QCRI ADVANCED TRANSCRIPTION SYSTEM: QATS

Ahmed Ali, Yifan Zhang, Stephan Vogel

Qatar Computing Research Institute

{amali, yzhang, svogel}@qf.org.qa

ABSTRACT

This demo paper describes a system overview for a state-of-the-art Modern Standard Arabic (MSA) speech recognition system. QCRI Advanced Transcription System (QATS) is integrated with Aljazeera's Arabic website www.aljazeera.net . The current system is using fMLLR-based speaker adaptation in a training scenario using the Minimum Phone Error (MPE) criteria combined with sequential Deep Neural Network (DNN) training. Acoustic Models (AM) have been trained using 400 hours that have been manually transcribed. Language Model (LM) has been built using a five-year archive of the Aljazeera website; the lexicon has 800K words with an Out Of Vocabulary (OOV) rate less than 4% on test data. The current system is phoneme level trained with 36 phones. We report results for two different types of test data: broadcast news reports, with a best Word Error Rate (WER) of 13.5%, and broadcast conversations with a best WER of 28.8%. The overall WER on this test set is 21.6%.

Index Terms: Arabic, ASR demo system, QATS

1. Introduction

Given the nature of the Arabic broadcast news data coming through Aljazeera Arabic news channel, we expect around 20% non-MSA speech, which can be classified into four dialects: Egyptian, Levantine, Gulf, and Maghrebi (Moroccan). The current version of QATS ASR is optimized for MSA; consequently we need classifiers for non-MSA speech. We use state-of-the-art language/dialect recognition systems based on the acoustics with no prior knowledge about the speech files. The current models use factor analysis similar to that of the i-vector framework. The obtained subspace vectors are then applied in conjunction with i-vectors to the problem. The evaluation show that the proposed adaptation method yields more accurate recognition results compared to three conventional weight adaptation approaches, namely maximum likelihood re-estimation, non-negative matrix factorization, and a subspace multinomial model. The dialect recognition system has been evaluated using percentage of incorrectly classified utterances (EIC); for development EIC 16% and for evaluation set EIC is 15%. The non-MSA segments are not transcribed at the moment and being collected to build robust dialectal ASR. Only MSA segments sent to QATS server on the cloud for ASR as illustrated in figure 1. The ASR is constantly monitoring Aljazeera traffic for transcription, and it is also connected to QCRI servers to use the latest AM, and LM. The rate of updating the LM is relatively faster than AM, as we are always crawling Aljazeera website and updating the LM. More details about the technologies used in QATS can be found in [1]–[3]

2. QATS

QATS is continuously monitoring Aljazeera.net, and fetches any video file being labeled by journalists to be transcribed. The current system is deployed on Azure platform using three different servers; master node, ASR nodes, and subtitle server, as illustrated below:

QATS master node: This is a basic server, mainly monitoring the high quality video server for Aljazeera on brightcove and fetches any new video; it runs Voice Activity Detection (VAD) and dialect identification. This machine decides the number of ASR Virtual Machines VM will be used for recognition, based on the amount of data queuing to be transcribed. The master node is A0 basic server, 1 core, 0.75GB RAM and 20GB disk.

QATS ASR nodes: These are machines capturing the image for the ASR VM A9; with 16 cores, 112 GB RAM, 382G disk. The server is mainly running the ASR based on the segmentation and VAD results from the master node. Most of the time there is only one ASR node running due to the relatively small daily traffic from Aljazeera. However, the system can scale up and start more VM to run ASR if needed. Once the ASR is done, Distribution Format Exchange Profile (DFXP) file is created and sent to Aljazeera server, and machines are closed to control the cost of the transcription.

QATS Archive Server: This is basic server like the master node, mainly used to keep the archive of the subtitle files, DFXP along with the version of the ASR used. This machine stores the modified subtitles, as sometime journalists make modification to the DFXP. We intend to use the modified DFXP files to improve the ASR, and potentially use it for an active learning setup.

Processing time is typically less than 1.5 Real Time (RT), with a minimum Turn Around Time (TAT) of 17

minutes. TAT is the exact time from downloading the video until DFXP is being made ready. QATS is currently being deployed on Aljazeera archive videos to allow retrieval capabilities using the video content as well as the metadata. The system has already transcribed more than 1500 hours from Aljazeera video archive.

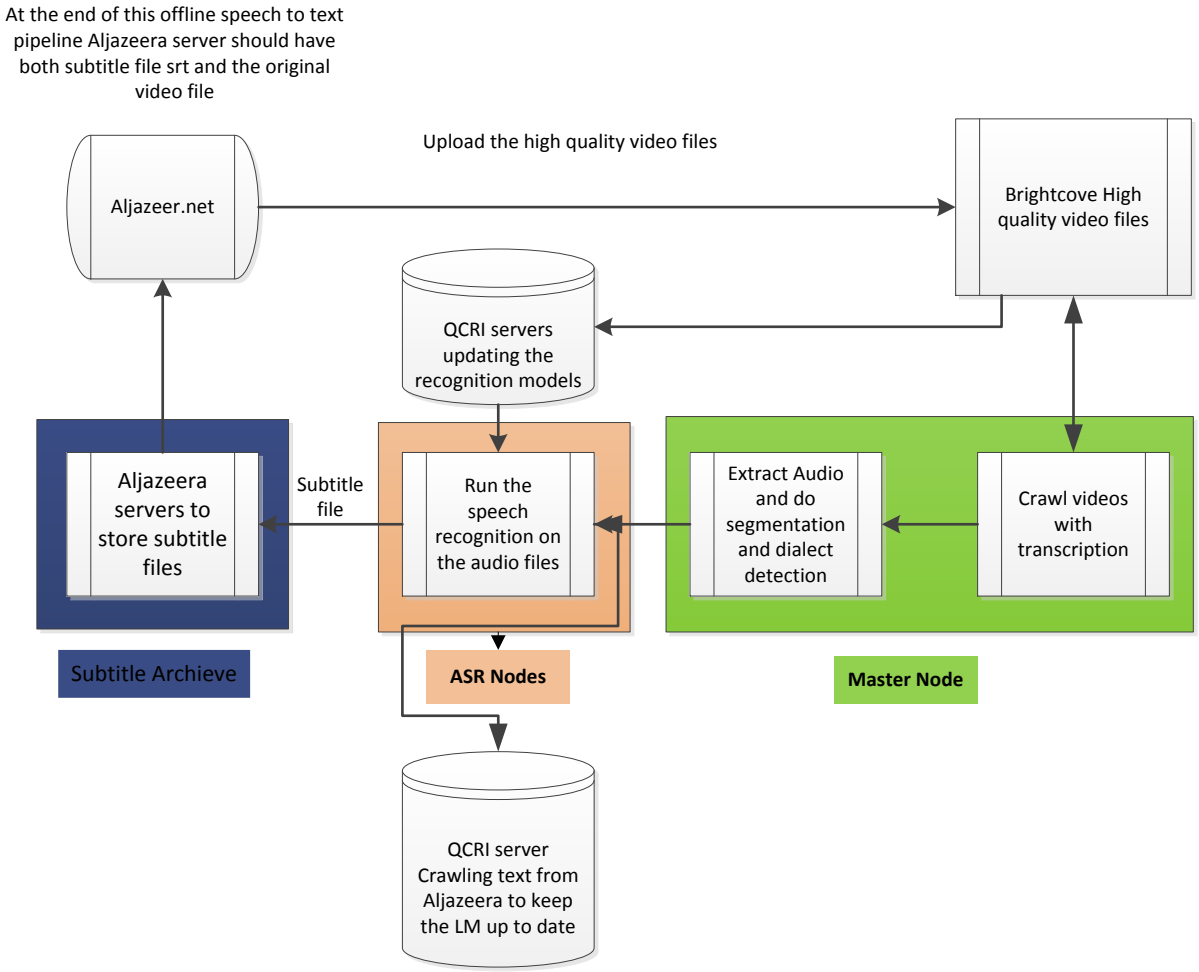

Figure 1: QATS Architecture

3. Conclusion

[1] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A COMPLETE KALDI RECIPE FOR BUILDING ARABIC SPEECH RECOGNITION SYSTEMS," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014.

[2] P. Cardinal, A. Ali, N. Dahak, T. Al Hanai, Y. Zhang, J. Glass, and V. Stephan, "Recent Advances in ASR Applied to an Arabic Transcription System for Al-Jazeera," in *To appear in Proceedings of 15th Annual Conference of the International Speech Communi- cation Association (Interspeech)*, 2014.

[3] M. H. Bahari, N. Dehak, L. Burget, A. Ali, J. Glass, and others, "Non-negative factor analysis for GMM weight adaptation," *IEEE Trans. Audio Speech Lang. Process.*, 2014.