Automated Radio Auditing Application for Positive Communication

Chenxi Zhong

Roberto Togneri
School of Electrical, Electronic and Computer Engineering
University of Western Australia

Pieter Lottering and Charday Williams
CEED Client: Global Integrated Operations

Abstract

Given the importance of radio auditing for mining industry radio communication in terms of safety and efficiency, and opportunities for performance evaluation and management in Integrated Remote Operations Centres, the process has not been well developed, relying on manual auditing. The project is to provide a partial solution to the automation of radio auditing, where a keyword spotting program will be developed with training using historical radio recordings obtained from industry. A completely unsupervised method is applied for keyword spotting from audio recordings, which serves as the foundation for automated auditing for radio data without transcription. In this application of Positive Communications analysis, the audio data contains strong accents and multiple speakers. Before applying the model to industry historical data, testing of algorithm has been conducted on clean TIMIT corpus data, the results of which are illustrated in this paper to show the proof of concept.

1. Introduction

Radio has been widely used as a major communication channel in the resource industry. Only since the recent implementation of Integrated Remote Operations Centres (IROCs), has radio communication been made remotely available in one central location for auditing and performance management purposes.

Specific content, such as Positive Communications, is of interest in the radio auditing process. Compliance with Positive Communications protocols is of significant importance to the safety of traffic in the industry, as it requires the intention and confirmation between drivers to be clearly stated and understood when interactions between vehicles take place.

Currently, auditing radio recordings for Positive Communications requires the auditing personnel to manually select relevant recordings from a large historical database and judge their adherence to the protocol. The process can be very time-consuming, and prone to inconsistency and human error, with only a small portion of interactions being audited as evaluation samples to infer the overall performance.

Automation of radio auditing would provide a solution to replace manual auditing, offering higher efficiency, standardisation, easier performance management and extended application.

One of the biggest challenges for the automation process is that the historical radio data has no transcription available. Early research and testing also revealed that Automatic Speech Recognition (ASR) followed by text mining is not feasible due to accent, jargon and identifiers that are specific to the domain.

Therefore, before the development of the decision-making process, the content of interest need to be automatically identified from the audio recordings, which also contain irrelevant data. Moreover, the appearance or absence of certain keywords and phrases can directly facilitate the identification of protocol compliance or breach. Due to time constraint, this project aimed to explore possible solutions to identify spoken keywords relevant to Positive Communications under a completely unsupervised environment without any annotated or transcribed speech data.

1.1 Literature Review

Formally, searching for spoken keywords relevant to specific content of interested is referred to as Keyword Spotting (KWS) or Spoken Term Detection (STD). Under circumstances where labelled training data is limited or absent, the query-by-example (QBE) STD method provides a solution to search audio databases using audio queries (Hazen, et al., 2009).

The QBE method originated from template matching, where examples of the target spoken terms are in the form of spoken queries (Mandal, et al., 2014). After both the query and test utterance are converted to template representations, dynamic time warping (DTW) is generally employed to measure the differences between templates (Mandal, et al., 2014).

Early template matching relied on a typical speech feature vector, Mel-frequency cepstral coefficients (MFCCs), but suffered in the presence of multiple speakers or environment changes (Hazen, et al., 2009). To address the problem, recent research proposed novel posterior features for representation of the speech signal with some promising results (Mandal, et al., 2014). Among various posterior features, Hazen and colleagues examined phonetic posteriorgram representation in 2009. Their work did not require transcription of the audio files, but required a phonetic recognizer to be independently trained (Hazen, et al., 2009). Zhang and Glass then proposed Gaussian posteriorgram from Gaussian Mixture Model (GMM) for template representation, which is speaker-independent and does not require any transcribed data (Zhang, 2013; Zhang & Glass, 2009). Using the Gaussian posteriorgram and segmental DTW to detect the occurrence of keywords in the utterance, the performance was evaluated on both TIMIT Acoustic-Phonetic Continuous Speech Corpus (referred to as TIMIT corpus) and MIT Lecture corpus, exhibiting comparable results to methods requiring some supervised training (Zhang & Glass, 2009).

2. Methodology

The project therefore adopted GMM posteriorgram for template representation of both audio queries and test utterances, followed by template comparison based on the DTW method. The frame work from signal input to DTW search is shown in Figure 1.

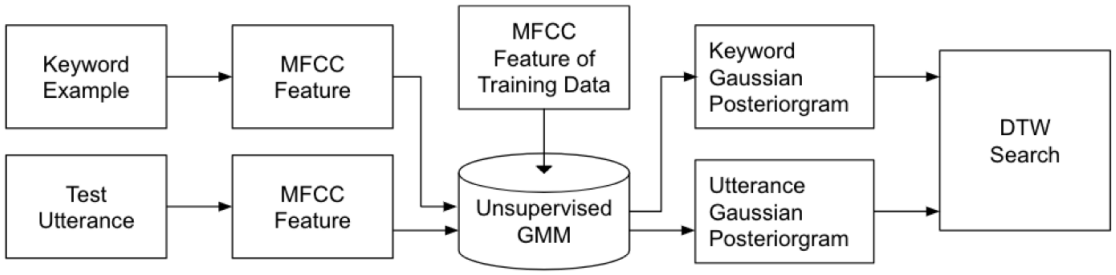

Figure 1 Framework of algorithm

A single GMM is firstly trained using all data, using MFCC features of the speech signal as input to model the probability distribution of speech data as a weighted linear combination of Gaussian components. The trained GMM is then used to produce a Gaussian posteriorgram for each keyword and test utterance. With a speech utterance of n frames denoted as $\boldsymbol{S} = (s_1, s_2, \dots, s_n)$, the Gaussian posteriorgram is defined as

$$\boldsymbol{GP}(\boldsymbol{S}) = (q_1, q_2, \dots, q_n)$$

$$q_i = (P(C_1|s_i), P(C_2|s_i), \dots, P(C_m|s_i))$$

where C_i is the i^{th} of m Gaussian components (Zhang & Glass, 2009).

Comparing the posteriorgram of N-frame query and that of the M-frame test utterance results in a $N \times M$ similarity matrix (Mandal, et al., 2014). Using DTW search, the optimal alignment path (warp path) through the similarity matrix can be found, which indicates the overall difference between the target keyword example and the test utterance by DTW distance (Hazen, et al., 2009). The extent of match, shown by the DTW distance falling within a threshold value, reflects the presence of the keyword. By calculating the similarity matrix and finding the path of minimum cost, the optimal path is able to identify the existence and location of the keywords.

To verify the feasibility of concept and validate the model, the constructed model was firstly applied to data from TIMIT corpus. The TIMIT corpus was designed and recorded at Texas Instruments, Inc. (TI), transcribed at Massachusetts Institute of Technology (MIT) and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST) (Garofolo, et al., 1993). With a total of 6,300 sentences recorded by 630 speakers of eight major dialects of American English, the TIMIT corpus is a rich collection of phonetic data, and was designed for development and evaluation of speech recognition systems (Garofolo, et al., 1993; Zue, et al., 1990).

3. Results and Discussion

3.1 DTW Comparison Between Keywords

The functionality of dynamic time warping algorithm was firstly verified by inputting a keyword and comparing it with itself. Since two inputs for the DTW are exactly the same, it is expected that the DTW cost should be 0. One example DTW comparison plot for keyword sample 1 from ('warm01.wav') is shown in the left figure in **Error! Reference source not**

found., where the symmetry along the diagonal warping path can be observed with total cost equal to 0. This result is consistent with expectation.

Comparing different keyword samples (sample 2 and sample 3) of the same keyword 'warm' yields the following similarity matrix plots. The costs are 22.966 and 11.681 respectively. On the right-hand side of Figure 2, the plots illustrate the DTW warping path of posteriorgram representations of keyword sample 1 comparing with that of keyword sample 2 and 3 respectively.

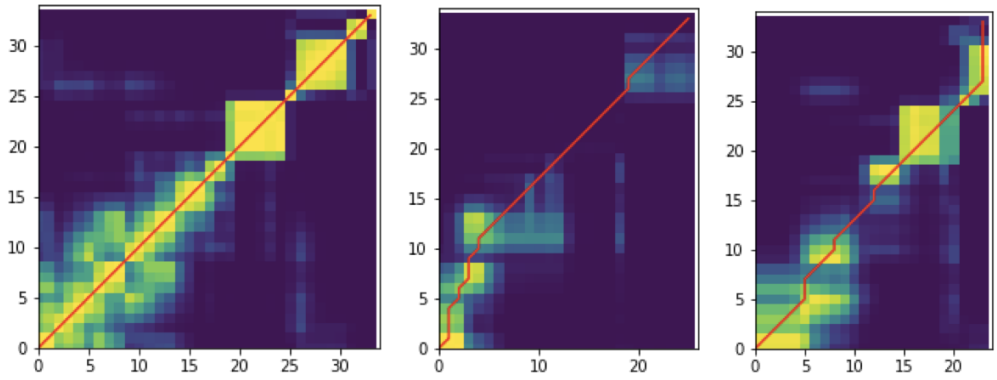

Figure 2 DTW comparison of GP representations of keyword sample 1. The left figure shows the DTW result for keyword 'warm' sample 1 compared with itself. The middle and right figure show the result for posteriorgram representations of keyword sample 1 compared with that of keyword sample 2 and 3.

3.2 DTW Comparison Between Keyword and Corresponding Utterance

After the functionality of DTW is verified by similarity comparison between keywords, the second step is to compare the keyword with its corresponding utterance. In other words, a certain part of the utterance is exactly the same as the keyword sample, which results in the same Gaussian Posterior (GP) representation. Therefore, that particular region in the cosine similarity matrix should be identified by the dynamic time warping path. One example test using keyword sample 1 from 'warm01.wav' and its corresponding test utterance 'u_warm01.wav' yields the DTW warping path as shown in Figure 3, with its enlarged identifying region on the right.

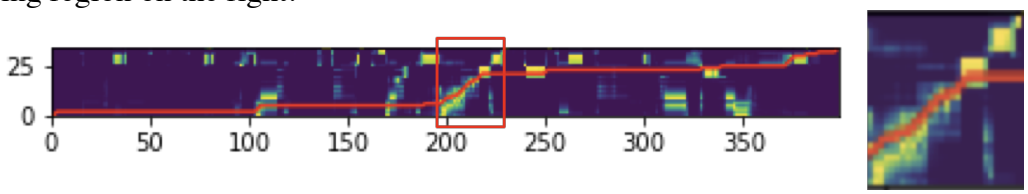

Figure 3 DTW Plot of keyword posteriorgram representation comparing with that of its corresponding test utterance

As can be seen from

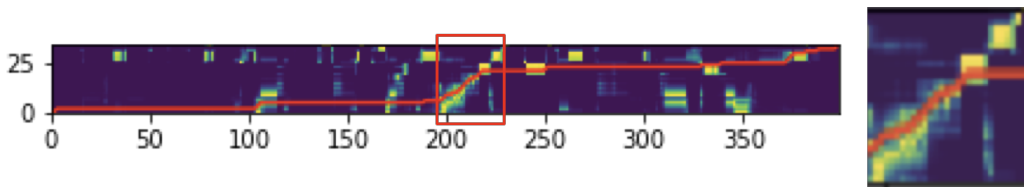

Figure 3 DTW Plot of keyword posteriorgram representation comparing with that of its corresponding test utterance, the corresponding keyword section in the utterance was

efficiently identified. Further verification was conducted by manually identifying the keyword region in the utterance using PRAAT.

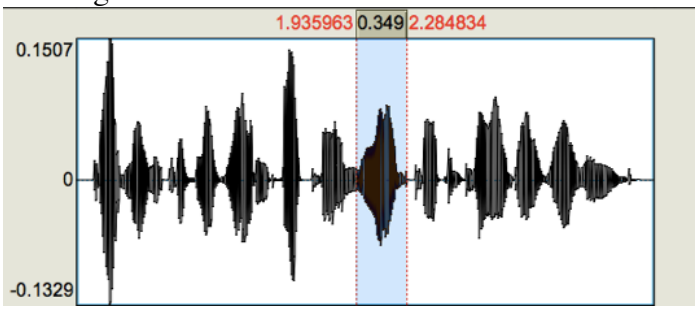

Figure 4 PRAAT plot of utterance clip with keyword region selected

Figure 4 shows that the keyword region (from about 1.93s to 2.28s was consistent with the identified region in DTW of the GP representations plot.

3.3 DTW Comparison Between Keyword and Different Utterance

The next step is to test keyword identification by using keyword query for a different utterance. The keyword sample was recorded by a different speaker having a distinct accent compared to that of the utterance sample. Since the GP representation should capture the feature of keywords regardless of the speaker, the same region in utterance-keyword similarity matrix plot is expected to be spotted. One example test between keyword sample 2 and test utterance 'u_warm01.wav' produced DTW plots as shown in the top figure of Figure 5. The bottom figure in Figure 5 shows the result for similar comparison between the test utterance with the third keyword sample.

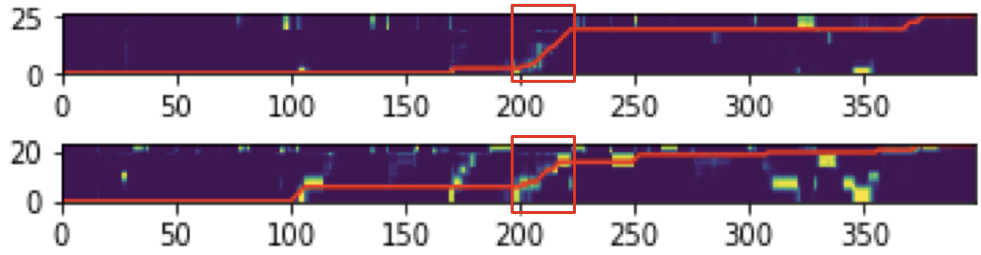

Figure 5 DTW plot between keyword sample 2 and test utterance 1 (top) and that between keyword sample 2 and test utterance 1 (bottom)

Both DTW plots in Figure 5 exhibit efficient spotting of the keyword region over the search of test utterance, which is consistent with expectations and verifies the functionality of algorithm.

3.4 Remaining Issue of the Algorithm

Despite the successful detection of keyword samples for some experiments, in some cases, the keyword region was missed by the warping path, due to the presence of other highly similar regions. One of the examples of such a case is shown in Figure 6.

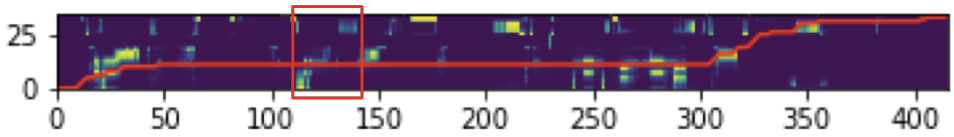

Figure 6 Example case where the keyword region (labelled with rectangle) is missed by DTW warping path

4. Conclusions and Future Work

Current progress has shown promising results for the identification of spoken query, which verified the concept and validated the algorithm model. To solve the current issue, a sliding window to restrict the search width will be applied to compare keywords template and part of utterance template sequentially.

After the issue is resolved, the constructed model is to be applied to the audio recordings of radio communication from industry. Once the algorithm properly identifies the keywords, an utterance based detection equal error rate can be employed to evaluate the keyword spotting outcome. Recommendations for future development of decision making process and implementation of the radio auditing automation will also be outlined.

5. Acknowledgements

The author would like to thank Pieter Lottering and Charday Williams for their offer of the project, kind support and generous advice. The academic support provided by supervisor Roberto Togneri has been invaluable and is also greatly appreciated. The author would also like to thank Jeremy Leggoe and Amanda Bolt for mentoring and administration matters throughout the project.

6. References

Garofolo, J. S. et al. (1993) TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download.

Hazen, T. J., Shen, W. & White, C. (2009) Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates. *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2009)*. Merano, Italy, IEEE, pp. 421-426.

Mandal, A., Kumar, K. P. & Mitra, P. (2014) Recent developments in spoken term detection: a survey. *International Journal of Speech Technology*, **17**(2), pp. 183-198.

Zhang, Y. (2013) *Unsupervised speech processing with applications to query-by-example spoken term detection* (Doctoral dissertation, Massachusetts Institute of Technology).

Zhang, Y. & Glass, J. R. (2009) Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2009)*. Merano, Italy, IEEE, pp. 398-403

Zue, V., Seneff, S. & Glass, J. R. (1990) Speech database development at MIT: TIMIT and beyond. *Speech Communication*, **9**(4), pp. 351-356.