

# **Various Approaches to Search for Words or Phrases in Audio Data**

A white paper, which compares and analyses the strengths and weaknesses of various approaches to detect words or phrases in audio streams or audio recordings

# 1 Table of Contents

<b>1</b>	<b>TABLE OF CONTENTS.....</b>	<b>2</b>
<b>2</b>	<b>INTRODUCTION.....</b>	<b>2</b>
<b>3</b>	<b>THREE DIFFERENT APPROACHES.....</b>	<b>3</b>
<b>3.1</b>	<b>KEY WORD SPOTTING.....</b>	<b>3</b>
3.1.1	KWS STRENGTHS.....	3
3.1.2	KWS WEAKNESSES.....	3
3.1.3	KWS APPLICATION AREAS.....	3
<b>3.2</b>	<b>PHONETIC INDEX SEARCH.....</b>	<b>4</b>
3.2.1	PIS STRENGTHS.....	4
3.2.2	PIS WEAKNESSES.....	4
3.2.3	PIS APPLICATION AREAS.....	4
<b>3.3</b>	<b>SPEECH TO TEXT.....</b>	<b>5</b>
3.3.1	S2T STRENGTHS.....	5
3.3.2	S2T WEAKNESSES.....	5
3.3.3	S2T APPLICATION AREAS.....	5
<b>4</b>	<b>STRENGTHS AND WEAKNESSES OF THE APPROACHES.....</b>	<b>6</b>
<b>5</b>	<b>CONCLUSION AND OUTLOOK.....</b>	<b>7</b>

## 2 Introduction

This paper lists and compares all possible approaches of how to search for keywords or phrases in audio streams or audio recordings.

We will analyze different approaches based on speech recognition such as keyword spotting, phonetic index search and speech to text. All of these approaches can be used to classify recordings as well as searching for specific keywords or phrases in online or stored audio data.

However these approaches are not equal and each of them has its strengths and weaknesses. We will show the different strengths and weaknesses as well as discuss how they can be used in the most efficient way.

## 3 Three different approaches

Various vendors offer technologies to search for words or phrases in online (streaming) or stored audio data. These technologies are typically based on one of the following approaches:

1. Key Word Spotting (KWS)
2. Phonetic Index Search (PIS)
3. Speech To Text (S2T)

All of these approaches are language dependent and based on standard speech recognition methods. The main differences are in the way when they actually convert from the audio domain to the phoneme domain and finally to the word domain.

### 3.1 Key Word Spotting

KWS searches for a given set of words or phrases directly in the audio data. This means, that the set of words to search for should be specified in advance. The technology converts the given words or phrases to the corresponding phoneme sequence, either by using a lookup lexicon or by statistical approaches and searches then for the occurrence of these phoneme sequences in the audio stream. It alerts the client application when one of the given phoneme sequences can be found.

#### 3.1.1 KWS Strengths

Since KWS searches directly in the audio stream the accuracy is very high. It allows searching for any words, even names. It is robust regarding spelling/typing errors since the search is performed on the phonetic level.

KWS is capable of alarming immediately as a given phrase was said in the audio.

#### 3.1.2 KWS Weaknesses

The main weakness of KWS lies in the relatively slow speed when you search directly in the audio data. Therefore the words should be known before the audio stream is analyzed. Analyzing and audio streams can be performed in approx. 10 times real-time (1 hour of sound can be analyzed in approx. 6 minutes on a single CPU). Processing speed is also slightly depending on the number of keywords specified.

Due to its low analysis speed it is not efficient when users want to find words or phrases in larger sound recordings. With other words: KWS is not efficient to implement a search engine to search in audio or video recordings.

KWS main disadvantage is its static behavior. Once you analyzed the sound data with a given set of keywords you cannot get more out from the analysis result.

#### 3.1.3 KWS Application Areas

KWS is mainly used in monitoring ("Alert when a certain keyword or phrase is mentioned") environments where the list of words to be monitored is typically known.

It is used in call-centers to trigger an alert or mark the position in a recording when a certain keyword is mentioned.

The weakness of relatively slow search speed limits the size of recordings that can be searched in. E.g. if you want to find a word in 10 hours of sound, you would need to wait for 1 hour until you get a list with all occurrences.

KWS can also be used very well to classify recordings / calls. E.g. for each category a set of keywords is specified and depending on the KWS results the most likely category is chosen.

## 3.2 Phonetic Index Search

PIS comprises a two-step approach: In the first step a phonetic index is generated from the audio stream. The audio data is converted into a phoneme graph, which can be used as an index for future searches. This phonetic index is language dependent but word or phrase independent.

In the 2<sup>nd</sup> step the index can be used to search for any word or phrase. Similar to the KWS approach the written word (grapheme) is converted into a phoneme sequence. This phoneme sequence is then looked up in the phonetic index, which was created in the first step. Phonetic search at its core doesn't search for words, but for phonetic patterns, e.g. "whether" and "weather" will produce the same results, which is not the case for S2T and sometimes also not for KWS. Words are just a convenient way to specify the phonetic patterns.

Searching in such phonetic index can be optimized so that words can be found immediately. The search speed in the phonetic index is approx. 100,000 times real-time (1 hour of sound can be analyzed in 36 milliseconds).

### 3.2.1 PIS Strengths

PIS is very accurate. Similar accuracy results as with KWS can be achieved. It allows searching for any words, even names. It is robust regarding spelling/typing errors since the search is performed on the phonetic level.

The very high search speed enables users to search precisely for words or phrases also in larger audio databases.

Therefore you can see PIS as very "dynamic" approach. Once you have created an index, you can search for any word or phrase very fast. The index is very general since it is word and topic independent. There is no need to update and maintain a lexicon and no need to make user specific training.

### 3.2.2 PIS Weaknesses

A weakness of PIS compared to KWS is the requirement to manage and store the phonetic index file. The index has to be loaded into memory so that it can be used for fast searches.

### 3.2.3 PIS Application Areas

PIS can be used for any small to mid-sized recordings. Typically users accept a response time of max. 3 seconds. Therefore PIS can be used to search in audio recordings of up to several hundred hours of recorded sound on a single PC.

PIS is typically used in Call-centers, (telephone-) recorders, Digital Asset Management, Media Asset Management, Knowledge Management, E-Learning, Audio/Video conferences and larger Audio/Video archives.

Some PIS implementations also contain KWS functionalities. During the creation of the phonetic index also words or phrases can be spotted directly without additional processing overhead.

## 3.3 Speech To Text

S2T comprises a two-step approach: In the first step the sound file is converted into a textual representation, which is a kind of rough transcript of the sound file. To automatically create the transcript typically a speech to text speech recognition engine is used, which is typically guided by a given lexicon and language model. The lexicon is a list of words, which can be recognized. The language model is a table, which defines the likelihood of how likely a specific word appears (Unigram) or one word follows the other word (Bigram, or for 3 words a Trigram). The speech recognition engine tries to select the correct word from the lexicon and uses the language model to make its choice.

In the 2<sup>nd</sup> step the transcript of a sound file can be used to search for words or phrases similar to any text search. Typically also additional algorithms are required to cope with the potential problem of transcripts with a low accuracy of e.g. less than 50%.

### 3.3.1 S2T Strengths

Since the search can be performed on a textual level the search speed is extremely high. Any text search engine can be used to perform the search on the transcript. The storage needed to store the transcript is lower than with the PIS approach.

### 3.3.2 S2T Weaknesses

Current state-of-the-art speech recognition engines create transcripts with an accuracy that might drop below 50%. Higher accuracy can only be achieved if the lexicon and language model are specifically adapted to the topic.

Since the recognition process is based on a lexicon only words that are in the lexicon can be found. If a word is not in the given lexicon (e.g. names or hardly used words) it will not be transcribed properly. Instead of the correct words, several wrong words are typically recognized. Therefore the text search engine of the 2<sup>nd</sup> step can never find the correct words again.

The Speech To Text engine is always limited by the vocabulary and statistical language model used during the transcription phase. If the language model (likelihood model of words) does not fit the content in the audio recording then the accuracy of the transcript will be very poor. E.g. if you use a language model for Medicine then the transcript will always contain medical terms no matter what was spoken.

### 3.3.3 S2T Application Areas

If the lexicon and language model exactly fit the topic very high accuracies can be achieved. Therefore S2T is very efficient for news broadcast or domain specific applications (e.g. medical reports).

The main strength of S2T is in the area of very large sound recordings (more than several 100 hours) since the output (= text) is can be easily managed and XML or text search engines can be used to find the words. S2T is used for very large audio archives, which typically can be found in broadcasting companies or telephone interception applications.

## 4 Strengths and Weaknesses of the Approaches

The following table provides an overview about the strengths and weaknesses of the different approaches:

	Keyword Spotting (KWS)	Phonetic Index Search (PIS)	Speech To Text (S2T)
Passes	<b>1</b>	<b>2</b>	<b>2</b>
Parameters needed	<b>Language</b>	<b>Language</b>	<b>Language Topic / ConText</b>
Limitations on Keywords or phrases which can be found	<b>Predefined set</b> (Words have to be specified before processing)	<b>No limitation</b> (Once the phonetic index was created, you can search for any word)	<b>Words have to be part of lexicon</b> (Only words which were part of the lexicon during phase 1 can be found)
Support for new ("unknown") words like names	<b>Yes</b>	<b>Yes</b>	<b>No</b> (All words have to be in the lexicon before transcription starts)
Search Speed	<b>10 times real-time</b>	<b>50,000-100,000 times real-time</b>	<b>Very fast</b> (Text search)
Amount of audio data analyzed in 3 seconds on a single CPU	<b>30 seconds</b>	<b>Up to 80 hours</b> (3x100, 000 seconds)	<b>1000s of hours</b> (Depending on speed of text search engine used)
Language Dependent	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
Accuracy	<b>Very high</b>	<b>Very high</b>	<b>Low</b> (Can be high with tuning, but even with tuning you likely miss some important words)
Additional storage needed	<b>None</b>	<b>4kB/sec</b> (Size of phonetic index is scalable and can be nearly as low as for S2T. Tradeoff accuracy vs. size)	<b>Low</b> (Only text)

## 5 Conclusion and Outlook

All of the 3 approaches have its advantages and disadvantages.

For monitoring applications (alerting when a given keyword is mentioned) we recommend to use KWS. It has a high accuracy and only minimal data management overhead.

For searching in small to medium size (one to several hundred hours of sound) recorded databases it is recommended to use PIS. It provides fast searching capabilities in combination with maximal accuracy.

For larger sound archives (several thousands hours of sound) we recommend to combine S2T with PIS. S2T together with a state-of-the-art text or XML search engine can be used to identify a sub-set of relevant recordings. In a 2<sup>nd</sup> step PIS can be applied additionally to identify the occurrences in the subset very accurately.

There is no general "best way". It is rather essential to select the most effective approach or even mix of these technologies depending on each individual use-case.