IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Cenatav Voice Group System for Albayzin 2018 Search on Speech Evaluation

Ana R. Montalvo, Jose M. Ramirez, Alejandro Roble and Jose R. Calvo

Voice Group, Advanced Technologies Application Center, CENATAV, Havana, Cuba
{amontalvo, jsanchez, aroble, jcalvo}@cenatav.co.cu

Abstract

This paper presents the system employed in the Albayzin 2018 "Search on Speech" Evaluation by the Voice Group of CENATAV. The system used in the Spoken Term Detection (STD) task consists on an Automatic Speech Recognizer (ASR) and a module to detect the terms. The open source Kaldi toolkit is used to build both modules. ASR acoustic models are based on DNN-HMM, S-GMM or GMM-HMM, trained with audio data provided by the organizers and other obtained from ELDA. The lexicon and trigram language model are obtained from the text associated to the audio. The ASR generates the lattices and the word alignments required to detect the terms. Results with development data shown that DNN-HMM model brings up a behavior better or similar to obtained in previous challenges.

Index Terms: Spoken Term Detection, Automatic Speech Recognition, Kaldi

1. Introduction

This is the first participation of the Voice Group of Cenatav in the Albayzin Challenges. We participated in the STD task of the Albayzin 2018 Search on Speech Evaluation, developing a system described in the next section. Spoken Term Detection (STD), is defined by NIST [1] as "searching vast, heterogeneous audio archives for occurrences of spoken terms". Iberian institutions have been conducted researches recently on this task, as shown in previous Albayzin Challenges [2, 3, 4, 5].

In STD task, a list of written terms must be detected in different audio files. Although many of the terms may belong to the train vocabulary (INV), the term list is known after processing the audio, which provokes a specific treatment of out-of-vocabulary (OOV) words. The words lexicon and language models of the ASR systems employed to detect OOV words, were obtained from texts where a specific OOV terms list, provided by the evaluators were removed, to force the systems to deal with real OOV words.

The evaluation of the system is performed with three different databases in native Spanish. The first one is the test partition of the MAVIR [6] corpus, the second one is the SoS test partition of RTVE database [7], while the third is known as COREMAH [8] database.

The system presented for the task consists of two modules: A Large Vocabulary Continuous Speech Recognition (LVCSR) module and a Spoken Term Detection (STD) module. The acoustic modeling and words decoding were implemented using the Kaldi toolkit [9], while the trigram language models were computed with SRILM [10]. Firstly, the test audio files are processed by a LVCSR based on Deep Neural Network. This produces word lattices which contain the most probable transcriptions of the audio files.

Section 2 describes the details of the implemented system. Section 3 shows the experimental results obtained using the development data set. Finally, Section 4 concludes the paper and describes the work being under development.

2. Cenatav Voice Group STD Kaldi System

The following section describes more detailed the proposed system.

2.1. Large Vocabulary Continuous Speech Recognition (LVCSR) module

The first module of the system is a Large Vocabulary Continuous Speech Recognizer implemented following an adaptation of s5 WSJ recipe of Kaldi. The acoustic features used are 13 Mel-Frequency Cepstral Coefficients (MFCC) with cepstral normalization (CMVN) to reduce the effects of the channel.

The training of the acoustic models begins with a flat initialization of context-independent phone Hidden Markov Models (HMM). Then several re-training and alignment of acoustic models to obtain context-dependent phone HMM, are done following the transformation of the acoustic features:

- The 13-dimensional MFCC features are spliced across +/-4 frames to obtain 117 dimensional vectors.
- Then linear discriminant analysis (LDA) is applied to reduce the dimensionality to 40, using context-dependent HMM states as classes for the acoustic model estimation.
- Maximum likelihood linear transform (MLLT) is applied to the resulting features, making them more accurately modelled by diagonal-covariance Gaussians.
- Then, feature-space maximum likelihood linear regression (fMLLR) is applied to normalize inter-speaker variability of the features.

Obtained phone models consist of three HMM states each in a tied-pdf cross-word tri-phone context This GMM-HMM model was trained with 15000 Gaussians and 2129 senones.

Then the GMM-HMM model is speaker adapted in a sub-space of Gaussian Mixtures (S-GMM), following [11], using fMLLR features and sharing the same Gaussian Model and same Gaussians quantity per state of the model. This S-GMM contains 9000 Gaussians and 7000 branches.

The GMM-HMM model is used too, as an alignment for training a DNN-based acoustic model (DNN-HMM) following Dan's DNN implementation, with 2 hidden layers with 300 nodes each. The number of spliced frames was nine to produce 360 dimensional vectors as input to the first layer. The output layer is a soft-max layer representing the log-posteriors of the context-dependent 2129 states. The total number of parameters is 1703600.

The LVCSR decoder generates word lattices [12] using any of the mentioned models, these lattices contain the most probable transcriptions of the test utterances where the term

10.21437/IberSPEECH.2018-53

search will be performed. These lattices and the transcriptions obtained are the primary input to the STD module.

The lattices are processed using the lattice indexing technique described in [13], where the lattices of all the test utterances 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 is stored as a 3-dimensional cost. This factor transducer is an inverted index of all word sequences seen in the lattices.

Thus, given a list of terms to detect, a simple finite state machine is created such that it accepts the terms and composes it with the factor transducer to obtain all its occurrences in the tests utterances, along with the utterance ID, start-time, end-time and lattice posterior probability of each occurrence. All those occurrences are sorted according to their posterior probabilities and a YES/NO decision is assigned to each instance.

2.2. Train and development data

Here is a description of the databases used to train and develop our system for the STD task.

- **TC-STAR:** We obtained free from ELDA-ELRA a set of audios and transcriptions of Spanish partition of TCSTAR 2005-2007 [14], corresponding to the Evaluation Package database. It contains 26:40 hours of audio and consists of 17163 expressions with 241412 words. This set was used for training the acoustic models and the language model.
- **MAVIR:** This corpus was provided by the challenge organizers, and corresponds to talks held by the MAVIR consortium in 2006, 2007 and 2008 [6]. The Spanish training data is contained in “SoS2018_training”, contains 4: 20 hours and consists in 5 talks segmented in 2400 expressions with 44423 words. This set was used for training the acoustic models and the language model. The MAVIR development data is contained in “SoS2018_development (1)” and is about one hour in two talks.
- **RTVE:** The Challenge organizers provided this corpus and its structure is explained in [15]. The corpus is divided in 4 partitions, a “train” one, two development “dev1”, “dev2” and one “test”. The audio files of the “train” partition do not have human-revised transcriptions. Partition “dev1” contains about 53 hours of audios and their corresponding human-revised transcriptions and can be used for either development or training. So, transcriptions (files “trn”) and audio (files “aac”) of twelve RTVE programs (four programs “20H” and eight programs “La Noche en 24H”), of this partition, were manually segmented and labeled by speaker, by us, in expressions less than 40 seconds to obtain a training dataset of 2013 expressions with 139,983 words with a duration of 13:45 hours. This set was used for training the acoustic models and the language model. The development data is in the partition “dev2” and is about 15 hours in twelve RTVE programs.

During the manual segmentation of transcriptions and audio of RTVE programs, we observed that the transcript files did not include many speech expressions that happen in the audio and also some speaker voices overlapping, typical of the spontaneity and level of improvisation during the conversations among the journalists participating in the programs. We eliminate the voices overlapping when segmenting the audio and its corresponding transcription, however only some not transcribed speech expressions, were transcribed by us when we detected them, listening carefully and replaying many times the audio file, during the segmenting process.

All the textual training material of the three databases was revised and corrected, carry to uppercase, substituting the numbers and acronyms for their transcription and finally grouped in a database of 23029 different words and 44:45 hours of duration.

2.3. Vocabulary and Lexicon

The dictionary used by the LVSCR is composed only by words from the transcriptions of the training data. Multilingual G2P transcriber [16] was used to obtain the phonetic transcription of each word. We obtain a general lexicon of 23029 different words.

2.4. Language models

To train the language model used by the LVCSR, we used only the transcriptions of the training data corpus. It consists of 21575 expressions and 23029 different words. This text has been supplied to the SRILM tool to create an Arpa format, trigram language model with 23002 unigrams, 156778 bigrams and 38628 trigrams.

2.5. INV and OOV Terms

This Challenge evaluation defines two sets of terms for STD task: an in-vocabulary (INV) set of terms and an out of-vocabulary (OOV) set of terms. The OOV set of terms will be composed by out-of-vocabulary words for the LVCSR system, so these OOV terms must be removed from the system dictionary and consequently from the lexicon and the language model.

Table 1: *STD scores for MAVIR and RTVE development corpus.*

Development set	Kaldi Models	MTWV	ATWV	Pfa	Pmiss
MAVIR	tri3b (GMM-HMM)	0.5705	0.5556	0,00006	0.373
MAVIR	sgmm2_4 (S-GMM)	0.6045	0.5897	0.00005	0.348
MAVIR	tri4_nnet (DNN_HMM)	0.5974	0.5946	0.00008	0.322
RTVE	tri3b (GMM-HMM)	0.2943	0.2939	0.00002	0.690
RTVE	sgmm2_4 (S-GMM)	0.2889	0.2868	0.00001	0.699
RTVE	tri4_nnet (DNN_HMM)	0.3303	0.3295	0.00002	0.648

3. Experimental results

Table 1 contains the STD scores obtained with the three proposed models (GMM-HMM, S-GMM and DNN-HMM), using the DEV set of MAVIR and RTVE corpus, evaluating with the NIST STDEval-0.7 tool, provided by the Challenge organizers.

This tool provides the probabilities of False Acceptances (Pfa) and Misses (Pmiss) of the STD system, and two metrics that integrates both probabilities [17]:

- Actual Term Weighted Value (ATWV) that integrates Pfa and Pmiss for each term, and averages over all the terms, representing the term weighted value for a threshold set by the system tuned on development data
- Maximum Term Weighted Value (MTWV) that is the maximum TWV achieved by the system for all possible thresholds, not depending on the tuned threshold, representing an upper bound of the system performance.

Comparing with results obtained with the same MAVIR dataset in Albayzin 2016 spoken term detection evaluation [5], our results, for all the proposed models, are similar to obtained by the best system with the DEV set (first line of table 9 of [5]) and the TEST set (first line of table 10 of [5]). Additionally, proposed DNN-HMM model surpasses the behavior of the best system in that evaluation.

However, results shown that the behavior of the evaluated methods is worst with the RTVE set, probably by:

- the differences between the speech of the training set and their transcriptions, explained above, and
- the differences in spontaneity and level of improvisation between MAVIR and RTVE sets.

4. Conclusions and future work

Due to the inexperience of the participants in the use of the Kaldi tool, the lack of the necessary time, and difficulties in the exploitation of the computational resources that we have, it was impossible for us to carry out, before the dead line, the evaluation of the proposed systems applying the Kaldi proxy method, for the DEV and TEST set.

Taking into account the best results of DNN-HMM model in the RTVE set, its lowest Pmiss and lowest performance gap between MTWV and ATWV metrics, indication of well calibrated term detection scores, we decided to send as a primary system the one obtained with the DNN-HMM model, and as contrastive systems 1 and 2, those obtained with the models S-GMM and GMM-HMM, respectively.

We consider the participation in this Challenge very useful. The experience acquired by all the participants will serve us for next competitions and more importantly, for the development of our research in the fields of ASR and STD.

We propose to continue and conclude the experiments evaluating the Kaldi proxy method and refining the transcripts of the RTVE database samples, to make them available to the Spanish Thematic Network on Speech Technology (RTTH).

5. Acknowledgments

Finally, we would like to thank the University of Zaragoza and the University of Vigo, in the persons of the professors Eduardo Lleida Solano and Carmen Garcia Mateo, for their support in the elaboration of the textual training data of our system.

References

[1] NIST. The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edn., 2006. http://www.nist.gov/speech/tests/std.

[2] Tejedor, J., Toledano, D.T., Anguera, X., Varona, A., Hurtado, L.F., Miguel, A., Colas, J.: Query-by-example spoken term detection Albayzin 2012 evaluation: overview, systems, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 2013(1):23.

[3] Tejedor, J., Toledano, D.T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., Cardenal, A., Echeverry-Correa, J.D., Coucheiro-Limeres, A., Olcoz, J., Miguel, A.: Spoken term detection Albayzin 2014 evaluation: overview, systems, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing, (2015) 2015 (1):21

[4] Tejedor, J., Toledano, D.T., Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Comparison of Albayzin query-byexample spoken term detection 2012 and 2014 evaluations. EURASIP Journal on Audio, Speech, and Music Processing, 2016(1):1.

[5] Tejedor, J., Toledano, D.T., Lopez-Otero, P., Docio-Fernandez, L., Serrano, L., Hernaez, I., Coucheiro-Limeres, A., Ferreiros, J., Olcoz, J., Llombart, J.: Albayzin 2016 spoken term detection evaluation: an international open competitive evaluation in spanish. EURASIP Journal on Audio, Speech, and Music Processing, 2017(1):22.

[6] Sandoval, A.M., Llanos, L.C.: MAVIR: a corpus of spontaneous formal speech in Spanish and English. In: Iberspeech 2012: VII Jornadas en Tecnología del Habla, 2012. \MAVIR corpus: http://www.lllf.uam.es/ESP/CorpusMavir.html".

[7] RTVE corpus: http://catedrartve.unizar.es/reto2018.html.

[8] COREMAH corpus: http://www.lllf.uam.es/coremah/".

[9] 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 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[10] Stolcke, A. et al.: SRILM-an extensible language modeling toolkit. In: ISCA Seven International Conference of Speech Technologies, ICSLP 2002, pp 901-904.

[11] Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Rastrow, A., Rose, R., Schwarz, P., and Thomas, S.: The subspace Gaussian mixture model: A structured model for speech recognition. Computer Speech & Language, 25(2):404–439, 2011.

[12] Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafiát, M., Kombrink, S., Motlícek, P., Qian, Y., Riedhammer, K., Veselý, K., Vu, N.T.: Generating exact lattices in the WFST framework. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 4213–4216, 2012.

[13] Can, D., Saraclar, M.: Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech and Language Processing 19(8), 2338–2347, 2011.

[14] TC-STAR Technology and Corpora for Speech to Speech Translation. http://www.tcstar.org/pages/main.htm

[15] Lleida E., Ortega A., Miguel A., Bazán V., Pérez C., Zotano M. and De Prada A.: RTVE2018 Database Description

[16] Multilingual Grapheme to Phoneme. https://github.com/jcsilva/multilingual-g2p

[17] Fiscus, J. G., Ajot, J., Garofolo, J. S., & Doddingtion, G.: Results of the 2006 spoken term detection evaluation. In Proc. of workshop on searching spontaneous conversational speech, pp. 45–50, 2007.