

PHONETIC SEARCHING OF DIGITAL AUDIO

Mark Clements, Sc.D.
Peter S. Cardillo, MSEE
Michael Miller, CTO

Fast-Talk Communications, inc.
Atlanta, Georgia

INTRODUCTION

As archives of digital audio and video expand, and people need to find specific information within those archives, it becomes clear that a highly efficient method of searching recorded media is required. The metadata that currently tag audio information (such as title, date of recording, subject, or person) are not sufficient for the accurate and rapid retrieval of specifically requested data.

This paper summarizes previous audio searching techniques and the development and deployment of a new high-speed search technology, the *phonetic search engine*. The potential current and future uses for the new search methodology are also discussed.

The Phonetic Search Engine (abbreviated as PSE, trademark pending) is an open-vocabulary retrieval system, which greatly reduces the time, and increases the accuracy of searches against large collections of recorded speech. Searches can be conducted at speeds up to 36,000 times faster than real-time playback of the recordings.

AUDIO SEARCHING TECHNIQUES

Early applications of speech searches have been in the areas of surveillance, and military command and control. In these applications, the search algorithms are executed in real time. Other approaches include a preprocessing step and produce results only after the preprocessing is completed.

Several different methods have commonly been applied to the speech retrieval problem. One approach is to employ a Large Vocabulary Continuous Speech Recognizer (LVCSR, also known as "speech-to-text"). Speech is converted to text that can be searched very quickly for occurrences of a specified keyword or keywords. However, there are significant problems with this method. Since the outputted transcripts are a concatenation of words in the recognizer's dictionary, words outside that lexicon will not be detected. This results in a closed vocabulary. The addition of new vocabulary requires reprocessing, which is difficult to accomplish faster than real time and therefore incurs significant overheads for large archives. Also, word error rates are unacceptably high (often above 40%).

Another disadvantage of speech-to-text is that hard decisions about each word's existence must be made during the recognition phase. Once a word is bound to a particular sequence of sounds, all other possible definitions are abandoned. However, the choice made may not be the correct one, and the other choices will not even be considered during the search phase.

In another retrieval technique, called *word spotting*, the search is performed on the speech after the keyword is presented. This produces an open vocabulary. The disadvantage to this process is the difficulty of searching much faster than real time. Approximations can be introduced to accelerate the search but usually at the expense of accuracy.

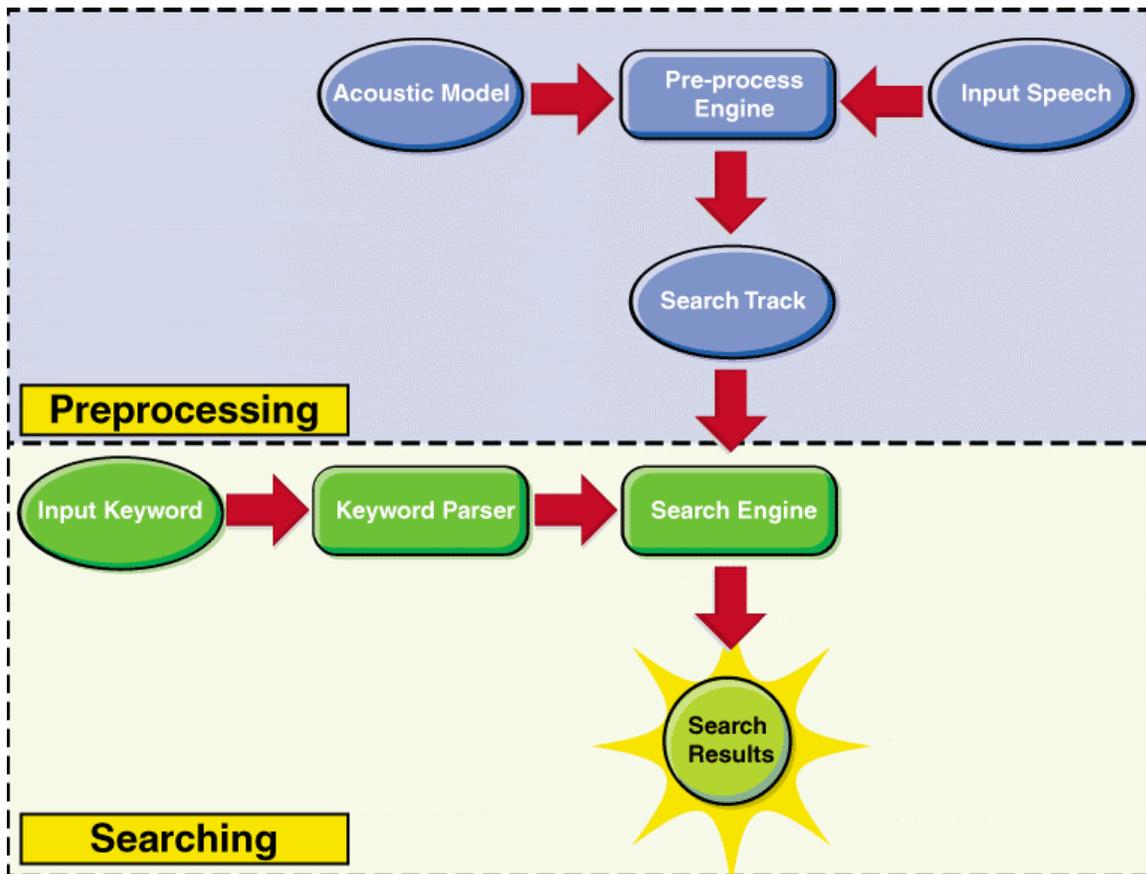


Figure 1 – Search Engine Architecture

THE NEW SEARCH ARCHITECTURE

In order to search the vast archives of available prerecorded speech in a useful and timely manner, a search engine must be able to hunt through hundreds of thousands of hours of speech, many orders of magnitude faster than real time. In order to do this, an entirely new class of algorithms must be applied to the recordings, preprocessing the speech to speed the retrieval when keywords are subsequently specified for searching (see *Figure 1*).

Language and Channel Issues

To accurately interpret speech, certain parameters must be considered. Each language has its own set of phonemes and grammatical structures. Even within one language there are variants (for example: British versus North American English, or Mexican versus Castilian

Spanish). Additionally, speech may be recorded on different channels of varying quality: broadcast quality audio, lower quality audio, landline telephones, wireless cellular, etc. Each of these languages and channels presents challenges when automating the process of identifying specific speech elements. The PSE uses multiple techniques to account for these variants when locating the requested information.

Acoustic Models

Acoustic models describe the expected characteristics of the audio files to be preprocessed. These parameters include non-linear frequency response, background noise, reverberation, and others. All of these characteristics are present, to some degree, in every audio recording. The best case is to record people who speak clearly, using high-quality

microphones with minimal background noise (broadcast quality audio).

When broadcast quality audio is preprocessed with a matching acoustical model, the resulting searches will be very accurate. Lower quality audio (such as landline or wireless telephony), recorded with inexpensive microphones, moderate background noise, etc., can still be preprocessed and provide usable results. This is only possible if the acoustic model accurately describes the characteristics of the original audio file. The PSE applies the most appropriate acoustic model to the media it is preprocessing.

Preprocessing and the Search Track

During the preprocessing phase, the PSE creates a search track, an auxiliary dataset that is subsequently searched when a retrieval request is made. The search track is a highly compressed representation of the *phonetic* content of the original digital speech. Unlike speech-to-text, no hard decisions about word bindings are made by the PSE during preprocessing.

The search track requires about 6.4 kilobytes of auxiliary storage for each second of recorded audio (around 22.5 megabytes per hour). Using commercially available secondary storage, this can amount to less than \$.10 per recorded hour of speech.

Preprocessing is easily accomplished in real time or better on standard PC hardware. See *Measuring the PSE* below for details.

Query Term Processing and Searching

To process queries, the PSE accepts text input from the user and applies an internal dictionary and a letter-to-sound algorithm to convert that text into a set of phonemes that are used in the search. Single words as well as word phrases may be specified. A Boolean grammar contains a temporal operator to specify time intervals between terms (for example, “brain cancer” spoken within 60 seconds of “cellular telephone”). In cases where spelling is uncertain (or mispronunciation is intended), phonetic renderings may be used. Future plans for the PSE include the ability to use digital audio (via desktop microphones, landline or wireless

telephones, or hand-held appliances) to specify the search terms.

Once query terms are specified and converted to phonemes, actual searching commences. Because search tracks are significantly smaller than the original digital speech, and because they are organized for quick retrieval, they can be scanned for search terms by conventional PC hardware at approximately 36,000 times real time (that is, the search track for 10 hours of speech can be scanned in one second).

Results are returned as 3-tuples:

- *Search_Track*
- *Time_Offset*
- *Confidence_Level*

sorted by decreasing *Confidence_Level*. Each result should be interpreted to mean that the PSE has the indicated *Confidence_Level* (between 0.0 and 1.0) that the specified search term was spoken in the media segment corresponding to the indicated *Search_Track* at the indicated *Time_Offset* (accurate to 0.01 seconds).

Because the PSE uses a phonetic model for searching and returns results based on the confidence level of the match, it does not have to make hard decisions about whether a piece of speech fits the search criteria. It simply reports its confidence in each result and sorts the complete list so that candidates with the highest level of confidence are at the top of the list. The user may peruse the list as deeply as desired.

Opportunities for Parallel Implementations

The PSE is structured to take full advantage of any parallel processing accommodations. For example, when a single-processor computer is upgraded to dual processors, preprocessing speed doubles. Also, if the search data are divided into multiple search tracks, banks of computers can be used to retrieve data in parallel. The results can be merged into one list, sorted in confidence order, just as if only one search were performed.

Opportunities for Distributed Preprocessing

The PSE search track provides an interesting opportunity for broadcasters or other providers of audio/video archives. If a centralized organization preprocesses media to enable searches of their archives, the preprocessed files can be distributed to remote installations where searches can be performed without further preprocessing (which can be done once using powerful computer systems that are only available to the large central sites).

Advantages of the PSE

There are compelling reasons why using the Phonetic Search Engine is preferable to using speech-to-text searches. The PSE has a completely open vocabulary. No base lexicon is required. In contrast, the speech-to-text method must map all words into lexicon entries. If an entry is not found, the speech cannot be recognized as a match. The PSE does use a “hints” dictionary, but updating that dictionary does not require an existing search track to be recreated (because the dictionary is used only during the high-speed searching phase, not during the relatively slower preprocessing phase).

Another advantage of the PSE is that accuracy is not compromised for speed. Speech-to-text must limit its search and must make hard decisions about word bindings – else searches are too slow to be of any use. This is why speech-to-text lexicons are never large enough and seldom contain enough key search terms, which are often proper names or unusual phrases.

Speech-to-text depends heavily on linguistic evidence (probability of word sequences), which can sometimes result in bad decisions. Phonetic searching emphasizes how things sound, not what the machine may infer they mean. This is especially evident when searching for proper names. Exact spelling is not required. For example, the PSE can find references to “Sudatenland” spelled properly, or even as “Sue Dayton Land.” This is an extreme example, but the utility of this kind of searching becomes obvious when you look at a name like “Qaddafi” that has been written with many different

spellings such as Khaddafi, Quadafy, and Kaddafi. This name could be input into the PSE as *KADOFFEE* and still be found.

The PSE returns results sorted by confidence so that users may decide their own depth of search. The most likely hits are returned at the top of the list, but other candidates, possibly interspersed with occasional incorrect matches, will still be returned, albeit lower in the list.

IMPLEMENTATION OF THE NEW SEARCH METHODS

Platform

The Phonetic Search Engine has been implemented in the Microsoft® Windows NT™/Windows 2000™ environment but is easily ported to other operating systems, such as Linux™, etc. Currently, the engine supports only North American English. However, the technology will work just as well with any other phonetic (as opposed to tonal) language. The search engine is capable of dealing with Unicode file systems, so file names using character sets other than standard ASCII can be recognized.

Core Technology

The kernel of the phonetic search engine is implemented as a library for direct linking from Microsoft Visual C++™. It is multi-threaded and multi-processor enabled, and can be freely integrated into multi-threaded applications on supported platforms. The technology also includes an interface layer compliant with Microsoft Component Object Model™ (COM). This supports integration into any languages that can access COM objects (Visual Basic™, Active Server Pages™, Java, and anything running under the new Microsoft.NET™ platform, etc.).

Plug-in Approach

The PSE will be provided as a plug-in for leading audio/video asset-management platforms. These vendors provide frameworks so that third-party developers can provide technology that offer additional features and benefits.

Sample Embodiments

The PSE can be used in several ways. It can be incorporated into a browser-based application for use via the World-Wide Web. This is especially applicable to users located far from the site where media are stored, who need to extract content remotely. Using this method, performance is obviously dependent on the amount of available bandwidth. However, remote entry of search terms and subsequent review of streaming audio clips are entirely feasible even in bandwidth-limited environments.

In a fast-paced production environment, a producer needs access to sound or video clips as quickly as possible. When the media are accessible through a LAN or WAN, a local application can be used to quickly retrieve and review clips in an extremely efficient way. Using optimized interface techniques, a producer can rapidly identify whether a selected clip is a match.

Measuring the PSE

Key characteristics of the Phonetic Search Engine include preprocessing speed, search speed, and search accuracy. These metrics, for the current implementation, are presented below.

Table 1 summarizes preprocessing speed for several of the latest processors on the market. Note that even inexpensive CPUs can preprocess faster than real time while state-of-the-art CPUs can preprocess *two* media streams in real time.

Preprocessing Speed (times real time)	CPU Clock Rate (Mhz)
1.56	Intel Pentium-III™ 800
1.99	Intel Pentium-III™ 1000
2.23	AMD Athlon™ 1000

Table 1 – Preprocessing Speed

Figure 2 summarizes search speed for both RAM-based and disk-based search tracks. It may be appropriate to dedicate RAM to certain tracks in scenarios where extremely high

performance is required, perhaps because a given media archive is subject to large numbers of simultaneous searches. In any case, the current implementation treats system RAM as a cache for most recently used search tracks to optimize their search speed.

Note that shorter queries run somewhat faster. To help put query length into context, *Table 2* lists phoneme counts for some sample query terms.

Phoneme Count	Sample Query Terms
3	mob, rear, bought, loose, cake, might, jet
6	crowded, withdraw, precious, ownership
9	save you money, February, someone says
12	the new standards, took away our rights, Washington today
15	maximum strength, special broadcasts, astounding profits

Table 2 – Sample Query Terms

Search accuracy was measured under various conditions to contrast current performance by the Phonetic Search Engine with text search of output from a popular speech-to-text system embedded within a leading video asset management platform.

Media used for this benchmark include 25 segments of television news programming listed in *Table 3* (over 15 hours of core material).

Segments	Program
8	PBS Newshour
9	ABC World News
8	ABC Nightline

Table 3 – Media in Accuracy Benchmark

All media segments were recorded by consumer VCR, professionally transcribed and truthed, logged and encoded by the asset management platform, and preprocessed by the PSE. Approximately 1900 words and short phrases (uniformly distributed between 2 and 20

phonemes each) were selected randomly from the transcripts along with over 175 proper names.

These query terms were submitted both to the PSE and to a simple search engine scanning the output of the speech-to-text system. Each result was scored automatically using time codes inserted into the transcripts approximately every five seconds.

Figure 3 illustrates average accuracy on the complete set of words and short phrases. The speech-to-text results comprise a single point on the graph whereas the PSE results form a curve that continually rises as more and more correct results are encountered, occasionally interspersed with incorrect results (“false alarms”), along the list of results that was sorted by decreasing confidence.

Note that the PSE successfully located over 80% of the search terms at 10 false alarms per hour (nominal basis for evaluating word spotting systems) whereas the speech-to-text system recognized fewer than half as many – with no hope of recovering the remaining instances.

Figure 4 provides more details on the same benchmark of words and phrases by averaging ranges of lengths of query terms. PSE accuracy increases dramatically with query length, almost immediately reaching (one false alarm per hour) over 90% accuracy for phrases such as “Supreme Court justice”, “gastrointestinal”, and “Social Security reform” whereas speech-to-text locates only one-quarter of all occurrences.

Figure 5 illustrates respective performance on proper names. There is simply no contest here, partly because names are fairly long strings of phonemes (hence easily located by the PSE), and partly because the closed vocabularies of typical speech-to-text systems often omit names and other specialized terminology (even though they are likely search targets in realistic scenarios).

CURRENT APPLICATIONS OF THE TECHNOLOGY

The ability to search audio by words or phrases enormously enhances the usability of otherwise opaque media. The basic preprocessing and

search technologies serve as the cornerstones of a wide variety of management tools, each serving a diverse set of needs. The applications fall into two distinct categories: audio/video and telephony. Although the latter is certainly a special case of the former, its variability and inconsistent quality mandate a different set of tools.

Audio/Video Applications

Video is usually accompanied by audio. In an environment where the content of the speech is the important component, using the video track alone for editing purposes (even in ultra fast-forward mode) is of marginal utility. Although speech can be accelerated up to three times actual speed and still retain intelligibility, content that lasts over an hour becomes difficult to manage.

Since the Phonetic Search Engine can be used to pinpoint the location of target words or phrases within ten milliseconds of each of their occurrences, a list sorted either by time or likelihood could be used for quick access to desired locations. The result is a valuable editing tool that provides fast retrieval of archived materials that otherwise may be totally inaccessible.

A related application of the PSE is content management. Ideally, audio/video content (regardless of its delivery channel) should be accompanied by extensive metadata. Accordingly, an active area of research is automatic generation of metadata so that the need for human labor is minimized. However, it is difficult to extract such tags without transcriptions. Fortunately, Phonetic Search technology can operate in environments where only limited amounts of metadata exist (for example, one-line labels identifying a piece of content). In many cases, the content can be managed without meta-tags at all, using only the search track to locate the key words and phrases that uniquely identify the video or audio passage.

Individual consumers may also need to access audio or video databases. This includes searching historical archives, looking for topics in talk shows or news broadcasts or commercials, or searching movies for key lines, and other such

uses. Because of the potential for many individuals to search a single database at one time, the PSE architecture was designed so that a given search track can be accessed by many simultaneous users at peak times.

Public relations, opinion gathering, advertising, political consulting, and media analysis firms can use PSE technology to determine the level at which various products and people are exposed within specified media. Any company that has access to preprocessed data can deliver speedy and precise reports to clients telling them exactly the frequencies and contexts of such exposures. For example, suppose that a firm was contracted to monitor ten major television networks broadcasting continuously, 24 hours a day for three months, for references to a specific product. In that time frame, the networks would produce roughly 20,000 hours of content. Manually searching that much content for references to a product would be unimaginable. Using the PSE, the firm could process such queries in roughly half an hour.

Market researchers often record focus groups in order to assess consumer reactions to various issues, products, and ideas. Real-time note taking could be supplemented by reviewing these recorded sessions. Today, such reviews are rarely performed because of the cumbersome procedure necessary to find relevant information. Using the PSE would greatly raise the value of these recordings by enabling the researchers to find specific references quickly.

Telephony Opportunities

Public network telephony adds new challenges as well as many potential applications for Phonetic Search technology. Unpredictable effects (such as bad channels, cellular compression distortions, dropped packets, noise, competing talkers, non-linearities from inexpensive electronics, and phase distortions) will always render accuracy below that obtainable with carefully controlled studio environments. PSE accuracy, human intelligibility, and perceived quality all suffer by allowing unrestricted telephony input. This lack of clarity has made searching telephonic speech difficult to accomplish, yet promising results from preliminary experiments suggest that

existing PSE algorithms are directly applicable to this problem. Current applications for telephone channels include voice-mail management and call center support.

In the past two years, inexpensive storage makes it feasible to save voice mail stored as digital audio as cheaply as it was, in the past, to save electronic mail stored as text. The big difference between voice mail and e-mail is accessibility. E-mail has subject lines, sender information, as well as a message body that can all be searched using text search tools. Currently, people save thousands of searchable e-mail messages for later reference. The PSE provides this capability for voice mail messages as well. If the caller leaves a name followed by a message, searching by "sender" and/or by "subject" or "body" becomes easy. Hundreds of hours of telephone messages could be searched in a matter of seconds and, since names of persons and places are searched phonetically, they can be found even if the spelling is not exact.

Large business call centers usually take messages for later callbacks. The messages might contain complaints, comments, and commendations; specific products or activities might be mentioned. Using Phonetic Search technology, a company could scan messages for key words and route them, without any human intervention, to individuals targeted for the selected type of messages. In cases where callbacks are not required, automated analysis of message characteristics could be used for a wide variety of tracking statistics.

Opportunities for Industry Standards

Phonetic Search tracks could become a standard entity, attached to any file format that includes audio, just like video can contain closed captioning. Since this technology is significantly cheaper to enable than closed captioning, it could be more universally applied.

FUTURE AUDIO SEARCH APPLICATIONS AND OFFERINGS

Search Track Compression

Ongoing research based on the core Phonetic Search technologies can be used to make performance tradeoffs. One example of this

effort is trading off search speed for reducing the size of the search track. Current experiments with search track compression allow a factor of two times more efficient storage with only a modest decrease in search speed.

Musical and Noisy Backgrounds

The current Phonetic Search Engine assumes at least a moderate level of signal-to-noise ratio. More difficult conditions approaching zero decibels render results to be of marginal utility. However, the PSE framework allows easy incorporation of both digital noise removal and robust pattern matching.

One partial solution to the high noise problem is that of specifying search strings with longer sequences of phonemes. Another alternative under investigation is analysis of the waveforms before converting to a search track to better characterize the level and descriptions of the background interference.

Real-Time Scanning

An interesting possibility for the PSE is the application of real-time scanning of large datasets. Suppose that an executive of a large corporation wants to know about all references to the company's product name in real time, anywhere in the world. A continuously running PSE could locate the occurrences and notify the executive (via e-mail, instant messaging, paging, machine telephony, or fax) each time the product is mentioned. It may even be possible to detect the tone of the reference (positive vs. negative reference).

Scanning for Lyrics and Nonverbal Material

Analyzing music with lyrics presents a difficult challenge, since the energy produced by the voice is often weaker than the rest of the signal. In many cases, however, if a long string of words is queried, the algorithm works well, even with no music-specific acoustic models. Formal evaluations of lyric searches are already underway.

The PSE could also be used for high speed searching of music for key non-speech events such as melodies, rhythms, themes, etc. This is

analogous to the way speech is searched for sounds. More details about this set of applications will be made available at a later time.

Multimodal Searching

Since the current preprocessing step can easily be performed over twice as fast as real time on a standard PC, there is plenty of excess capacity for performing other tasks in parallel on a dedicated machine. If other metadata are available, the Phonetic Search algorithm could make use of such information (multimodal searching). In each case, a strategy would be formulated for dealing with the specific type of metadata. As an example, a closed captioning signal may give an approximate verbatim rendering of dialog, but it may not be exact, and it may be many seconds off in time synchronization. Using this signal to delimit the search range will allow the Phonetic Search algorithm to pinpoint exact locations with unprecedented accuracy.

Phonetic Searching of Text

One of the inherent advantages of the Phonetic Search Engine is its reliance on sounds of speech rather than exact orthography. The established framework will also allow the searching of text based on sounds, with search queries entered by voice or by typing. This is particularly useful for text entries of proper nouns that could be misspelled by either the searcher or the text originator.

CONCLUSION

The vast archives of prerecorded digital audio cannot be rapidly and effectively mined without significantly faster search technology. The ability to search phonetically presents one way that high-speed retrievals can be accomplished. The technology is new, but it is available today. The future holds even more possibilities. The Phonetic Search Engine, built on years of research, provides a valuable tool for anyone who needs to look for information stored within the volumes of available digital sound.

REFERENCES

Rabiner, L. R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 1989.

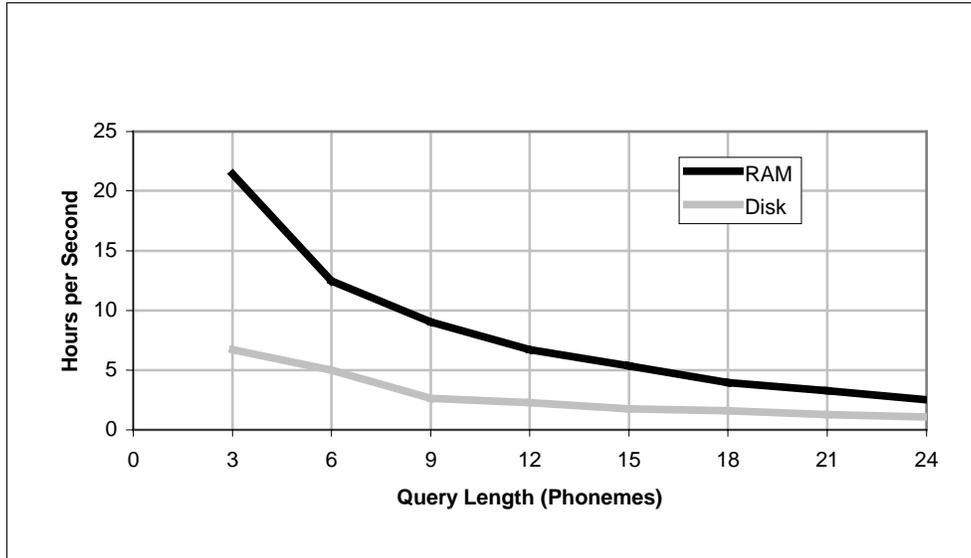


Figure 2 – Search Speed

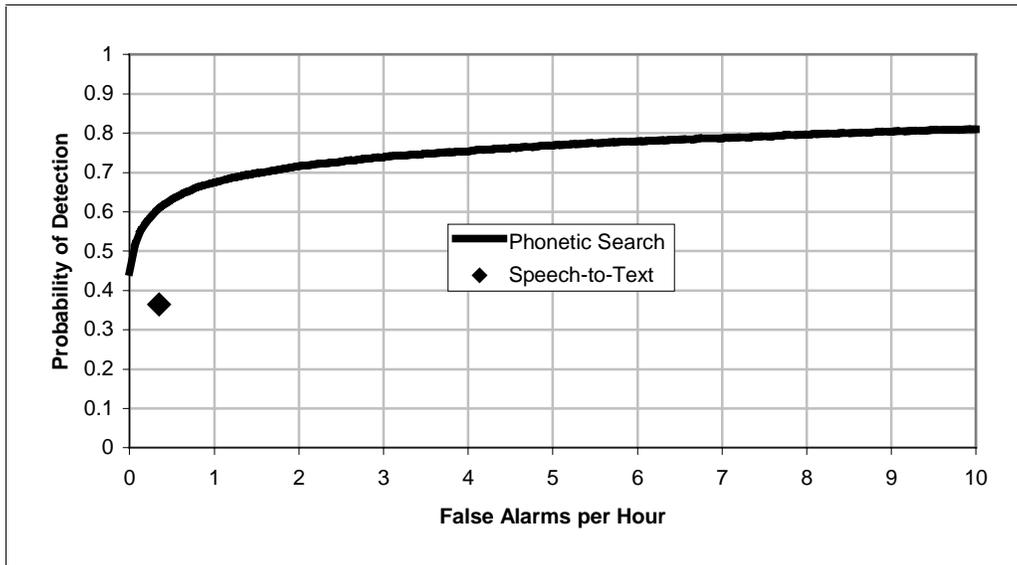


Figure 3 – Search Accuracy (Words & Phrases, Average)

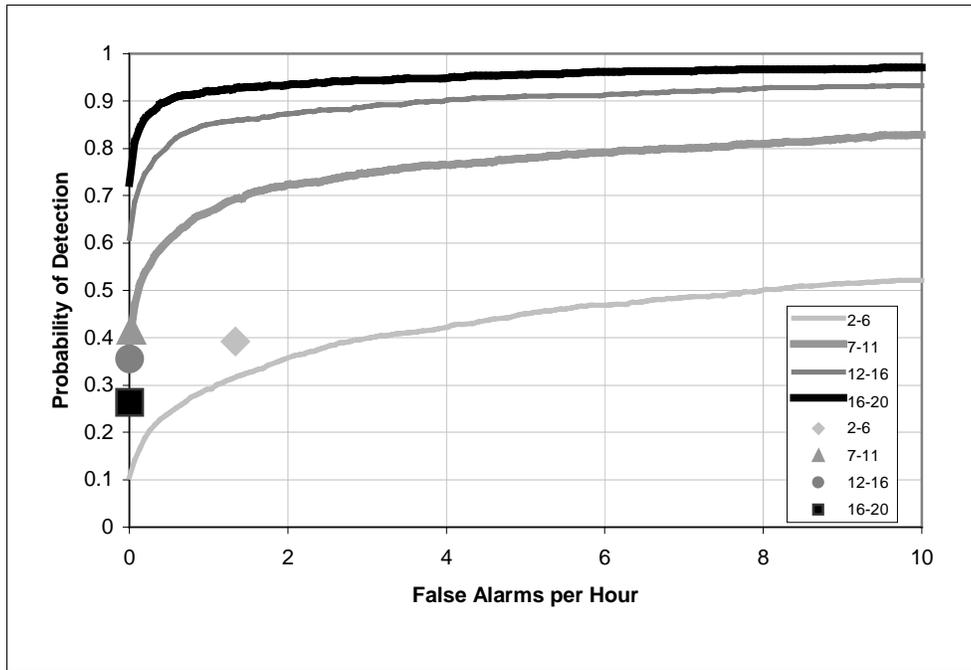


Figure 4 – Search Accuracy (Words & Phrases, by Query Length)

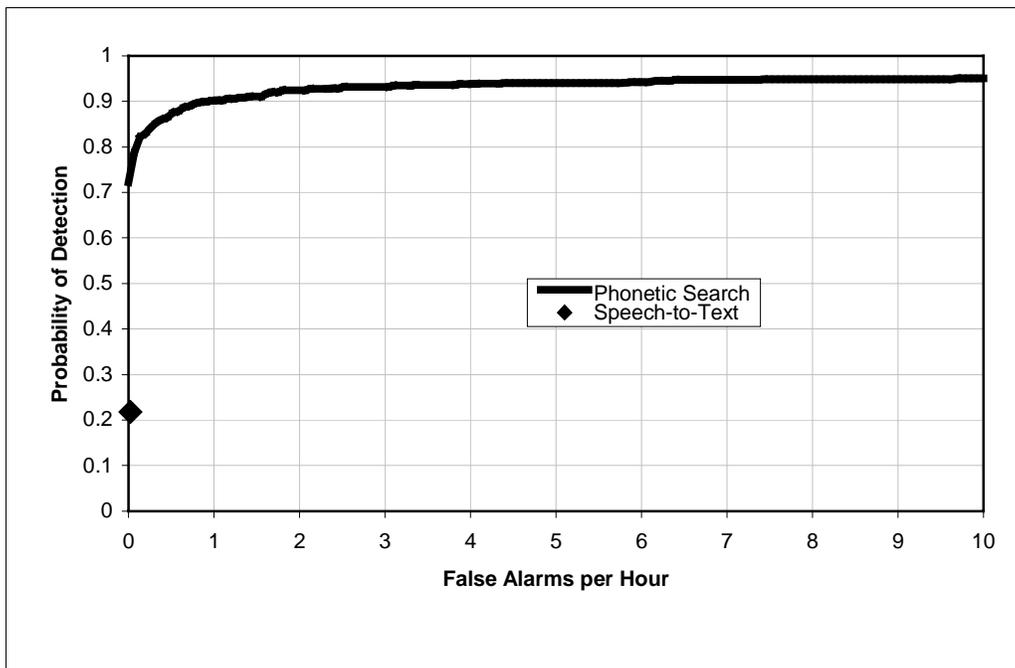


Figure 5 – Search Accuracy (Proper Names)