Uniform Speech Recognition Platform for Evaluation of New Algorithms

Andrej Žgank, Tomaž Rotovnik, Zdravko Kačič, and Bogomir Horvat

Institute for Electronics,Faculty of EE & CS, University of Maribor,
Smetanova 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

Abstract. This paper presents the development of speech recognition platform, which main area of use is the evaluation of different new and improved algorithms for speech recognition (noise reduction, feature extraction, language model generation, training of acoustic models, ...). To enable wide use of the platform, different test configurations were added – from alphabet spelling to large vocabulary continuous speech recognition. At the moment, the speech recognition platform was implemented and evaluated with a studio (SNABI) and a fixed telephone (SpeechDat(II)) speech database. ...

1 Introduction

In the process of developing a new algorithm or method for speech processing, evaluation and comparison of new results with the standard approach is very important. To simplify and standardize this step in the development procedure, we decided to build a uniform speech recognition platform.

The system is based on similar "COST 249 refrec" project [1] with many new and different features. There were also other similar projects in the past (e.g. SAM [2]), but none of them was designed for broad spectrum of tests with different database types. The speech recognition platform consists of perl scripts and programs in C language, and is implemented for Unix (tested on Linux and HP-UX). In the future, port to the Windows platform is planned. The scripts use the public domain HTK toolkit [3] for designing and testing of acoustical models.

2 Architecture overview

2.1 Training part

To achieve the uniform operation with different databases, the speech recognition platform is divided into three parts. The first part is the interface between the database and the data format of speech material that is needed in the second part. In the case when the data preparation part can not be completed with the platform, the user must prepare the data for the second part manually.

Because the interface handles the data processing, the second part can be universal for all databases and presents the core training procedure of the platform. The last part consists of different speech recognition test configurations. The currently presented implementation of uniform speech recognition platform is developed for two different acoustic environments: for studio (Slovenian SNABI-studio database [4], 13779 utterances in the training set) environment and for fixed telephone lines (Slovenian FDB 1000 SpeechDat(II) [5], 28938 utterances in the training set).

In database interface, audio signal is converted into features with the use of 12 mel cepstral coefficients and energy. With the first and second derivative, the size of the feature vector is 39. The user can simply add his own different frontend to the interface, so it is expected that different feature extraction methods will be tested in the future.

In the second part, the speech recognition system is constructed with speaker independent HMM acoustic models. Ordinary acoustic models for phonemes are generated with 3 state left-right topology, while there are additional models for acoustic events (silence, noise, cough, ...) with different, more complex topology. The user can choose between training of context independent or context dependent models. This way system accuracy, complexity and speed can be modified. The best models generated by the speech recognition platform are context dependent models with 8 Gaussian mixtures per state.

2.2 Testing part

Due to extensive spectrum of testing configurations (from small vocabulary isolated words to large vocabulary continuous speech) different language models are used. The most simple one is the word-loop model and the most complex are the bigram and trigram backoff language models [6]. The large vocabulary continuous speech recognition with trigram language model is performed with the two pass decoder. When testing different language models or recognition vocabularies with context dependent acoustic models, new unseen triphones can occur. This problem is solved in such a way, that unseen triphones are added to the acoustic models, without retraining the whole system. The n-gram language models, used in the SNABI implementation of the platform, are trained on 50M words text corpus from a Slovenian newspaper Večer.

3 Evaluation of implementation

First part of the platform evaluation was performed on SpeechDat(II) database [5]. The same test set (A – words, I – isol. digits, BC – conn. digits, Q – yes/no, O – city names, W – phon. rich words [7]) was used as in the "COST 249 refrec" project [1] to enable a comparison of the results. The results are presented in Table 1. The best performance for both types of acoustical models: monophones (8.67% WER) and triphones (1.45% WER), was achieved with the yes/no (Q1-2) test set. This test configuration was the easiest, due to only 2 words in the

Table 1. Number of utterances and word error rate (WER) for different SpeechDat(II) test sets

	A1-6	I1	B1,C1	Q1-2	O2	W1-4
Num. of utter.	1070	193	380	346	194	749
Monophones	21.59	11.92	13.40	8.67	40.72	52.87
Triphones	7.57	7.25	7.96	1.45	13.02	23.29

recognition vocabulary [7]. The worst speech recognition result was for the W1-4 (phonetically rich words) test set with 52.87% WER for monophones and 23.29% WER for triphones. The number of words in this recognition vocabulary was 1500, which also represents the hardest recognition task for SpeechDat(II) database. If we compare these results with the results from the "COST 249 refrec" project [1], we can see that WER for some test sets are similar or equal. The I1 test set in the "COST 249 refrec" project also achieved 7.25% WER and the Q1-2 test set 1.16% WER. The greatest difference is observable with the W1-4 test set, where the COST249 refrec project achieved WER of 34.31% for monophones. The main reason for this distinction in word error rate is the fact, that the training procedure in our platform is currently much more simple than in the "COST 249 refrec" project.

Table 2. Number of utterances and word error rates (WER) for SNABI test sets with word-loop language model

	ABC	Connected dig.	Isolated dig.	Words
Num. of utter.	52	100	48	271
Monophones	65.38	10.33	2.08	2.95
Triphones	19.23	4.00	2.08	0.74

Table 3. Number of utterances and word error rates (WER) for SNABI test sets with 2-gram and 3-gram language model

	MMC	Lingua 1	Lingua 2
Num. of utter.	1092	518	671
Monophones, 2-gram	28.08	42.04	65.05
Monophones, 3-gram	22.24	32.62	62.59
Triphones, 2-gram	9.87	18.32	45.53
Triphones, 3-gram	6.69	12.09	42.85

Second part of the platform evaluation was performed on SNABI models and database. The first subpart of tests (Table 2) was completed with the word-loop language model. The word error rate (WER) for comparable test configurations (isolated digits, connected digits, words) was significantly better than in the case of SpeechDat(II) database, due to studio quality of speech in the SNABI database. The hardest task with the word-loop language model was the alphabet spelling test set with 65.38% WER for monophones and 19.23% for triphones. The second subpart of tests (Table 3) with the SNABI database was performed with the use of n-gram language models. Lingua 1 test set is only acoustically independent from the training set, but Lingua 2 test set is acoustical and textually independent from the training set. The improvement of results when triphone acoustic models or 3-gram language model were used is obvious for all test sets in the second subpart. The best performance (6.69% WER) was achieved with MMC test set. The average number of words in a sentence for this test set was 5.9. The average number of words in the test set Lingua 1 was 10.2 – this fact is reflected in the increase of WER to 12.09% in comparison to MMC test set. The 42.85% WER for test set Lingua 2 is mainly caused by different topics of speech and text corpus used.

4 Conclusion

This paper describes the design and evaluation of uniform speech recognition platform, which main area of use will be evaluation of new algorithms. Implementation with Speechdat(II) and SNABI database was presented. The results were compared to previous published results with the same database. In the future, a part for database segmentation will be added to this platform. The platform and the results of implementations with different databases will be available on our home page.

References

1. Lindberg, B., Johansen, F.T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., Salvi, G.: A noise robust multilingual reference recogniser based on SpeechDat(II). ICSLP 2000, Beijing, China, 2000.
2. Pols, L.C.W.: Evaluating the performance of speech input/output systems. A report of the ESPRIT-SAM project. Proc. DAGA '91, 139- 150, Bochum, Germany, 1991.
3. Young, S.: The HTK Book (for HTK version 3.1). Cambridge University, 2001.
4. Kačič, Z., Horvat, B., Zögling A.: Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language. Proc. LREC-2000, Athens, 2000.
5. Kaiser, J., Kačič, Z.: Development of the Slovenian SpeechDat database. Proc. LREC-1998, Granada, Spain, 1998.
6. Clarkson, P.R., Rosenfeld, R.: Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proc. of the Eurospeech '97, Rhodes, Greece, 1997.
7. van den Heuvel, H., Boves, L., Moreno, A., Omologo, M., Richard, G., Sanders, E.: 2001. Annotation in the SpeechDat Projects. International Journal of Speech Technology, 4(2):127 – 143.