

Chapter 2

Keyword Spotting Methods

This chapter will review in detail the three KWS methods, LVCSR KWS, Acoustic KWS and Phonetic Search KWS, followed by a discussion and comparison of the methods.

2.1 LVCSR-Based KWS

Performing KWS on textual databases is relatively straightforward. The text is perused for a given list of words and the location of the words is tagged within the text. Translating this method for use in speech databases is a two-stage process. First, an LVCSR engine is employed to transform the entire speech signal into text. The LVCSR engine performs the search for the most probable sequence of words based on the Viterbi search algorithm, using acoustic models, a large lexicon of words and a language model. In the second stage, the KWS mechanism utilizes established text-based search methods to locate the keywords within the text. An indexing phase can be performed on the resulting text in order to accelerate the search response time. This method will be referred to as LVCSR-based KWS.

Figure 2 illustrates the two sequential stages involved in LVCSR-based KWS.

2.2 Acoustic KWS

Another common KWS method is Acoustic KWS. Using this method, the engine does not attempt to transcribe the entire stream of speech. Like the LVCSR-based method, this method employs the Viterbi search. That is, the system employs a speech recognition engine on the speech. However, rather than a large vocabulary which is intended to cover all potentially spoken words, a smaller set of designated keywords is used as the recognition vocabulary (Thambiratnam 2005) and general speech models (as part of the acoustical models) are used to model

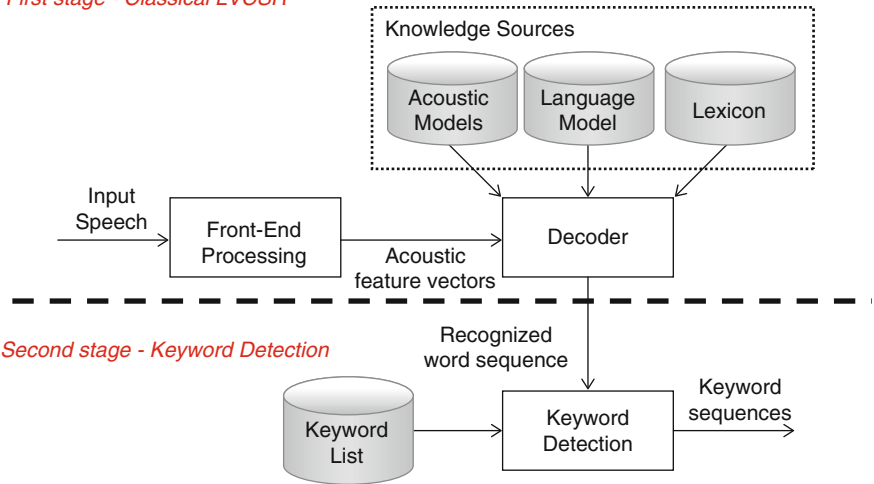
First stage - Classical LVCSR

Fig. 2 An LVSCR keyword spotting system – one-time transformation of a speech database (DB) into a textual word DB and KWS engine

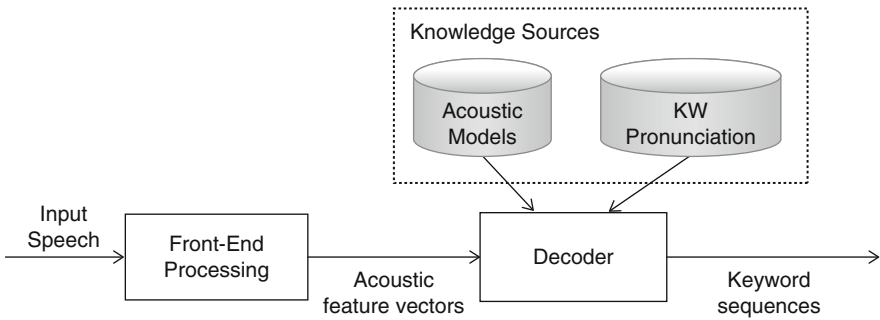


Fig. 3 An acoustic keyword spotting system

non-keyword speech (Szöke et al. 2005). Thus, acoustic KWS can be performed in only one stage, as illustrated in Fig. 3.

2.3 Phonetic Search KWS

As its name suggests, phonetic search KWS utilizes a phonetic search engine. In the first stage, a phoneme decoder is employed once to transform the speech input into a textual sequence. However, rather than producing a string of words, the decoder transforms the speech signal into a string (or lattice) of phonemes (Amir et al. 2001;

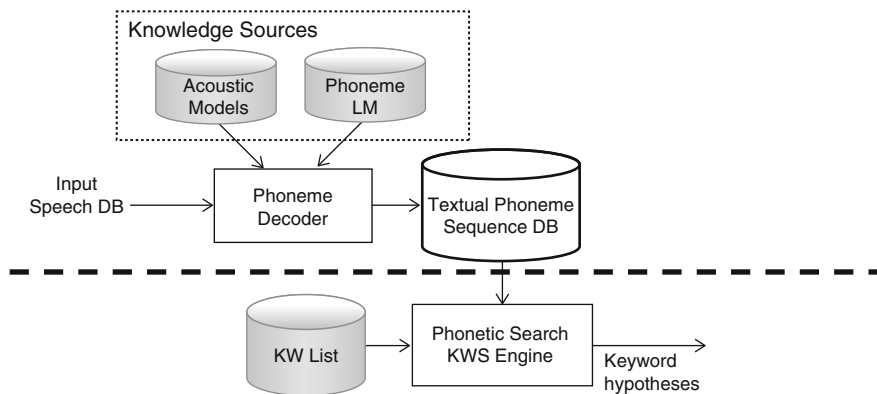


Fig. 4 A phonetic search system – one-time transformation of a speech DB to a textual phoneme DB and KWS phonetic search engine

Yu and Seide 2004; Thambiratnam and Sridharan 2005). In the second stage, the phonetic search engine employs a distance measure to compute the textual distance between the phoneme sequences that correspond to the keyword vocabulary and the phoneme sequences within the phoneme string (Alon 2005).

As shown in Fig. 4, the phonetic search engine uses two types of input data: a list of keywords, where each word is represented by a sequence of phonemes, and a speech database which has been run through a phoneme decoder to produce a sequence of recognized phonemes.

2.4 Discussion: Why Phonetic Search?

Each of the three KWS methods presented above has advantages and shortcomings. The crucial parameters to evaluate are response time, KWS performance, and keyword flexibility (James and Young 1994; Dharanipragada and Roukos 2002; Mamou et al. 2007; Thambiratnam and Sridharan 2007; Schneider 2011).

2.4.1 Response Time

In terms of overall computational complexity, LVCSR-based KWS and phonetic search KWS both implement a double stage process: (1) transformation of speech to text (word sequences in the case of LVCSR and phoneme sequences in the case of phonetic search) and (2) a keyword search (word-based in the case of LVCSR and phoneme-based in the case of phonetic search). Acoustic-based KWS, on the other hand, is performed in one stage and operates on the speech itself with no textual

transformation. Although a keyword search that is implemented on fully transcribed text in the LVCSR method is fast (particularly if the text has also been indexed), it is usually at a disadvantage in comparison to the phonetic search and acoustic methods due to the fact that an LVCSR engine demands a large vocabulary and a complex language model to produce recognition results, thus resulting in a high level of complexity during the pre-processing stage.

The phonetic search method performs phoneme recognition using phoneme transition probabilities (di-phones) with no lexicon or word level language model. During the search stage, however, phonetic search KWS uses a textual sequence distance measure that requires more computation. This is because the phonetic search must generate word-level hypotheses based on phoneme sequences, while in LVCSR-based KWS the textual output is already word-level (Burget et al. 2006).

In contrast, the acoustic-based KWS uses a vocabulary consisting only of the keywords and does not require a language model at all. Because the acoustic-based method operates on the speech itself and requires only a small vocabulary, it is appropriate for real-time keyword spotting or KWS in small speech databases. However, this means that general speech must be well-modeled (Thambiratnam 2005) to avoid extensive over detection (false alarms).

2.4.2 KWS Performance

The spontaneous speech and poor recording quality of speech databases often leads to deficient LVCSR performance (Butzberger et al. 1992; Cardillo et al. 2002). The large number of disfluencies, including mispronounced words, false starts, filled pauses, overlapping speech, speaker noises and background noise found in spontaneous speech (Butzberger et al. 1992; Gishri and Silber-Varod 2010) often results in outputs strewn with word insertions, deletions and substitutions. Thus the “most probable” word sequences produced by the engine may not adequately reflect the actual input speech. This, in turn, affects the reliability of the keyword search.

The same is true with regard to phonetic search results. Poor phoneme recognition may yield lower keyword recognition performance in comparison with the acoustic KWS method, which works on the speech itself by searching for a specific sequence of phonemes without textual transformation.

2.4.3 Keyword Flexibility

In comparison to the phonetic search method, which runs on sequences of phonemes rather than words, the LVCSR method is at a disadvantage when it comes to keyword flexibility (Cardillo et al. 2002; Burget et al. 2006; Wallace et al. 2007). Using the phonetic search method allows application users total freedom in changing the designated keywords, since the textual transformation into phonemes

is not restricted by a vocabulary. Adding new keywords is a simple procedure that entails re-running the phonetic search on the phoneme sequences, but does not require re-running the phoneme decoder.

The textual transformation produced by an LVCSR engine, on the other hand, is constrained by the recognition vocabulary and the language model employed. Thus, unless the designated keywords were part of the original recognition vocabulary, they cannot be changed without repeating the recognition process (Clements et al. 2001; Cardillo et al. 2002; Szöke et al. 2005; Mamou and Ramabhadran 2008). Since keywords are in many cases names or domain-specific vernacular, they are often not found in standard lexicons (Wallace et al. 2007; Gishri and Silber-Varod 2010). This is a substantial shortcoming of the LVCSR method.

Acoustic-based KWS also represents an impractical solution for searching large databases that require rapid and flexible searching capabilities. Because it consists of only one stage, the entire process needs to be re-run on the speech database each time a new keyword dictionary is introduced.

The majority of applications require keyword flexibility, as well as the shortest possible response time when searching very large speech databases, making the phonetic search KWS method more attractive than the LVCSR and acoustic-based options when searching very large speech databases. Thus, the focus of the following chapters will be on phonetic search KWS, and the implementation of an algorithm for the reduction of computational complexity in the phonetic search KWS process.

Phonetic Search Methods for Large Speech Databases

Moyal, A.; Aharonson, V.; Tetariy, E.; Gishri, M.

2013, X, 53 p. 21 illus., 6 illus. in color., Softcover

ISBN: 978-1-4614-6488-4