ABSTRACT

The problem of monolingual Query-by-Example Spoken Term Detection (QbE-STD) is studied in the beginning. Posteriorgrams have been widely used in the research on QbE-STD. Features based on articulatory classes are known to be robust to phonemic variations. The articulatory features like voicing and place of articulation are the main distinguishing features among some plosives and fricatives. These properties inspire the study of posteriorgrams based on articulatory classes for QbE-STD. Most of the previous works based on articulatory features have defined a large number of articulatory classes making it difficult to use them directly for pattern matching. Also, most of the works have completely ignored the uniqueness of the phonemes having transitory places of articulation eg. diphthongs and approximants. These issues have been addressed in this work while carefully selecting low dimensional articulatory motivated (LDAM) posteriorgrams on the basis of detailed experiments. The three LDAM posteriorgrams used in this work are named as vow, cpos and cman posteriorgrams based on the vowels, consonant position and consonant manner. It is shown that LDAM posteriorgrams can perform significantly better than widely used phonemic posteriorgram for the task of based QbE-STD in English speech corpus. Various methods for assimilating the information in these three posteriorgrams are studied. Different distances to compute distances between LDAM posteriorgrams are also explored.

Changes are done in LDAM posteriorgrams generated by English trained MLPs to bring out the minute details which are significant in multilingual scenario. With these enhancements, we show that articulatory posteriorgrams can give comparable or slightly better performance than the state-of-the-art phonemic posteriorgrams in the context of three Indian languages viz. Hindi, Telugu and Bangla. In addition, we also show that the combination of phonemic and articulatory posteriorgrams trained on English language can perform between 4.6 to 5.9 % (absolute) better than phonemic posteriorgrams alone for the task of Multilingual QbE-STD without using diversity of any other language or acoustic features. However, it is found that some phonemes unique to Indian languages can not be represented uniquely by this approach.

Different strategies like transfer training and joint training are therefore studied to adapt English trained neural networks with smaller amount of Bangla data. LDAM framework is designed to handle typical features like aspirated plosives, nasalized vowels, combined letters, unvoiced retroflex plosive which are found in majority of Indian languages. It is possible to train such a network with data of multiple Indian languages. It is observed that joint training not only improves QbE-STD in the language of same language family like Hindi but also other Indian languages like
Tamil and Telugu. While cross lingual adaptation of neural networks with a language specific softmax layer has been studied earlier in the context of speech recognition, this work is the first to report a completely language independent network which can be further adapted with a large number of languages. It is also noticed that many classes of Indian phonemes and English phonemes get well represented simultaneously with this training which is desirable for mixlingual scenario.

A novel technique of text query search in speech is proposed for mixlingual and multilingual speech in the framework of spoken query search which is described above. It is based on LDAM posteriorgrams. The capability of spoken query search enables it to search out-of-vocabulary (OOV) words. The capability of searching text queries enables it to perform semantic search. An advanced application of searching keyword translations in mixlingual speech corpus is also possible within posteriorgram framework with this system. It is shown that the performance of text queries is comparable or better than the performance of spoken queries if the language of the keyword is included in the training languages. Also, a technique for combining information from text and spoken queries is proposed which further enhances the search performance.