Comparative evaluations in the domain of automatic speech recognition

Alex Trutnev, Martin Rajman

Artificial Intelligence Laboratory
Institute of Core Computing Science
School of Computer and Communication Sciences
Swiss Federal Institute of Technology
IN (Ecublens), CH-1015 Lausanne (Switzerland)
{alex.trutnev, martin.rajman}@epfl.ch

Abstract

The goal of this contribution is threefold: (1) to present the results of a comparative evaluation of different, academic and commercial, speech recognitions engines; (2) to study relative performances of Hidden Markov Model and hybrid technologies, as used in state-of-the-art systems; and (3) to study the impact of different linguistic resources, such as simple word spotting, statistical and grammar-based language models, on the speech recognition accuracy. All the evaluations were made on the basis of the same test data sets and conclusions derived from the obtained Word Error Rate scores. The evaluated speech recognition engines are all speaker independent, continuous speech recognition engines, either academic systems widely used in the research community or commercial tools currently available on the market. In this work, we considered three academic systems (HTK, Sirocco, and Strut/DRSpeech) and two commercial ones (for the confidence reasons, we name these systems SRE1 and SRE2). The main obtained results are that (1) the Hidden Markov Model (HMM) based technology performs better than the hybrid approach in the case of unconstrained continuous speech, and (2) the academic systems perform better in the case of continuous speech in French, while the commercial systems show better recognition accuracy for continuous speech in German.

1. Introduction

The growing availability of different algorithms, technologies and systems in the domain of automatic speech recognition leads to a quite natural question: among the available options, what is the most adequate for a given application. Different evaluation campaigns, for example, NIST[1], can be mentionned, aiming among others at the evaluation of conversational telephone speech or the one of speaker recognition. As to the presented work, the main interest was not to try to provide an exhaustive overview of all available academic and commercial speech recognition engines (SREs) and state-of-the-art technologies. The aim was to present the results of the evaluation work that we had to carry out to choose the SRE to be used in the Inspire project[2] aiming at the developpement of a dialogue-based control of different home appliances in the framework of a SmartHome environnement.

The rest of the paper is organized as follows: section 2. describes the evaluation framework; section 3. the used SREs, section 4. the used train and test data; the results of evaluation are reported in section 5.; and finally, section 6. proposes the discussion of the obtained results.

2. Evaluation framework

The framework used for the evaluation of the different SREs was similar to the one proposed by NIST: all the considered SREs are used to recognize the same test data set, in our case conversational telephone speech acquired during the two field tests of the InfoVox project (Rajman et al., 2003). The obtained Word Error Rate (WER) scores (Gillick, 1989; Pallett, 1990) are then used as a measure of the global recognition performance of the SREs. For these evaluation performances, the acoustic models (AM) used by the all academic SREs were trained on audio data of the same quality as test data, the Swiss French Polyphone database of conversational telephone speech. The AMs of the commercial systems were used as provided. As to the language model (LM) training, the same data (transcriptions of dialogues recorded over the phone during Wizard-of-Oz experiments in the InfoVox project) was used for all the SREs.

The second task considered in this work was the comparison of HMM based and hybrid[3] technologies. For this purpose, the academic SREs using HMM on one hand and hybrid AM on the other hand were all used to recognize digits. Digits recognition was selected because no language model is necessary, and the obtained WER scores can thus be used as direct measures of the recognition performances of the underlying technologies. Data used to train the AM, as well as the testing data, were both extracted from the Swiss French Polyphone database.

Finally, the impact of the use of different linguistic resources (such as word spotting grammar, statistical and grammar based language models) on the recognition performances was studied. This was done for only one SRE, the commercial system SRE1, and the same data set was used to produce the required language models.

All the results presented hereafter were computed with *sclite* evaluation tool (http://www.nist.gov/speech/tools).

[1] http://www.nist.gov/speech/tests/index.htm

[2] http://www.inspire-project.org

[3] This technology consists in the combination of artificial neural networks (in our case, *MultiLayer Perceptron* (MLP)) for estimating phonemes probability distributions, and of HMMs for the modeling of the phonetic lexicon and the representation of sequences of words

3. Speech Recognition Engines

The systems used for the evaluations were the following:

HTK : (http://htk.eng.cam.ac.uk) an academic system using the HMM technology; supports the whole speech recognition process; includes numerous algorithms for features extraction and the possibility to train AM and LM. The version 3.1 of the software was used;

DRSpeech/STRUT : academic systems using a hybrid technology; support the whole speech recognition process; include numerous algorithms for features extraction and the possibility to train AM. The implementation of the softwares available at IDIAP[4] was used, and the version 2.9p1 of the *Noway* acoustic decoder was used;

Sirocco : (http://www.irisa.fr/sirocco) an academic acoustic decoder integrating estimation of phonemes probabilitiy distribution on the basis of feature vectors extracted by another system, *HTK* for example. The version 1.2.1 of the software was used in the combination with HTK;

SRE1 : a (HMM based) commercial system, supports the whole recognition process; includes the possibility to train LM;

SRE2 : a commercial system, supports the whole speech recognition process.

4. Test and train data

4.1. Test databases

Details about the used test data are presented in the tables 1 and 2 hereafter. In these tables, column "*Type*" indicates the type of audio input data, "*Volume*" provides the size of the databases in recordings and the size of the phonetic vocabularies in words, and "*Quality*" characterizes the coding of audio input data.

Database	Language	Type	Volume
InfoVox	French	Free speech	876/988
Inspire	German	Read speech	1370/235
SFP	French	Digits	420/30
SDGe	German	Digits	500/14

Table 1: Details about the used test data

Database	Quality
InfoVox	*a-law*, 8 KHz, 8 bits
Inspire	*wav*, 32 KHz, 16 bits
SFP	*a-law*, 8 KHz, 8 bits
SDGe	*a-law*, 8 KHz, 8 bits

Table 2: Details about the quality of the used test data

Some additional notes about the used test databases:

[4]http://www.idiap.ch

Inspire : 10 dialogues read by 10 speakers and recorded over a microphone for evaluation purposes during the Inspire project;

SFP : digits subset extracted from the Swiss French Polyphone database, 420 recordings, telephone speech, from 2 to 15 digits per recording;

SDGe : digits subset extracted from the German SpeechDat(II) database, 500 recordings, telephone speech, 1 digit per recording;

InfoVox 70 telephone dialogues recorded in French speaking Switzerland during the two field-tests of the InfoVox project. 236 sentences were recorded during the first (internal) field-test, 640 sentences were recorded during the second (external) field-test.

4.2. Acoustic Models

The table 3 presents the AMs used by the SREs. In this table, the column "System" indicates the system used to prepare the AM, the column "Language" the language of the AM, and the column "Model" the type of audio input the model can recognize.

System	Language	Model
HTK	Fr	Continuous telephone speech
STRUT	Fr	Continuous telephone speech
SRE1	Fr/Ge	Unconstrained speech
SRE2	Fr	Unconstrained speech

Table 3: Acoustic models used by the SREs

Additional notes on the table:

1. the *Sirocco* system uses the same AM as *HTK*, only a format conversion is necessary;
2. the AM used by *SRE1* and *SRE2* are delivered with the systems;
3. for the *HTK* AM, 26 Mel Frequency Cepstral coefficients (12 coefficients, energy, delta coefficients and delta energy) were used. Context-independent models were trained (36 phonemes for French and 48 phonemes for German), and each of them had 3 states and 24 gaussians per state;
4. for the *STRUT* AM, a MultiLayer Perceptron (234 input units, one hidden layer with 600 units, 36 output units) was trained. The input units correspond to 234 coefficients (a context of 9 consecutive feature vectors, each vector contains 26 coefficients, 12 RASTA-PLP coefficients along with their first derivatives 12ΔRASTA-PLP, as well as Δ-log- and $\Delta\Delta$-log-energy);
5. the training of the *HTK* and *STRUT* AM was made on the basis of the same data: 3000 sentences from Swiss French polyphone database used for training, and 200 sentences for cross-validation. For more details about the training of these AMs, see (Andersen et al., 1997).

4.3. Language Models

The SRE1 commercial system is provided with a specific tool to train language models in its proprietary formats. For the other systems, the CMU Toolkit[5] was used to produce one language model that has then been converted in the formats supported by each of the concerned systems.

Table 4 presents the data used to train the LMs. The column "Database" indicates the sources, and "Volume" the size of the database in sentences and words.

Language	Database	Volume
French	InfoVox WoZ	2'183/22'790
German	German SpeechDat	26'831/295'586
German	Inspire WoZ	137/757

Table 4: Data used to train language models

Additional note on the table: *InfoVox WoZ* stands for the Wizard-of-Oz experiment carried out in the InfoVox project; 255 dialogues involving a total of 100 persons were recorded during the experiment.

5. Results of the comparative evaluations

5.1. Comparative evaluation of the SREs

The results of evaluation of the SREs are presented in the tables 5 for the InfoVox data and 6 for the Inspire data.

System	**WER**, %
HTK	63.3
SRE2	65.0
SRE1	66.3
Sirocco	68.9
DRSp/STRUT	76.6

Table 5: Evaluation of the SREs on the *InfoVox* data

System	**WER**, %
SRE1	35.8
HTK	81.0

Table 6: Evaluation of the SREs on the *Inspire* data

The first conclusion that can be drawn from the obtained results is that the academic systems using the HMM technology have the similar performances as the commercial ones. As to the hybrid technology, it performs worth than the HMM based technology. The poor performance of HTK on the Inspire data is due to the fact that the used AM is of weak recognition quality. This is shown in the next section.

The huge difference in the recognition of the InfoVox and Inspire databases observed for the SRE1 commercial system can be explained by the quality of the used LMs: in the case of the InfoVox database, the train corpus is composed of only 22'790 words, in the case of the evaluation on the Inspire data, the size of train corpus is more than 290'000 words.

[5] http://mi.eng.cam.ac.uk/p̃rc14/toolkit.html

5.2. Comparative evaluation of the acoustic models

The results of evaluation are presented in the table 7 for French and 8 for German.

System	**WER**, %
SRE2	8.0
SRE1	20.6
HMM	22.1
ANN	36.0

Table 7: Evaluation of the AMs on the *SwissFrench Polyphone* data

System	**WER**, %
SRE1	0.4
HMM	26.6

Table 8: Evaluation of the AMs on the *German SpeechDat* data

The obtained results show that the HMM techonology performs better than the hybrid one. Then, the substantial difference between the performances of digits recognition with commercial and academic SREs may be due to the fact that the commercial SREs have often the possibility either to switch between several AMs tuned to specific applications, or to tune the used AM online during the recognition process.

Finally, the huge difference in the recognition of Swiss French Polypphone and SpeechDat (II) databases observed for the SRE1 can be explained by the complexity of the tasks: the complexity of the first task (from 2 to 15 digits per recording) is much higher than in the case of SpeechDat (II) (only 1 digit per recording).

5.3. Comparative evaluation of the linguistic resources

The linguistic resources we wanted to evaluate are those that can be used by any state-of-the-art speech recognition engine. In our experiments, we concentrated on the LM and the grammar (either a Context-Free Grammar (CFG) or a Word-Spotting Grammar (WSG)). The goal of this evaluation was to measure the impact on the recognition performance of the different linguistic resources and their varying adaptation to the domain of application. To achieve a controlable variability of the resource adaptation, the resources were derived from a corpus containing a set of sentences only characteristic for the application (the Inspire WoZ database) and progressively extended with a growing number of sentences not characteristic for the application (extracted from the SpeechDat (II) database). The derived CFGs were of course not very realistic for real life applications, but they allowed to keep comparable evaluation conditions. For the CFGs, the sentences contained in the training corpora were directly

used as rules in the grammar, therefore yielding a grammar of the following form:

$$< S >= (sentence_1|sentence_2|\ldots|sentence_X);$$

To produce the LMs, the sentences of the training corpora were used in the traditional way, i.e. for each training corpus, SATCA estimated the 3-grams LM.

For the WSG, we used a regular grammar accepting any sequence of any of the phonetic words extracted from Inspire WoZ (this grammar is also sometimes referred to as the "ergodic language model"). Formally, the definition of such a grammar is the following:

$$< S >= (word_1|word_2|\ldots|word_N)+;$$

The results of the evaluation of the linguistic resources are reported in the table 9 for LMs evaluation, table 10 for CFGs evaluation, and table 11 for WSG evaluation. Discussion of the obtained results is provided in the next section. In these tables, column "Training" explains the data used to train the models: the lines '10K + Inspire WoZ', '5K + Inspire WoZ', '1K + Inspire WoZ' and '500 + Inspire WoZ' means that 10'000, 5'000, 1'000 and 500 sentences were randomly chosen in the German SpeechDat (II) database and added to the Inspire WoZ database.

Performance variation associated with the different resources is displayed in the figure 1.

Training	WER, %
SpeechDat	33.6
SpeechDat + Inspire WoZ	5.0
10K + Inspire WoZ	5.6
5K + Inspire WoZ	6.2
1K + Inspire WoZ	7.9
500 + Inspire WoZ	9.0
Inspire WoZ	11.4

Table 9: Evaluation with the LMs

Training	WER, %
SpeechDat	106.9
SpeechDat + Inspire WoZ	11.9
10K + Inspire WoZ	11.0
5K + Inspire WoZ	10.1
1K + Inspire WoZ	8.7
500 + Inspire WoZ	8.2
Inspire WoZ	8.0

Table 10: Evaluation with the CFGs

Training	WER, %
Inspire WoZ	28.9

Table 11: Evaluation with the WSG

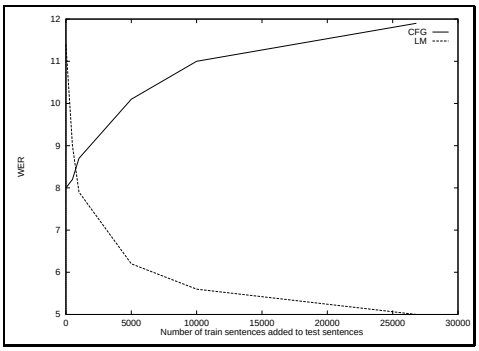

Figure 1: Evaluation with LMs and CFGs

6. Discussion

Evaluation shows that the academic SREs using the HMM based techonology have the recognition performances similar to the ones obtained with the commercial SREs if a "reasonnable" AM is available (compare the "HMM" lines in tables 7 and 8 and the corresponding global recognition performance of HMM based academic SREs in tables 5 and 6).

Concerning the evaluation of the linguistic resources, a very interesting observation is that the LMs act in the opposite way as CFGs (figure 1). The reason for this can be the following: the perplexity of the trained CFGs increases with the increase of the train material. This is not the case for LM models: more train data leads to more accurate estimation of the same N-grams structures.

Finally, one interesting surprising conclusion concerns the fact that the use of a simple word spotting grammar can lead to better recognition performances than "traditional" LMs trained on huge corpora (table 11 and the first line in tables 6, 9, 10). This means that if no sufficiant corpora are available to train the LMs, one can always use the WSGs.

7. Acknowledgements

We would like to acknowledge the anonymous reviewers that provided us with the clever indications about the content of the contribution.

8. References

Andersen, J. M., G. Caloz, and H. Bourlard, 1997. Advanced vocal interfaces services. Technical Report 392, SwissCom Avis Project.

Gillick, Cox. S., L., 1989. Some statistical issues in the comparison of speech recognition algorithms. In *ICASSP*.

Pallett, et al., D., 1990. Tools for the analysis of benchmark speech recognition tests. In *ICASSP*, volume 1.

Rajman, M., A. Rajman, F. Seydoux, and A. Trutnev, 2003. Assessing the usability of a dialogue management system designed in the framework of a rapid dialogue prototyping methodology. In *Proc. of the 1st ISCA Workshop on Auditory Quality of Systems*.