GTH-UPM System for Search on Speech Evaluation

Julian Echeverry-Correa, Alejandro Coucheiro-Limeres, and Javier Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{jdec,a.coucheiro,jfl}@die.upm.es

Abstract. This paper describes the GTH-UPM system for the Albayzin 2014 Search on Speech Evaluation. The evaluation task consists of searching a list of terms/queries in audio files. The GTH-UPM system we are presenting is based on a LVCSR (*Large Vocabulary Continuous Speech Recognition*) system. We have used MAVIR corpus and the Spanish partition of the EPPS (*European Parliament Plenary Sessions*) database for training both acoustic and language models. The main effort has been focused on lexicon preparation and text selection for the language model construction. The system makes use of different lexicon and language models depending on the task that is performed. For the best configuration of the system on the development set, we have obtained a FOM of 75.27 for the keyword spotting task.

Keywords: keyword spotting, spoken term detection, query by example, automatic speech recognition

1 Introduction

The search of information on speech has found many applications in the field of automatic speech recognition (ASR) in recent years. For applications such as dialog managers, conversational agents or spoken information retrieval systems, spotting significant keywords could be more important than recognizing the whole content of an utterance.

The tasks proposed in the 2014 Albayzin Search on Speech Evaluation entail several difficulties that must be taken into account in order to develop an optimal system. Besides the specific conditions and requirements of each task, there are some common features inside the MAVIR corpus (the one used for the evaluation) which demand to be studied at the early stages of the system design. These are mainly related to the acoustic conditions of the audio files. Due to the diversity of the recording conditions, the complexity of the task is increased and the robustness of the system must be optimized. On the one hand, the audios in the MAVIR corpus have several Spanish speakers, including both men and women. This means the acoustic models need to be trained covering this variety. And on the other hand, the quality of the audios changes, in terms of noise and different conditions between recording sessions. This characteristic is

crucial in order to look for supplementary material corpus, whose audios should have similar acoustic conditions.

There are four tasks proposed in the Search on Speech Evaluation: Keyword Spotting (KWS), Spoken Term Detection (STD), Query-by-Example Spoken Term Detection (QbE STD) and Query-by-Example Spoken Document Retrieval (QbE SDR). They are very similar between them, since for all we have a list of terms (written or spoken) that we must search on the input speech, outputting the timestamps and a score of trust. We briefly describe their particularities. For KWS, the list of written terms (keywords) is known before processing the audios, so we can prevent the system to listen carefully for them. For STD, we pursue the same goal as for KWS except that the list of terms is known after processing the audios, so no prevention can be made. For QbE STD, we have the same conditions as for STD except that the search list is made by spoken terms and then an initial stage of recognition of these terms has to be made before searching. And for QbE SDR, as for QbE STD, the list is provided by spoken terms (with the possibility of more instances per term, all in Basque language), but now the output is a score of confidence for a spoken term appearing in a spoken document.

There are several approaches to each task in the state-of-art of Automatic Speech Recognition. As a first gross division, specially for the KWS and STD detection tasks, we can distinguish between systems based on Large Vocabulary Continuous Speech Recognition (LVCSR) and systems based only on keywords and non-keywords models. LVCSR systems allow a simple word-level search, but they need a complete training of the models in order to make possible the recognition of such a large vocabulary. Besides, they also need a proper language model for the correct connection between words in a continuous speech recognition system. Only terms in the vocabulary may be recognized, so any term out-of-vocabulary will never be recognized. For open vocabulary systems where no information about the set of keywords is provided while training the models, it may be required a large amount of training data in order to increase the probability of modeling the probable keywords [1]. And even using an extremely large corpus, we can never accurately model all possible strings of words. In this sense, the most common probabilistic approach for building language models in ASR applications is based on N-grams. This approach models the probability of finding ordered sequences of N words. Nevertheless, in order to face the data sparsity when modeling language, regarding to its variety and complexity, we can employ a smooth variation of N-gram, that is skip-grams. Skip-grams allow us to form new N-grams by skipping one or more words in a word sequence, so the context can be obtained widely around a word. This may overcome the data sparsity problem and may reduce the need of larger corpus. In [2], skip-grams are proven to outperform the standard N-grams for different test documents by using less amount of training data.

Within non-LVCRS systems we find variety depending on the purpose. For the KWS task, it is extended the use of systems based on filler models (also called garbage models). These systems make a phonetic decoding and look for

the phonetic sequence that best fits the phonetic transcription of each keyword, making use of a confident measure based on word segments or on the proportion of correct phonemes. In order to minimize the number of false alarms, these systems do not only model the keywords, but also the non-keyword parts of speech. This background model is referred as filler model, and it is also based in phonetic models. One advantage of these systems over LVCSR systems is the higher speed due to their simplicity. Besides, phonetic-based systems do not depend on a large vocabulary like LVCSR, so the problem with out-of-vocabulary terms is avoided, and can be used as well for the STD [3] and QbESTD [4] tasks. Some systems, as in [5], are hybrid systems of LVCSR and phonetic engines. LVCSR is reserved for in-vocabulary terms due to its robustness and phonetic search and alignment is employed with out-of-vocabulary terms, so no query is uncovered by the system.

The systems mentioned above often make use of Hidden Markov Models. However, other approaches have been developed based on neural networks, on discriminative learning procedures, or on graphical models (GM). As an example, GM makes use of the graph theory in order to describe the time evolution of speech statistically. In [1], GM was used to perform a KWS task with a non-LVCSR system, with the particularity of being vocabulary independent and without require the training of an explicit filler model.

In the next section, we will describe the system submitted by our group for this Search on Speech Evaluation. We have attempted to perform tasks 1 to 2, with a LVCSR system as described below.

2 System description

As we previously said, the system developed for this evaluation consists of a LVCSR system. The feature vectors we used for the acoustic model training consisted of the first 13 PLP coefficients, as well as their first and second order time derivatives. The phoneme models were composed of three hidden states each. We used cross-word triphone models in order to account for contextual information and we consider up to 16 Gaussians per state during training.

We used the transcriptions of the training/development data set, which are available in the MAVIR web page [1], for training the models and for testing the performance of the ASR. These transcriptions are composed of 2878 sentences and a vocabulary size of 5309 words. We also used the transcriptions of the Spanish Parliament partition of the EPPS database (this database is described in section 2.1) to compose the training corpora for the language models. This database is composed of 16514 sentences and a 17.5k vocabulary.

To enrich the vocabulary and the robustness of the language models, we performed a manual data search based on the topics found in the training dataset of the MAVIR corpus. For instance, we searched for data related to *language technologies* and from the obtained results we selected texts on various topics, like

[1] http://cartago.lllf.uam.es/mavir/index.pl?m=videos

sentiment analysis, data crawling, etc. We also guided our data search through the websites of the companies that are mentioned in the audio files (for instance: daedalus, bitext, isoco, etc.). We collected nearly 2000 sentences, composed of a 7.2k vocabulary. These complementary data have been used in the training of language models.

As a first step for the recognition stage, we used a voice activity detector (VAD) to segment the speech signal and perform ASR on the segments of detected speech. The VAD that we used is included in the Voicebox toolbox [6]. We tuned the VAD for splitting the audio in segments with a length under 30 seconds.

For the KWS task and in order to boost the probability of keywords, we repeated twice the sentences in the LM training corpora that contained any keywords and we also repeated the keywords that were missing in the initial vocabulary from the training corpora. Also for this task we added to the initial vocabulary the pre-specified keyword terms so that there were no OOV keywords during ASR search. Multi-term keywords were added as separate words (each of these keywords is treated as a set of single words during recognition).

Regarding the implementation issues, the HTK Toolkit [7] was used for training acoustic models and for the ASR decoding stage. The SRILM Toolkit [8] was employed for creating the language models that the system uses. We use trigram models.

2.1 Databases description

We have used two databases:

– MAVIR corpus is a collection of audio and video recordings, with their corresponding orthographic transcriptions. The audio recordings come from lectures and talks held by the MAVIR consortium. The corpus is made up of 13 recordings in Spanish and English language (nevertheless for this evaluation, only the Spanish partition is available for training, development and evaluation purposes). Data were collected during the I, II, and III MAVIR Conference held in Madrid in 2006, 2007 and 2008 respectively. The details of this database are shown in Table 1. We used all the training audio

Table 1. Details of the MAVIR database

Partition	Files	Length
Training	MAVIR 2, 3, 6, 8, 9 and 12	4h56m
Development	MAVIR 7	0h21m
Evaluation	MAVIR 4, 11 and 13	2h0m

files (except for MAVIR 2 and MAVIR 9) for training the acoustic models. We decided to remove MAVIR 2 and MAVIR 9 files because of the poor

acoustic conditions in which they were recorded. Nevertheless, we use the transcriptions of the all training files in order to train the language models.

– EPPS (*European Parliament Plenary Sessions*) is a database developed by the project TC-STAR (Technology and Corpora for Speech to Speech Translation) [9]. It consists of 61 hours of audio recordings with their corresponding orthographic transcriptions. These recordings were collected between 2004 and 2007. Most of the speakers are interpreters, nevertheless there are also native Spanish speakers. This database also includes 38 hours of audio recordings of the Spanish Parliament (PARL) collected between 2004 and 2006. All the speakers in this group are native Spanish speakers. We selected this database because its acoustic conditions can be similar to those encountered in MAVIR corpus. We use the audio files of both EPPS and PARL partitions to train acoustic models and we use the texts provided by the PARL partition to enrich the vocabulary of the system and the robustness of the language models.

3 Evaluation metrics

For the keyword spotting task, the Figure-of-Merit (FOM), as defined in [7], will be the primary metric for the evaluation. The FOM is defined as the detection rate averaged over the range of 0 to 10 false alarms per hour, and in its calculation it is assumed that the total duration of the test speech is T hours. For each keyword, all of the spots must be ranked in score order. The percentage of true hits p_i found before the $i'th$ false alarm is then calculated for $i = 1 \ldots N+1$ where N is the first integer $\geq 10T - 0.5$. The FOM is defined as

$$FOM = \frac{1}{10T}\left(p_1 + p_2 + \ldots + p_N + a p_{N+1}\right)$$

where $a = 10T - N$ interpolates to 10 false alarms per hour. Table 2 show the results obtained in the development set for the keyword spotting task. The results are presented in terms of the `Hits`, `FA` (false alarms) and `FOM`.

4 Final results

For the final evaluation we present the results obtained on the development and training sets. These results are shown in Tables 2 and 3. The only difference between the principal (PRI) and the contrastive system (CON1) is that the contrastive system employs a language model trained by using the transcriptions of the training dataset of the MAVIR corpus combined with data from the PARL partition of the EPPS database, and in contrast, the principal system does not use the resources from the PARL partition. This will allow a wider coverage for the keyword spotting in the CON1 system but also may introduce a higher

Table 2. Final results with the development set (mavir 07)

System	Task	Hits	FA	Act.	FOM
PRI	KWS	241	73	296	75.27
PRI	STD	227	38	296	72.78
CON1	KWS	231	52	296	72.45

number of false alarms to the system. Next, we show the results obtained in the training set of the database. We are aware that these results do not reflect the performance of the system, since they are obtained over the same dataset for which the system was trained. Nevertheless, these results may offer an oracle approximation of the performance of the system.

Table 3. Results with the training set

File	System	Task	Hits	FA	Act.	FOM
mavir 02	PRI	KWS	599	281	1016	55.31
mavir 03	PRI	KWS	596	52	653	87.71
mavir 06	PRI	KWS	427	20	446	94.12
mavir 08	PRI	KWS	197	10	200	93.66
mavir 09	PRI	KWS	106	186	910	11.26
mavir 12	PRI	KWS	637	41	671	92.70
mavir 02	PRI	STD	426	203	1016	40.06
mavir 03	PRI	STD	551	38	653	82.47
mavir 06	PRI	STD	412	19	446	91.36
mavir 08	PRI	STD	189	10	200	89.99
mavir 09	PRI	STD	39	94	910	4.16
mavir 12	PRI	STD	609	38	671	88.65
mavir 02	CON1	KWS	577	227	1016	54.22
mavir 03	CON1	KWS	598	38	653	89.05
mavir 06	CON1	KWS	433	20	446	95.88
mavir 08	CON1	KWS	196	9	200	93.65
mavir 09	CON1	KWS	94	147	910	10.08
mavir 12	CON1	KWS	636	34	671	92.78

5 Conclusions

In this paper we have presented the description of the system submitted for the 2014 Albayzin Search on Speech Evaluation. The proposed system is based on a LVCSR system. We have used not only MAVIR corpus but also EPPS database to train both acoustic and language models. From the experiments conducted on the development dataset we can conclude that including complementary texts

for the training of the language models may improve the keyword spotting but may also introduce a higher number of false alarms.

From this evaluation it is clear that for developing a proper system for a concrete task it is necessary to study the corpus under study so we can collect the adequate training data that best fits that corpus. This applies not only for the acoustic conditions and variety of speakers but also for the topics discussed in the audio recordings.

Acknowledgements. The work leading to these results has received funding from TIMPANO (TIN2011-28169-C05-03), and MA2VICMR (Comunidad Autonóma de Madrid, S2009/TIC-1542) projects. It has also been supported by the European Union under grant agreement number 287678. Julian Echeverry-Correa also acknowledges the support from Colciencias and from the Universidad Tecnológica de Pereira, Colombia.

References

[1] Wöllmer, M. et al. "Robust vocabulary independent keyword spotting with graphical models". Automatic Speech Recognition & Understanding (ASRU 2009. IEEE Workshop on, 2009, 349-353), 2009.

[2] Guthrie, D. et al. "A closer look at skip-gram modelling". In Proc. of the 5th international Conference on Language Resources and Evaluation (LREC-2006), 1-4, 2006.

[3] Wallace, R. et al. "A phonetic search approach to the 2006 NIST spoken term detection evaluation". International Speech Communication Association (ISCA), 2007

[4] Shen, W. et al. "A comparison of query-by-example methods for spoken term detection". DTIC Document, 2009

[5] Mamou, J. et al. "Vocabulary independent spoken term detection". In Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 615-622, 2007.

[6] Brookes, M. "Voicebox: A Speech Processing Toolbox for Matlab". Department of Electrical and Electronic Engineering, Imperial College, London.

[7] Young, S. et al. "The HTK Book". Engineering Department of Cambridge University, 2006.

[8] Stolcke, A. "SRILM-An extensible language modeling toolkit". In 3rd International Conference on Speech and Language Technology (INTERSPEECH02), 2002.

[9] Mostefa, D. and Hamon, O. and Moreau, N. and Choukri, K. "Evaluation Report for the Technology and Corpora for Speech to Speech Translation (TC-STAR Project). Deliverable N. 30"