GTM-UVigo systems for Albayzin 2014 Search on Speech Evaluation

Marta Martinez, Paula Lopez-Otero, Rocio Varela, Antonio Cardenal-Lopez, Laura Docio-Fernandez, Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{mmartinez,plopez,rvarela,cardenal,ldocio,carmen}@gts.uvigo.es

Abstract. This paper describes the systems developed by the GTM-UVigo team for the Albayzin 2014 Search on Speech evaluation. The primary system for the spoken term detection task consisted on the fusion of two different large vocabulary continuous speech recognition systems that differed in almost all their components: front-end, acoustic modelling, decoder and keyword search approach. An isolate word recognition system was fused with the two aforementioned speech recognition systems for the keyword spotting task. For the query by example spoken term detection task, a fusion of three systems was presented: one of them followed one of the aforementioned continuous speech recognition approaches, with the difference that in this case it was necessary to obtain a transcription of the queries; the other two systems performed a dynamic time warping search, being the use of fingerprints as feature vectors the main novelty of the presented approach.

Keywords: Keyword spotting, spoken term detection, query by example spoken term detection

1 Introduction

In this paper, the systems developed by the GTM-UVigo team for the Albayzin 2014 Search on Speech evaluation are described. Specifically, systems for the tasks keyword spotting (KWS), spoken term detection (STD) and query by example spoken term detection (QbESTD) are presented.

In the KWS task, a fusion of three systems was submitted: two of them rely on large vocabulary continuous speech recognition (LVCSR) systems, while the other one is an isolate word recognition system. One of these LVCSR systems was built using the Kaldi tools [11] to train a set of acoustic models, to generate the output lattices and to perform lattice indexing and keyword search [5]. The second system used the UVigo LVCSR [9] to extract a list of N-Best hypothesis, which were converted to word meshes using the SRILM tool [15]. The first of the aforementioned systems was also submitted as a contrastive system. The other contrastive system submitted consisted on the fusion of the two LVCSR systems.

The systems presented for the STD task were equal to the LVCSR systems submitted for the KWS task, being the only difference the language model: in KWS, the keyword terms were included in the language model, while they were not included in the case of the STD task, as the terms are supposed to be unknown beforehand.

For the QbESTD task, the proposed primary system consisted on a fusion of three different QbESTD systems. The first system was the lattice-search system used in the KWS and STD tasks, with the difference that, in this case, a transcription of the queries had to be performed. The other two systems were based in dynamic time warping search, and they differed in the feature representation of the audio documents and queries: in one of them, a fingerprinting approach was used to obtain a binary representation of the audio [8], and in the other one the audio was represented by means of phoneme posteriorgrams obtained using an English phoneme recognizer based on long temporal context [14].

The rest of this paper is organized as follows: Sections 2, 3 and 4 describe the systems for the KWS, STD and QbESTD tasks, respectively; Section 5 presents the preliminary results obtained for the different tasks on the development data; and Section 6 presents some conclusions extracted from the experimental validation of the different systems.

2 Systems for keyword spotting

The primary system presented for the KWS task consisted on the fusion of three different systems: two different large vocabulary continuous speech recognition (LVCSR) based systems, which are described below, and an isolate word recognition system that includes all the search words in its grammar. One of the LVCSR systems was built using Kaldi while the other one was based on the UVigo LVCSR system. The Kaldi-based system was submitted as a contrastive system. A second contrastive system was submitted, which consisted on the fusion of the two LVCSR systems. The fusion strategy used in this system is also described in this Section.

2.1 Kaldi-based LVCSR System Description

A large vocabulary continuous speech recognition (LVCSR) system was built using the Kaldi open-source toolkit [11]. This system uses standard perceptual linear prediction (PLP) analysis to extract 13 dimensional acoustic features, and follows a state-of-the-art maximum likelihood (ML) acoustic training recipe, which begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMMs), and ends with a speaker adaptive training (SAT) of state-clustered triphone HMMs with Gaussian mixture model (GMM) output densities. The ML stage is followed by the training of a Universal background model (UBM) from speaker-transformed training data, which is then used to train a subspace GMM (SGMM) that will be used in the decoding stage.

The Kaldi LVCSR decoder generates word lattices [12] using the above SGMM models. These lattices are processed using the lattice indexing technique

described in [5] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token is stored as a 3-dimensional cost. This factor transducer is actually an inverted index of all word sequences seen in the lattices. Thus, given a list of keywords or phrases, we then create a simple finite state machine that accepts the keywords/phrases and composes it with the factor transducer to obtain all occurrences of the keywords/phrases in the search collection.

The data used to train the acoustic models of this Kaldi-based LVCSR system was extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign [1]. Specifically, the training data from the European Parliamentary plenary sessions and the Spanish Parliament sessions, which was manually transcribed, was used for this purpose [7]. It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences and short speech utterances were discarded, so in the end the training material consisted of 2 hours and 36 minutes.

The language model (LM) was trained using a text database of 160 MWords composed of material from several sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses and the transcriptions of the Mavir sessions included in the development set[2] [13]. For the development dataset, a different LM was created for each Mavir session, using the transcription of the session to obtain the optimum mixture of the partial LMs. For the evaluation set, the LM was generated using a normalized average of the weights obtained with the development sessions. In this task, the keywords were added to the language model. Note that the vocabulary was selected at the last stage of the LM training, once the partial LMs and its weights were computed. We used a trigram-based LM with a vocabulary of 60K words and a Kesser-Ney discount strategy.

2.2 UVigo-based LVCSR System Description

In this system, we used the UVigo LVCSR described in [9] and the lattice tool provided by the SRILM toolkit [15]. The recognition was performed in three stages. First, an energy-based voice activity detector was used to segment the audio stream in manageable chunks. In the second stage, an acoustic model was selected for each segment. Finally, the UVigo decoder was applied to extract the N-Best hypothesis.

The employed LMs were the ones described in previous Section. For the acoustic modelling we used two state demiphones, with 12 Mel frequency cepstrum coefficients (MFCCs) plus energy and their delta and acceleration values.

[1] http://www.tc-star.org
[2] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

We used acoustic models trained with the TC-STAR training database described in section 2.1, but a set of new additional models was adapted using the Mavir database material. The acoustic model selection was performed using a simple phonetic recognizer, selecting the model which provided the best acoustic scores.

Finally, the N-Best lists were post-processed using the SRI-LM toolkit [15] and converted to word meshes with posterior probabilities. The posterior probabilities were used as a confidence measure for the detected keyword.

2.3 Isolated word recognition system

This system consists on a decoder with a simple grammar composed of a set of N keywords interconnected in parallel. A free phoneme network is employed to obtain a hypothesis test. The acoustic modelling was the same as described in Section 2.2.

2.4 Fusion

Discriminative calibration and fusion was applied in order to combine the outputs of different KWS systems, aiming at taking advantage of the strengths of the individual KWS approaches [2]. First, a global minimum zero-mean and unit-variance normalization was applied, in order to prevent the scores of the individual systems to be in different ranges and also to obtain keyword-independent scores. The global minimum score produced by the system for all queries was used to hypothesize the missing scores. After normalization, calibration and fusion parameters were estimated by logistic regression on a development dataset in order to obtain improved discriminative and well-calibrated scores [3]. This calibration and fusion training was performed using the Bosaris toolkit [4].

3 Systems for spoken term detection

In this task we applied the LVCSR systems used for keyword spotting, which were described in Sections 2.1 and 2.2, i.e. the Kaldi-based and the UVigo-based LVCSR systems. Both systems were combined using the same techniques described in Section 2.4. The only differences between the strategy used in this task and in the keyword spotting task were that, for obvious reasons, the search terms were not included in the vocabulary nor in the LM, and the isolated word recognizer was not used here. Note that, apart from this fact, the LM training was the same, including the partial LM weights employed to compose the final model.

3.1 System fusion

In this task the fusion strategy described above was used. The difference was that, before applying the fusion step, the optimal operating point calculated in the development set was applied to each of the individual systems.

4 Systems for query by example spoken term detection

The primary system for the QbESTD consists on a fusion of three systems:

- MFCC-Fingerprint: a dynamic time warping (DTW) based system that uses audio fingerprints as feature vectors.
- Phoneme posteriorgrams: a DTW based system that uses phoneme posteriorgrams as feature vectors.
- Kaldi-LVCSR: the Kaldi-based LVCSR system described in Section 2.1 was used for QbESTD. To do so, first a transcription of the query was obtained, and then the aforementioned strategy was employed to find all the occurrences of the query.

A contrastive system was also presented, which consisted on the fusion of the MFCC-Fingerprint and the Kaldi-LVCSR systems.

A detailed description of the DTW systems mentioned above is presented in the rest of this Section, as well as a brief description of the fusion strategy.

4.1 Dynamic time warping systems

Two language-independent systems were developed for this task, which perform search on the audio by means of DTW. The search procedure is the same for both of them, but they differ on the feature vectors that are used. We developed an approach inspired by [1], which follows three main steps:

- Feature extraction. Acoustic features are extracted from the waveforms, both from the audio and from the queries.
  - MFCC-fingerprint. We used a fingerprint representation of the acoustic features, motivated by the fact that this representation removes the perceptual irrelevancies of the audio: we are not looking of exact matches, so the information about the speaker or the channel is negligible [8]. The fingerprints corresponding to the acoustic features of each frame were obtained as described in [10]. A convolution mask was used to binarize the acoustic features, specifically a mask for finding negative slopes on the spectrogram in two consecutive frames was applied. Given a set of acoustic features $S \in \Re^{I \times J}$ where $S_{i,j}$ is the feature corresponding to energy band i and frame j, the value $F_{i,j}$ of the frame-level fingerprint corresponding to frame j obtained after applying the convolution mask is

$$F_{i,j} = \begin{cases} 1 \text{ if } S_{i,j} - S_{i,j+1} + S_{i-1,j} - S_{i-1,j+1} > 0 \\ 0 \text{ if } S_{i,j} - S_{i,j+1} + S_{i-1,j} - S_{i-1,j+1} \leq 0 \end{cases} \quad (1)$$

    After running several tests with different features, we finally chose to use MFCCs with delta, acceleration and C_0 coefficients.

  - Phoneme Posteriorgrams. Phoneme posteriorgrams [6] were extracted using a phoneme recognizer based on long temporal context [14] developed at the Brno University of Technology; specifically, the English system of the ones provided by them was used, as it was the one that achieved the best performance on the development data.
- Coarse search. We first perform a coarse search for candidate matches for each query and audio file by following the approach described in [1]. The Euclidean distance matrix between all the vectors of the query and the match audio is computed and the minimum distance per audio vector is selected. Then, the average of these minima in a window of the size of the query is used as an approximation of DTW. We also used a sliding window with 50% overlap.
- Fine search. After selecting those candidates that obtained the smallest distances (the number of candidates is the length of the audio divided by 100, with a minimum of 100 candidates), DTW is computed for all of them. Those candidates whose DTW distance is less than a threshold are confirmed and considered as matches, while the rest of them are discarded.

4.2 System fusion

The strategy used to fuse the different QbESTD systems was the one described in Section 2.4 for KWS but, in this case, a per-query zero-mean and unit-variance normalization (q-norm) was applied. In this task, the discriminative calibration and fusion was trained using all the training and development data.

5 Preliminary results

Table 1 shows the performance obtained with the individual and the fused systems on the development dataset for the KWS task, measured in terms of the Figure of Merit (FOM). As mentioned in Section 2, the primary system consists on the fusion of the Kaldi-LVCSR, UVigo-LVCSR and UVigo-IWR systems, while the constrastive2 system consists on the fusion of the Kaldi-LVCSR and the UVigo-LVCSR systems. This preliminary results show that the best performance was achieved by the Kaldi-LVCSR system (also submitted as contrastive system) but, as the difference between this system and the fusion of the three systems is negligible, we decided to submit the fusion as the primary system, because we rely that the combination of different systems will result in a better performance on the evaluation dataset.

Table 2 shows the performance obtained with the individual and the fused systems on the development dataset for the STD task in terms of the actual term weighted value (ATWV), the false alarm probability P_{fa} and the miss probability P_{miss}. In this case, the difference in performance between the Kaldi-LVCSR and the primary system is more noticeable than in the KWS task, but we decided to keep on with the same criterion, so we submitted the fusion as the primary system.

Table 1. KWS systems: results on the development data

System	FOM
Kaldi-LVCSR (Constrastive1)	84.08%
UVigo-LVCSR	46.75%
UVigo-IWR	44.70%
Primary	83.95%
Contrastive2	83.65%

Table 2. STD systems: results on the development data

System	ATWV	P_{fa}	P_{miss}
Kaldi-LVCSR (Constrastive1)	0.581	0.00008	0.341
UVigo-LVCSR	0.215	0.00017	0.620
Primary	0.568	0.00007	0.363

Table 3 shows the performance obtained with the individual and the fused systems on the development dataset for the QbESTD task, in terms of ATWV, P_{fa} and P_{miss}. It can be seen that the fusion of different systems clearly enhanced their individual performance, achieving an ATWV of 0.3026 when fusing the three proposed systems.

Table 3. QbESTD systems: results on the development data

System	ATWV	P_{fa}	P_{miss}
MFCC-Fingerprint	0.1787	0.00002	0.801
Phoneme posteriorgrams	0.1580	0.00001	0.834
Kaldi-LVCSR	0.1819	0.00006	0.758
Primary	0.3026	0.00009	0.607
Contrastive1	0.2995	0.00009	0.611

6 Conclusions and future work

This paper presented different systems used to perform keyword spotting, spoken term detection and query by example spoken term detection in the framework

of Albayzin 2014 Search on Speech evaluation. The preliminary results obtained for the two first task on the development data are encouraging, as a good performance was achieved in spite of the quality of some recordings where background noise is present, there are different speakers per recording and there is a big amount of pronunciation errors, which makes this scenario challenging for speech recognition based approaches. In future work, a strategy to deal with the out-of-vocabulary issue will be incorporated to the continuous speech recognition systems, as in the spoken term detection task, the out-of-vocabulary terms are completely ignored. Specifically, we intend to implement a strategy that, whenever an out-of-vocabulary word appears, similar words are used as search terms: if these similar words are spotted in the audio document we will consider that our out-of-vocabulary word is present in this document.

With respect to the query by example spoken term detection task, the presented preliminary results outperformed those obtained in the Albayzin 2012 Search on Speech evaluation. The novelty presented in this task consisted on the use of audio fingerprints as feature vectors, motivated by the idea that this representation removes perceptual irrelevancies from the audio; further experiments will be run in order to ensure the validity of this representation for query by example spoken term detection.

Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (CN2011/019, 'Consolidation of Research Units: AtlantTIC Project' CN2012/160), the Spanish Government (FPI grant BES-2010-033358 and 'SpeechTech4All Project' TEC2012-38939-C03-01) and 'TecAnDAli Research Network' of the Consellería de Educación e Ordenación Universitaria, Xunta de Galicia.

References

1. Abad, A., Astudillo, R.F., Trancoso, I.: The L2F Spoken Web Search system for Mediaeval 2013. In: MediaEval'13 (2013)
2. Abad, A., Rodríguez-Fuentes, L.J., Peagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. In: Proceedings of Interspeech. pp. 20–24 (2013)
3. Brümmer, N., van Leeuwen, D.: On calibration of language recognition scores. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop. pp. 1–8 (2006)
4. Brümmer, N., de Villiers, E.: The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing. Tech. rep. (2011), `https://sites.google.com/site/nikobrummer`
5. Can, D., Saraclar, M.: Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech and Language Processing 19(8), 2338–2347 (2011)
6. Hazen, T.J., Shen, W., White, C.M.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: 2009 IEEE Workshop on Automatic

Speech Recognition & Understanding, ASRU 2009, Merano/Meran, Italy, December 13-17, 2009. pp. 421–426 (2009)
7. Laura Docio-Fernandez, A.C.L., Garcia-Mateo, C.: Tc-star 2006 automatic speech recognition evaluation: The uvigo system. In: TC-STAR Workshop on Speech-to-Speech Translation (2006)
8. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Introducing a framework for the evaluation of music detection tools. In: 9th Language Resources and Evaluation Conference. pp. 568–572 (2014)
9. Martinez, M., Cardenal, A.: Experiments on keyword spotting over the transcrigal database. In: Iberspeech 2014: VIII Jornadas en Tecnologa del Habla and IV SLTech Workshop (2012)
10. Neves, C., Veiga, A., Sá, L., ao, F.P.: Audio fingerprinting system for broadcast streams. In: Proceedings of Conference on Telecommunications - ConfTele. vol. 1, pp. 481–484. Santa Maria da Feira, Portugal (2009)
11. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
12. Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafit, M., Kombrink, S., Motlcek, P., Qian, Y., Riedhammer, K., Vesel, K., Vu, N.T.: Generating exact lattices in the wfst framework. In: ICASSP. pp. 4213–4216. IEEE (2012)
13. Sandoval, A.M., Llanos, L.C.: MAVIR: a corpus of spontaneous formal speech in Spanish and English. In: Iberspeech 2012: VII Jornadas en Tecnologa del Habla and III SLTech Workshop (2012)
14. Schwarz, P.: Phoneme Recognition based on long temporal context. Ph.D. thesis, Brno University of Technology (2009)
15. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing. pp. 901–904 (2002)