Preliminary Evaluation of Slovenian Mobile Database PoliDat

Andrej Žgank, Zdravko Kačič, Bogomir Horvat

Institute of Electronics, Faculty of Electrical Engineering & Computer Science,
University of Maribor
Smetanova ul. 17, SI-2000 Maribor
Slovenia
{andrej.zgank, kacic, bogo.horvat}@uni-mb.si

Abstract

The following paper describes the preliminary speech recognition evaluation of PoliDat database. This new database contains Slovenian speech captured over mobile telephones. The design of database is modeled according to the SpeechDat(II) specifications. The recording of speech material and the format of the database are shortly described. The speech recognition experiment is based on slightly modified COST 249 refrec0.96 script. Acoustic HMM speech models are trained on the fixed telephone Slovenian 1000 FDB SpeechDat(II) database. 40 speakers were taken from mobile PoliDat database, 20 for test set and 20 for adaptation set. First the signal to noise ratio of all recordings was calculated, then the speech recognition with unadapted acoustic models was performed. In the next step the retraining of acoustic models and maximum likelihood linear regression procedure were used for adaptation. In the last step, the adapted acoustic models were used for speech recognition with the PoliDat database. The adaptation procedures significantly improved the mobile speech recognition with fixed acoustic models. The overall word error rate decreased from 46.5% for unadapted models to 19.1% and 5.2% for adapted models.

1. Introduction

This paper presents the first speech recognition evaluation of Slovenian mobile phone PoliDat database[1]. Different speech resources for the Slovenian language (Kaiser and Kačič, 1998; Kačič, et.al., 2000; Žibert, et.al., 2000) already exist, but there is no large speech database collected over the mobile phone. Further reason to start the project of collecting the PoliDat database lies in the wide circulation of mobile phones in Slovenia, which is also reflected in the large number of mobile phones that are used in different speech recognition applications. At the end of the year 2001, the penetration rate of mobile phones in Slovenia was approximately 67% (Stergar, 2002; www.ris.org, 2001).

The first preliminary evaluation of PoliDat database was done on a set of 40 speakers whose captured speech recordings were edited, verified and transcribed. Because the quantity of speech material in this set of PoliDat database isn't large enough to train reliable acoustic models, another approach was utilized in the evaluation process.

The speech models were designed with the use of the most similar database that exist for the Slovenian language. In this case, this was the fixed telephone 1000 FDB SpeechDat(II) database (Kaiser and Kačič, 1998). The generated fixed phone speech models were used for speech recognition of mobile PoliDat database. Due to the fact that different phone types and environments from where each call was performed, significantly influence the quality (signal to noise ratio, analog vs. digital transmission, background noise, ...) of captured speech material, two different acoustic adaptation procedures were applied. It is expected that these adaptation procedures would compensate the deficiency of mobile speech material. Also the signal to noise ratio for all mobile PoliDat recordings used in this experiment was calculated and the values were compared to those of the fixed SpeechDat(II) database.

This paper is constructed as follows: the new Slovenian mobile PoliDat database is presented in Section 2. The speech recognition system constructed with fixed SpeechDat(II) database is described and tested in Section 3. The evaluation with unadapted acoustic models and analysis of quality is done in Section 4. The adaptation process and tests with adapted acoustic models are presented in Section 5. The discussion and conclusion of database evaluation are given in Section 6.

2. PoliDat database

When the PoliDat database format was planned, the SpeechDat databases (Höge, et.al, 1997; van den Heuvel, et al., 2001) were already widely accepted and proven in different tasks and configurations. Further important thing was the fact that the Slovenian version of FDB SpeechDat(II) (Kaiser and Kačič, 1998) was also created at the University of Maribor in cooperation with Siemens AG. When all these factors were considered, the decision was taken that the PoliDat database should be generated according to the SpeechDat(II) specifications (van den Heuvel, et al., 2001). This decision ensures the high quality of database and the possibility to compare the results with other SpeechDat databases (Refrec home, 2002; Johansen, et.al., 2000).

The PoliDat database speakers were recruited from employees and students from University of Maribor and from employees of Slovenian national telecommunication provider Telekom d.d., whose branches are distributed over the whole country. The speakers were chosen in such a way that different demands were taken into consideration:

- age,
- gender,

[1] This work was supported by the Slovenian Ministry of Education, Science and Sport under contract number PP-0796/99-0796-503.

- dialect,
- education, ...

The number of finally collected speakers will be around 1400, then 1000 of them will be selected for the final version of PoliDat database according to the quality of recordings. Each speaker read the speech material from prompt sheet, that was sent to him. Small part of speech material is also spontaneous, obtained as answers to different questions (e.g.: Are you older than 80 years?). More than 1600 different templates were used for the prompt sheets. The template contains 43 different utterances:

- spelling,
- digits and numbers,
- application words,
- yes/no answers,
- different names,
- phonetically balanced isolated words and sentences,
- other types, according to the SpeechDat(II) specifications (van den Heuvel, et al., 2001).

The speech signal from mobile phone was captured with an ISDN interface card, which was connected to a host computer. The use of ISDN card ensures the digital transmission channel between the phone and ISDN interface card. The speech is captured directly to host computer's hard disc in ISDN 711 file format.

Each recording was verified and transcribed with the proprietary developed tool ”PoliDat Label”. Different acoustic events (e.g.: static noise, cough, breath, ...) where marked in the final orthographic transcription, to improve the quality of acoustic modeling. The ISDN 711 file format was converted into A-law file format, as recommended in the SpeechDat(II) specifications (van den Heuvel, et al., 2001). At the end of database generation procedure, all accompanying files (e.g.: lexicon, test set, list of sessions, ...) according to the SpeechDat(II) standards (van den Heuvel, et al., 2001) will be generated.

3. Evaluation system

3.1. Training procedure with fixed database

The first step in the preliminary evaluation procedure of mobile PoliDat database was the preparation of suitable acoustic models. Due to the fact that there is no large mobile telephone database for the Slovenian language available at the moment, the fixed telephone Slovenian SpeechDat(II) database (Kaiser and Kačič, 1998) was an appropriate substitute. To enable the comparison of results, the noise robust COST 249 SpeechDat task force refrec0.96 script (Lindberg, et.al., 2000) with different frontend was used in the development process. This script was used to automatically train HMM acoustic models with SpeechDat(II) databases.

The frontend module was used to convert the speech signal into 24 mel cepstral coefficients and high pass filtered energy. 13 first and 13 second derivatives of mel cepstral coefficients and energy were also added to the feature vector. The linear discriminant analysis was used to reduce the complexity of feature vector to 24 elements. Feature extraction procedure also included the maximum likelihood channel adaptation (Haunstein and Marschall, 1995) and linear discriminant analysis (Haunstein and Marschall, 1995) to improve the robustness of the system to more degraded speech over mobile phones.

According to the SpeechDat(II) database specifications (van den Heuvel, et al., 2001), 800 speakers were in the training set and 200 in the test set. Some recordings were excluded from the training set due to unintelligible and truncated speech. After this procedure, the final training set consists of 28790 different utterances. Special acoustic models for noise events during the speech are also added to the phoneme inventory in the training process (Lindberg, et.al., 2000), because the frequency of such events in the mobile PoliDat database is significantly higher than in the fixed SpeechDat(II) database. The acoustic models used in the speech recognizer are standard, 3 state left – right hidden Markov models with continuous Gaussian mixtures densities. In the first part of training process the monophone acoustic models are generated from scratch (Lindberg, et.al., 2000). In the second part of this process the acoustic models were refined with the use of context – the triphone acoustic models. At the end, the number of Gaussian densities per HMM state was increased to 32, so that acoustical diversity of speech was better modeled.

3.2. Testing with fixed database

The quality of final acoustical models was checked and compared to the word error rates (WER), attained in the COST 249 experiment (Refrec home, 2002). For this purpose, six different test sets were used (van den Heuvel, et al., 2001):

- A1-6, application words, 31 words in the recognition vocabulary,
- Q1-2, yes/no answers,
- I1, isolated digits,
- B1,C1, connected digits,
- O2, city names, 597 words in the recognition vocabulary,
- W1-4, phonetically balanced isolated words, 1491 words in the recognition vocabulary.

Word error rate for all six test configurations and the overall result are presented in Table 1. As can be seen, the best result with fixed test set is achieved with the yes/no answers (Q set) with 0.87% WER, which is also the simplest test configuration, with regard to the recognition vocabulary. The hardest task is the W test configuration (14.95% WER), with almost 1.500 words in the recognition vocabulary. The comparison with the COST 249 results in the

Acoust. Models	A1-6	I1	B1,C1	Q1-2	O2	W1-4	Overall
COST 249	3.46%	5.70%	4.73%	1.73%	8.61%	18.56%	7.78%
modif. COST 249	2.99%	4.15%	4.73%	0.87%	11.34%	14.95%	6.65%

Table 1: Comparison of word error rate (WER) for original COST 249 script (Refrec home, 2002) and for modified COST 249 script with different frontend, which was used in the experiment as fixed reference

first line shows that different frontend improved the speech recognition results for fixed SpeechDat(II) database in almost all cases. The overall word error rate improved from 7.78% to 6.65%. The major improvement was observable in the case of W test set, where the word error rate decreased for more than 3% absolute. The WER increased only in the case of O test set.

4. Evaluation with unadapted speech models

The second step in the preliminary speech recognition evaluation process was the recognition of mobile speech from PoliDat database with the fixed speech models from SpeechDat(II) database. Prior to the process of evaluation, the acoustic quality of all new material from PoliDat database was checked with the calculation of signal to noise ratio (SNR). All recordings were classified into five different classes, presented in Table 2, with SNR values from under 10 dB to over 40 dB.

The average signal to noise ratio of all recordings used in this preliminary speech recognition evaluation was 38.69 dB. The majority of all recordings belong to class 3 (43.06%) and class 4 (46.77%), their SNR value was over 30 dB. The recording with the best acoustical quality has SNR of 49.21 dB. The worst recording in this preliminary evaluation was with SNR of 9.83 dB. The average SNR value for all utterances from the fixed SpeechDat(II) database, that was used for training of acoustical HMM models was 25.77 dB. The probable cause for this large difference in the SNR value between both databases lies in the fact, that at the time when the SpeechDat(II) database was collected, the majority of local exchanges in Slovenia were analog. The Slovenian national fixed telephony provider Telekom, d.d., has modernized all local exchanges to digital type in the meantime.

The set of recordings from PoliDat database was selected in such a way, that chosen recordings belonged to equal test sets, that were chosen in Section 3.2 for fixed SpeechDat(II) database. 20 speakers were used for testing and other 20 for acoustical adaptation of fixed telephone speech models to mobile telephone acoustic environment and speech quality. The first results for speech recognition with unadapted acoustic models are presented in Table 3.

The word error rate in the case of unadapted acoustic models was much higher than in the fixed reference configuration (see Table 1). The contrast in the WER between less and more complex test configurations (Q, I, BC, versus A, O, W) is still noticeable. The best result was achieved by the Q test set (yes/no answers) with 21.6% WER. The test sets with digits (I and BC) followed with the 30.0% and 38.9% word error rate. The worst result was again for the most complex W test set (91.9% WER). The overall WER of unadapted acoustic models was 46.5% - significantly more than in the fixed reference system, where it was 6.65%.

5. Evaluation with adapted speech models

5.1. Adaptation

We decided to apply two different acoustic adaptation procedures to the final speech models trained on fixed database, to improve the performance of the recognizer with the speech over mobile phones:

- retraining: the mean values and the mixture weights of Gaussian probability functions of HMM acoustic models were retrained on small adaptation set.
- maximum likelihood linear regression: the method described in (Leggetter, 1995) was used for acoustical adaptation of speech models.

The adaptation set consisted from speech of 20 speakers, with 313 utterances from the same database sets as were used in the tests.

5.2. Testing

The last step of PoliDat database evaluation was the speech recognition with both versions of adapted speech models. The test set was the same as for unadapted acoustic models from Section 4. The results of this last database evaluation step are presented in Table 4.

The middle row (PoliDat_RETR) in Table 4 presents the speech recognition results of retrained acoustic models on small adaptation set. Significant improvement for all test sets, except for the O set, was achieved. The improvement of system accuracy was bigger for less complex test configurations. The word error rate for digits test sets (I and BC) improved from 30.0% and 38.9% to 10.0% and 5.6%. Observable is that the WER for isolated digits (I), which is simpler task, is greater than for the connected digits (BC). In the case of yes/no answers (Q set), all utterances were recognized correctly. The possible cause for aberration of performance in the case of O test set is the small number (19) of utterances in the test set. So small test set does not entirely reflect the accuracy of acoustic models. The overall word error rate after retraining of acoustic models was improved from 46.5% to 19.1%.

The last row (PoliDat_MLLR) in Table 4 shows the speech recognition results with the MLLR adapted acoustic models. This adaptation procedure improved the system accuracy even more than the retraining of acoustic models with small adaptation set. The decrease of word error rate was observable in all cases. All utterances were recognized correctly in two test sets – yes/no answers (Q) and isolated

	< 10dB	10-20dB	20-30dB	30-40dB	> 40dB
	cl. 0	cl. 1	cl. 2	cl. 3	cl. 4
all	0.16%	0.32%	9.68%	43.06%	46.77%

Table 2: All recordings classified into 5 different SNR classes and the ratio between these classes

Acoust. Models	A1-6	I1	B1,C1	Q1-2	O2	W1-4	Overall
PoliDat_UNAD	52.5%	30.0%	38.9%	21.6%	65.2%	91.9%	46.5%

Table 3: PoliDat word error rate (WER) with fixed unadapted acoustic models for different mobile test sets

digits (I). The word error rate for connected digits improved from 38.9% for unadapted models (Table 3) to 1.5% for MLLR adapted models. In the case of retrained acoustic models, the result improved to 5.6%. Similar improvements of speech recognition accuracy were also achieved with the A test set. The biggest decrease of word error rate attained with the W test set, where the WER decreased from 91.9% to 18.9% – more than 70% absolutely. The best overall word error rate of only 5.2% was also reached with MLLR adapted acoustic models.

The side effect of speech recognition accuracy improvement was also noticed in the decrease in the number of observations of special acoustic models (speaker noise) in the recognized utterances, although these special acoustic models were not included in the calculation of system accuracy. Both adaptation procedures significantly improved the mobile speech recognition accuracy in comparison to unadapted fixed acoustic models (see Figure 1), as was expected. If we compare both adaptation procedures we can see that the maximum likelihood linear regression outperformed the retraining of acoustic models. The advantage of retraining procedure lies in the fact that it can be performed with the same tool as is used for HMM training, but on the other side the MLLR method requires separate tool.

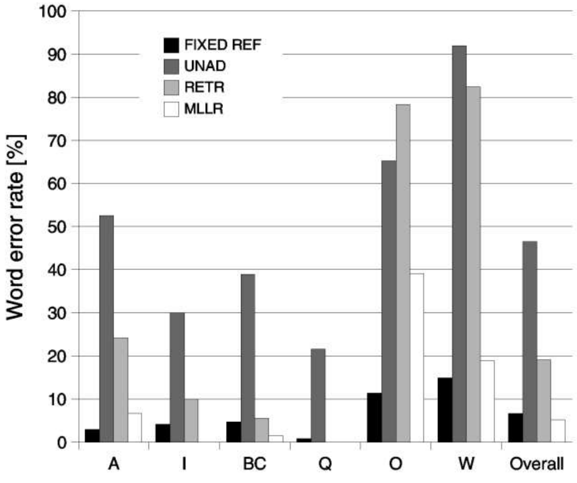

Figure 1: Extracted from Table 3 and Table 4: word error rate for different acoustic models.

6. Conclusion

The preliminary speech recognition evaluation of a new mobile Slovenian PoliDat database was presented in this paper. The fixed acoustic models were first used for recognition of mobile speech without adaptation. Due to differences between fixed and mobile telephone lines, the system performance degraded in comparison to fixed reference system. Later, two acoustical adaptation procedures were applied to fixed acoustic models. Both adaptation procedures significantly improved the results and confirmed the hypothesis that fixed acoustic models can be used for mobile speech recognition in the case of lack of appropriate mobile speech material. If we compare the achieved mobile PoliDat word error rates with other similar mobile databases (mobile SpeechDat(II)) for other languages, we can see, that the results are in the similar range(Refrec home, 2002).

In the future, when the mobile PoliDat database will be completed, the speech recognition evaluation will be performed on all 1000 speakers.

7. References

A. Haunstein, E. Marschall. 1995. Methods for improved speech recognition over the telephone lines. In: *Proc. ICASSP '95*.

H. van den Heuvel, L. Boves, A. Moreno, M. Omologo, G. Richard, E. Sanders. 2001. Annotation in the SpeechDat Projects. *International Journal of Speech Technology*, 4(2):127 – 143.

H. Höge, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach. 1997. European speech databases for telephone applications. In: *Proc. ICASSP '97*, pages 1771 – 1774, Munich.

F. T. Johansen, N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, G. Salvi. 2000. The COST 249 SpeechDat Multilingual Reference Recogniser. In: *Proc. LREC'2000*, Athens.

Z. Kačič, B. Horvat, A. Zögling. 2000. Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. In: *Proc. LREC-2000*, Athens.

J. Kaiser, Z. Kačič. 1998. Development of the Slovenian SpeechDat database. In: *Speech Database Development for Central and Eastern European Languages*, Granada.

C. J. Leggetter, P. C. Woodland. 1995. Flexible Speaker Adaptation using Maximum Likelihood Linear Regression. In: *Proc. ARPA Spoken Language Technology Workshop*, pages 104–109, Austin, Texas.

Acoust. Models	A1-6	I1	B1,C1	Q1-2	O2	W1-4	Overall
PoliDat_RETR	24.2%	10.0%	5.6%	0.0%	78.3%	82.4%	19.1%
PoliDat_MLLR	6.7%	0.0%	1.5%	0.0%	39.1%	18.9%	5.2%

Table 4: PoliDat word error rate (WER) for retrained (PoliDat_RETR) and adapted (PoliDat_MLLR) acoustic models with different mobile test sets

B. Lindberg, F. T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, G. Salvi. 2000. A noise robust multilingual reference recogniser based on SpeechDat(II). In: *Proc. ICSLP 2000*, Beijing, China.

A. Stergar. 2002. V Mobilkomu zadovoljni. In: *Newspaper Delo, 2.3.2002*, Ljubljana, Slovenia.

J. Žibert, F. Mihelič. 2000. Govorna zbirka vremenskih napovedi. In: *Information Society multi-conference: Language Technologies*, Ljubljana, Slovenia.

RIS, http://www.ris.org/ict/mob_tel_junij2001.htm

Refrec home, http://www.telenor.no/fou/prosjekter/-taletek/refrec/