Spoken Keyword Rescoring and Document Retrieval for Low-resource Languages

Min Ma
Graduate Center, The City University of New York
mma@gradcenter.cuny.edu

Abstract
For languages that have adequate data for automatic speech recognition (ASR), many keyword search (KWS) and document retrieval (SDR) systems have been developed with near-optimal performance. However, lacking of sufficient training data to produce high accuracy transcript, identification and retrieval of queries in speech data from low-resources languages remains challenging. To compensate for these shortcomings, we extracted signals from the data that are useful in KWS/SDR. These signals take multiple forms: word/morpheme burstiness, rescoring confusion network posteriors, acoustic/prosodic qualities, phoneme recognition results, etc. Different strategies have been proposed: 1) a four-way classification of keyword hypotheses; 2) learning-to-rank algorithms; 3) fusion of symbolic subsequence matching and information-based dynamic time wrapping; 4) prosody-based language model. We evaluated these strategies to the speech data from IARPA-BABEL and MediaEval-QUESST Workshop. The paper summarizes the previous work of author in spoken keyword rescoring and document retrieval, as well as discusses the ongoing exploration.

Index Terms: keyword search system, spoken document retrieval, IARPA-BABEL, QUESST

1. Introduction
1.1. KWS and SDR Tasks
The KWS task is to find all of the occurrences of a keyword, a sequence of one or more words in an original languages orthography, in a corpus of un-segmented speech data[1]. SDR, which may take more complicated phrase/sentence as query, provides content-based retrieval of excerpts from archives of recordings of speech[2]. For rich-resource languages, such as English, both of the two tasks combines the efforts from ASR and information retrieval (IR) fields.

However, low-resource languages present a special challenge for KWS/SDR because the performance of automatic speech recognition is poor due to a paucity of training data. To enhance the performances of KWS/SDR systems for low-resource languages, we propose to elaborate on standard ASR-based approaches by extracting new signals from KWS hypotheses to rescore hypotheses. The paper summarizes the previous studies the author participated in BABEL Program and MediaEval QUEST Workshop, as well as describes the ongoing work.

1.2. BABEL Program
Babel Program aims to develop methods to build speech recognition technology for low-resource languages, which are in the majority of human languages[3]. The author used to contribute to the keyword candidate rescoring module of the Program, based on the KWS result data we use is generated by the IBM Speaker-Adapted Deep Neural Network (SA-DNN) system[6]. All data for this project is disseminated by the National Institute of Standards and Technology (NIST) on behalf of the Intelligence Advanced Research Projects Activity (IARPA). The language packs are named IARPA-babel104b-v0.4bY, IARPA-babel105b-v0.4, IARPA-babel106b-v0.2g, IARPA-babel107b-v0.7, and IARPA-babel206b-v0.1e, for Pashto, Turkish, Tagalog, Vietnamese and Zulu, respectively.

We use results from an ASR system trained on the full language pack of 40 speech hours for the first four languages and on the limited language pack of 10 speech hours for Zulu. An additional set of development and evaluation data, with a distinct list of keyword queries for each data pack, is released by IARPA. Development queries number approximately 300; evaluation queries number approximately 3,000. A keyword query may comprise one or more words.

1.3. MediaEval QUESST Workshop
The Query by Example Search on Speech Task (QUESST) involves searching for audio content within audio content using an audio content query[4]. This task is particularly interesting for speech researchers in the area of spoken term detection or zero/low-resource speech processing. In 2015, six low-resources language packs have been released: Albanian, Romanian, Slovak, Portuguese, and code switched Mandarin/English. The task consists in determining how likely it is that a query appears within an audio file. Given an audio file and a spoken query, systems will have to produce a score. The higher the score the more likely is that the query appears in the audio file. Note that the task involves verifying the presence of the query anywhere in the file, and not finding the exact time point of query occurrences.

Three different types of query searches are considered in the generic retrieval system: exact match, re-ordering and small lexical variations, conversational queries in context(e.g. silent pauses, filled pauses and irrelevant words within the query). The primary metric used this year will be the cross entropy score (Cnxe), and the Actual Term Weighted Value (ATWV) metric will be used as a secondary metric.

2. Features
2.1. Word/Morpheme Burstiness
The assumption of word burst, or burstiness, analysis is that once a word or phrase has been uttered in a conversation, the probability of that phrases repetition is higher than its marginal probability [5]. We integrated the burstiness information into the KWS system at its last stage, i.e. rescoring the deci-
sions made by the system. Specifically, we extracted 25 burst-related features from each target hit hypothesis $t$ in a posting list. These features involve calculations regarding the number, strength, and proximity of neighbor hypotheses $n$ within a conversation[7].

Besides, for the languages that are highly agglutinative (e.g., Zulu, Turkish, etc.), a word would rarely appear without prefixes, infixes, or suffixes that depend on the words grammatical context. This complicates our burstiness assumption, especially when keyword queries are multi-word phrases, which they often are. A potential solution is taking the morpheme as basic units so that word burstiness could be extended to morpheme (stem, prefix, suffix) burstiness.

### 2.2. Consensus Net Features

Given KWS result data generated by the IBM SA-DNN system for speech recognition[6], we utilize the pruned lattice of word hypotheses termed a consensus net (CN). In a lattice, multiple hypotheses of word boundaries are entertained, but a CN forces consensus on word boundaries. The arcs between boundary nodes are potential transcriptions, with each arc assigned a confidence value, and it is this score that we target with our rescoring strategies[9].

### 2.3. Acoustic Features

We extract the pitch contour of the of the posting list entry and compute its median, as well as the pulses, duration, jitter and shimmer, etc. All acoustic features are normalized at the segment level. Additionally, intonational phrase boundaries and pitch accents had been detected using prosodic event detectors trained in cross-language corpus[7].

### 2.4. Prosodic Features

Prosody, or intonation, is a critically important component of spoken communication[7]. Automatic detection and classification of prosodic events specifically, pitch accents and prosodic phrase boundaries, plays a critical role in capturing many aspects of the manner in which these words are spoken. Although the keyword query varies in word composition, the speaker tends to use the same habitual modal to highlight his/her key points. We are trying to take advantage of the kind of patterns by embedding the prosodic type of each word, so that might help the keyword rescoring.

### 3. Keyword Rescoring Strategies

#### 3.1. Rule-based Rescoring

The rule-based method of rescoring is an algorithm for selectively boosting the scores of hits that depends on three parameters: a score threshold $\tau$, an increment size $\epsilon$, and window size $\omega$, expressed in seconds. For a target hit $t$ in a given conversation file, we perform the following steps:

1. Assemble a list $N_t$ of neighbor hits for the same keyword within the time window.
2. Find $\max_{N_t}$, the highest-scoring hit in $N_t$, and compare its confidence score to $\tau$.
3. If $\max_{N_t}$ is higher than $\tau$, then boost the score of $t$ by adding to it $\epsilon \times$ the score of $\max_{N_t}$.

For tuning, we searched a 0 to 1 range for the increment and threshold parameters, and we search a range of 0 to 600 seconds (the duration of a conversation file) for the window parameter. Experiments on Pashto, Turkish, Tagalog, Vietnamese language packs from BABEL shown 1.52% average improvement in MTWV, and 0.85% performance gain for Zulu.

#### 3.2. Four-class Classification For Rescoring

In [7], we proposed burst-based machine learning models to bolster KWS performance. First, we assigned to each hit one of four class labels: low-scoring correct hits, low-scoring false alarms, high-scoring correct hits, and high-scoring false alarms, with an intention to target low-scoring correct hits for rescoring. We then train a logistic regression model to generate predictions with confidence scores over the four class labels. We take a weighted average of these four class confidences and interpolate that average with the score of a given hit using the coefficient $\eta$, yielding the formula $R(t) = (1 - \eta)s(t) + \eta \sum_{k \in C} w_k \cdot c_k$, where $R(t)$ is the rescoring value, $s(t)$ is the original score of a target hypothesis $t$, $C$ is the label set $\{\text{LowCORR, LowFA, HighCORR, HighFA}\}$ of class confidences, and $W$ is the set of co-indexed weights for those confidences.

In tuning, we seek optimal values for each of the four class weights as well as the interpolation coefficient $\eta$. We find that setting non-zero weights for all high-scoring hits improves our rescoring results except on Zulu. It may due to the agglutinative nature of Zulu, and the author are working to overcome this obstacle by decomposing Zulu keywords into morphological segments.

#### 3.3. Learn-to-rank For Rescoring

We sought to reorder the scores in a posting list not by classifying each hit individually, but by learning and predicting an ideal ranking of the entire list[10]. To rank our training data accordingly, we balanced the input of reference labels and ASR confidence. For each keyword in a posting list, we ordered its correct hits by CN score, followed by the a similar ordering of false alarm hits, such that the highest-scoring false alarm is ranked directly below the lowest-scoring correct hit.

Therefore, we trained several ranking models using the learn-to-rank algorithms, then interpolate the normalized rank scores with posting list scores. The method promoted the MTWV of 1.10% in average for all languages.

#### 3.4. Combination and Ensemble

We validated all the three rescoring strategies based on word burstiness, consensus net, acoustic features and their combination; also, we assembled the results produced by single strategy. Experiment results showed that for nearly every strategy assessed and every language studied, improvement had been gained in MTWV. The ensemble method, though it does not improve over the highest-scoring strategy, proves a reliable way to choose a strategy.

#### 3.5. Prosody-based Language Model for Rescoring

The intuition behind the idea is people tend to emphasize the keywords in a habitual modal, which could be characterized by the prosodic features. First, we recognize the acoustic regions of the consensus nets, then split the audio by word and extract prosodic features; second, we do a k-nearest-neighbor clustering based on the features; then we incorporate the cluster information in language model, which will be used for keyword
candidate rescoring. The work currently is under working.

4. Spoken Document Retrieval

4.1. Fuzzy Sequence Matching

An intuitive way to spot the occurrences of a query in the audio documents is first transcribe both of them into text, then apply the textual information retrieval techniques. However, lacking of both annotated transcripts and pronunciation dictionaries, the quality of ASR for the under-resource languages would not be high. Thus the next best thing would be phoneme recognition. Using the phoneme recognizer provided by BUT[11], we get the phoneme sequence for both the query and reference audio. Apply textual fuzzy subsequence matching to these sequences could assist determining the occurrence/absence of the query in each reference audio document.

4.2. Information-based DTW

Information retrieval-based dynamic time warping ([12][13]) performs subsequence matching in two steps. First, it identifies the set of possible matching frames from the search collection for every query frame; then the pairs of very similar frames in both time series) will be connected together to form matching paths between query and reference time series. For each query frame, the two steps are executed sequentially.

5. Future Work

The author did some work on detecting speaker states (such as emotion, nativeness, Parkinson’s Disease Rating, etc.) [18, 16, 17] based on acoustic, prosodic and phonotactic features. Since these features are independent from the ASR outputs and can be extracted from the audio itself, extending these researches to low-resource spoken languages should help understanding the languages, and build more accurate and robust information extraction and retrieval systems.

6. References

[16] Min Ma, Keelan Evanini, Anastassia Loukina, Xinhao Wang, Using F0 Contours to Assess Nativeness in a Sentence Repeat Task. Accepted by INTERSPEECH, 2015, Dresden.