# Abstract

Spoken term detection (STD) or keyword spotting provides an efficient means for content based searching of speech. Here only a few spoken query word exemplars are provided for detecting their presence in a target speech utterance. Achieving high detection performance, faster speed, detecting out-of-vocabulary words in the context of STD on low resource languages are some of the major research challenges. The problem can be better addressed using template based methods rather than a full-fledged large vocabulary automatic speech recognizer (ASR). The templates are derived from limited number of spoken word exemplars provided as reference. These are matched against templates derived from segments of the test utterance having similar representation. The thesis presents techniques for representation and classification of templates obtained using only spoken word exemplars without any labeled speech data or linguistic resources. The focus is on robust classification performance and faster matching with limited number of reference word exemplars. This includes non-availability of labeled speech resource for training full-fledged ASR and noisy environments. It brings out relevance and need of template based approaches for low resource scenarios such as zero (labeled) resource frameworks. Then a review of the approaches available in literature is presented that are suitable for scenarios where large volume of labeled speech corpora for supervised training of acoustic models is not available. Three new template representations are proposed next. The first is based on ranks of the mixture components of a Gaussian mixture model (GMM) trained in an unsupervised setting without labeled data. The resulting GMM is used to obtain frame level posterior probabilities of each component. The top ranking component per frame is used to derive the template representation. The second representation is based on weights of atoms of a learnt dictionary. Two formulations of over complete dictionary namely *dict-concat* and *dict-mixed* are presented. The third template representation is based on quantized spectral peaks and the process of extraction of these peaks from speech signal for template representation is presented. The proposed representations are mostly sparse and invariant.

The proposed representations are used by parametric and non-parametric classifiers for template classification. First, two non-parametric techniques for template classification are developed. The non-parametric techniques perform direct matching of the test and the reference templates without building models of spoken word classes. Locality Sensitive Hashing (LSH) and Dictionary Activation Grouping (DAG) based classifiers are proposed for classification. Classification time for these algorithms are much lesser compared to DTW based approaches. Next, a technique is proposed for parametric modeling of temporal dynamics of feature elements constituting the templates. The approach is based on Point Process modeling (PPM) of the arrival events associated with feature elements over the duration of the template. These elements can be Gaussian mixture components, spectral peaks or dictionary atoms depending on the representation. All the above feature elements are shown to have linguistic correlates. The performance evaluation of different combinations of the template representations and classification (both parametric and non-parametric) techniques for classification of spoken word templates is presented for both clean and noisy speech. The experiments are performed using isolated spoken word utterances from TI46 database. The database consists of twenty isolated words (10 isolated digits from 0-9 and 10 English words) spoken by 16 speakers (8 males and 8 females) for each of the train (reference) and test sets. First, a comparative analysis of the template recognition performance of the proposed techniques on clean speech is presented. A relatively high recognition accuracy with fewer reference templates, lower classification time, and robust recognition performance with variation in number of reference templates is achieved with the proposed techniques compared to related literature. The performance of the proposed template representations and classifiers is studied next on noisy exemplars. The robustness of recognition performance in presence of additive noise is reported. The study is conducted on noisy speech data contaminated with additive white Gaussian noise resulting in templates obtained from speech with different signal-to-noise ratio. The proposed techniques are found to provide superior performance as compared to related techniques.

**Key words:**
Template Representation, Template Classification, Gaussian Posteriors, Dictionary Learning, Spectral Peaks, Locality Sensitive Hashing, Point Process Models.