Abstract

A large volume of spoken content stacked without annotation is challenging while the desired information is to be retrieved. The conventional Automatic Speech Recogniser (ASR) based approach accomplishes the retrieval task by converting the speech into text, and text-based matching was employed to retrieve the desired spoken contents. The ASR-based retrieval system requires a large volume of annotated content to train the system and demands a lot of manpower to prepare the annotated content. Spoken content belongs to low-resource languages (limited annotation), and zero-resource languages (without annotation and transcriptions) are not considered for retrieval tasks under the ASR-based retrieval framework. Hence, in this research work, we aim to overcome the resource constraint problem using an unsupervised pattern discovery approach. In the proposed approach, the spoken content indexing and retrieval technique was formulated in the zero-resource constraint by directly mining the pattern matches from the acoustic feature representation of the speech signal itself.

The Spoken Content Retrieval (SCR) task in the zero-resource scenario was achieved by computing the pattern matches between the spoken query and the spoken content based on their acoustic feature representations. Despite the feasibility, the variabilities generated by the natural speech signal produce a lot of false alarms during the retrieval and degrade the performance of the system. In the

proposed approach, a four-stage SCR technique was developed to overcome the challenges offered due to the inherent speech variabilities and resource constraints.

At first, the speaker variability challenge was addressed at the acoustic feature level by disentangling the speaker-specific characteristics from the spoken content. In the second stage, repeated patterns within the speech corpus were discovered by applying the appropriate pattern-matching techniques to the acoustic feature representation. In the third stage, the repeated patterns discovered are grouped and indexed using the Community discovery approach. Finally, given a spoken query, the occurrences of query content were retrieved with the help of indices. In all stages, the SCR task was achieved in an unsupervised way without using any linguistic resources. In this research work, the contributions made to achieve the SCR in the zero-resource constraint are listed as follows:

(i) The speaker-independent acoustic feature representation has been derived by the proposed Deep Convolutional Encoder-Decoder neural network that disentangles the speaker-specific characteristics from the spoken content.

(ii) Affinity kernel propagation and Heuristic pattern matching methods were proposed to discover the repeated patterns in the presence of acoustic feature variabilities.

(iii) The proposed community discovery algorithm groups the repeated patterns discovered based on their similarity propagation and temporal locality information. In addition, the best acoustic feature representation that matches its members was promoted as an index.

(iv) Finally, the overall SCR framework was designed to suit the zero-resource scenario in which the retrieval was viable for multiple languages.

Based on the evaluation using Microsoft Low-Resource speech corpus, it is inferred

that the proposed approach reduces the false alarms by 36.1% and improves the hit rate by 14.3% in comparison with the other state-of-the-art methods.

Keywords: acoustic feature representation, affinity kernel propagation, acoustic feature map, heuristic pattern match, community discovery, spoken document indexing, spoken content retrieval.