Development of a Bilingual Spoken Dialog System for Weather Information Retrieval

Janez Žibert†, Sanda Martinčić-Ipšić*, Melita Hajdinjak†, Ivo Ipšić*, France Mihelič†

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

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

Abstract

In this paper we present a strategy, current activities and results of a joint project in designing a spoken dialog system for Slovenian and Croatian weather information retrieval. We give a brief description of the system design, of the procedures we have performed in order to obtain domain specific speech databases, monolingual and bilingual speech recognition experiments and WOZ simulation experiments. Recognition results for Croatian and Slovenian speech are presented, as well as bilingual speech recognition results when using common acoustic models. We propose two different approaches to the language identification problem and show recognition results for the both acoustically similar languages. Results of dialog simulations, performed in order to gain user behaviors when accessing a spoken dialog system, are also presented.

1. Introduction

The paper describes the development of a bilingual spoken dialog system for weather information retrieval1 (Fig. 1). Using such a system a user can ask questions about weather conditions and forecasts in two different languages. The dialog system would provide information about weather in different regions of Slovenia and Croatia and for different time periods in the two languages, collecting the information from the available web sites over the Internet.

One of the possible reasons to build such an information system is a great number of tourist exchange between both countries during winter and summer periods and the fact that from the time when Slovenia and Croatia became different countries, less and less citizens in these countries speak and understand both languages.

Since the first step in spoken dialog system design is the collection of speech material, we first describe the process of speech data acquisition, segmentation and transcription of the Croatian and Slovenian domain related speech databases. In the next sections we present the language and speech recognition experiments, and show some results. Next we describe the experiments, we have performed in order to collect information about user behavior, possible dialogs and spontaneous speech material. We give a brief description and simulation results of the Wizard-of-Oz (WOZ) system for the Slovenian language using the weather forecast database and the set of predefined system answers. We conclude with the description of current activities and future plans in spoken dialog system development.

2. Speech databases

To develop a spoken dialog system a large collection of speech material is needed in order to build statistical models for the word recognition process. We are currently collecting weather forecasts spoken within news broadcast of national TV, radio programmes and spontaneous speech using WOZ simulation.

2.1. Croatian weather forecasts

The Croatian speech database consists of weather forecasts read by professional speakers within the news broadcasts of the national radio. The recordings have been performed since March 2002. The weather forecasts were recorded four times a day using a PC with an additional hardware. 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 are used for speech transcription and for training of a language model for weather forecast speech recognition experiments. The transcription is performed on the word level with some additional marks for silence, breath, paper noise, coughs and restarts.

Speech database contains speech files of 11 male and 12 female speakers and consists from 644 weather forecast recordings lasting 6 hours and 26 minutes. The vocabulary consists of 1160 words.

2.2. Slovenian weather forecasts

The Slovenian speech database consists of television weather forecasts. The recordings were performed from October 1999. Speech database consists of weather reports captured three times a day on the national TV program.

Television weather forecasts were recorded using a PC with a built-in TV receiver. The recordings were sampled with the same sampling rate and format as in the Croatian database.

---

1This study is partly founded with the grant of Slovenian Ministry of education, science and sport and Croatian Ministry for science and technology, contr. nr. SLO–CRO 23.
In Slovenian weather forecast recordings data is characterized as planned speech. Forecasts have been prepared in advance and have been read with some spontaneous speech effects as disfluencies and/or hesitations.

The database contains 5 speakers (1 female + 4 male) and was collected from one to three minutes long television weather forecasts. We collected and transcribed 233 forecasts of approximately 5 hours and 22 minutes of speech material. The vocabulary includes 3312 words.

The transcription process was semi-automatic. When enough speech material was collected we built a speech recognition system which was used for the initial transcription of the recorded speech data. Transcription was later corrected by a human expert. All final transcriptions were made with the Transcriber tool [1].

2.3. Slovenian spontaneous speech queries

A total of 76 speakers (38 female, 38 male) were chosen for the Wizard-of-Oz experiments. The selection of speakers was made according to their ages, education and their dialects; the types of telephone lines (analog line, ISDN line, GSM) and the background environments that simulate the actual scenarios. The average age of speakers is 34 years and their average education is secondary school level. All 8 major Slovenian dialects were covered.

Speech data was recorded using a PC with a built in ISDN server card, with the same sampling rate and format as in previous databases. The collected speech data was transcribed using the Transcriber tool [1]. Beside the word labels, special labels for dialectal words were used, together with additional several marks for non-speech events.

The corpus consists of 685 queries. The vocabulary includes 949 different words extracted from the corpus of 5103 words in the database. This vocabulary enriched Slovenian weather forecasts dictionary with additional 402 new words.

The collected query corpus revealed some interesting information. In our experiments the majority of speakers started their questions with "I am interested in ..." or similar. Another interesting effect is the adaption of the speaker’s behavior to the expected language abilities of the counterpart. In several dialogues the first introductory question was much longer than the following ones. In the case of repetitions requested by the system the speech mode became more articulated, slower and/or louder.

3. Speech recognition experiments

The speech recognition system is based on continuous hidden Markov models of crossword triphone units. The training of acoustic and language models was performed using the HTK toolkit [2].

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 8 ms for a speech signal frame length 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 [3]. The selected phonemes are derived according to [4]. Machine readable phonetic alphabet (MRPA) symbols [5] were used for the transcription of the Slovenian speech database.

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

<table>
<thead>
<tr>
<th>speaker</th>
<th>% correct</th>
<th>% accuracy</th>
<th>% sentence corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRO</td>
<td>93.86</td>
<td>88.95</td>
<td>41.73</td>
</tr>
<tr>
<td>SLO</td>
<td>93.66</td>
<td>90.99</td>
<td>56.47</td>
</tr>
</tbody>
</table>

Table 1: Croatian and Slovenian speech recognition results in terms of sentence and word correctness and word accuracy.

In Table 1 results of separated Croatian and Slovenian word recognition scores in terms of correctness and accuracy are depicted. The recognition results were obtained using crossword triphone models with mixture of 3 Gaussian density functions per state. Additional models for silence, breath noise, paper noise and restarts were used in all experiments. Bigram language models with perplexity 19.57 and 10.73 respectively were used for Slovenian and Croatian recognizers.

The experiments have shown similar recognition results for the Croatian and Slovenian speech data. 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 have been already used for transcription purposes of new speech material, for the development of HMM based text-to-speech systems for both languages [6] and initialization of acoustic models for the spontaneous speech recognition system.
3.1. Language identification

There exist several techniques for simultaneous multilingual speech recognition [7]. In the development of a bilingual Croatian and Slovenian spoken dialog system for weather information we explored two methods for the simultaneous bilingual speech recognition.

In the first approach the language identification was performed first, and after that the recognition using monolingual recognizers, Fig. 2. To detect the spoken language a monophone based recognizer was used. The monophone inventory contains 30 Croatian and 32 Slovenian acoustic units and additional models for silence, breath noise, paper noise and breaks. The recognizer was built using 3-state HMMs with one Gaussian probability density function per state.

In the language identification process after a few seconds or after a fixed number of recognized monophones of the spoken utterance the model chose one language by counting the number of recognized units of each language. In our experiments using the test part of both speech databases the language identification module worked with an error of false acceptance and false rejection rate of less than 1%. Thus, it had small influence on the speech recognition results which remained almost the same as in the monolingual speech recognition tasks.

The second approach combines the lexicons, the acoustic and language models to one bilingual recognition system. There we used one set of context-dependent acoustic units shared across the languages and a concatenated language model. In this approach language identification is done implicitly.

3.2. Simultaneous bilingual speech recognition

Sharing of acoustic units in the second approach can reduce the number of parameters and guarantee better training conditions, if we can use similarities between the involved languages. The process of joining the parameters and models into one bilingual speech recognizer was done in several stages.

The language model was built by concatenating Slovenian and Croatian back-off bigram language models, Fig. 2 (inside dotted-line area). This was done in a way that only words from one language were allowed. Thus each word chain hypothesis contains words from only one language and no transitions between the languages were possible. Perplexity of this joint bigram language model was estimated to 16.08.

The language model was built by concatenating Slovenian and Croatian back-off bigram language models, Fig. 2 (inside dotted-line area). This was done in a way that only words from one language were allowed. Thus each word chain hypothesis contains words from only one language and no transitions between the languages were possible. Perplexity of this joint bigram language model was estimated to 16.08.

In the language identification process after a few seconds or after a fixed number of recognized monophones of the spoken utterance the model chose one language by counting the number of recognized units of each language. In our experiments using the test part of both speech databases the language identification module worked with an error of false acceptance and false rejection rate of less than 1%. Thus, it had small influence on the speech recognition results which remained almost the same as in the monolingual speech recognition tasks.

The second approach combines the lexicons, the acoustic and language models to one bilingual recognition system. There we used one set of context-dependent acoustic units shared across the languages and a concatenated language model. In this approach language identification is done implicitly.

Table 2: Bilingual speech recognition results using different acoustic models: monophones, triphones with less similarities (ls) and triphones with more similarities (ms).

<table>
<thead>
<tr>
<th></th>
<th>% correct</th>
<th>% accuracy</th>
<th>% sent. corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>monophones</td>
<td>91.45</td>
<td>89.90</td>
<td>43.65</td>
</tr>
<tr>
<td>triphones (ls)</td>
<td>94.84</td>
<td>91.59</td>
<td>55.43</td>
</tr>
<tr>
<td>triphones (ms)</td>
<td>94.69</td>
<td>91.33</td>
<td>55.39</td>
</tr>
</tbody>
</table>

The next stage of the merging process was performed on the context-dependent acoustic models. We used the crossword triphone units. Merging was applied in the parameter-tying part of the training process, where we used similar phonetic characteristics of Slovenian and Croatian speech to form common phonetic rules for tree-based clustering of the context dependent acoustic units. We experimented with two kind of phonetic rules, where we used less or more phonetic similarities between involved languages. In the first case we applied separate phonetic rules of both languages, thus clustering was performed on acoustic units from both languages separately (Table 2 second row). In the second case similar phonetic rules for both languages were merged and applied to tree-based parameter tying to form common triphone classes (Table 2 third row). As can be seen from Table 2 there are no significant differences in speech recognition results when comparing both approaches.

4. Dialog system simulation

We decided to use the Wizard-of-Oz (WOZ) experiments [8] to collect data for speech recognition purposes, speech understanding analysis and construction of the dialog manager for developing a dialog system for Slovenian and Croatian weather information retrieval.

The WOZ system was designed as an internet application based on the graphical interface, HTML code and CGI scripts in order to minimize the necessary typing and to maximize the uniformity of the system’s responses.

All the wizard’s utterances were produced by a speech synthesizer in the Slovenian language [9]. For this purpose the wizard’s graphical interface includes facilities for the playback of predefined spoken commands and general answers, as well as forms, image fields and could handle keyboard shortcuts. The wizard communicated with information database through CGI scripts.

The information database was automatically updated several times a day. The data was extracted from the web pages of the national meteorological institute EARS, which is the only weather information provider in Slovenia. One part of this data (the weather forecasts) was in text form and needed to be semantically analyzed [10] and transformed into language independent semantic table data structure. Our database contains only table structured information, which is different from the
approach used in the JUPITER system [11].

The language generation process followed by the completion of the database search. The generated answers were then sent to the TTS system and transmitted as speech utterances to the user over the telephone line.

The participant in the WOZ experiment had to accomplish two tasks. The first one was to find certain information about weather. The second task was a given situation and the participant had to gather specific information about the weather. An example of such situations was “You are planning a trip to... What are you interested in?” After completing both tasks participant were told to ask questions about weather conditions and forecasts in general.

Simulating speech understanding and speech recognition errors is a hard task in WOZ experiments, so we decided not to model this part of system behavior.

In order to evaluate user satisfaction rate a statistical survey was performed where users were asked to evaluate the system’s performance. The user was asked to rate the task they had completed the task or not. The rest of the questionnaire probed a number of different aspects of the user’s perceptions of their interaction with the system in order to focus the user on the task of rating the system, as in [12].

The answers were given with a five class ranking scale between 1 (strong disagreement) and 5 (strong agreement). Each question emphasized the user’s experience with the WOZ system in the current conversation. The primary goal of our WOZ study was to record and analyze user’s input and wizard’s output. At the same time we examined if WOZ system with a human operator could perform the domain task well enough in order to use the collected data in the future as a gold standard [13] in building the dialog system.

The average user rating of the interaction pace was 3.94. This was rather good if we consider that the user had to wait for the answer for 4.80 seconds on average. An important information was the average user opinion of the future usage of the system for weather information retrieval, which is 3.99. All the other average ratings (TTS Performance, ASR Performance, Task Ease, User Expertise, System Response, Expected Behavior), that totally or partly referred to the wizard in these experiments, lied between 4.23 and 4.51. So we could conclude that the wizard performed reasonably well in the domain task.

5. Conclusion

In this paper we have presented the work we have done within a joint project in the development of a bilingual spoken dialog system providing weather information data. The paper gives some statistical information about the so far collected speech data, describes transcription procedures and speech recognition experiments. 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. Bilingual speech recognition and language identification experiments show that there is no big loss of accuracy when combining two languages into one bilingual speech recognition system.

Preparation of a spontaneous bilingual speech recognizer is under development. Our research experience from previous projects on spoken dialog systems [14] will be used in the development of a semantic speech analysis module and a dialog manager. WOZ experiments for Slovene language give some promising results providing spontaneous speech data, users behavior estimations and quite high scores concerning users experiences with the system. Extension of these experiments to the Croatian language and bilingual test are planned for the future. For this purpose weather forecast data retrieval, semantic analysis and TTS for the Croatian language will be developed.

6. Acknowledgements

The authors would like to thank the Environmental Agency of the Republic of Slovenia for kindly providing the weather reports for transcription purposes and allowing us to use their data in our WOZ experiments.

7. References


