection of allophone units for speech recognition and speech synthesis systems for the Slovene language

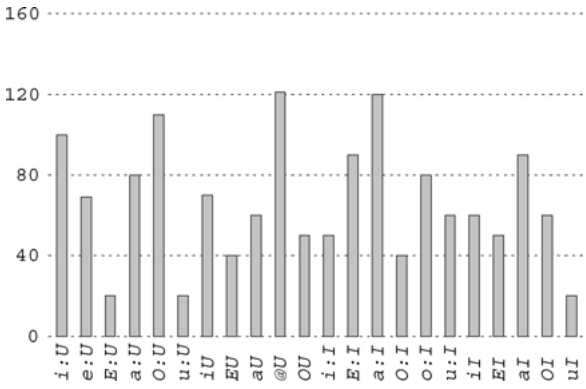

Figure 2. Frequency of diphthongs in the K211d speech corpus.

(Dobrišek et al., 1998a; Dobrišek, 2001). As a phonetically rich database, it was used also in the evaluation process of various configurations of Slovene speech recognition systems concerning acoustic modeling (Dobrišek, 2001) and feature extraction (Žibert et al., 2002).

The K221d speech corpus is freely available for research purposes.[2]

3. Read Speech-Corpus GOPOLIS

The GOPOLIS[3] corpus is a large, multi-speaker speech database that was derived from real situation dialogues used in airline timetable information services. This corpus was used as the Slovene speech database in the international SQEL project[4] for building a multi-lingual speech recognition and understanding dialog system, capable of passing information over the telephone line to a client in one of four European languages—German, Czech, Slovak and Slovene (Aretoulaki et al., 1998).

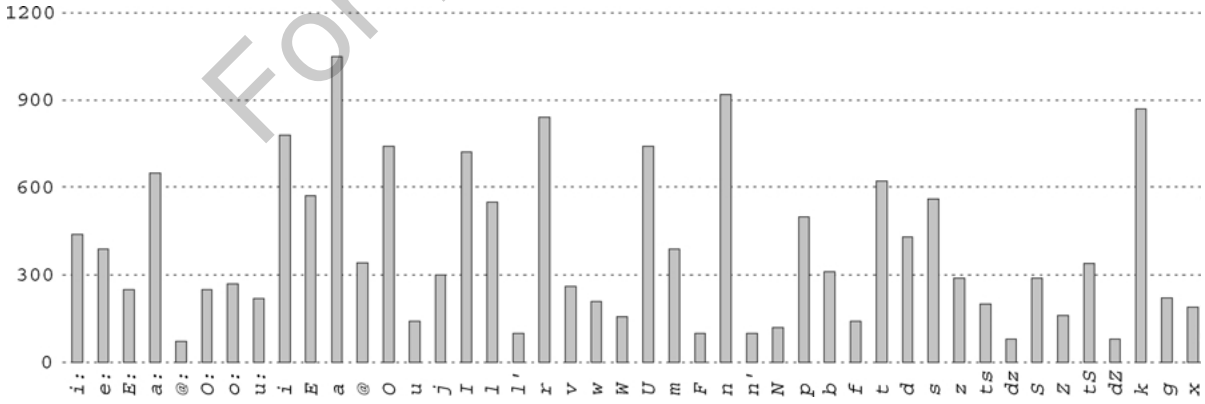

Figure 1. Frequency of allophones in the K211d speech corpus.

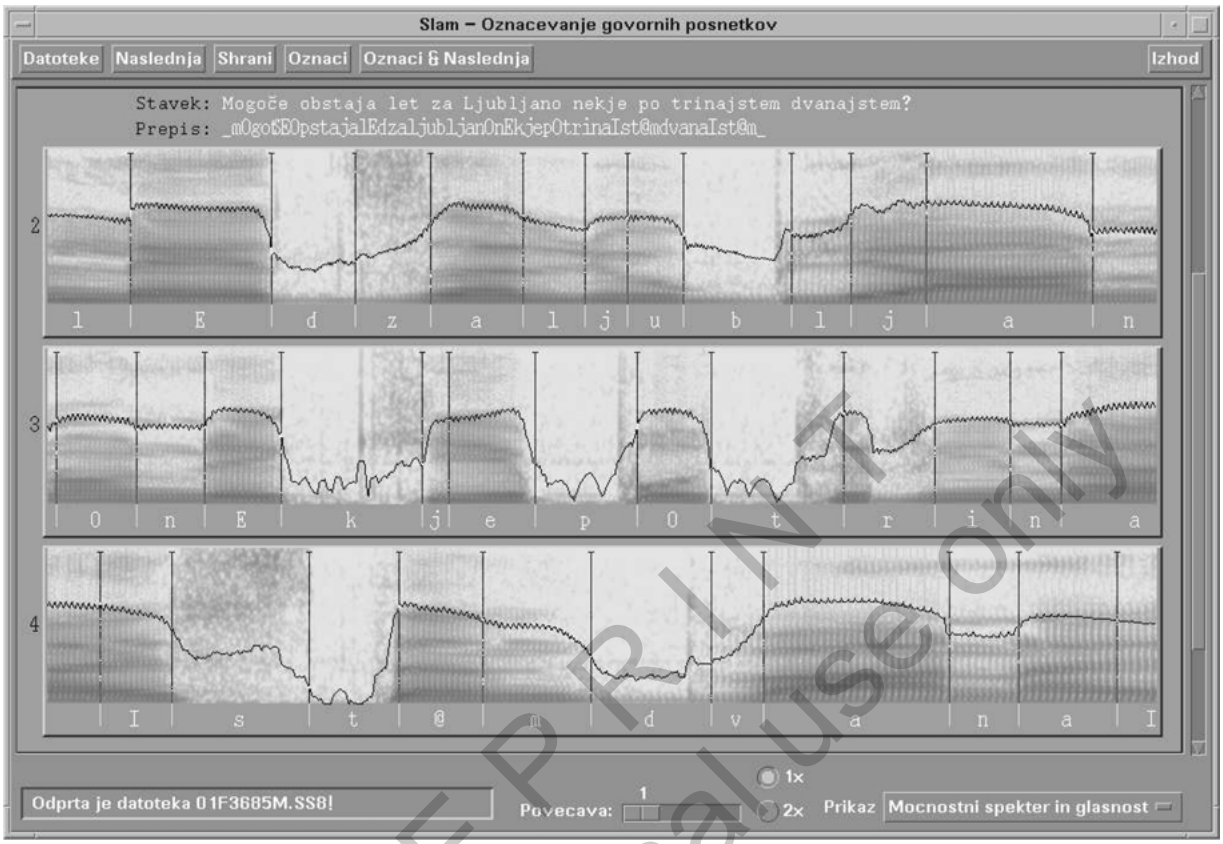

Figure 3. Screen snapshot of our software for segmentation and labeling of speech signals. Top to bottom: orthographic and phonetic transcription of the utterance, graphic representation of the segmentation of the log amplitude spectrum with short-time loudness curve, segment borders and allophone notations for segments.

3.1. Corpus Description

The sentence corpus was drawn from listening to recordings of real situation dialogues between anonymous clients and telephone operators at the Adria Airways information center (15 hrs of speech stored on audiotapes). The selected 300 typical sentences were compiled into the form of rewrite rules to obtain a generative sentence pattern grammar (Gros et al., 1995). Using this grammar we produced 22,500 different sentences for short introductory inquiries, long inquiries and short confirmations. The final sentence corpus was comprised of 5077 of those sentences. Each of the total of 50 speakers (25 female and 25 male) read about 100 randomly selected unique corpus sentences; added to this were 71 sentences of welcome greetings, introductory phrases, short affirmations and farewell greetings that were common to all of the speakers. Each session has a list of attributes with speaker and recording-session descriptors.

3.1.1. Recording Conditions. The recording sessions were conducted in a laboratory with a normal acoustic environment. Additional noise, such as background speech or slamming doors, was avoided. The utterances were acquired simultaneously with a close-talk microphone and a telephone. Thus, an additional analysis of both audio devices was possible. A set of recording environment programs was developed for the HP 9000 workstation platform in the HP Unix environment.[5] The user interface program that was built for the recording communicates with two audio servers over the network, displays and saves the acquired signals, and displays the sentences that the speaker should utter so the acoustic realizations are in the form of continuously read speech. The program is also equipped with loudness detection and an acoustic messaging system that takes care of the correct maximum loudness level, and the begin and end pauses, and synchronizes the acquired telephone- and microphone-speech signals. The audio servers use the HP 9000/735 common audio

hardware components and the additional Gradient Technology DeskLab hardware with a full telephone interface (DeskLab is a data-acquisition and play device, that communicates via SCSI with a workstation). A sampling rate of 16 kHz and a 16-bit data format with MSB-LSB byte order for both the microphone and telephone signals was chosen. The same sampling frequency was chosen because of the unification of the speech signals representation in the corpus. However, telephone signals can be down-sampled to 8 kHz, if necessary.

3.2. Annotations and Labeling

The GOPOLIS corpus is encoded in accordance with TEI recommendations (Sperberg-McQueen and Burnard, 1994); it includes the base tagset for Transcriptions of Speech, the additional tag-sets for Simple Analytic Mechanisms and Language Corpora, and some local modifications. The corpus contains the TEI header, which gives the File, Encoding and Profile descriptions. The header also provides general information about the corpus, including speaker descriptions. The body of the corpus consists of the 5077 sentences (utterances), each marked with an ID and references to its speakers. The utterances are segmented into words and punctuation marks, and each word is given in its orthographic form, as well as in the automatically derived phonetic transcription. Furthermore, the words were automatically tagged for their lemma and morphosyntactic description. The text corpora also contain prosodic and semantic annotations. Two lexica of 978 words accompany the corpus: a pronunciation dictionary with segmental information and a word-form lexicon with morphosyntactic descriptions. The perplexity of the corpus, as estimated by a bigram language model trained on an artificially derived text corpus (derived by the generative sentence pattern grammar in 3.1.), is 5.7.

3.2.1. Transcription and Tagging. The phonetic transcriptions in the corpus are based on the Slovene Machine Readable Phonetic Alphabet (MRPA) set (Dobrišek et al., 1996), containing machine-readable phonetic symbols equivalent to the Slovene IPA symbols (Šuštaršič et al., 1999). The morphosyntactic descriptions and lexicon are based on the MULTEXT-East (Slovene) tag-set and lexicon (Dimitrova et al., 1998; Ide et al., 1998). The lexicon contains lemmas, their full inflectional paradigms and the morphosyntactic descriptions of the word forms (Erjavec, 1998). The descriptions have a feature structure like format and encode information such as part-of-speech, number, case, etc. These descriptions and lexicon were then used to tag and lemmatize the corpus automatically. The tagger used is TnT (Brants, 2000), which had been trained on the Slovene MULTEXT-East corpus.

3.2.2. Speech Segmentation and Alignment. The automatic segmentation of the recorded speech signals was performed using the dynamic time-warping (DTW) based approach described in Dobrišek et al. (1997), in which the speech material is automatically segmented on the allophone level and labeled using DTW alignment of a natural utterance with a synthesized speech signal. The synthesis of speech signals was achieved simply by concatenating labeled diphone speech signals using a simplified TD-PSOLA technique. The diphone inventory described in Section 5 was borrowed from the Slovene text-to-speech system called S5 (Gros et al., 1997). A conventional DTW alignment of the utterance with the synthesized speech signal was performed with two sequences of feature vectors derived from both speech signals. One part of the automatically segmented corpora was also hand-checked by two human experts. Comparisons and segmentation results are described in detail in Dobrišek et al. (1998).

3.3. Applications

The GOPOLIS corpus was widely used in several speech technology experiments for the Slovene language. It was used for training and recognition purposes during acoustic modeling (Ipsic et al., 1998), language modeling (Gros et al., 1995; Žibert et al., 1999) and semantic parsing (Pepelnjak et al., 1996) within a dialogue system for flight information retrieval in a SQEL project (Aretoulaki et al., 1998). This database was also used for extracting duration parameters for Slovene speech synthesis (Gros et al., 1997a), for a study of some new approaches in acoustic modeling using diphones as basic recognition units (Dobrišek et al., 1999; Dobrišek, 2001) and for studies of prosodic events recognition (Mihelič et al., 2000).

The GOPOLIS corpus, which is freely available for research purposes,[6] includes the following: telephone speech recordings, text transcriptions and a lexicon with different phonetic transcriptions.

4. Broadcast Weather-Forecast VNTV Corpus

The transcription of radio and television news broadcasts poses a number of challenges for large-vocabulary transcription systems. The data in the broadcast news is not homogeneous and includes a number of data types for which speech recognition systems trained on read-speech corpora produce high error rates. A typical news broadcast may include data with different speech styles (read, spontaneous and conversational) and high- or low-bandwidth channels (e.g., telephone speech) with or without background music or other background noise.

4.1. Corpus Description

We decided to collect speech data from weather forecasts broadcast on television news. This is the first organized attempt to collect broadcast news speech data for the Slovene language. Other Slovene speech databases are restricted to read speech, as described in the previous sections and in Kacic et al. (2000). There are several reasons for collecting speech data from weather forecasts. The speech database of this restricted domain has a small vocabulary and a small number of speakers.

4.1.1. Speech Data Collection. The decision to collect speech data from weather forecasts is only the first step in the process of collection of a large amount of speech data from broadcast news, similarly to Garofolo et al. (1997). The recording started at the end of October 1999. The VNTV[7] consists of recordings of weather reports captured three times a day on the national TV program TVSLO1. The television weather forecasts were recorded using a PC with an ATI All in Wonder graphical card with a built-in TV receiver. The recording conditions were kept the same at all times and, consequently, the quality of the recorded data is consistent. The recordings were sampled at 22,050 Hz and stored as 16-bit PCM-encoded mono-waveform sample files.[8] All the weather forecasts were divided into individual sentences that were stored separately.

4.2. Transcription and Labeling

The speeches in the television weather forecasts were prepared in advance and were read with some spontaneous interruptions that were identified as planned or quasi-spontaneous.

Orthographic transcriptions of the recorded weather reports were made in two stages. In the first stage we received texts of the TV weather forecasts from the Environmental Agency of Slovenia, who prepared the broadcasts for the TV. The texts were not exact transcriptions, and we had to correct them, but they were a good start.

For each speech file there is one document (transcript) file containing orthographic transcriptions for each sentence, the consecutive number of the sentence and the start and end frame at the sentence boundaries. At the beginning of the transcriptions there are attributes to identify the file name, the speaker and the date of the recorded data. We also used special symbols enclosed in brackets <·> to represent disfluencies or hesitations in the speech signal (e.g., <inhalation>, <smack>, <break>, . . .).

All the transcriptions (Fig. 4) were made using the Transcriber tool[9] (Barras et al., 2001). The tool allows the user to segment, label and transcribe speech signals manually for later use in automatic speech processing. The transcriptions are in XML format for easier automatic processing and exchange.

The first stage was finished by the middle of January 2000, by which time we had collected enough data to build a speech recognizer based on continuous HMM using the HTK toolkit (Young et al., 2000). Later, speech recognition was used in the next stage of the transcription process (Žibert and Mihelič, 2000). Then the transcription methods became semi-automatic. We also tried to build special silence models to tag the sentence boundaries, so we only had to make certain that the transcriptions were exact. Thus we minimized the time required for the transcription process, which was one of our goals.

4.2.1. Speech Database Statistics. The corpus contains recordings from five speakers (one female and four males). The data in the VNTV database are collected from television weather forecasts that were one to two minutes in length. We collected 178 forecasts with approximately 252 minutes of speech material. The sentence corpus consists of 3882 (different) sentences. The current vocabulary includes 2857 words extracted from a corpus of 41,277 words (all the words in the database) with a bigram language model perplexity of 22.7. Language model perplexity was derived on a set consisting of the first 2493 sentences.

We also divided the database into a test part and a training part. The test set includes 1389 sentences

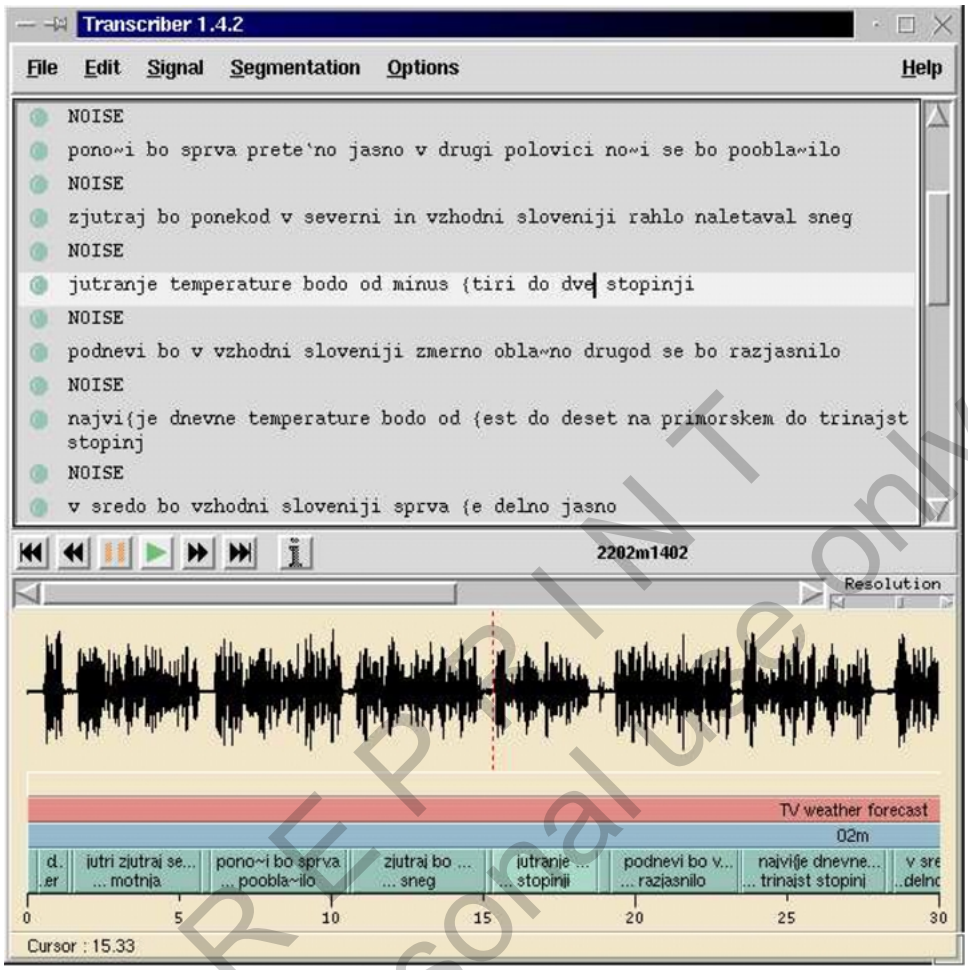

Figure 4. Screen snapshot of the transcription tool Transcriber.[7] Top to bottom: sentence transcriptions, speech waveform, sentence alignments.

(93 minutes of speech data) and represents 36% of the data.

The basic characteristics of the collected material per speaker are shown in Table 2. The collected speech material represents a strictly domain-oriented speech database with a mid-sized vocabulary, a large sentence corpus and a small number of speakers.

Table 2. Statistics of the VNTV speech corpus.

Speaker	Forecasts	Sentences	Words	Duration
01f	36	789	7609	51 min
01m	43	1078	12008	70 min
02m	32	578	6398	39 min
03m	39	965	10041	59 min
04m	28	472	5221	32 min
Overall	178	3882	41277	252 min

4.3. Applications

The speech recognition system build for the transcription process purposes (Section 4.2) was also used within a system for automatic subtitling of TV weather broadcasts (Žibert et al., 2000).

The application is designed so that deaf and partially deaf people can receive on-line information about the weather during a TV or radio broadcast. The subtitling of weather forecasts is a real-time process: The synchronized transcriptions are shown in a separate window on the screen or on teletext for radio broadcasts (Fig. 5).

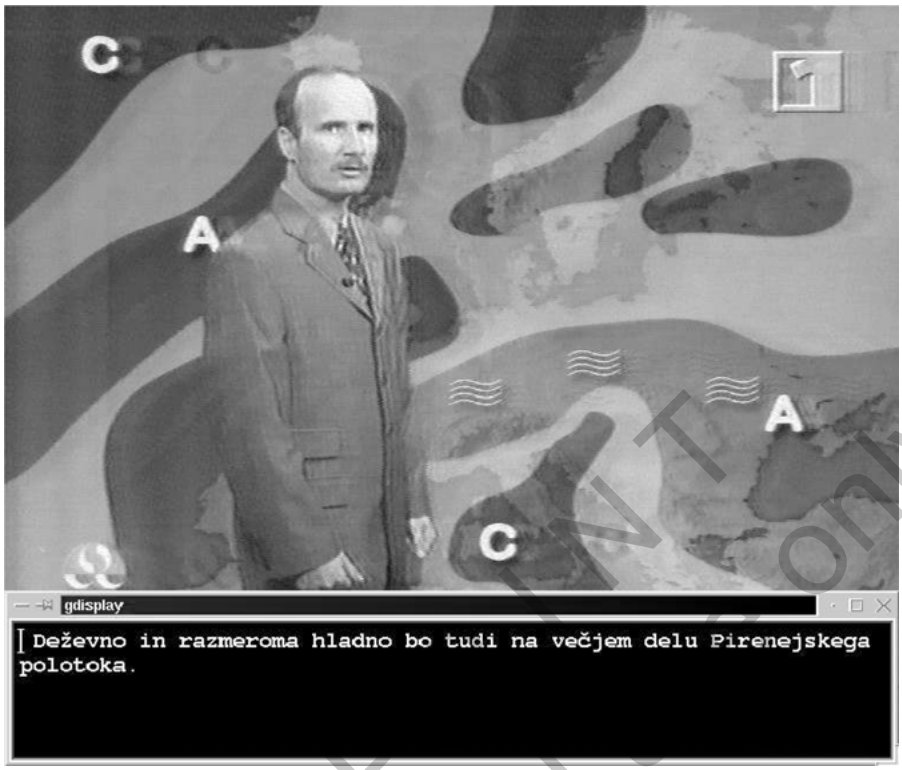

Figure 5. System for automatic real-time subtitling of TV weather.

The application runs on a PC with Pentium II processor and a minimum of 32 MB of RAM, with a Linux operating system. Additionally, it requires a video card with a built-in TV receiver. The recognition system was realized using HAPI speech recognition engine (Odell et al., 1998). The extension of this application to other types of TV broadcasts is planned for the future.

The speech corpus was also designed and organized in such a way that it can be used as a database for the corpus-based text-to-speech synthesis in dialogue systems providing weather forecast data. Preliminary studies on this type of speech synthesis have just begun (Vesnicer et al., 2001).

5. LUKS Diphones

Diphone units are adopted most commonly as a compromise between the size of a unit inventory and the quality of synthetic speech. A diphone is, generally speaking, a unit that starts in the middle of a phone, passes through the transition to the next phone and ends in the middle of this phone. Thus, the transition between two phones is encapsulated and does not need to be calculated. In nonrestricted speech synthesis one diphone is required to model the transition of every possible allophone combination in a given language.

5.1. Corpus Description

Primarily for speech synthesis purposes, a Slovene diphone inventory comprised of 1027 diphones pronounced by one male speaker was created. In order to guarantee optimum synthesis quality, a neutral phonetic context in which the diphones needed to be located was specified. Unfavorable positions, like inside stressed syllables or in over-articulated contexts, were excluded. The diphones were placed in the middle of logatoms—meaningless words that were to be pronounced with a steady intonation. The speech signals were recorded with a close-talk microphone using a sampling rate of 16 kHz and a 16-bit linear A/D conversion.

5.2. Annotation and Labeling

After the recording phase the logatoms were hand-segmented, and the center of the transition between

the phones was marked using information from both the temporal and spectral representation of the speech signal. Finally, pitch markers were manually set for the voiced parts; a special user-friendly interface was developed for this purpose.

While concatenating diphones into words, we discovered that there was a large discrepancy between the duration of the allophones, as suggested by the prosody model, and the actual corresponding diphone duration stored in the diphone inventory. This happened because of the exaggerated eagerness of the speaker trying to pronounce the meaningless logatoms in a correct and clear way (Gros et al., 1996).

The diphone inventory has been updated recently. Most of the database was rerecorded, and new diphones were added. Special attention was paid to diphones containing stressed and non-stressed vowels. An automatic segmentation procedure and pitch determination was used to speed up the data preparation process. However, final segment borders determination and pitch marker positioning are still controlled by a human expert. A new diphone inventory already is being used in the HOMER III information retrieval system (Dobrišek et al., 2002).

5.3. Applications

The diphone database is being used in a speech synthesis system, S5 (Gros et al., 1997) and other automatic information retrieval systems using speech synthesis (Aretoulaki et al., 1998; Dobrišek et al., 2003) developed in LUKS. The diphone database was also used for the automatic segmentation of the GOPOLIS speech corpus (Dobrišek et al., 1998), as mentioned in Section 3. In studies of different types of acoustic modeling for speech recognition we used this database for initializations of HMM model parameters in the training process, where we used diphones as the basic units (Dobrišek et al., 1999; Dobrišek, 2001).

6. Read- and Spontaneous-Speech Corpus—VINDAT

The VINDAT[10] corpus was designed for studies of the effects of different alcohol levels in the blood on speech characteristics. The information gained can prove to be an important factor for developing systems for speech recognition and understanding speakers who are under the influence of alcohol. Furthermore, the system can be used for detecting inebriated persons suspected of crimes in cases when a blood analysis can not be performed in time, but speech samples are available. A well-known example of such an analysis was reported by Brenner and Cash (1991).

6.1. Corpus Description

The VINDAT speech corpus includes recordings of ten Slovene speakers, five female and five male. The average age of the speakers was around 35 years and all of them occasionally drink alcoholic beverages. The corpus consists of two parts comprised of read and prompted speech.

The read-speech corpus was designed to study inherent phone duration for speakers with various alcohol levels in their blood. Fourteen words with neutral intonation were chosen for this purpose. The words were selected in such a way that they are not easy to pronounce (for instance, the word "otorhynolaryngologist"). To avoid the influence of sentence intonation and the changes in speaking rate at the borders of phrases, these words were nested within carrier sentences.

In the prompted-speech part of the corpus, the speakers were asked to repeat the sentences previously read to them by the operator. In this way we expected to get a more natural pronunciation of the selected text. This part of the corpus consisted of 18 sentences; each speaker recorded an average of five sentences.

6.1.1. Recording Conditions. The speech corpus was recorded in digital form using a notebook computer with a built-in sound-card and a close-talking microphone. The sampling frequency was 16 kHz, and we used a 16-bit linear quantization. The alcohol level in the speaker's exhaled air was measured using a handheld indicator.[11] We ensured there was a minimum of 15 minutes between the last alcoholic drink and the recording session, in accordance with the instructions for the alcohol-level indicator. The speakers were drinking red and white wines from the coastal region of Slovenia, with an average alcohol content of 12.4 vol. percent. Each speaker was recorded three times, initially in the sober condition (Table 3).

6.2. Annotation and Labeling

The VINDAT lexicon consists of 902 words with MRPA (Dobrišek et al., 1996) phonetic transcriptions.

Table 3. Wine consumption and blood alcohol levels of the speakers.

	Recording sessions					
	1		2		3	
Speaker	Liters	‰	Liters	‰	Liters	‰
01m	0.0	0.00	0.4	0.28	1.0	0.64
02m	0.0	0.00	0.7	0.73	1.5	1.10
03m	0.0	0.00	0.6	0.35	1.3	1.00
04m	0.0	0.00	1.1	0.51	1.6	0.95
05m	0.0	0.00	0.8	0.73	1.3	1.00
01f	0.0	0.00	0.6	0.77	1.0	1.14
02f	0.0	0.00	0.5	0.24	0.9	0.67
03f	0.0	0.00	0.4	0.27	1.2	1.00
04f	0.0	0.00	0.8	0.61	1.2	0.58
05f	0.0	0.00	0.6	0.49		

The lexicon contains all the transcriptions of the pronunciations that were recorded during the sessions. All pronunciations, some of them being rather distant from the regular transcriptions, were also included.

When listening to the recordings we found approximately 50 (10%) of the records were not in accordance with the proposed text. The clearest differences appeared during the prompted part of the corpus. There were several spontaneous-speech phenomena like repetitions, hesitations, non-lexical words, inhalations and laughter during the pronunciation of the sentences. All these effects were carefully transcribed for later analysis.

The speech recordings were automatically segmented on the phone level using a speech recognition system (Dobrišek, 2001) that distinguishes between 32 different Slovene allophones. We performed a forced alignment using text transcription and HMM acoustic models for the allophones as well as models for silence, different manifestations of laughter and other non-lexical phenomena.

6.3. Applications

A short statistical analysis was performed to measure the influence of alcohol levels in the blood of the speakers on vowel duration (Skerl et al., 2001). Average vowel duration has been measured, as well as changes in vowel duration (at different alcohol levels) in a pairwise analysis. The analysis yielded some interesting results. Female speakers tended to accelerate their speech rate with higher rates of alcohol consumption, whereas male speakers systematically decreased their speech rate. Further analyses of other prosodic and acoustics features of the VINDAT speech signals will be performed in the future.

7. Conclusion

In this paper we have discussed some Slovene spoken language resources available at LUKS together with their usage in various applications. The broadcast news collection is continuing well as is collection of other types of speech data. Currently, in a joint Slovene-Croatian project, a bilingual speech database consisting of different broadcast news is being recorded (Mihelič et al., 2002). We also plan to record a multi-speaker diphone speech database for speech synthesis purposes in the near future.

As with many other languages with a relatively small population speaking the language,[12] spoken language resources and research and development in the area of speech technologies will play a very important role in preserving and developing the Slovene language. With this paper we hope to encourage and show guidelines for further studies, including data collection, comparisons of results, and methods in speech technologies for Slovene and other Slavic languages.

Acknowledgment

This work was partially supported by the Slovene Ministry of Education, Science and Sport, contr. no. J2-5113-0781, 3411-97-22-7634 and T2-0409.

Notes

1. The name for this database comes from its Slovene title "Korpus 2 krat 11 sklopov diftongov" meaning "Corpus of 2 times 11 diphtongs".
2. Please contact http://luks.fe.uni-lj.si/eng/research/speech.
3. The name of the database has been derived from "GOvorjena POizvedovanja o Letalskih Informacijah v Slovenskem jeziku", meaning "Spoken Flight Information Queries in the Slovene Language".
4. Copernicus project COP-94, contract No. 01634.
5. Programs were recently ported successfully to the Linux environment.
6. Please contact http://luks.fe.uni-lj.si/eng/research/speech.
7. The name of the database has been derived from "Vremenske Napovedi na TV" meaning "Weather forecasts on TV".

8. Later recordings were down-sampled to 16,000 Hz and no significant differences in results were noticed when performing different recognition tasks.
9. Available at http://www.etca.fr/CTA/gip/Projets/Transcriber/.
10. The name was derived from the words "VINe DATa".
11. ALCOQUANT 3020 from ENVITEC-WISMAR GmbH. This type of indicator is generally used in police traffic-control actions in Slovenia.
12. In case of the Slovene language a population of approximately 2,000,000 speakers.

References

Aretoulaki, M., Harbeck, S., Gallwitz, F., Nöth, E., Niemann, H., Ivanecky, J., Ipšić, I., Pavešić, N., and Matoušek, V. (1998). SQEL: A multilingual and multifunctional dialogue system. *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, pp. 855–858.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Use of a tool for assisting speech corpora production. Speech Communication special issue on *Speech Annotation and Corpus Tools*, *33*(1/2):5–22.

Brants, T. (2000). TnT-A Statistical Part-of-Speech Tagger. *Proceedings of the ANLP-NAACL*, Seattle, pp. 224–231.

Brenner, M. and Cash, J.R. (1991). Speech analysis as an index of alcohol intoxication—The Exxon Valdez Accident. *Aviation, Space and Environmental Medicine*, *62*(9):893–898.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.J., Petkevič, V., and Tufis, D. (1998). Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. *COLING-ACL'98 Proceedings*, Montreal, pp. 315–319.

Dobrišek, S., Kačič, Z., Gros, J., Horvat, B., and Mihelič, F. (1996). Initiative for the standardization of machine readable phonetic alphabet for Slovene speech. *Proceedings of the ERK'96 Conference*, Portorož, Slovenia, vol. B, pp. 247–250.

Dobrišek, S., Gros, J., Mihelič, F., and Pavešić, N. (1997). Automatic segmentation and labeling for the GOPOLIS speech database. *Proceedings of the 2nd SQEL Workshop on Multi-Lingual Information Retreival Dialogs*, Plzen, Czech Republic, pp. 37–46.

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

Dobrišek, S., Gros, J., Mihelič, F., and Pavešić, N. (1998a). Analysis of speech formant characteristics for selection of basic Slovene speech units. *Proceedings of the Scientific Conference Artificial Intelligence in Industry from Theory to Practice and 3rd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, High Tatras, Slovakia, pp. 399–406.

Dobrišek, S., Mihelič, F., and Pavešić, N. (1999). Acoustical modeling of phone transitions: Biphones and diphones—What are the differences? *Eurospeech'99: Proceedings*, Hungary, vol. 3, pp. 1307–1310.

Dobrišek, S. (2001). Analysis and recognition of phones in speech signal. Ph.D. Thesis, University of Ljubljana.

Dobrišek, S., Gros, J., Vesnicer, B., Mihelič, F., and Pavešić, N. (2003). Homer III—Evolution of the information retrieval system for blind and visually impaired people. *International Journal of Speech Technology*, vol. 6, pp. 301–309.

Erjavec, T. (1998). The MULTEXT-East Slovene Lexicon. *Proceedings of the ERK'98 Conference*, Portorož, Slovenia, pp. 189–192.

Garofolo, J., Fiscus, J.G., and Fisher, W.M. (1997). Design and preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora. *Proceedings of DARPA Speech Recognition Workshop*, Chantilly, pp. 15–21.

Gibbon, D., Moore, R., and Winski, R. (1997). *EAGLES Handbook, Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gros, J., Mihelič, F., and Pavešić, N. (1995). Sentence hypothesisation using Ng-grams. *Proceedings of the Eurospeech95*, Madrid, pp. 1759–1762.

Gros, J., Ipšić, I., Mihelič, F., and Pavešić, N. (1996). Segmentation and labeling of Slovene diphone inventories. *COLING'96*, Copenhagen, Denmark, pp. 298–303.

Gros, J., Pavešić, N., and Mihelič, F. (1997). Text-to-speech synthesis: A complete system for the Slovene language. *Journal of Computing and Information Technology*, *5*(1):11–19.

Gros, J., Pavešić, N., and Mihelič, F. (1997a). Speech Timing in Slovene TTS. *EUROSPEECH'97, Proceedings of the 5'th European Conference on Speech Communication and Technology*, Rodos, Greece, vol. 1, pp. 323–326.

Ide, N., Tufis, D., and Erjavec, T. (1998). Development and assessment of common Lexical specifications for six central and Eastern European languages. *Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98*, Granada, pp. 233–240.

Ipšić, I., Mihelič, F., Dobrišek, S., Gros, J., and Pavešić, N. (1998). An overview of the spoken queries in European languages: The Slovene spoken dialog system. *Proceedings of the Scientific Conference Artificial Intelligence in Industry from Theory to Practice and 3rd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, High Tatras, Slovakia, pp. 431–438.

Kačič, Z. and Horvat, B. (1998). Setting up the speech resources needed for the development of speech technology for Slovene language. *Proceedings of the Conference on Language Technologies for the Slovene Language*, Ljubljana, pp. 100–104.

Kačič, Z., Horvat, B., and Zögling, A. (2000). Issues in design and collection of large telephone speech corpus for Slovene language. *Proceedings of LREC 2000, 2nd International Conference on Language Resources & Evaluation*, Athens, Greece, pp. 943–946.

Kaiser, J. and Kačič, Z. (1998). Development of Slovene SpeechDat database. *Proceedings of the Workshop on Speech Database Development for Central and Eastern European Languages*, Granada, Spain.

Mihelič, F., Gros, J., Noeth, E., and Warnke, V. (2000). Recognition and labeling of prosodic events in Slovene speech. *Text, Speech and Dialogue (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence, 1902)*. Berlin, Heidelberg: Springer, pp. 165–170.

Mihelič, F., Ipšić, I., Žibert, J., and Martinčić–Ipšić, S. (2002). Development of a SLO-CRO Bilingual Speech Database. *International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2002, Proceedings*, pp. 577–581.

Odell, J., Kershaw, D., Ollason, D., Valtchev, V., and Whitehouse, D. (1998). *The HAPI Book*. Entropic, Cambridge, Great Britain.

Pepelnjak, K., Mihelič, F., and Pavešić, N. (1996). Semantic decomposition of sentences in the system supporting flight services.

CIT—Journal of Computing and Information Technology, Zagreb, *4*(1):17–24.

Sperberg-McQueen, C.M. and Burnard, L. (Eds). (1994). Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford, ACH/ACL/ALLC Text Encoding Initiative.

Škerl, M., Mihelič, F., Gros, J., and Dobrišek S. (2001). Speech corpora VINDAT—The influence of the psychophysical condition of the speaker on speech characteristics. *Proceedings of 10th Electrotechnical and Computer Science Conference ERK 2001*, Portorož, Slovenia, pp. 261–264.

Šustaršič, R., Komar, S., and Petek, B. (1999). *Slovene IPA Symbols. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press, pp. 135–139.

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

Young, S., Kershaw, D., Odell, J., Ollason, D., Vatchev, V., and Woodland, P. (2000). *The HTK Book*. Cambridge, Great Britain: Cambridge University Engineering Department.

Žibert, J., Gros, J., Dobrišek, S., and Mihelič, F. (1999). Language model representations for the GOPOLIS database. *Text, Speech and Dialogue, (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence*, vol. 1692). Berlin [etc.]: Springer, pp. 380–383.

Žibert, J. and Mihelič, F. (2000). Slovenian weather forecast speech database. *International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2000, Proceedings*, vol. 1, 199–206.

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

Žibert, J., Mihelič, F., and Pavešić, N. (2002). Speech Features Extraction Using Cone-Shaped Kernel Distribution. *Lecture Notes in Artificial Intelligence*, vol. 2448. Berlin: Springer, pp. 245–252.