Audient:
An Acoustic Search Engine

Ted Leath

Supervisor: Prof. Paul Mc Kevitt

First year report

Date: August, 2005

Presented as a requirement for Ph.D. in

School of Computing and Intelligent Systems
Faculty of Engineering
University of Ulster, Magee
Email: ta.leath@ulster.ac.uk

Abstract

Most current Spoken Document Retrieval (SDR) systems involve the production of intermediate text for the purposes of storing, indexing, searching and retrieval. These systems are predominantly lexical (word-based) in nature, performing their operations on the textual representations of words. The work described in this report proposes *Audient*, an acoustic search engine that will store, index, search and retrieve the spoken audio portion of both audio and video files. The Audient architecture has an SDR system core that uses standards-based phonogrammic streams for internal data representation. This should avoid the time penalties, overheads and errors introduced through the production of intermediate text while also avoiding some of the problems inherent in current lexical systems (i.e. handling closed vocabulary, mispronunciations, unintelligible and truncated speech, longer search terms and queries dependent on spelling rather than phonetics). Audient has a wide range of potential storage, indexing, search, retrieval and monitoring applications and also provides tools for philosophical and cognitive investigation.

While the Audient architecture does not require lexical, syntactic, grammatical, semantic or pragmatic contextual information in the fundamental processes, the examination of the performance of compound contextual strategies for the abstraction and refinement of standards-based phonogrammic streams is proposed. Also proposed are the potential of mimicry in spoken document retrieval and the movement of the *man-machine boundary* for spoken audio information retrieval. It is also suggested how Audient research results may be compared and demonstrated against other existing architectures.

Current and previous SDR systems are surveyed, and overviews on Automatic Speech Recognition (ASR) and Information Retrieval (IR) are provided. A sampling of non-speech audio retrieval systems is also included for comparison and convenience. Previous sub-word based approaches are surveyed along with a brief explanation of the various sub-word units used in these approaches. The textual annotation of audio files is discussed, and a look at word based transcription, phonetic transcription, markup languages (including SMIL, SSML, VoiceXML, SALT and XHTML + Voice) and the emerging MPEG-7 standard is included.

Audient will be evaluated by established evaluation criteria. This will provide a direct comparison with the performance of many previous systems. Mimetic *Audient Parrots* are proposed as additional tools for evaluation, testing and refinement, and are to be immediately useful in demonstrating the relative accuracy and speed of differing combinations of potential system configurations.

Audient core modules are to be developed using various available speech recognition engines, The Hidden Markov Model Toolkit, CSLU Toolkit, speech corpora, various programming and scripting languages, the CMU Pronouncing Dictionary, the Festival Speech Synthesis System, Apache HTTP Server and SSML.

Keywords: artificial intelligence, information retrieval, mimicry, speech recognition, spoken document retrieval.

Acknowledgements

Firstly, I’d like to thank my supervisor Prof. Paul Mc Kevitt for his systematic, experienced and knowledgeable supervision. I have worked in higher education for many years, and have observed in the work of others the contribution that good academic supervision can make.

Prof. Norman Black has approved my continuance within the Ph.D. programme each year since 2002, and I very much appreciate his continued support. Prof. Sally McClean and Prof. Mike McTear have both provided helpful comments and suggestions.

I’d also like to thank fellow postgraduate students Minhua Eunice Ma, Dimitrios Konstantinou and Tony Solon. Eunice’s Ph.D. confirmation report contributed significantly in the formation of the outline structure of this report. Dimitrios is always enjoyable and stimulating to talk to, and I enjoy discussing new ideas (particularly philosophically based ones) with him. Both Eunice and Dimitrios have significant experience in Natural Language Processing, Computational Linguistics and Multimodal Computing – all areas in which I am a relative novice. I have appreciated their occasional advice. Tony is further along in his Ph.D. studies than I am, and is regularly drawing my attention to items of potential interest.

I have never met Corrina Ng in person, but I am grateful for her helpful and encouraging correspondence regarding SDR sub-word research and the Ph.D. research process generally. I also have appreciated the e-mail feedback, redirection and interest of Yoshi Gotoh, Phil Green, Paul Martin, Ji Ming and Steve Renals.

My professional and experienced work colleagues Paddy McDonough, Pat Kinsella and Bernard McGarry contribute to making my “day job” progress as smoothly as possible. This is a great contribution to my continuing research.

Finally and probably most importantly, I would like to acknowledge the help and support of my family. Time spent in completing this work usually meant time not spent with them, time that will never be available again. To date I do not recall one serious complaint from my wife Melanie, or any of our four children Kirstin, John, Laura and Michael. My daughter Kirstin, currently embarking on her own Ph.D., even critiqued parts of the report for me!

While I owe a debt of gratitude to all of those listed who have helped and supported me in the production of this report, I claim any errors or deficiencies as exclusively my own.

Contents

1. Introduction: the motivation for Audient

Within the broad area of information discovery and retrieval, there is an ever increasing requirement for an effective means of indexing, searching and retrieving audio information. Since most video material also contains an audio portion, any developments within the audio area also have implications for video indexing, search and retrieval. Most current audio retrieval systems process spoken content and are lexical (relating to words) in nature, involving the production of intermediate lexical text as an internal data representation and input to an Information Retrieval (IR) system. In existing systems, this intermediate text is derived either from an audio stream via Automatic Speech Recognition (ASR) (Jones et al., 1997), manually transcribed (Takeshita et al. 1997), or partially derived from associated artefacts like metadata, closed captioning (in the case of video with an audio stream) or by other textual annotations (Hauptmann and Witbrock, 1997, Mani et al. 1997, Maybury, 1997).

By employing a speech-centric model using an abstraction of the phonetic information derived from the sequential speech stream rather than lexical text for internal data representation and input to an IR system, the following indexing problems inherent in exclusively lexical architectures might be avoided:

- *Closed vocabulary prohibiting recognition of Out of Vocabulary (OOV) words (particularly proper nouns) and new words*. Current Large Vocabulary Continuous Speech Recognition (LVCSR) systems typically have a vocabulary of around 5,000 to 60,000 words (Jurafsky and Martin 2000) while the Oxford English Dictionary currently has in excess of 290,000 entries.

- *Mispronunciation, unintelligible and truncated speech within the audio stream*. A system based on smaller sub-word units rather than words may minimise the problems of mispronunciation. The same should be true with unintelligible and truncated speech.

- *Less success with longer search terms*. Larger units of evaluation and a fixed vocabulary reduce the chances of success with longer queries.

- *Queries dependent on spelling rather than phonetics*. Queries for lexically-based systems depend on correct identification of the words in question to match against the recognition output derived from the recogniser vocabulary. The Audient architecture has no lexical vocabulary, and matching is performed on sub-word units.

- *Lack of granularity for user determined search parameters*. Sub-word units are more fine-grained than lexical units for indexing, search and retrieval functions.

The architecture for Audient suggests that it may be possible to effectively index, search and retrieve audio material avoiding the time penalties and errors introduced through the ASR phase by taking a non-lexical approach.

Audient also proposes to move the man-machine boundary for spoken audio information. Early Structured Systems Analysis and Design Method (SSADM) had the concept of the determination of a man-machine boundary for each system being analysed and/or planned. This is a notional boundary distinguishing between those processes of the system that would be carried out by human beings, and those processes that were best automated (DeMarco, 1978). Some tasks are better automated for a range of benefits (speed, cost, volume, tedium, etc.), while other tasks are more easily and competently performed by human beings. Most human beings have unparalleled inherent strategies and facilities for speech recognition, including speech sources that are fragmented or partially obscured by a noisy environment.

1.1. Objectives of the research

The primary objectives of this research are to:

- Explore the efficacy of using standards-based phonogrammic streams as an internal data representation for storing, indexing, searching and retrieving spoken audio information.
- Compare the performance of optional compound strategies for the abstraction and refinement of standards-based phonogrammic streams.
- Design, implement, refine and test Audient.
- Present and demonstrate research results, comparing Audient with other existing system architectures.

1.2. Features of Audient

As illustrated in Figure 1.1, Audient will take as input raw audio speech files and produce phonetic and temporal abstractions in the form of standards-based phonogrammic streams and related temporal information. This data will be indexed and stored in a database. Users may interact with Audient using either speech or text queries. Text queries will be translated to phonogrammic streams by means of a pronunciation dictionary, table or rule-based method. Queries will be applied against the database of indexed data, and a query response output to the user.

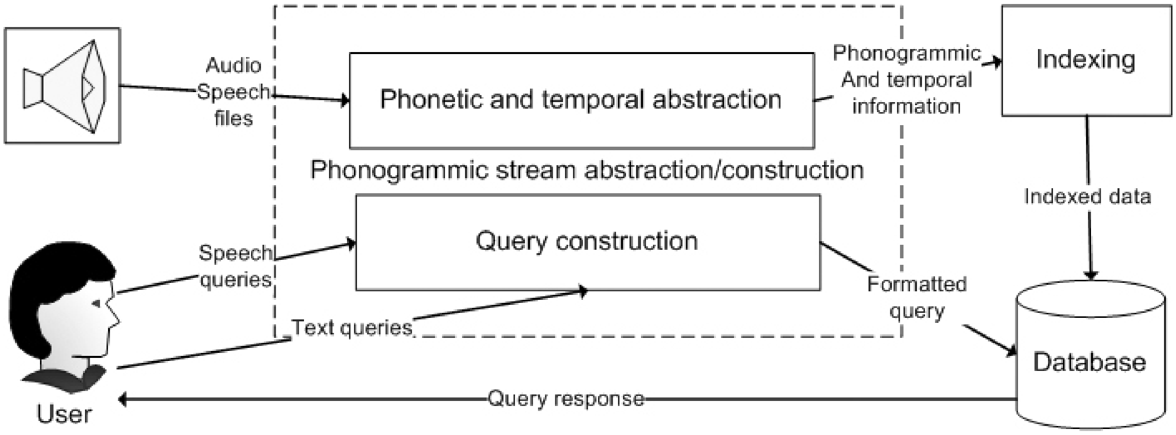


Figure 1.1 System architecture of Audient core modules

1.3. Areas of contribution

Fundamental to the core of any information retrieval system are the concepts of how information is to be represented, and the interpretation of that information's structure (Meadow, 1992). An area of contribution proposed by Audient research is the use of standards-based phonogrammic data as an internal data representation. Audient proposes to generate standards-based phonogrammic streams along with temporal information from audio speech files. A phonogrammic stream is a succession of orthographic symbols representing phonetic values. Unlike most existing systems, the Audient architecture does not necessarily consider the original concept of the speech originator (see Figure 2.2), or the words (tokens or symbols) used to convey the message. It also does not attempt to derive words and/or associated meaning, but attempts to generate a phonogrammic representation of the speech stream for internal data representation and input to an IR system.

For example, a user could speak the query “speech recognition” which would be then be constructed into a phonogrammic representation like “S P IY CH R EH K AH G N IH SH AH N” or “S P IY CH R EH K AH G N IH SH AH N” (an alternative pronunciation) using phonetic recognition. This would then be compared to the system database. Likewise, text queries would be converted to phonogrammic streams using a text translation table, pronouncing dictionary or rule-based method and compared in the same way.

The standard initially proposed for Audient phonogrammic streams is the Speech Synthesis Markup Language (SSML). SSML is designed to integrate with other markup languages, and may be regarded as a subset of the Voice Extensible Markup Language (VoiceXML), Speech Application Language Tags (SALT), XHTML+Voice (X+V) and Synchronized Multimedia Integration Language (SMIL). Some commercial Automatic Speech Recognition (ASR) and audio mining products now provide VoiceXML and SALT output, but these are lexically based.

Another contribution is the allowance of compound contextual strategies to allow for refinement of phonogrammic streams. One of the important features of the Audient architecture is that lexical refinement is unnecessary. However, the architecture allows for the evaluation of compound contextual strategies to further refine the phonogrammic stream. It is envisaged that these strategies may be used to improve the accuracy of phonogrammic streams, but it remains to be seen how significant an improvement may be made. It is proposed that the performance of compound strategies for the abstraction and refinement of standards-based phonogrammic streams be compared against a baseline system using no such strategies for stream production.

Human beings employ multiple strategies in speech understanding and perception to help cope with homophony (words or speech segments with the same pronunciation but different in meaning, derivation or spelling), OOV words and missing portions of speech due to noise or fragmentation (Greenberg, 1996). Contextual strategies employed include:

Lexical strategies – a lexical strategy involves matching the speech stream against words in a finite vocabulary.
Syntactic or grammatical strategies – a syntactic or grammatical strategy could be described as reference to a given language's syntax (the rules for constructing phrases or clauses) and grammar (that which is to be preferred and avoided in inflection and syntax) to the present point in speech stream in relation to the surrounding context.
Semantic strategies – A semantic strategy would involve an attempt to derive possible meaning for the present point in the speech stream in relation to the meaning of the surrounding context.
Pragmatic strategies – A pragmatic strategy would involve an attempt to derive what the speaker meant to communicate rather than the strict meaning of what was said. Pragmatics focuses on the relationship between words and interpretations, while Semantics focuses on the actual objects or ideas that a word refers to.

A mimetic method for adequacy evaluation, diagnostic evaluation and demonstration called an *Audient Parrot* is to be developed and employed. An Audient Parrot is a system that takes as it's input an audio speech file and applies to the file a specified speech recognition engine, optionally with specified compound strategies (lexical, syntactic, grammatical, semantic or pragmatic) to produce a phonogrammic stream and produces and audible reproduction of the original speech.

The Audient architecture also allows for multimodal queries in supporting both unconstrained text and speech queries.

2. Literature review

2.1. Information Retrieval

As the volume of information continues to grow worldwide, finding specific information has become a bit like searching for the proverbial “needle in a haystack”. The problem of information storage and retrieval has received increased attention since the 1940s and the term *information retrieval* (IR) is said to have been coined by Calvin N. Mooers in the early 1950s (Mooers, 1951). IR emerged as a distinct research community near the end of the same decade (Spärck Jones and Willett, 1997) and has subsequently encompassed a wide area of research and applications including the representation, storage, access and organisation of information items (Salton and McGill, 1983). IR shares some elements in common with the field of *data retrieval*. Where data retrieval normally looks for an exact match against a query, IR looks for items that are a partial match, and tries to identify the best match. Queries for data retrieval tend to have a restricted syntax and vocabulary, while IR queries tend to be more natural and open. While data retrieval systems return matching items against a query, IR systems endeavour to return all *relevant* items (van Rijsbergen, 1979).

Typical IR tasks involve the retrieval of relevant information items from various types of *documents* by matching a user request or query. Documents may be thought of as objects or computer files of various formats that contain thoughts and/or information usually represented by means of symbols. Early IR dealt almost exclusively with text documents and was often regarded as synonymous with *document retrieval* and *text retrieval*. In more recent years, IR has encompassed other media types containing different types of information like images, video and audio information. The terms *image retrieval*; *speech retrieval*, *video retrieval* and *multimedia retrieval* are all used to denote specific areas within IR. Audio recordings of speech can be referred to as *spoken documents*.

Figure 2.1 shows three main components of a typical IR system: input, processor and output. Most automated IR systems store abstractions or representations of original documents (and queries). During a search session, it is usually possible for a user to change their query in the light of a previous result. This is referred to as feedback. The output is often document references and positional or temporal information relating to the information within documents. Most IR research to date has focused on the tasks of *indexing* and s*earching* to systematically manipulate information to allow it to be easily and selectively retrieved according to relevance and requirement. The effectiveness of IR systems is measured primarily in terms of *precision* (the proportion of retrieved information that is relevant) and *recall* (the proportion of relevant information retrieved).

2.2 ASR and Spoken Document Retrieval

ASR attempts to mimic the human capacity for recognising speech by enabling a computer to identify spoken words and/or sub-word units.

Most current ASR systems are lexical in nature, and conceptually follow the processes of encoding and decoding introduced in Figure 2.2.

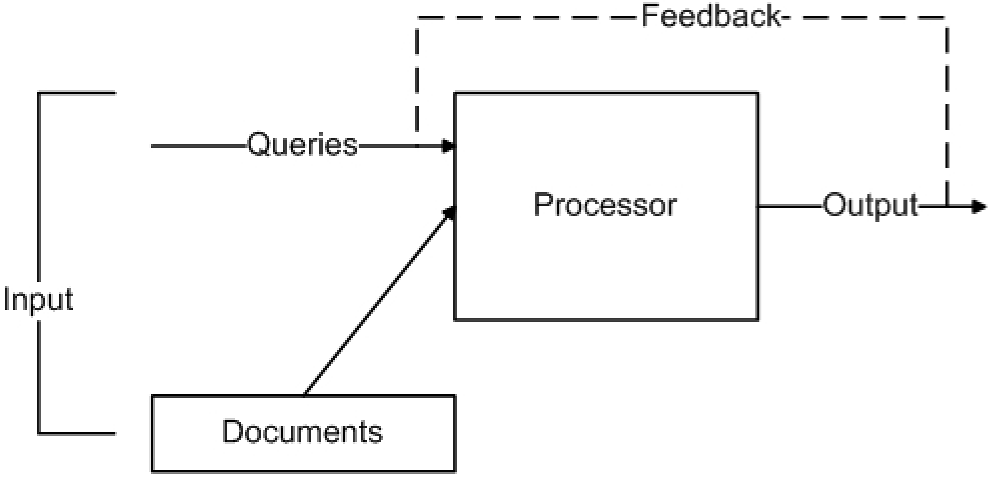


Figure 2.1 A typical IR system (van Rijsbergen, 1979)

This model begins with the speaker constructing a sequence of words (could also be referred as *symbols* or *tokens*) which are in turn generated as a speech signal from the speaker. The speech signal is converted into digital samples suitable for processing by the computer by means of *Digital Signal Processing* (DSP). Helpful tools can be employed at this stage to help improve eventual recognition accuracy. Volume level may be increased or decreased, and perhaps background noise filtered out. DSP attempts to extract the acoustic features of the speech by taking a "*window*" of samples and producing feature vectors from them. A variety of recognition processes and strategies may then be employed to attempt to reconstruct the word sequence originally spoken.

Speech modelling attempts to match the derived feature vectors to a textual unit. There can be a variety of matching strategies employed in the modelling process and speech modelling can match speech to be recognised against different levels of a language model. At the lowest level, matching usually begins with matching feature vectors against sub-word units based on acoustics. Sub-words units of recognition for ASR include various types of phonetic sequences like *phones* and *phonemes* (Ng, 2000), *Vowel-Consonant-Vowel (VCV) features* (3 concatenated sequences of consecutive vowels or consonants - Wechsler, 1998) and syllables. Phones are speech sounds considered as physical events without regard to their place in the sound system of a language. By comparison, phonemes may be regarded as representing all the speech sounds needed to distinguish one word from another in a given language. Phonemes are the smallest unit capable of inducing a minimal meaning difference between two utterances (Keller, 1994) and can vary in their physical manifestation depending on their context. These phonemic variances are called *allophones*.

Once the matching of acoustic features to sub-word units is complete, contextual information is used to improve matching at the lexical level. A recognition vocabulary defines the words that may be recognised by concatenated strings of sub-word units. Further improvement in matching may be achieved by using a *language model* or *grammar* which defines allowable word sequences.

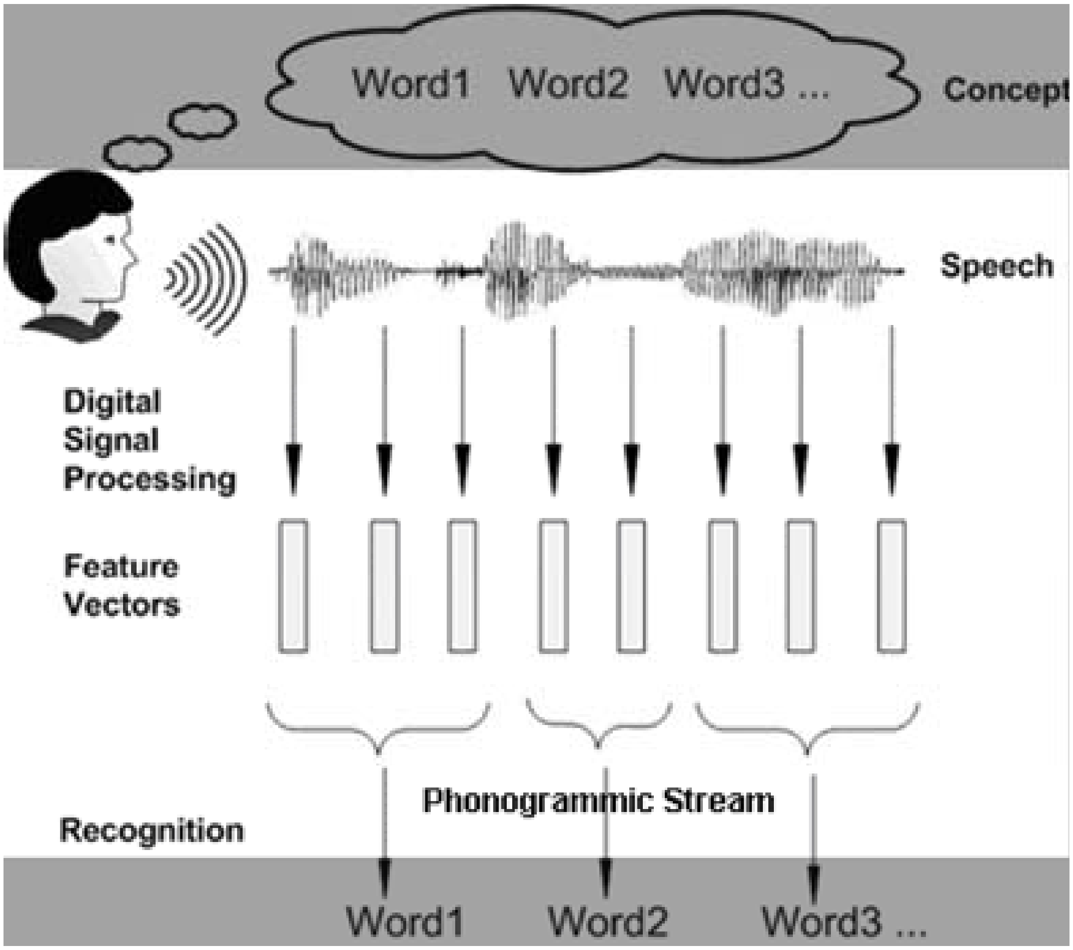


Figure 2.2 Audient decoding domain (adapted from Young et al., 2002)

The encoding and decoding processes indicated within the grey bands in Figure 2.2 are typically constituent parts of lexically based systems, but are not required in Audient.

Most ASR systems currently employ Hidden Markov Models (HMMs) to sort through immense arrays of plausible alternatives during the recognition process. HMMs are a means of representing finite sets of states, each of which is associated with a (usually multidimensional) statistical probability distribution. Figure 2.3 illustrates the linear flow of HMM state transitions. At some point in time, the sampled speech stream changes state from the state it has been in to the next state. Transitions between states are governed by a set of *transition probabilities*. In a given state an outcome or *observation* can be generated according to statistical probability, but it is only the outcome, not the state. The state is said to be “hidden” to external view, hence the name Hidden Markov

Model. In Figure 2.3, the *begin state* is designated state “1”, and marks entry to the model while state “N” represents the *end state* and exit from the model. These two states are *non-emitting states* and exist to facilitate the construction of composite models.

One advantage of using HMMs for speech recognition is that the models and methods may be cascaded to model multiple (acoustic, lexical and language) levels. On the lowest level, HMMs can model the acoustic properties of speech to phones in terms of pitch, intensity and frequency. Around 48 distinct phones are typically used for recognition. Acoustic models can be *context-independent* or *context-dependent*. Monophone models do not consider the phones surrounding the phone being examined. A biphone model may be used to model the effects of *coarticulation*. A left-context biphone model considers the current phone and the previous phone in the speech sequence, while a right-context biphone model considers the current phone and the next sequential phone. A triphone model considers the current phone and the phones on either side in sequence. The acoustic model used affects the speed and accuracy of recognition along with the training complexity and size of training samples required. On the lexical level, a dictionary can be created with word models consisting of concatenated strings of sub-word units.

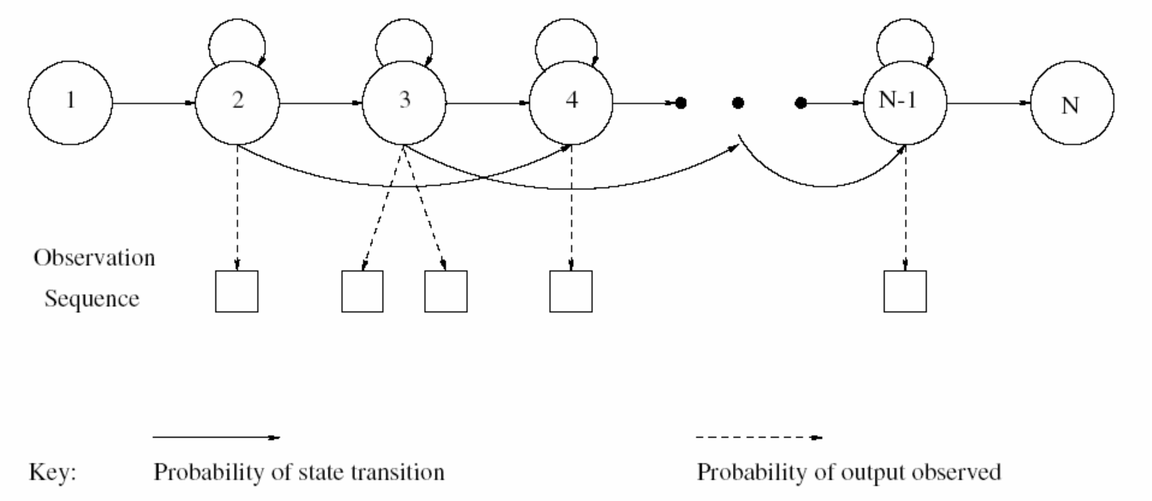
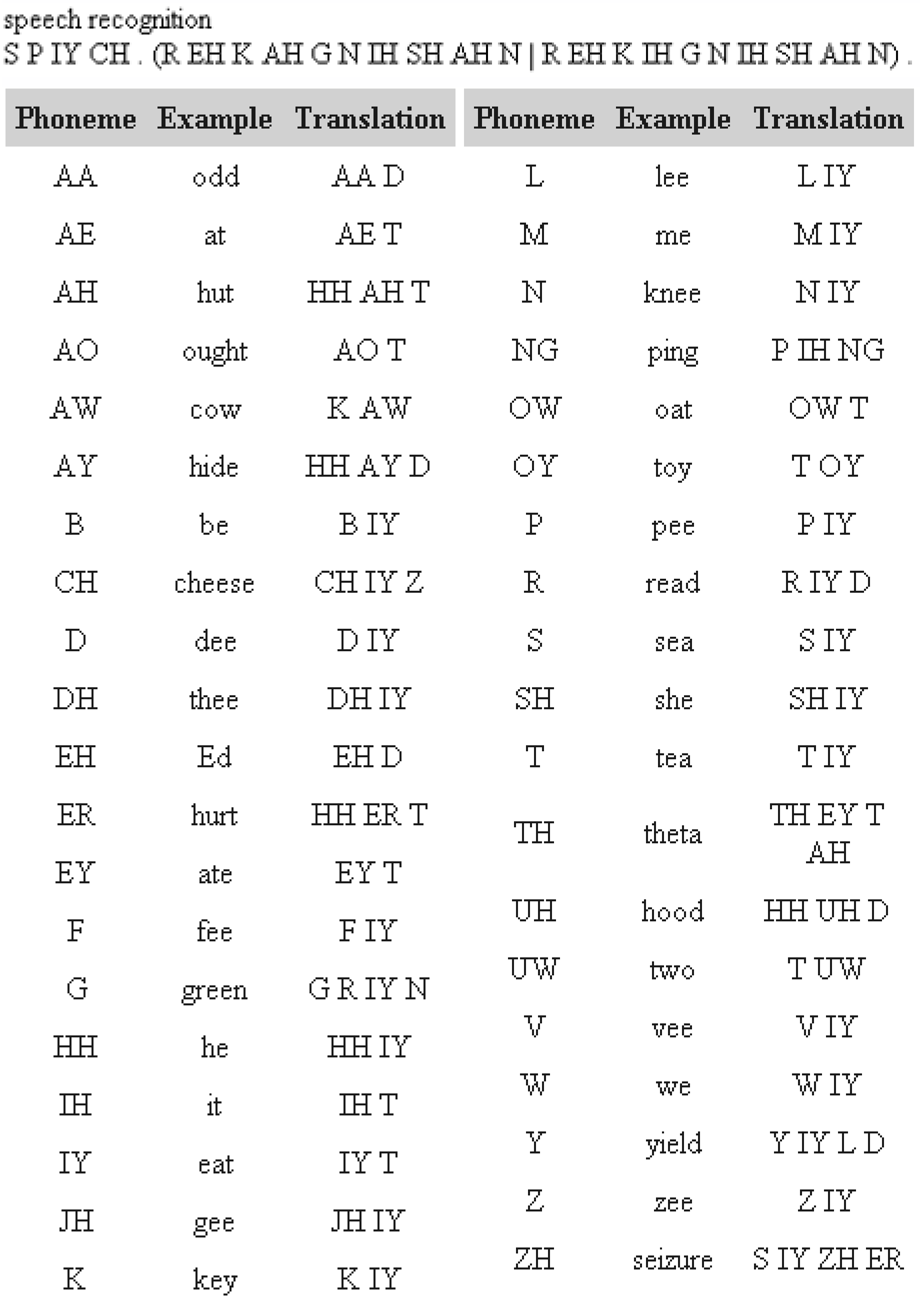

1
2
3
4
N-1
N
Observation
Sequence
Key:
Probability of state transition
Probability of output observed

Figure 2.3 An illustration of a Hidden Markov Model (HMM) (Ng, 2001)

The example in Figure 2.4 shows the North American English phoneme translation of the words “speech recognition” using the freely available Carnegie Mellon University Pronouncing Dictionary (CMUPD, 2005). Note that there are two alternative pronunciations given for the word “recognition”. Sub-word units may be used at this recognition stage as well. Instead of the recognised units being words, various sub-word units (i.e. phonemes, VCV features, syllables, etc.) may be used.

As with the acoustic and lexical levels, HMMs may also be used on the language level. N-Grams can be used in what are called *stochastic grammars*. The term “stochastic” suggests estimation of probability and/or guesswork. An N-Gram grammar can model the probability of a word based on the prior occurrence of N-1 other words. Unigrams examine only the current word while bigrams rely on the likelihood of word pairs, trigrams on word triples and so on.



speech recognition
S P IY CH . (R EH K AH G N IH SH AH N | R EH K IH G N IH SH AH N) .

Phoneme	Example	Translation	Phoneme	Example	Translation
AA	odd	AA D	L	lee	L IY
AE	at	AE T	M	me	M IY
AH	hut	HH AH T	N	knee	N IY
AO	ought	AO T	NG	ping	P IH NG
AW	cow	K AW	OW	oat	OW T
AY	hide	HH AY D	OY	toy	T OY
B	be	B IY	P	pee	P IY
CH	cheese	CH IY Z	R	read	R IY D
D	dee	D IY	S	sea	S IY
DH	thee	DH IY	SH	she	SH IY
EH	Ed	EH D	T	tea	T IY
ER	hurt	HH ER T	TH	theta	TH EY T AH
EY	ate	EY T			
F	fee	F IY	UH	hood	HH UH D
G	green	G R IY N	UW	two	T UW
HH	he	HH IY	V	vee	V IY
IH	it	IH T	W	we	W IY
IY	eat	IY T	Y	yield	Y IY L D
JH	gee	JH IY	Z	zee	Z IY
K	key	K IY	ZH	seizure	S IY ZH ER

Figure 2.4 Phoneme translation of “speech recognition”

In ASR, the size of the system vocabulary affects the complexity, processing requirements and accuracy of the system. A small recognizer vocabulary may consist of as few as ten to thirty words. A small vocabulary restricts the type of application that the recogniser can be used for, but can improve recognition accuracy. The larger the vocabulary is, the greater the possibility of confusion between similar sounding words. Large vocabulary systems typically possess vocabularies of 5,000 to 60,000 words. The number of phones in a language is much smaller than the number of words. The commonly used vocabulary of an English speaker may be 10^5 words while the number of phones is around 45 (Manjunath et al. 2002).

One convenient way of classifying most current ASR systems is by asking the following questions:

- Does the recogniser operate on continuous speech or isolated words?
- Is the recogniser speaker dependent, speaker adaptive or speaker independent?
- What is the recogniser vocabulary size?

Each of these questions/parameters affects the performance of the recogniser with regard to speed and accuracy.

ASR systems can operate on the basis of *continuous speech* or *isolated words*. Continuous speech is more difficult to process because there are often no pauses between words making word boundaries difficult to detect. The variability of speed and the coarticulation of adjacent speech elements present problems for continuous speech recognition. Isolated-word systems require a pause between each word making word boundaries easy to detect and keeping speed and pronunciation more uniform. Speaker dependent systems are designed to perform for a single speaker. These systems are ordinarily easier to develop and more accurate than speaker adaptive or speaker independent systems. Speaker adaptive systems are more difficult to develop, and are designed to adapt to the characteristics of new speakers. Speaker independent systems are the most difficult to develop, and are designed to perform for any speaker of a particular language and dialect. ASR and human speech recognition are similar in that they are both parametric in nature, and may improve in performance with parameter tuning (Becchetti and Ricotti, 1999). *Spoken Document Retrieval* (SDR) involves the search and retrieval of excerpts from spoken audio recordings using a combination of ASR and IR technologies (Garfolo et al., 2000). In most SDR systems, ASR techniques are used for the conversion of speech into text, and IR techniques are used to find the relevant documents.

2.3. Current and previous research in SDR systems

A significant amount of research has been conducted in SDR, and periodic performance evaluations of SDR systems and components have been used successfully in the USA to encourage development and share information. One of these conferences is the Text REtrieval Conference (TREC).

2.3.1. TREC

TREC began in 1992, and is jointly sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) (TREC 2004). Its purpose is to support research within the information retrieval community. TREC conferences run annually and consist of a set of "tracks" - areas of focus in which particular retrieval tasks are defined. One of these tracks which ran from TREC-6 (1997) through TREC-9 (2000) was the Spoken Document Retrieval (SDR) track.

The 1997 TREC-6 SDR track included 13 participating groups and involved a first evaluation of retrieval of broadcast news excerpts using ASR and IR techniques (Garfolo et al., 1998). The 1998 TREC-7 SDR track also had 13 groups participating. In TREC-7 as in TREC-6, the Linguistic Data Consortium (LDC) Broadcast News (BN) corpus consisting of radio and television broadcast news recordings was used for testing and evaluation. In TREC-6 it was found that known-item retrieval tasks could be successfully implemented using broadcast news. In TREC-7 it was found that ad-hoc retrieval tasks could also be successfully implemented using a larger subset of the LDC BN corpus (Garfolo et al., 1999). Ten groups participated in the 1999 TREC-8 SDR track. This time, the larger LDC TDT-2 corpus was used for testing and evaluation. The LDC TDT-2 corpus contains nearly 600 hours of broadcast news recordings containing evenly sampled broadcasts over a 6 month period. The LDC TDT-2 corpus was originally collected for the DARPA Topic Detection and Tracking (TDT) programme. It was found that that the systems evaluated were robust with the larger corpus, and that the systems evaluated had also improved significantly since TREC-7.

Figure 2.5 below represents a typical TREC SDR process. Starting with the audio corpus, transcripts are produced by the participant's speech recognition engine. The transcript is then indexed and searched by a retrieval system. The result returned for a query is a list of temporal pointers to the audio stream ordered by similarity between the content of the speech being pointed to and the query.

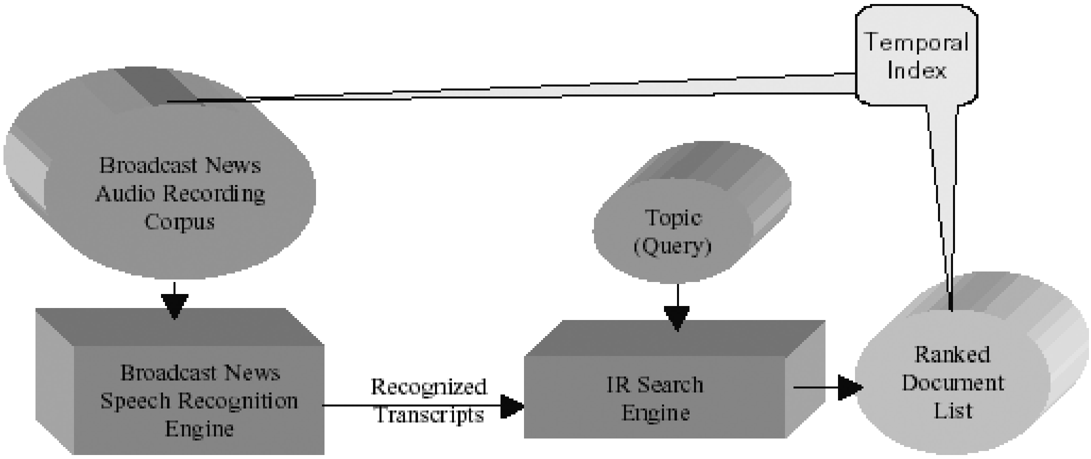


Figure 2.5 A typical TREC SDR process (Garfolo et al., 2000)

Three groups participated in the 2000 TREC-9 SDR track. The retrieval results were judged to be excellent, and retrieval results from each participant's recognizer transcripts were comparable to retrieval results from human produced reference transcripts (Voorhees and Harman, 2000).

Several notable research efforts in SDR have been participants in the TREC SDR track, including:

- The Informedia projects at Carnegie Mellon University (Hauptmann and Witbrock, 1997, Informedia, 2004)
- The Video Mail Retrieval and Multimedia Document Retrieval projects at Cambridge University (Jones et al., 1997, Video Mail, 1997, Tuerk et al., 2000)
- The SCAN system at AT&T Research (Choi et al., 1999)
- The THISL project at Sheffield University (Abberley et al., 1998, THISL, 1998)

2.3.2. CMU Informedia I, Informedia II and Sphinx Projects

The Informedia initiatives at Carnegie Mellon University (CMU) endeavor to build and implement technology for the searching, retrieval, visualization and summarization of various types of media (Hauptmann and Witbrock, 1997, Informedia, 2004). The base technology developed under Informedia I used speech and image recognition along with natural language processing to automate the transcription, segmentation and indexing of video for search and retrieval. Informedia II seeks to improve the speed and accuracy of information extraction. The Informedia projects use the CMU Sphinx-2 and Sphinx-3 Large Vocabulary Continuous Speech Recognition (LVCSR) engines to automate the transcription of narratives and dialogues.

The Sphinx Group at Carnegie Mellon University is endeavouring to release the DARPA-funded Sphinx projects widely in order to stimulate the creation of speech tools and applications, and to advance the state of the art both in speech recognition and related areas including dialogue systems and speech synthesis. The CMU Sphinx family includes LVCSR engines with associated tools. Sphinx-2 is intended to be a real-time engine and is regarded as appropriate for handheld, portable, and embedded devices, and telephone and desktop systems that require short response times. Sphinx-3 is slower but potentially more accurate and can be used for applications like broadcast news transcription (CMU Sphinx, 2005). Sphinx-4 is a Java implementation created via a joint collaboration between the Sphinx Group, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), and Hewlett Packard (HP), with contributions from the University of California at Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT) (CMU Sphinx-4, 2004).

2.3.3. Video Mail Retrieval and Multimedia Document Retrieval projects

The Video Mail Retrieval Project developed at Cambridge University in collaboration Olivetti Research Laboratory, was organised into 3 stages, culminating in a prototype demonstration system (Video Mail, 1997 and Jones et al., 1997). The first stage prototype was completed in September 1994 and successfully demonstrated message retrieval from known speakers using a set of 35 predefined keywords. The second stage, completed in 1995, extended this to allow for unknown speakers. In 1996 the final stage demonstrated open-keyword video document retrieval from arbitrary speakers, as well as a video mail browser allowing random access to video documents.

Figure 2.6 shows the architecture of the Video Mail Retrieval system. New video mail is passed to the ASR engine which creates a phone lattice for the message. To search for messages, the user inputs words that indicate the information required. A query/message correlation score is then computed between the query and each of the messages depending on the frequency of the word in the message, the number of messages in which the word appears as well as the length of the message. The user is then presented with a ranked list of the best message matches.

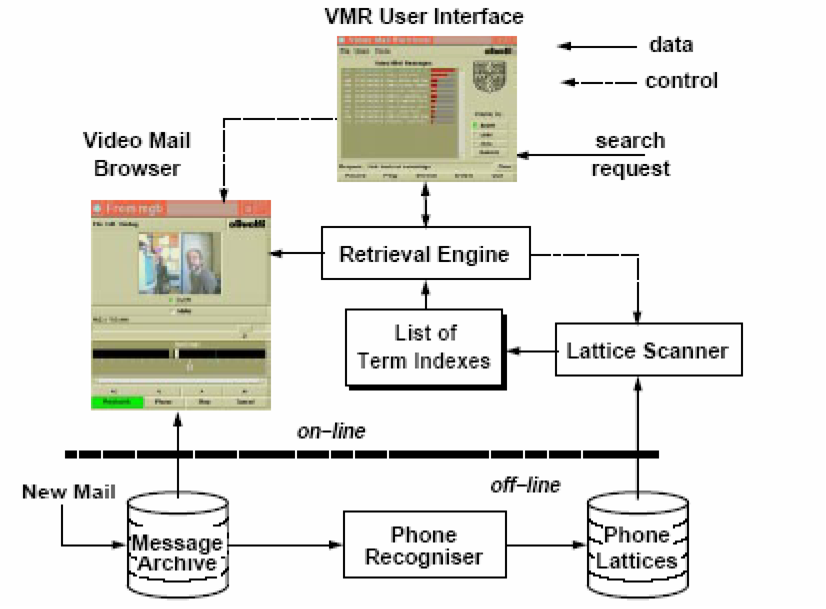


Figure 2.6 Architecture of Video Mail Retrieval system (Video Mail, 1997)

The system ultimately demonstrated that IR methods developed for searching text archives could be used to accurately retrieve audio and video data using index terms dynamically generated from phone lattices with an open vocabulary (Brown et al., 1997).

Staff from Cambridge University's Engineering Department and Computer Laboratory joined by staff from Entropic and AT&T followed on from the Video Mail Retrieval project in 1997 with the Multimedia Document Retrieval Project (MDR, 2001, Spärck Jones et al., 2001) which ran until 2000. The focus of the MDR Project was on retrieval and speech tests directly related to retrieval rather than speech recognition itself. Results from the MDR Project experiments demonstrated that retrieval from automatically transcribed files could match the retrieval performance from files transcribed manually. It was also demonstrated that query expansion during retrieval can be valuable.

2.3.4. SCAN

SCAN (Spoken Content-based Audio Navigator) is a system for retrieving and browsing speech documents from large audio corpora (Choi et al., 1998 and Choi et al., 1999). SCAN uses intonational structure to segment spoken documents into units for browsing and retrieval while using automatically produced transcriptions in parallel to increase effectiveness. Figure 2.7 illustrates the SCAN system architecture. SCAN consists of three primary components – a speaker-independent LVCSR engine which segments the input from the speech corpus and generates transcripts, an IR engine which indexes the transcriptions and determines suggested relevance in response to queries and a GUI for navigation. The first phase of the LVCSR engine is segmentation by intonational phrase boundaries. A two-pass speech recognition process is then performed on the resulting segments, ultimately producing a word lattice.

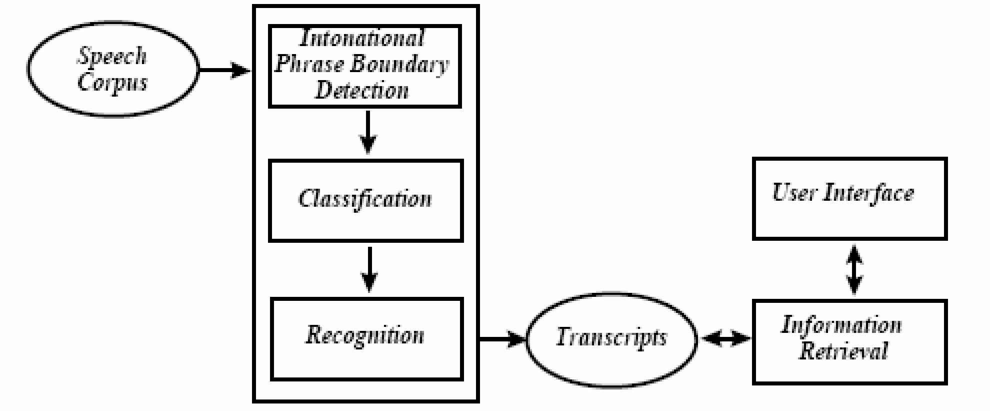


Figure 2.7 SCAN system architecture (Choi et al., 1998)

2.3.5. THISL and Abbot

THISL (THematic Indexing of Spoken Language) is a spoken document retrieval system for Broadcast News which allows multimodal queries (THISL, 1998 and Abberley et al., 1998). THISL uses the Abbot LVCSR system (Abbot, 1999) to produce approximate transcriptions of the audio documents, and then to treat the task as a text retrieval problem, relying on well-understood techniques to perform indexing and retrieval of the transcribed data. In addition to the usual keyboard/mouse interface, a spoken query interface allows users to interact verbally with the system. Abbot is an LVCSR system developed by Cambridge University, Sheffield University and SoftSound that uses hybrid artificial neural networks and Hidden Markov Models. At recognition time, the recognizer uses a vocabulary of around 65,000 words producing a single best transcription, a word graph (containing other possible hypotheses) and word and phone level confidence measures.

2.3.6. Taiscéalaí

Taiscéalaí (Smeaton et al., 1998) is a web based system that provides content based retrieval on radio news archives using streams of phones recognised from raw audio input. In 1998 the system was operational on over 4,500 minutes (nearly 80 hours) of audio news. Taiscéalaí developers used the Hidden Markov Model Toolkit (Young et al., 2002) originally developed at the Speech Vision and Robotics Group (now the Machine Intelligence Laboratory) of the Cambridge University Engineering Department to develop the recognition system. The speech recognition process produces streams of recognised phones and was based largely on work done at the Swiss Federal Institute of Technology (Schauble and Weschler, 1995). Taiscéalaí queries are entered as text. But rather than textual queries being matched lexically, the queries undergo phonemic translation before being further refined to triphones for searching against the phone streams.

Current and previous research efforts in SDR systems have brought together speech recognition and information retrieval communities and have established the feasibility of both the implementation and evaluation of retrieval from spoken audio recordings.

2.4. Public access SDR systems

Some SDR systems have been designed to be publicly accessible, primarily over the Internet.

2.4.1. SpeechBot

SpeechBot (HP SpeechBot, 2004) claims to be the first Internet search site indexing streaming spoken audio on the Web (Quinn, 2000). Figure 2.8 below represents the SpeechBot system architecture (Van Thong et al., 2001).

SpeechBot is currently applied to broadcast radio shows from public web sites such as Broadcast.com, PBS.org and InternetNews.com. Some shows have been indexed as far back as July 1996. As of the end of March, 2004 there were over 17,517 hours of shows indexed. SpeechBot can also index video and processes only the audio track for video streams.

The SpeechBot transcoders fetch and decode video and audio files from the Internet. For each item, they extract the metadata (sample rate, copyright, the story title, and possibly a short description) if available, and pass formatted audio files to the speech decoder. For speech decoding, SpeechBot uses Calista, an in-house speech recogniser derived from the Sphinx-3 system to produce automated transcripts. The librarian database manages workflow and stores metadata and other information required by the user interface including transcripts in 10 second segments. The librarian module is the main controller for the system. The indexer catalogues documents based on the transcriptions received from the speech decoder (using a modified version of the AltaVista engine). Users can interactively search an index of transcribed audio files using an ordinary web browser over the Internet.

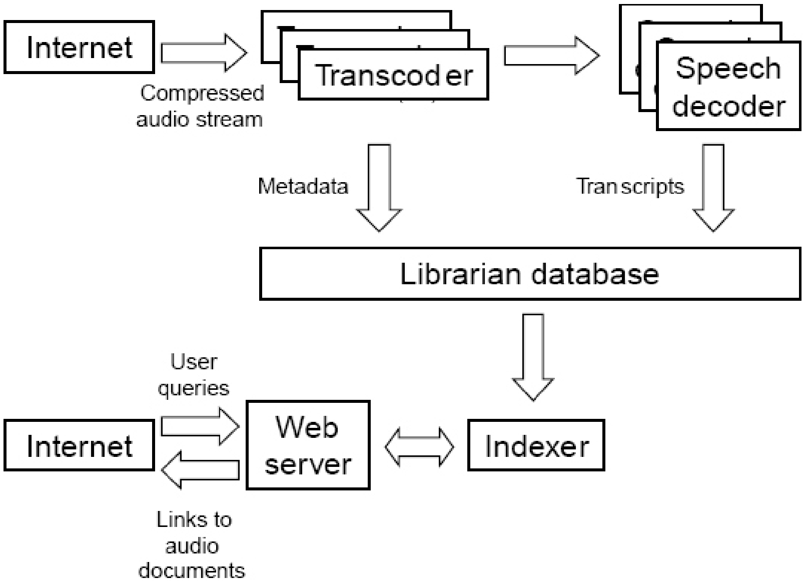


Figure 2.8 SpeechBot system architecture (Van Thong et al., 2001)

2.4.2. National Public Radio (NPR) Online

NPR Online in the USA offers a hybrid choice of streaming special events and breaking news, on-demand streaming of archived programs and archive retrieval (NPR, 2000). NPR Online has manual transcripts for radio shows stretching as far back as 1990 and the NPR Online archive search (NPR Archives, 2004) allows retrieval of both textual and audio information through an index derived from these transcripts. The preferred method of recording is a monaural stream via a direct feed from radio studios. NPR Online have developed software to automate recording and encoding. Newscasts, promotion and underwriting credits are removed for most of the radio shows to be made available online.

NPR is also a participant in the Public Broadcasting Metadata Dictionary Project (PBMDP, 2005), participating in the definition of PBCore (Public Broadcasting Metadata Dictionary). It is envisaged that within public broadcasting, the application of a shared metadata dictionary will facilitate the exchange and delivery of content and data (including both program elements and completed programs) throughout multiplatform production teams, interconnected licensees and broadcast and Internet constituents. It is regarded as an important step as public broadcast networks and individual stations begin to acquire and use asset management systems to organize their content. PBCore was deployed in a number of test implementations in May 2004. As of July 2004 in response to consistent feedback to make metadata standards easy to use, the number of metadata elements was reduced to 48 from the original set of 58 developed by the Metadata Dictionary Team. Also, efforts are ongoing to provide more focused metadata examples that are specific to TV and radio.

2.4.3. SpeechFind and The National Gallery of the Spoken Word

The US National Science foundation funded a project beginning in September 1999 called *The National Gallery of the Spoken Word* (NGSW, 2005). Part of the NGSW has involved the development of SpeechFind (Hansen et al., 2004, SpeechFind, 2003), an experimental audio index and search engine. Figure 2.9 below represents the SpeechFind system architecture (Hansen et al., 2004, Zhou and Hansen, 2002).

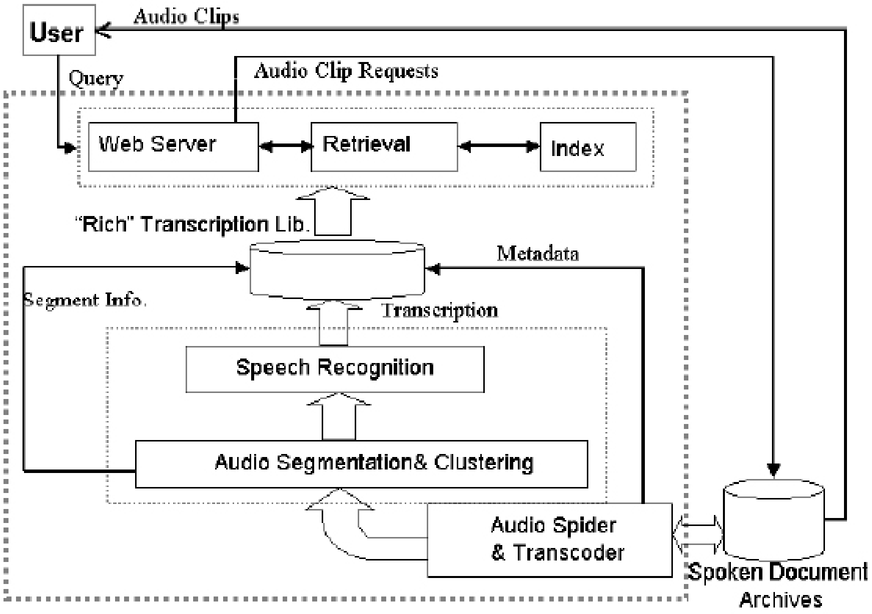


Figure 2.9 SpeechFind system architecture (Hansen et al., 2004)

The objective of NGSW is to make historically significant voice recordings freely available and easily accessible via the Internet. The University of Colorado at Boulder is the key collaborator in the engineering of the NGSW project data storage and retrieval (Hansen et al. 2001).

Speechfind includes an audio spider and transcoder, spoken documents transcriber, "rich" transcription database, and an on-line publicly accessible search engine. The audio spider and transcoder automatically fetch available audio archives from a range of available servers and transcode the incoming audio files into uniform format. For documents with metadata labels, this module also extracts relevant information into a "rich" transcript database for guiding the future information retrieval. The spoken document transcriber includes two components, the audio segmenter (audio segmentation & clustering) and the transcriber (speech recognition). The audio segmenter partitions audio data into smaller segments by detecting speaker, channel and environmental change points. The transcriber then automatically decodes these segments into text. An on-line search engine is responsible for IR tasks and includes a web-based user interface on the client-side and search and index engines on the server-side. The audio spider, transcoder and indexer run periodically and are activated in an event-driven manner (i.e., indexing the current database when new transcripts or metadata are available). The local system does not store audio archives, due to both copyright and disk space issues. Instead, SpeechFind only fetches related audio clips on request.

Despite the progress in SDR research, publicly available SDR systems are still few in number, and operate on limited domains particularly in comparison to publicly available textually-based search engines.

2.5. Commercial ASR and audio mining products

Several companies have released commercial audio mining software, and industry observers expect the number of products to increase during the next few years. Currently, accuracy levels are relatively low and some products expensive with high-end software packages costing in excess of $100,000 US dollars for full scale deployment (Leavitt, 2002).

2.5.1. BBN Rough ‘n’ Ready

The BBN *Rough'n'Ready* system (Rough'n'Ready, 2004, Kubala et al., 1999) produces rough transcriptions of audio files using large-vocabulary speech recognition, topic spotting and relationship extraction. Rough ‘n’ Ready is meant to be used as a meeting recorder and browser that will automatically produce a “rough” transcription of what was said, along with a content-based structural outline of the audio recording that is “ready” for browsing.

Rough ‘n’ Ready transcriptions include the following features:

- Segmented continuous audio input into stories, passages, or sections based on topic. Topic classification is automated using HMMs trained on as many as 10,000 different topics. A sorted list of the most likely topics is produced for each section.

- Speaker demarcation. This consists of 2 operations – speaker change detection, to locate the boundaries between speakers – and speaker identification, to specify the set of utterances belonging to each speaker. Where speaker identities are not known, a unique label is assigned to each distinct but as yet unknown speaker. Later, speaker segments can be labelled with their true identity if a system user labels one exemplar segment manually.

- Text transcript. The text transcript of the spoken content is produced by the BBN Byblos LVCSR engine (BBN Byblos, 2005).

- Information denoting the speaker’s designation within the organisation. Where a known speaker is identified, and their place within the organisation is known, information about the speaker’s designation within the organisation may be added to the transcript information.

- Indexing by speaker, topic, or concept. Once the various transcription components are determined, they may be indexed and retrieved using text-based approaches to information extraction and retrieval.

2.5.2. Nexidia Fast-Talk and Convera RetrievalWare

Most commercial ASR and audio mining products (as in current research systems) use intermediate text for the purposes of indexing, searching and retrieval. An exception is the Fast-Talk system from Nexidia Inc. (Clements et al., 2001a) which is referred to as a *phonetic search engine*. It does not employ intermediate text, but rather uses an approach called *high-speed phonetic searching* (Clements et al., 2001b). In Fast-Talk a *search track* is created in the pre-processing phase. This is comprised of a highly compressed, proprietary representation of the phonetic content of the original digitised speech. Convera Corporation has partnered with Nexidia to add phonetic searching capability to their RetrievalWare search and categorization platform (Convera, 2004).

2.5.3. ScanSoft

Scansoft Inc. has developed and acquired a range of speech technologies for home, business, enterprise and development use. In 2003, ScanSoft acquired the Speech Processing Telephony and Voice Control business units and related intellectual property from Royal Philips Electronics. Since then, development of the Philips SpeechPearl ASR engine has been overseen by ScanSoft (Philips Speech Processing, 2005). Appendix A provides a more detailed listing of ScanSoft speech technologies by product.

2.5.4. Virage AudioLogger

Virage AudioLogger (Virage, 2004) appeared in 1999 as a PC-based application that automatically converted audio content of a video into searchable text in real time. AudioLogger combined three unique audio processing engines to automatically generate keyword, speaker identification and audio classification indices from a raw audio signal. In AudioLogger, keyword indices are produced using IBM's ViaVoice technology for Broadcast Speech Transcription. This speech recognition engine handles continuous speech in real time and is speaker independent, eliminating the need for it to be pre-trained for individual speakers. The engine also incorporates special filtering to eliminate background noise and other signal contamination. Speaker voices are identified from a user-defined library of many speakers per session, regardless of the words or the language spoken. By simply providing a short speech sample, users can easily add new speakers to the library. Multiple libraries can be created to support different content types and sources. AudioLogger also generates an audio classification index that allows users to locate specific audio cues. For example, a segment might be classified as speech, music, ambient noise or silence. Users may program their own classifications and there is speaker change detection. AudioLogger combines these various audio indices with the video indices created by Virage's VideoLogger.

2.5.5. Nuance

Nuance Communications Inc. produces commercial solutions for speech related technologies including speech recognition, speaker recognition and speech synthesis (Nuance, 2005). Nuance speech recognition features a distributed client-server architecture enabling separation of light client processing from CPU-intensive server processing. Alternatively, for small configuration or for prototyping, the client and server side applications can run in a single-tier configuration. Primarily developed for telephony-based applications, Nuance speech recognition software accepts speaker-independent, continuous speech and supports very large vocabularies. Included is a "template matching" natural language capability for identifying the meaning of speech. A toolkit is available for use in developing a wide variety of speech recognition applications.

2.5.6. AT&T SCANMail

SCANMail (Hirschberg et al., 2001 and SCANMail, 2003) follows the general paradigm for audio search systems like the Video Mail Retrieval project described earlier, and those systems involved in the TREC SDR track. SCANMail is a set of ASR, IR and Information Extraction (IE) processes that use ASR to automatically transcribe voice mail messages, IR to index messages in a user's mailbox for future search, and IE to extract information from each message.

As illustrated in Figure 2.10, voice messages are first received from the *Audix* commercial voicemail system via a POP3 mail server that polls the voicemail server.

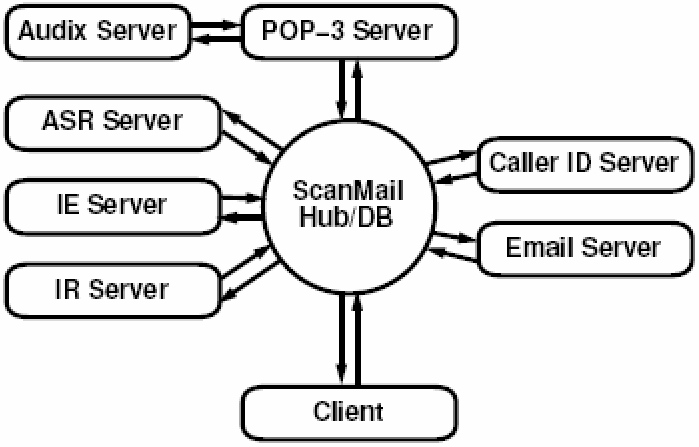


Figure 2.10 SCANMail system architecture (Hirschberg et al., 2001)

These voice messages are processed by the ASR server, which provides a written transcript. This transcript and the original audio are passed to the IR server, which indexes all of the messages in a user's mailbox to support content-based search. The transcriptions are then presented to users in a graphical user interface (GUI) similar to an email client, allowing users to browse and search their voice mail messages by content. Users can then play and/or read the portions of interest. The Email server sends the original voice message along with it’s transcription to the user along with the sender’s Caller ID information.

2.5.7. Microsoft Speech Server (MSS)

MSS is a platform for supported integrated speech services including telephony (voice-only) and multimodal (voice/visual) applications (MSS, 2005). MSS combines Web technologies, speech-processing services, and telephony capabilities within a single system performing speech recognition and speech synthesis for applications that can be accessed by telephone, cell phone, Pocket PC, Tablet PC and other devices. MSS includes the Microsoft Speech Recognition Engine but also supports third-party options like the ScanSoft/SpeechWorks OpenSpeech Recognizer.

2.6. Sub-word based approaches to SDR

Various sub-word approaches have been previously proposed for both text and spoken document retrieval (Wechsler, 1998, Ng, 2000, Larson, 2001, Ng, 2001). As mentioned previously, a range of sub-word phonetic sequences derived from phonetic transcriptions have been used in related work.

The most basic units used have been *phone n-gram* (James and Young, 1994, Warnke et al., 1997, Ng and Zobel, 1998, Wechsler, 1998, Ng, 2000). Phone n-grams are phone sequences generated by post-processing the output from a phonetic speech recognizer. In addition to individual phones and phone sequences, phones may be grouped into *broad phonetic classes* (Ng and Zue, 1997, Ng, 2000) categorised by the characteristics of the sounds involved like acoustic similarity, place of articulation (bilabial, labiodental, interdental, etc.), manner of articulation (stop, affricate, fricative, nasal) or a combination of characteristics. Figure 2.11 shows a hierarchical clustering of phones according to class with segmentation boundaries at differing discrimination levels.

Another sub-word approach has been to use *phone multigrams*. Phone multigrams are phone sequences of a variable length. While the use of phone n-grams has involved overlapping phonetic sequences, the use of phone multigrams has involved non-overlapping, variable length phonetic sequences where length has been determined algorithmically. *Syllable* sub-word units may be derived from linguistic rules applied to phonetic sequences without regard to word boundaries (Larson and Eickeler, 2003) and syllable-like units (VCV features) have also been used (Glavitsch and Schäuble, 1992, Wechsler, 1998). Examples of phone n-grams, broad phonetic classes, phone multigrams and syllables used as retrieval indexing terms are given below in Table 2.1.

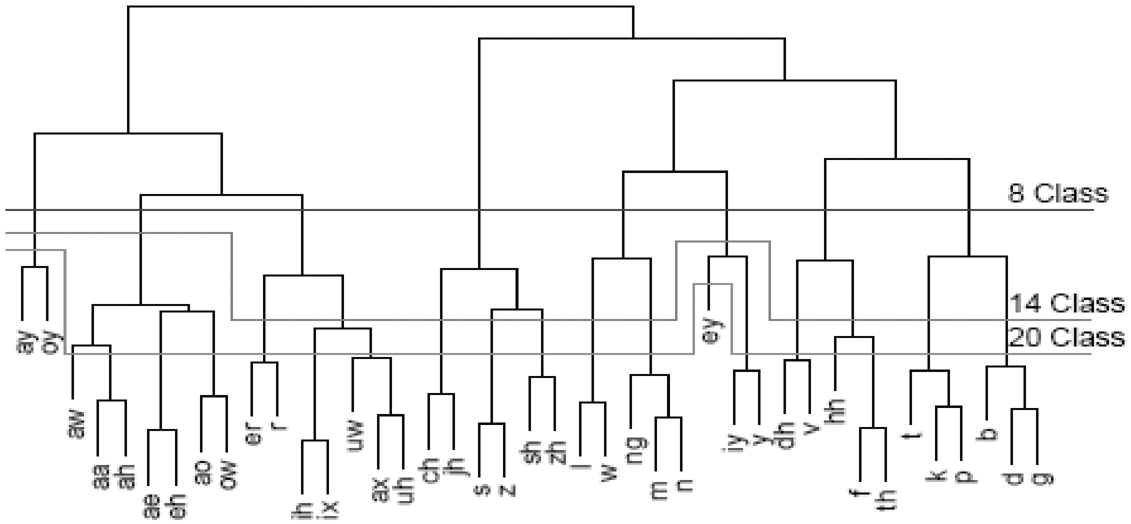
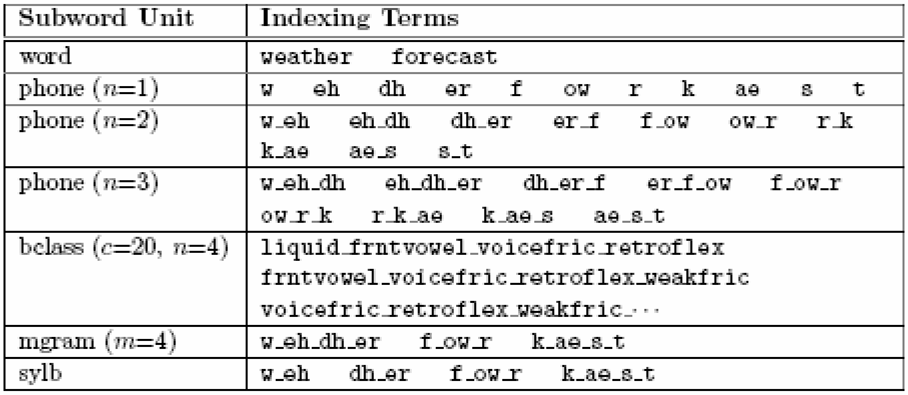


Figure 2.11 Segmentation of broad phonetic classes (Ng and Zue, 1997)



Subword Unit	Indexing Terms
word	weather forecast
phone (n=1)	w eh dh er f ow r k ae s t
phone (n=2)	w_eh eh_dh dh_er er_f f_ow ow_r r_k k_ae ae_s s_t
phone (n=3)	w_eh_dh eh_dh_er dh_er_f er_f_ow f_ow_r ow_r_k r_k_ae k_ae_s ae_s_t
bclass (c=20, n=4)	liquid_frntvowel_voicefric_retroflex frntvowel_voicefric_retroflex_weakfric voicefric_retroflex_weakfric_···
mgram (m=4)	w_eh_dh_er f_ow_r k_ae_s_t
sylb	w_eh dh_er f_ow_r k_ae_s_t

Table 2.1 Examples of indexing terms for various sub-word units (Ng, 2000)

The rules for syllable segmentation are based generally on the pattern of vowels and consonants. *Phonotactics* is the term given to restrictions on a given language which define the admissible syllable structure, consonant clusters and vowel sequences. Syllable-like sub-word units derived from delimited sequences of vowels and consonants have also been used for SDR. Various combinations of sub-word units have been used as query terms against phonetic transcriptions or phone lattice representations of the speech messages (Moreau et al.. 2004, James and Young, 1994, Jones et al., 1996). Research suggests that better phonetic transcription results may be obtained by using transcriptions in the training phase that best match actual pronunciations by modelling regional and other variations from read speech (Kessens and Strik, 2004).

2.7. *Transcripts, annotation and phonogrammic streams*

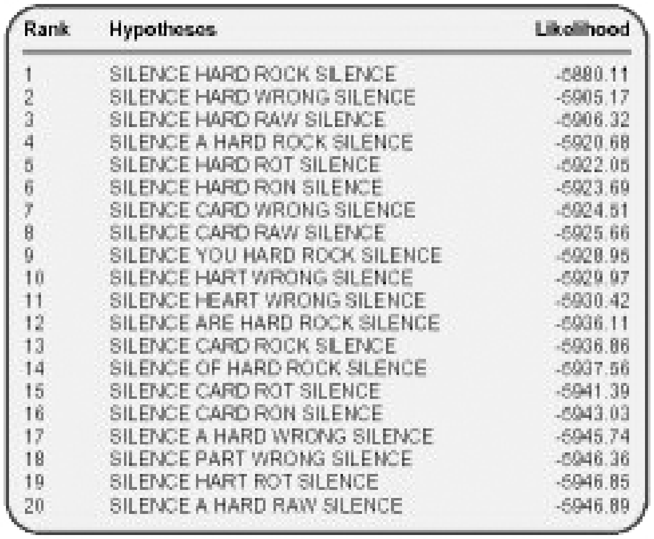
One of the intermediate goals of any SDR system is the transcription of spoken documents into an intermediate representation that allows for effective storing, indexing, searching and retrieval. Transcriptions may be phones or words either in a lattice or graph (probability network), n-best list (multiple individual transcriptions), or more typically, a 1-best transcription (the most probable transcription as determined by the recognizer). When a transcription becomes attached to the original spoken document, it becomes an *annotation*.

2.7.1. 1-best transcriptions

A 1-best transcription is the best single hypothesis that can be derived from a recognition process and may be used at the acoustic (phone), lexical (word) and language (phrase or sentence) levels.

2.7.2. N-best transcriptions

N-best transcriptions begin with the best hypothesis followed by other possibilities in decreasing likelihood. Figure 2.12 shows an example of an N-best list at the language (phrase) level. N-best transcriptions are used by multi-pass recognition engines or by various post-recognition natural language tools for further or later refinement (reordering).



Rank	Hypotheses	Likelihood
1	SILENCE HARD ROCK SILENCE	-5880.11
2	SILENCE HARD WRONG SILENCE	-5905.17
3	SILENCE HARD RAW SILENCE	-5906.32
4	SILENCE A HARD ROCK SILENCE	-5920.68
5	SILENCE HARD ROT SILENCE	-5922.05
6	SILENCE HARD RON SILENCE	-5923.69
7	SILENCE CARD WRONG SILENCE	-5924.51
8	SILENCE CARD RAW SILENCE	-5925.66
9	SILENCE YOU HARD ROCK SILENCE	-5928.95
10	SILENCE HART WRONG SILENCE	-5929.97
11	SILENCE HEART WRONG SILENCE	-5930.42
12	SILENCE ARE HARD ROCK SILENCE	-5936.11
13	SILENCE CARD ROCK SILENCE	-5936.86
14	SILENCE OF HARD ROCK SILENCE	-5937.56
15	SILENCE CARD ROT SILENCE	-5941.39
16	SILENCE CARD RON SILENCE	-5943.03
17	SILENCE A HARD WRONG SILENCE	-5945.74
18	SILENCE PART WRONG SILENCE	-5946.36
19	SILENCE HART ROT SILENCE	-5946.85
20	SILENCE A HARD RAW SILENCE	-5946.89

Figure 2.12 Example of an N-best list for a phrase (Fundamentals, 2005)

2.7.3. Lattices or graphs

Lattice or graph transcriptions are similar to N-best transcriptions in that are commonly used by multi-pass recognition engines or by various post-recognition natural language tools for further or later refinement through pruning the graph. Figure 2.13 shows an example of a word graph (or lattice) at the language (phrase) level for the same phrase used for the N-best list in Figure 2.12. Each node in the graph represents a temporal point in the speech signal. Links between nodes correspond to a given recognition hypothesis.

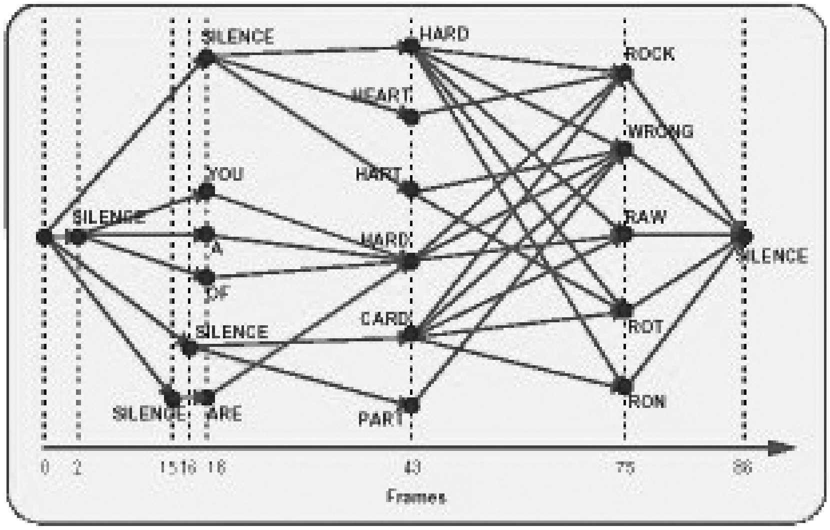


Figure 2.13 Example of a lattice (or word graph) for a phrase (Fundamentals, 2005)

2.7.4. SGML, W3C and markup languages

Markup languages have been used for many years to tag additional information onto text. The Standard Generalized Markup Language (SGML) emerged in 1986 as ISO standard 8879 and is a system for defining markup languages. Authors mark up their documents by representing structural, presentational, and semantic information alongside content. The markup process serves two primary purposes – to separate the logical elements of the document and to specify the processing functions to be performed on those elements (Goldfarb, 1990). Examples of well known and widely used markup languages defined by SGML include eXtensible Markup Language (XML), eXtensible HyperText Markup language (XHTML), HyperText Markup Language (HTML), SALT, SMIL and VoiceXML.

2.7.5. SSML

The World Wide Web Consortium (W3C) is a forum for developing interoperable technologies (specifications, guidelines, software, and tools) to further develop the World Wide Web to its full potential (W3C, 2005). The proliferation of mobile telephony, portable computers, PDAs, tablet computers and other devices has increased the requirement for multimodal, multimedia interaction with the World Wide Web. One of these means of interaction is via *voice browsers*. Voice browsers allow users to access the World Wide Web utilizing speech synthesis, pre-recorded audio, and speech recognition which can also be supplemented by keypads and small displays. The W3C Voice Browser

Working Group has been defining a suite of markup languages to facilitate dialogue, speech synthesis, speech recognition, call control and other aspects of interactive voice response applications (VBWG, 2005). SSML is one of these markup languages, and is a proposed standard (SSML, 2004) for speech synthesis markup within the SGML/XML families of markup languages. The primary role of SSML is to provide authors of documents a standard way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different synthesis-capable platforms. SSML is specifically designed to fully represent elements of speech. SSML is designed to integrate with other markup languages, and may be used within other XML markup languages including VoiceXML, SALT, XHTML+Voice profile and SMIL. Figure 2.14 shows alternative examples of the SSML *phoneme* element.

```
<phoneme ph="t&#252;m&#251;to&#28A;"> tomato </phoneme>

<!-- This is an example of IPA using character entities -->



<phoneme ph="tümûto"> tomato </phoneme>

<!-- This example uses the Unicode IPA characters. -->

<!-- Note: this will not display correctly on most browsers -->
```

Figure 2.14 Example of the SSML *phoneme* element

The first example uses the International Phonetic Alphabet (IPA) character set, and the second example uses the Unicode character set which includes the complete IPA character set as symbols U+0250 to U+02AF plus certain additional Latin and diacritic (accents, tilde, circumflex, etc.) characters. The SSML phoneme element provides a phonetic pronunciation for the contained text. The phonetic string is provided in the required *ph* attribute. The textual representation between the phoneme element start and end tags may be empty, but often contains human-readable text for non-spoken rendering.

2.7.6. VoiceXML

VoiceXML is another markup language being actively developed by the W3C Voice Browser Working Group. VoiceXML has been designed for creating audio dialogs that feature synthesized speech, digitized audio, recognition of spoken and Dual Tone Multi-Frequency (DTMF) telephone key input, recording of spoken input, telephony, and mixed initiative conversations. Its major goal is to bring the advantages of Web-based development and content delivery to interactive voice applications (VoiceXML, 2004).

2.7.7. SALT

Speech Application Language Tags (SALT) consists of set of extensions to existing SGML markup languages (in particular HTML and XHTML) that enables multimodal and telephony access to information, applications and Web services from PCs, telephones, tablet PCs and PDAs (SALT, 2005 and SALT 1.0, 2002). Multimodal access enables users to interact with applications in a variety of ways: input with speech, a keyboard, keypad, mouse and/or stylus; and output as synthesized speech, audio, plain text, motion video and/or graphics (SALT FAQ, 2005). Each mode can be used independently or concurrently. SALT browsers must implement SSML for speech synthesis to achieve SALT compliance. SALT and VoiceXML both describe speech interfaces but have different application niches. VoiceXML is developed for telephony applications to allow Interactive Voice Response (IVR) applications that access documents on the World Wide Web. SALT targets speech applications across a spectrum of devices.

2.7.8. XHTML+Voice profile

The XHTML+Voice (X+V) profile brings spoken interaction to standard web content by integrating the mature XHTML and XML-Events technologies with XML vocabularies developed as part of the W3C Speech Interface Framework. The profile includes voice modules that support speech synthesis, speech dialogs, command and control, and speech grammars. Voice handlers can be attached to XHTML elements and respond to specific events, thereby reusing the event model familiar to web developers. Voice interaction features are integrated with XHTML and CSS and can consequently be used directly within XHTML content. X+V is a Web markup language for developing multimodal applications. Like VoiceXML, X+V meets the increasing user demand for voice-based interaction in small and mobile devices. Unlike VoiceXML, X+V uses both voice and visual elements, bringing a world of new potential to the field of wireless user interface development (X+V, 2004).

2.7.9. MPEG-7 and spoken content

The Moving Picture Experts Group (MPEG) was established in 1988, and has made significant contributions in compression standards for digital audio and video (MPEG, 2005). MPEG-7 (also known as Multimedia Content Description Interface) is a wide-ranging standard for describing multimedia content. MPEG-7 creates a standard multimedia framework that enables searching indexing, filtering and access of multimedia through content description (metadata). A wide range of abstraction levels for metadata is supported, from low-level signal characteristics to high-level semantic information (Manjunath et al. 2002). MPEG-7 provides 4 elemental tools or structures for multimedia metadata: Descriptors, Description Schemes, a Description Definition Language (DDL) and Coding Schemes.

- Descriptors – allow the description of individual features of multimedia content.
- Description Schemes – allow the combination of multiple individual descriptors and multiple description schemes into more complex structures.
- DDL – is an extension of XML which defines MPEG-7 descriptions and description schemes.
- Coding Schemes – are compression schemes to allow textual XML descriptions to be compressed to satisfy application requirements.

MPEG-7 proposes a spoken content description scheme, appropriately named the SpokenContentDescriptionScheme. The design of the SpokenContentDescriptionScheme is based on the premise that a single textual transcript representing spoken content is insufficient. With this in mind, the scheme allows for both word and/or phone lattices along with the following additional components:

- Word lexicons
- Phone lexicons
- Confusion matrices
- Additional metadata (speaker, language, ASR system used, etc.)

The purpose of a word lexicon is to store the vocabulary utilized during recognition. A phone lexicon contains the phone set used during recognition. A confusion matrix contains statistics that allow evaluation of the probability of phone decoding errors. Experimental systems using phone-based retrieval methods along with the MPEG-7 SpokenContentDescriptionScheme are now emerging (Moreau et al.. 2004). Appendix B provides an example of an automatically generated MPEG-7 spoken content XML file generated in this case from a small audio speech file with the message content “mein name ist Ted” (MPEG-7 Demonstrator). Sections of the phone lexicon, confusion matrix and phone lattice have been removed for the sake of brevity.

2.8. Speech and non-speech audio

Research suggests that processing of speech is handled differently by humans than non-speech acoustic information (Liberman, 1982). This is perhaps not so surprising given the disparate requirements for feature abstraction between speech and non-speech audio. Others are examining the retrieval of non-speech acoustic information like music and sound effects. While it is not within the scope of this work to address non-speech audio retrieval, a sampling of non-speech audio retrieval work is provided for convenience and comparison.

2.8.1. Sampling of non-speech audio retrieval work

Several systems use humming, whistling or playing a melody as a query to retrieve music. *MELDEX* is designed to retrieve melodies from a database on the basis of a few notes sung, hummed or played into a microphone. It accepts acoustic input from users and transcribes it into music notation and then searches a database for tunes that contain the sung pattern, or patterns similar to it (McNab et al., 1997). MELDEX has now been integrated into the Greenstone Digital Library Software (Greenstone, 2005) which aims to empower universities, libraries, public service institutions and other users to build digital libraries. *Musipedia* (formerly known as *Melodyhound*) is a melody recognition system developed in 1997. It was originally known as *Tuneserver*. A melody can be found based on whistling or by typing in a simple up-down-repeat pattern referred to as *the Parsons Code* (Prechelt and Typke, 2001 and Musipedia, 2005). *Sonoda* is a WWW-based melody-retrieval system that uses a melody sung or hummed by a user as a query to retrieve the song's title from a music database of standard MIDI files (Sonoda et al., 2003). *Super MBox* is a content-based music retrieval system that allows retrieval by humming, singing or playing queries (Jang et al., 2001b). *MIRACLE* (Music Information Retrieval Acoustically with Clustered and paralleL Engine), can take a user's acoustic input (about 8 seconds) and perform a similarity comparison on a group of clustered PCs (Jang et al., 2001a). *SMILE* is a system designed for content-based musical retrieval. Two types of retrieval modes are provided - a querying function based on a virtual keyboard played by the user, and a browsing function to navigate an automatically constructed hyper-music (Melucci and Orio, 2000).

Some systems use samples of specific audio files to retrieve music. *Shazam* is a commercially available music recognition service that allows people to identify tunes using their mobile phones. When a song is heard on the radio, over a public address system, in a bar or on television the user dials a code number, points the phone at the source of the sound and holds it there for 15 seconds. Within a few minutes, the service returns a text message giving the name of the song and the artist (Harvey, 2003 and Shazam, 2005). *Name That Clip* is a system which is capable of identifying a song, given any short clip from it, such as a five to ten second radio sample. The system does not attempt to identify the main melody, but rather tries to extract a summary of the sound stream on a more basic level. When tested on a database of 1500 songs, spanning a variety of genres and with a query set of 500 microphone-recorded samples, a near-perfect success rate was achieved (Gibson, 1999).

The *Humdrum Toolkit* is a set of software tools intended to allow researchers to encode, manipulate, and output a wide variety of musically-pertinent representations (Humdrum Toolkit, 1999). It is used primarily for answering research questions about musical pieces and collections. *Themefinder* provides a web-based interface to the Humdrum "thema" command, which in turn allows searching of databases containing musical themes. Users may search using a variety of search-keys, including pitch contour, scale degree, date-of-composition, etc. The results for each theme consist of a text header that reports basic information such as composer and title, and notation used. In addition, users can request that a MIDI sound file be downloaded for listening (Themefinder, 2001).

Some systems attempt to classify songs or sounds (like applause, coughing and laughter). *Boogeebot* is an indexing search and retrieval system that generates a list of similar songs from a seed song or songs based solely on audio properties. The primary technique employed uses a distance measure which captures information about the frequency and rhythmic novelty of music. Conceptually, this corresponds to matching the type of instruments playing, including the singer or singing style, and the rhythm (Boogeebot, 2002). *Muscle Fish*'s content-based retrieval technology allows the searching of audio files on the basis of how they sound. Muscle Fish’s audio analysis, search and classification are based on reducing sounds to perceptual and acoustical features. Users may search and retrieve sounds by specifying previously learned classes based on these features or by selecting or entering reference sounds and requesting that similar or dissimilar sounds be retrieved. The Muscle Fish audio retrieval technology is used across several of their applications (Muscle Fish, 2005).

Most current non-speech audio retrieval techniques rely on relational aspects of notes within music melodies for retrieval. Some use the invariant extracted features of recorded songs for retrieval. Some systems use extracted features for general classification of sound types, while others seek to classify songs by frequency and rhythmic novelty. Given the significant differences, additional parameters and problems involved with continuous speech audio files (multiple speakers, variable speed, mispronunciation, unintelligible speech, truncated speech, coarticulation, etc.); the techniques currently being applied to non-speech audio do not immediately appear to be appropriate to SDR.

This section began with a brief introduction to the general field of IR and the primary components of a typical IR system. The fundamentals of ASR and SDR were then explained, followed by a listing of current and previous SDR research along with system origins and descriptions. Similar information was provided on publicly accessible SDR systems, commercially available ASR systems and audio mining products. Previous non-lexical sub-word based SDR approaches were described. Existing approaches for the annotation and transcription of spoken audio were explained along with SGML markup languages with the capability of phonetic expression. Finally, a sampling of non-speech audio retrieval systems along with their origins and descriptions was given.

3. Project proposal

In the following section, the Audient acoustic search engine for digitised audio streams is proposed. The primary development effort involves the design, construction, integration and testing of several core modules toward the overall goal of a demonstrable search engine initially optimised for spoken English. Each of these core modules is outlined in the section along with data flow diagrams. Several developments combine to provide a unique research contribution: (1) the internal standards-based data representation to be used is a 1-best orthographic representation of the speech stream at the phonetic level (phonogrammic stream), (2) the allowance of compound contextual strategies for the refinement of phonogrammic streams is to be available optionally, (3) a mimetic method for adequacy evaluation, diagnostic evaluation and demonstration is to be developed and employed, and (4) multimodal queries are to be available supporting both unconstrained text and speech queries.

3.1. Proposed architecture of Audient

Each of Audient's core modules is outlined below along with data flow diagrams. Figure 3.1 is a top level context diagram for the core modules, and Figure 3.2 is a more detailed level 1 data flow diagram showing the interaction of the core modules.

Recognition and Abstraction Module: Module for the conversion, abstraction and storage of digitised audio streams into abstracted phonogrammic streams with associated temporal information (the possibility of capturing prosodic information will also be investigated). Phonogrammic streams will be orthographical representations of phonemic streams.

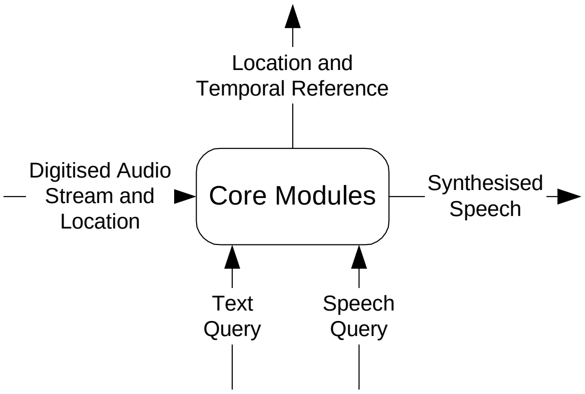

Location and Temporal Reference
Digitised Audio Stream and Location
Core Modules
Synthesised Speech
Text Query
Speech Query

Figure 3.1 Context diagram of core modules

It is desirable to construct phonogrammic streams with the minimal amount of semantic and syntactic interpretation, modelling a human behavioural "first pass" type of recognition (unconscious perception and intelligent action). However, lexical, syntactic, grammatical, semantic and pragmatic evaluation may assist in achieving higher levels of accuracy. This could be thought of as modelling what Dennett refers to as the "Multiple Drafts" model of consciousness (Dennett, 1991) in which speech comprehension seems to occur in a continuous temporal stream, but is actually being revised imperceptibly. A sub-task in the construction of this module is the definition of the internal data structures required for abstraction, storage and effective indexing.

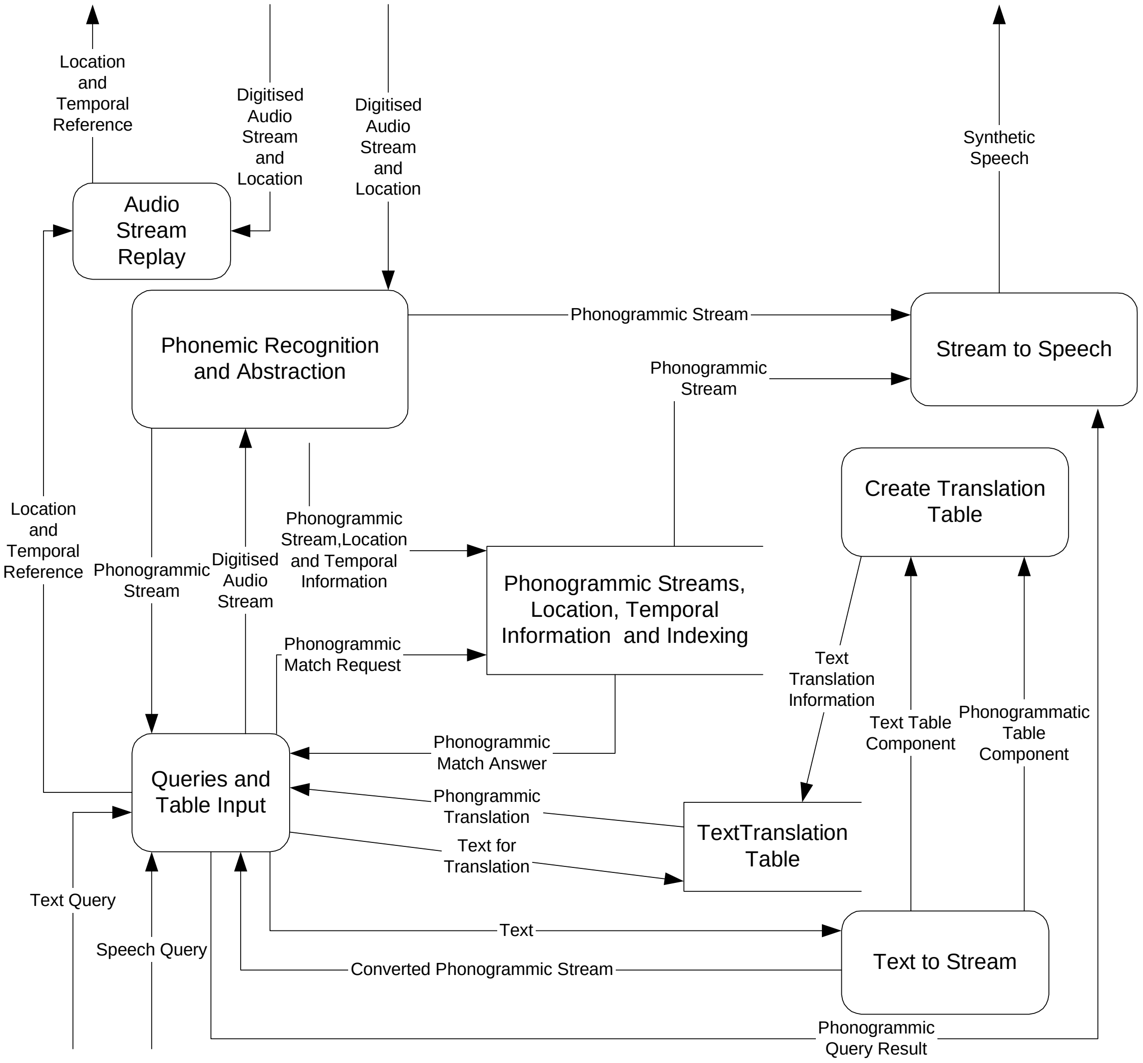

Location and Temporal Reference
Digitised Audio Stream and Location
Digitised Audio Stream and Location
Synthetic Speech
Audio Stream Replay
Phonemic Recognition and Abstraction
Phonogrammic Stream
Stream to Speech
Phonogrammic Stream
Create Translation Table
Location and Temporal Reference
Phonogrammic Stream
Digitised Audio Stream
Phonogrammic Stream,Location and Temporal Information
Phonogrammic Streams, Location, Temporal Information and Indexing
Phonogrammic Match Request
Text Translation Information
Text Table Component
Phonogrammatic Table Component
Phonogrammic Match Answer
Queries and Table Input
Phongrammic Translation
TextTranslation Table
Text for Translation
Text Query
Text
Speech Query
Converted Phonogrammic Stream
Text to Stream
Phonogrammic Query Result

Figure 3.2 Level 1 Data Flow Diagram of Core Modules

Stream to Speech Module: Module to produce synthesised speech from phonogrammic streams. For the purposes of this research, this module is required in the first instance for the development of the Phonemic Recognition and Abstraction Module. It is planned that full advantage be taken of the human aptitude for the evaluation of speech in fine-tuning this module. Later, the Stream to Speech Module should provide the output for query results.

For the purposes of future research, these first two modules should allow for a kind of computer "parrot" – the computer having the ability to "hear" spoken audio information, and repeat the information in a synthesised voice.

Text to Stream Module: Module for producing phonogrammic streams from plain text incorporating TTS conversion tools. This will be used for text queries and provide input for another module for the automated production of a table pairing text search terms and keywords with their phonogrammic translation. When it comes to implementation, it may well be that the text translation activities might best be covered by an existing pronunciation dictionary, or rule-based system

Queries and Table Input Module: Module to service queries originating in either textual or spoken form by reducing them to phonogrammic stream segments, accessing storage and presenting query results to the user. This module also provides text input to populate the Text Translation Table. As mentioned previously, when it comes to implementation, it may well be that the text translation activities might best be covered by an existing pronunciation dictionary, or rule-based system.

Audio Stream Replay Module: Module to fetch audio files, and to replay files from specific temporal reference points.

Create Translation Table Module: Module to create pairs of text with their phonogrammic equivalents for the Text Translation Table. As mentioned previously, when it comes to implementation, it may well be that the text translation activities might best be covered by an existing pronunciation dictionary, or rule-based system.

Creation and integration of these modules provides the core functions for Audient.

Having created and integrated the modules providing the core functions, modules will be tested. The first phase of testing will examine the efficacy of the conversion and abstraction functions. The next phase of testing will involve Information Retrieval (IR) functions being tested against audio corpora used in the evaluation of other IR systems. Iterative testing results will be compared throughout, and where possible, modules will be improved and optimised. Finally, search engine crawler elements are then to be integrated with the core functions, and features and interface further refined.

3.2. Evaluation, testing and refinement

In the evaluation of speech and Natural Language Processing (NLP) systems, three broad areas of evaluation with different goals have been identified (Hirschman and Thompson, 1997):

- *Adequacy evaluation* – The determination of the fitness of a system for a specific purpose.
- *Diagnostic evaluation* – The production of system outputs against possible inputs to detect possible errors.

- *Performance evaluation* – The measurement of system performance in specific areas. Used to compare alternative implementations.

In terms of performance evaluation, it is planned that Audient eventually evaluated by the evaluation criteria applied to the TREC SDR track participants. This will ultimately provide direct comparison with the performance of many previous systems. But because of the different emphasis in system architecture, an additional method is proposed with regard to adequacy and diagnostic evaluation.

Figure 3.3 illustrates functional processes, inputs and outputs of an Audient Parrot. The Audient Parrot is a system that takes as it's input an audio speech file and applies to the file a specified speech recognition engine, along with specified compound strategies (lexical, syntactic, grammatical, semantic, pragmatic or none) to produce a phonogrammic stream.

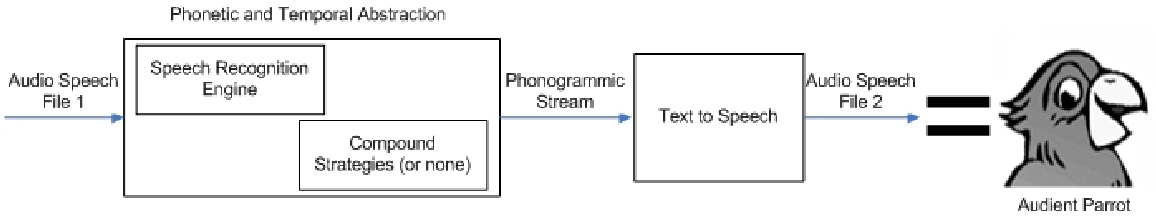

Phonetic and Temporal Abstraction
Audio Speech File 1
Speech Recognition Engine
Compound Strategies (or none)
Phonogrammic Stream
Text to Speech
Audio Speech File 2
Audient Parrot

Figure 3.3 Functional diagram for an Audient Parrot

The phonogrammic stream is in turn passed to a TTS process which outputs a second audio speech file. This allows a human listener to compare the known input to the TTS output. This should provide a relatively short feedback loop hopefully allowing the listener to quickly form hypotheses regarding performance and potential improvements. A simpler version of this proposed type of diagnostic has been used for the subjective comparison of automated and hand-labelled annotations (Cox et al., 1998).

Figure 3.4 illustrates how the Audient Parrot may be used to evaluate recognition differences by having a reader read a text document to a given Audient Parrot then have a writer record the audio speech from the Audient Parrot to a text document and compare the documents.

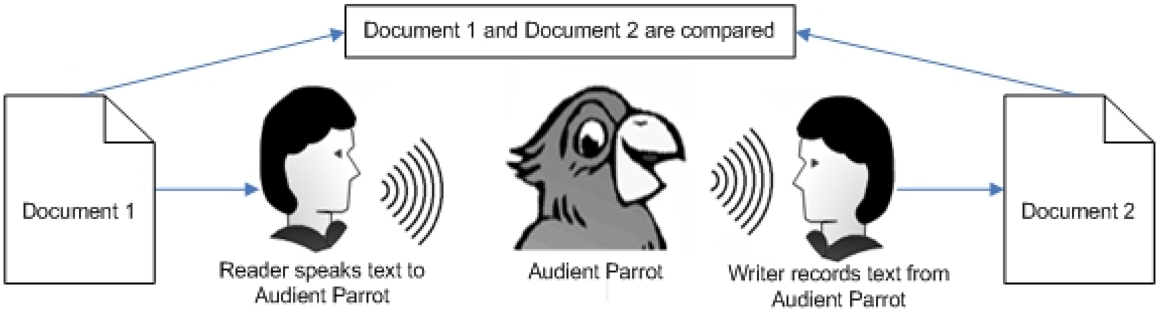

Document 1 and Document 2 are compared
Document 1
Reader speaks text to Audient Parrot
Audient Parrot
Writer records text from Audient Parrot
Document 2

Figure 3.4 Determining recognition differences

Multiple Audient Parrots can exist with different combinations of engines and strategies. One of the perceived weaknesses in this approach is the variability within the writers in terms of vocabulary, spelling and grammar. However, even given the problems in variability of vocabulary, spelling, grammar and homophony, it is still possible to convey phonetic sequence, and even allowing for many lexical mistakes in transmission, it is still arguably possible by this means to convey the original semantic message. Take the following example in Table 3.1:

Document 1	*Document 2*
She sells sea shells by the seashore.	She cells C shels bye the sea shore

Table 3.1 Comparing text documents

While the example documents differ in spelling, homophones and punctuation, many who read Document 2 will derive the meaning intended in Document 1. This is because most humans have highly developed language skills. Audient Parrots are to be intrinsically useful in demonstrating the relative accuracy and speed of differing combinations of speech recognition engines and compound strategies (or none).

With regard to Audient as an eventual complete IR system, Salton and McGill (1983) have identified six critical retrieval evaluation criteria which will be used in system adequacy evaluation:

1. The recall, that is, the ability of the system to present all relevant items
2. The precision, that is, the ability to present only the relevant items
3. The effort, intellectual or physical, required from the users in formulating the queries, conducting the search, and screening the output
4. The time interval which elapses between receipt of a user query and the presentation of user responses
5. The form of presentation of the search output which influences the user’s ability to utilize the retrieved materials
6. The collection coverage, that is, the extent to which all relevant items are included in the system

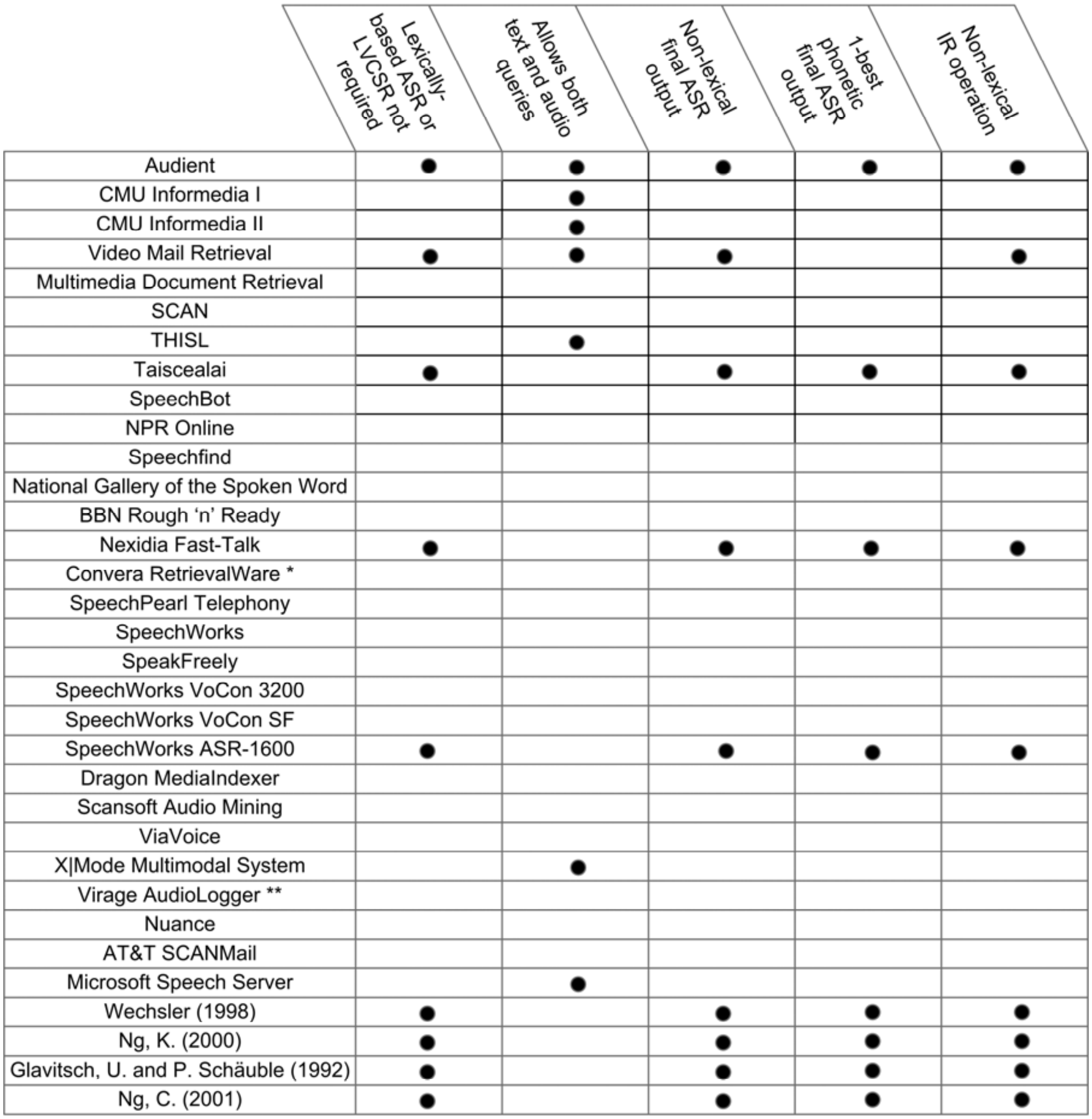
Audient will be evaluated by established evaluation criteria allowing for direct comparison with the performance of previous systems.

3.3. Comparison with previous work

Table 3.1 below presents a comparison of Audient with the approaches of the ASR, SDR and audio mining systems outlined in the literature review section of this report. These systems are diverse in size, speed, architecture and application areas. The approach used in any given system can be difficult to derive with precision since many systems have a plethora of optional configurations, and intermediate steps. The approaches compared are limited and were primarily chosen to highlight some of the unique features of Audient. Comparison columns are as follows:

Lexically-based ASR or LVCSR not required

This column indicates whether or not lexically-based ASR or LVCSR is fundamentally required for operation of the system. That is, essentially whether it is essentially a lexically-based system or not. Most of the systems considered are lexical in nature.



	Lexically-based ASR or LVCSR not required	Allows both text and audio queries	Non-lexical final ASR output	1-best phonetic final ASR output	Non-lexical IR operation
Audient	●	●	●	●	●
CMU Informedia I		●			
CMU Informedia II		●			
Video Mail Retrieval	●	●	●		●
Multimedia Document Retrieval					
SCAN					
THISL		●			
Taiscealai	●		●	●	●
SpeechBot					
NPR Online					
Speechfind					
National Gallery of the Spoken Word					
BBN Rough ‘n’ Ready					
Nexidia Fast-Talk	●		●	●	●
Convera RetrievalWare *					
SpeechPearl Telephony					
SpeechWorks					
SpeakFreely					
SpeechWorks VoCon 3200					
SpeechWorks VoCon SF					
SpeechWorks ASR-1600	●		●	●	●
Dragon MediaIndexer					
Scansoft Audio Mining					
ViaVoice					
X|Mode Multimodal System		●			
Virage AudioLogger **					
Nuance					
AT&T SCANMail					
Microsoft Speech Server		●			
Wechsler (1998)	●		●	●	●
Ng, K. (2000)	●		●	●	●
Glavitsch, U. and P. Schäuble (1992)	●		●	●	●
Ng, C. (2001)	●		●	●	●

* Convera RetrievalWare has Nexidia Fast-Talk as an indexing option

** Virage AudioLogger can use MuscleFish technology to index non-speech sounds

Table 3.1 ASR, SDR and audio mining systems comparison

Allows both text and audio queries

Many of the systems considered may only be queried using text queries typed by users, or generated by GUI interfaces. Some of the telephone oriented systems only allow audio queries by telephone. Only a few of those systems considered allow both text and audio queries.

Non-lexical final ASR output

Many systems can and do produce sub-word data as an intermediate step, but only a few of those listed but only a few produce non-lexical final output.

1-best phonetic final ASR output

Those few systems considered that do produce non-lexical final output could produce 1-best transcriptions, N-best transcriptions or lattices. This column indicates those systems that produce 1-best phonetic final output.

Non-lexical IR operation

After the result of the recognition process, most SDR systems use lexically-based IR functions for storing, indexing, searching and retrieval rather than sub-word units or other abstractions.

The sub-word research systems have more in common with Audient than most of the other research, commercial and publicly accessible systems. It is probable that Audient is unique in its proposed use of standards-based phonogrammic streams as an internal data representation. It is difficult to determine the precise internal data representations of each of the systems listed in Table 3.1 from the available literature, and so this feature has not been listed for comparison.

3.4. Project schedule and status

The work proposed requires several stages to achieve the desired objectives of Audient. Tasks include a literature survey of the subject area, written literature review, selection of software tools, installation and integration of software tools, construction of initial Audient Parrots for testing, construction of the core modules, integration and testing of the core modules, demonstration and finalization of the PhD thesis. Appendix C gives a representation of the project schedule.

4. Software analysis

To meet the objectives of the Audient research as outlined in this document, it is both possible and preferable to use as many existing components as possible. These components will include:

- Speech recognition toolkits, APIs and SDKs
- Speech recognition engines
- Speech corpora with transcripts
- Audio I/O APIs and SDKs
- TTS toolkits, APIs and SDKs
- Scripting and programming languages
- Web server software
- Tools to implement contextual strategies
- XML languages
- Web browsers (XML interpreters) and voice browsers

4.1 Hidden Markov Model Toolkit (HTK)

The Hidden Markov Model Toolkit (HTK) was originally developed at the Speech Vision and Robotics Group (now the Machine Intelligence Laboratory) of the Cambridge University Engineering Department (HTK History, 2002) and contains a set of library modules and tools available in C source form used primarily for speech recognition research. It is anticipated that HTK be used at least for phone level transcription and also for optional contextual strategies HTK also contains editing and re-estimation tools (Young et al., 2002).

4.2 LVCSR and CSLU Toolkit

The CSLU (Center for Spoken Language Understanding) LVCSR project started in 1997 and participated in the DARPA 1997 HUB-4E Broadcast News Evaluation. HUB-4E was an evaluation and scoring specification designed for the purpose of fostering research into the problem of accurately transcribing broadcast news speech, and to objectively measure the state of the art (HUB-4E, 2000). This project was CSLU’s first attempt at LVCSR research (Yan et al., 1998).

The CSLU Toolkit provides tools to build investigate and use interactive language systems. The toolkit includes speech recognition, natural language understanding, speech synthesis and facial animation technologies (CSLU Toolkit, 2005).

4.3 Sphinx-2, Sphinx-3, Sphinx-4

The Sphinx Group at Carnegie Mellon University make these three open source speech recognition engines generally available in order to stimulate the creation of speech tools and applications and to advance the state of the art both in speech recognition and related areas. As mentioned previously, Sphinx-2 is meant as a real-time engine and is regarded as appropriate for systems that require short response times. Sphinx-3 is slower but potentially more and Sphinx-4 is a Java implementation. The Sphinx group also makes available acoustic and language models for those wishing to skip aspects of training and data preparation, and tools for acoustic and language model production (Sphinx Resources, 2005).

4.4 TIMIT

The TIMIT corpus is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of 8 major dialects of American English, each reading 10 phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a speech waveform file for each utterance. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). The speech was recorded at TI, transcribed at MIT and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST) (TIMIT, 2004).

4.5 Linux and C++

While it should be possible to build all of the main HTK tools on any machine supporting ANSI C and either X-Windows or MS-Windows, it is currently planned that the Linux operating system be used, along with either the Intel C++ or GNU C++ compilers. Both of these compilers are ANSI compliant. The Linux operating system is a very rich environment for systems integration. Being open source, it is also very accessible, and has extensive X-Windows management and development facilities.

4.6 Perl and PHP

Perl was initially designed as a “glue language” for the UNIX operating system and its many variants and is now available for most major operating systems. Perl’s process, file and text manipulation facilities make it well suited for systems integration, text processing and rapid prototyping (Wall et al., 1996)). PHP is an HTML-embedded client-side scripting language that operates through a web server like Apache. Much of PHP’s syntax is borrowed from C, Java and Perl (PHP FAQ, 2005).

4.7 Festival

Festival is a speech synthesis system developed at The Centre for Speech Technology Research, University of Edinburgh (Festival, 2005). Festival offers a full text to speech capability. It is written in C++. This tool is to be used in the Stream to Speech core module of Audient, in the Text to Stream module where text input is converted to phonemic representation and in the Audient Parrots.

4.8 The CMU Pronouncing Dictionary

The freely available Carnegie Mellon University Pronouncing Dictionary is a machine-readable pronunciation dictionary for North American English that contains over 125,000 words and their transcriptions (CMUPD, 2005). This tool can be used in contextual strategies and could be used functionally as the foundation of the Text Translation Table.

4.9 SSML, VoiceXML, SALT and X+V

Significant compression should be achieved from the abstraction of phonemic and temporal information from the spectral features of the initial audio stream. Phonemic information is to be translated into a phonogrammic stream, preferably in an existing non-proprietary, standards-based form. VoiceXML, SALT and X+V contain elements from SSML allowing for the encoding of phonemic, prosodic and other information relating to speech synthesis (VoiceXML, 2004, SALT, 2005 and X+V, 2004) which may be suitable for these purposes. The use of SSML should allow the leveraging of currently available software, particularly with regard to browsing and speech synthesis elements of Audient.

4.10 The Apache Web Server

The Apache HTTP Server Project is a collaborative software development effort which has created an efficient and extensible HTTP server whose source code is freely available (Apache Web Server, 2005). The project is jointly managed by a group of volunteers located around the world, using the Internet and the Web to communicate, plan, and develop the server and its related documentation. The Apache HTTP Server is currently the most widely used HTTP server in the world. After the development of modules for the core functions of Audient, it will be necessary to allow users to interface with the modules. The Apache HTTP Server will provide the engine for this interface.

5. Conclusion and future work

In conclusion, this report proposes a programme of research that addresses several specific problems inherent in current SDR systems by implementing a novel approach and architecture. The research aims to explore the efficacy of using standards-based phonogrammic streams as a data abstraction, compare the performance of compound contextual strategies for the refinement of phonogrammic streams, develop mimicry-based means for evaluation and demonstration and provide comparative performance evaluation results.

This report briefly looked at the broad topics of IR, speech and non-speech audio, non-speech audio retrieval systems, ASR and SDR. Also included in the literature review section is a survey of current and previous notable SDR systems arranged under the categories of TREC participants, public access SDR systems and commercial ASR and audiomining products.

The Audient research programme suggests several innovations and contributions to existing knowledge and practice. Standards-based phonogrammic streams are proposed as a fundamental data structure, obviating contextual (lexical, syntactic, grammatic, semantic and pragmatic) requirements of lexically-based systems. While on the most fundamental level, context can be immaterial to the simplest instance of the Audient architecture; contextual strategies may be employed to improve the accuracy of the phonogrammic streams. The Audient architecture supports unconstrained multimodal queries. The new “Audient Parrot” mimicry-based method for evaluation and demonstration is also proposed and movement of the man-machine boundary is proposed for Audient SDR to allow more effective partitioning of tasks between the human and the machine portions of the system.

The software tools to be used in the Audient research include speech recognition toolkits, APIs and SDKs, speech recognition engines, speech corpora with transcripts, audio I/O APIs and SDKs, TTS toolkits, APIs and SDKs, scripting and programming languages, web server software, tools to implement contextual strategies, XML languages, web browsers (XML interpreters) and voice browsers. Specific examples of these include The Hidden Markov Model Toolkit, the CSLU Toolkit, the CSLU LVCSR ASR engine, the CMU Sphinx-2, Sphinx-3, Sphinx-4 ASR engines, the TIMIT corpora and accompanying transcriptions, Linux, C++, Perl, PHP, Festival, the CMU Pronouncing Dictionary, SSML, VoiceXML, SALT and the Apache Web Server.

Audient has a wide range of potential applications in the fields of indexing, search, retrieval and monitoring. Specific applications could include the indexing search and retrieval of Internet audio files, indexing search and retrieval of broadcast media, services for the blind, library services, surveillance and intelligence gathering, voice mail, audio mining and trend analysis (topic detection and tracking).

Audient also holds potential for philosophical and cognitive investigation. Since Audient is modelled in part on what is understood of human speech perception, it has the potential for facilitating research into artificial self-learning systems, philosophical investigations of speech-centric versus text-centric methods, research models for cognitive science and consciousness theories and examination of behaviourist versus cognitive semantic recognition of speech. Audient may also allow exploration of philosophical views on the differences between verbal, non-verbal and written communication (Palmer, 1997, Powell and Howell, 1996).

References

Abberley, D., S. Renals, G. Cook and T. Robinson (1998) The THISL Spoken Document Retrieval System. In *Proceedings of the Sixth Text REtrieval Conference* (TREC-6), E. M. Voorhees and D. K. Harman (Eds.), 747 – 752, Gaithersburg, Maryland, USA .

Abbot (1999) Speech Recognition Using Abbot
http://homepages.inf.ed.ac.uk/srenals/pubs/1999/esca99-thisl/node4.html Site visited 6/7/2005.

Apache Web Server (2005) Welcome! - The Apache HTTP Server Project
http://httpd.apache.org/ Site visited 18/2/2005.

AudioMining (2005) ScanSoft - AudioMining Development System
http://www.scansoft.com/audiomining/developers/ Site visited 18/2/2005.

BBN Byblos (2005) BBN Technologies
http://www.bbn.com/For_Government_Customers/Speech_Recognition/ Site visited 15/7/2005.

Becchetti, C. and Ricotti, L. P. (1999) *Speech Recognition*. Chichester, England, United Kingdom: John Wiley & Sons.

Boogeebot (2002) hp labs - cambridge research lab - speech and audio research
http://www.crl.hpl.hp.com/speech/boogee.htm Site visited 18/2/2005.

Brown, M. G., J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young (1997) Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of ACM Multimedia 96*, Boston, Massachusetts, USA, 307 – 316.

Choi, J., D. Hindle, J. Hirschberg, I. Magrin-Chagnolleau, C. Nakatani, F. Pereira, A. Singhal and S. Whittaker (1998) SCAN - speech content based audio navigator: a systems overview. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 2867 – 2870.

Choi, J., D. Hindle, F. Pereira, A. Singhal and S. Whittaker (1999) Spoken Content-Based Audio Navigation (SCAN). In *Proceedings of the ICPhS-99* (International Congress of Phonetics Sciences).

Clements, M., P. S. Cardillo, M. S. Miller (2001a) Phonetic Searching vs. LVCSR: How to Find What You Really Want in Audio Archives. In *Proceedings of the 20th Annual AVIOS (Applied Voice Input/Output Society) Conference*, San Jose, California, USA.

Clements, M., P. S. Cardillo, M. S. Miller (2001b) Phonetic Searching of Digital Audio. In *Proceedings of the 2001 Broadcast Engineering Conference*, Las Vegas, Nevada, USA.

CMU Sphinx (2005) CMUSphinx: The Carnegie Mellon Sphinx Project
http://cmusphinx.sourceforge.net/html/cmusphinx.php Site visited 29/3/2005.

CMU Sphinx-4 (2004) Sphinx-4 - A speech recognizer written entirely in the Java(TM) programming language
http://cmusphinx.sourceforge.net/sphinx4/ Site visited 18/2/2005.

CMUPD (2005) The CMU Pronouncing Dictionary
http://www.speech.cs.cmu.edu/cgi-bin/cmudict/ Site visited 18/2/2005.

Convera (2004) RetrievalWare Phonetic Search Server Product Brochure, Convera Corporation ,Vienna, Virginia, USA.

Cox, S.J., R. Brady and P. Jackson (1998) Techniques for accurate automatic annotation of speech waveforms. In *Proceedings of The International Conference on Spoken Language* (ICSLP '98), 1947 – 1950, Sydney, Australia.

CSLU Toolkit (2005)
http://cslu.cse.ogi.edu/toolkit/docs/index.html Site visited 18/2/2005.

DeMarco, T. (1978) *Structured Analysis and System Specification*. New York, New York, USA: Yourdon Inc..

Dennett, D.C. (1991) *Consciousness Explained*. London, England, United Kingdom: Penguin Books.

Embedded Speech (2004) ScanSoft - Embedded Speech
http://www.scansoft.com/embedded/ Site visited 18/2/2005.

Festival (2005) Festival
http://www.cstr.ed.ac.uk/projects/festival/ Site visited 18/2/2005.

Fundamentals (2005) Automatic Speech Recognition: Fundamentals of Speech Recognition
http://www.isip.msstate.edu/projects/speech/software/tutorials/production/fundamentals/current/ Site visited 18/2/2005.

Garfolo, J. S., E. M. Voorhees, V. M. Stanford and K. Spärck Jones (1998) TREC-6 1997 Spoken Document Retrieval Track Overview and Results. In *Proceedings of of the Sixth Text REtrieval Conference* (TREC-6), E. M. Voorhees and D. K. Harman (Eds.) , 83 – 91, Gaithersburg, Maryland, USA.

Garfolo, J. S., E. M. Voorhees, C. G. P. Auzanne, V. M. Stanford and B. A. Lund (1999) 1998 TREC-7 Spoken Document Retrieval Track Overview and Results. In *Proceedings of the Seventh Text REtrieval Conference* (TREC-7), E. M. Voorhees and D. K. Harman (Eds.) , 79 – 89, Gaithersburg, Maryland, USA.

Garfolo, J. S., C. G. P. Auzanne and E. M. Voorhees (2000) The TREC Spoken Document Retrieval Track: A Success Story. In *Proceedings of the Eighth Text REtrieval Conference* (TREC-8), E. M. Voorhees and D. K. Harman (Eds.) , 107 – 129, Gaithersburg, Maryland, USA.

Gibson, D. (1999) Name that clip: Music retrieval using audio clips Presentation at SIGIR 1999 Workshop on Music Information Retrieval, Berkeley, California, USA. Abstract at:
http://www.cs.berkeley.edu/~dag/NameThatClip/ Site visited 18/2/2005.

Glavitsch, U. and P. Schäuble (1992) A System for Retrieving Speech Documents. In *Proceedings of the 15th ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 168 - 176.

Goldfarb, C. F. (1990) *The SGML Handbook*. Oxford, England, United Kingdom: Oxford University Press.

Greenberg, S. (1996) Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception,* Keele, England, United Kingdom, 1 – 8.

Greenstone (2005) Greenstone Digital Library Software
http://www.greenstone.org/ Site visited 18/2/2005.

Hansen, J., J.R. Deller and M. Seadle (2001) Engineering Challenges in the Creation of a National Gallery of the Spoken Word: Transcript-Free Search of Audio Archives. In *Proceedings of the IEEE and ACM JCDL-2001: Joint Conference on Digital Libraries*, Roanoke, Virginia, USA, 235 – 236.

Hansen, J. H. L., R. Huang, P. Mangalath, B. Zhou, M. Seadle and J.R. Deller, Jr (2004) SPEECHFIND: Spoken Document Retrieval for a National Gallery of the Spoken Word. In *IEEE NORSIG-2004: Nordic Signal Processing Symposium, Plenary Symposium Paper*, 1 – 4, Espoo, Finland.

Harvey, F. (2003) Name That Tune. In *Scientific American*, June 2003, 84 – 86.

Hauptmann, A. G. and M. J. Witbrock (1997) Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In *Intelligent Multimedia Information Retrieval*, M. T. Maybury (Ed.) , 215 – 239, Menlo Park, California, USA: AAAI Press/MIT Press.

Hirschman, L. and H. S. Thompson (1997) Overview of Evaluation in Speech and Natural Language Processing. In *Survey of the State of the Art in Human Language Technology*, R. Cole (Ed.) , 409 – 414, Pisa, Italy: Giardini Editori (also Cambridge University Press).

Hirschberg, J., M. Bacchiani, D. Hindel, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker and G. Zamchick (2001) SCANMail: Browsing and Searching Speech Data by Content. In *Proceedings of European Conference on Speech Communication and Technology* (Eurospeech) 2001, Aalborg, Denmark.

HP SpeechBot (2004) HP SpeechBot
http://speechbot.research.compaq.com/ Site visited 18/2/2005.

HTK History (2002) "HTK Speech Recognition Toolkit"
http://htk.eng.cam.ac.uk/docs/history.shtml Site visited 18/2/2005.

HUB-4E (2000) 97 HUB-4E Evaluation Plan
http://www.nist.gov/speech/tests/bnr/hub4e_97/current_plan.htm Site visited 18/2/2005.

Humdrum Toolkit (1999) The Humdrum Toolkit: Software for Music Research
http://csml.som.ohio-state.edu/Humdrum/ Site visited 18/2/2005.

Informedia (2004) Informedia Home Page
http://www.informedia.cs.cmu.edu Site visited 18/2/2005.

James, D. A. and S. J. Young (1994) A Fast Lattice-Based Approach to Vocabulary Independent Wordspotting. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, Australia, 377 - 380.

Jang, J. –S. R., J. –C. Chen and M. –Y. Kao (2001a) MIRACLE: A music information retrieval system with clustered computing engines. In *Proceedings of the Second Annual International Symposium on Music Information Retrieval:ISMIR 2001*, J.S. Downie and D. Bainbridge (Eds.) , 11 – 12, Bloomington, Indiana, USA.

Jang, J. -S. R., H.-R. Lee and J. –C. Chen (2001b) Super MBox: An Efficient/Effective Content-based Music Retrieval System. In *Proceedings of the Ninth ACM International Conference on Multimedia*, Ottawa, Ontario, Canada, 636 – 637.

Jones, G. J. F., J. T. Foote, K. Spärck Jones and S. J. Young (1996). Retrieving Spoken Documents by Combining Multiple Index Sources. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, 30 – 38.

Jones, G., J. Foote, K. Spärck Jones and S. Young (1997) The Video Mail Retrieval Project: Experiences in Retrieving Spoken Documents. In *Intelligent Multimedia Information Retrieval*, M. T. Maybury (Ed.), 191 – 214, Menlo Park, California, USA: AAAI Press/MIT Press.

Jurafsky, D. and J. H. Martin (2000) *Speech and Language Processing*. Upper Saddle River, New Jersey, USA: Prentice-Hall.

Keller, E. (Ed.) (1994) *Fundamentals of Speech Synthesis and Speech Recognition*.

Chichester, England, United Kingdom: John Wiley & Sons.

Kessens, K. and H. Strik (2004) On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. In *Computer Speech & Language*, Vol. 18, Issue 2, 123 – 141.

Kubala, F., S. Colbath, D. Liu and J. Makhoul (1999) Rough 'n' Ready: A Meeting Recorder and Browser. In *ACM Computing Surveys*, Vol. 31, Issue 2es (June 1999).

Larson, M. (2001) Sub-Word-Based Language Models for Speech Recognition: Implications for Spoken Document Retrieval. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Larson, M. and S. Eickeler. (2003) Using Syllable-based Indexing Features and Language Models to improve German Spoken Document Retrieval. In Proceedings of The European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, 1217 – 1220.

Leavitt, N. (2002) Let's Hear It for Audio Mining. In *IEEE Computer*, Vol. 35, 23 – 24.

Leman, M. (2002) Musical Audio Mining In *Dealing with the Data Flood: Mining data, text and multimedia*. J. Meij (Ed.), Rotterdam, Netherlands: STT Netherlands Study Centre for Technology Trends.

Liberman, A. M. (1982) On Finding That Speech Is Special. In *American Psychologist*, Vol. 37, No. 2, 148 – 167.

Manjunath, B. S., P. Salembier and T. Sikora (2002*) Introduction to MPEG-7: Multimedia Content Description Interface*. Chichester, England, United Kingdom: John Wiley & Sons Ltd.

Mani, I., D. House, M. Maybury and M. Green (1997) Towards Content-Based Browsing of Broadcast News Video. In *Intelligent Multimedia Information Retrieval*, M. T. Maybury (Ed.), 241 – 258, Menlo Park, California, USA: AAAI Press/MIT Press.

Maybury, M.T. (Ed.) (1997) *Intelligent Multimedia Information Retrieval*. Menlo Park, California, USA: AAAI Press/MIT Press.

McNab, R.J., L. A. Smith, D. Bainbridge, and I. H. Witten (1997) The New Zealand Digital Library MELody inDEX In *D-Lib Magazine*
http://www.dlib.org/dlib/may97/meldex/05witten.html Site visited 18/2/2005.

MDR (2001) MultiMedia Document Retrieval (1997-2000)
http://mi.eng.cam.ac.uk/research/Projects/Multimedia_Document_Retrieval/ Site visited 18/2/2005.

Meadow, C.T. (1992) *Text Information Retrieval Systems*. San Diego, California, USA: Academic Press.

MediaIndexer (2004) ScanSoft - Dragon MediaIndexer
http://www.scansoft.com/mediaindexer/ Site visited 9/9/2004.

Melucci, M. and N. Orio (2000) SMILE: a System for Content-based Musical Information Retrieval Environments. In *RIAO 2000 Conference proceedings*, Vol. 2, Paris, France, 1261 -1279.

Mooers, C. N. (1951) Information retrieval viewed as temporal signalling. *In Proceedings of the International Conference of Mathematicians*, Cambridge, Massachusetts August 30th – September 6th, 1950. 572 – 573, Providence, Rhode Island, USA: American Mathematical Society.

Moreau N., H.-G. Kim and T. Sikora (2004) Phone-based Spoken Document Retrieval in Conformance with the MPEG-7 Standard , 25th International Audio Engineering Society Conference (Metadata for Audio), London, England, United Kingdom.
http://www.nue.tu-berlin.de/publications/papers/Moreau_AESAudioMetadata2004.pdf
Site visited 16/2/2005.

MPEG (2005) MPEG Home Page
http://www.chiariglione.org/mpeg/ Site visited 18/2/2005.

MPEG-7 Demonstrator (2004) TU-Berlin: MPEG-7 Spoken Content Demonstrator
http://mpeg7spkc.nue.tu-berlin.de/ Site visited 21/2/2005.

MSS (2005) Microsoft Speech – Home Page
http://www.microsoft.com/speech/ Site visited 21/2/2005.

Muscle Fish (2005) Muscle Fish - Consultants in audio, music, MIDI, digital signal processing (DSP), and audio search technology. SoundFisher sound file database management software.
http://www.musclefish.com/ Site visited 21/2/2005.

Musipedia (2005) Musipedia: The Open Music Encyclopedia
http://www.musipedia.org/ Site visited 18/2/2005.

NaturallySpeaking (2005) ScanSoft - Dragon NaturallySpeaking 8
http://www.scansoft.com/naturallyspeaking/ Site visited 21/2/2005.

Network Speech (2004) ScanSoft - Network Speech - Automatic Speech Recognition
http://www.scansoft.com/network/asr/ Site visited 21/2/2005.

Ng, C. and J. Zobel (1998) Speech Retrieval using Phonemes with Error Correction. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval, Melbourne, Australia, 365 – 366.

Ng, C. L. (2001) Content Based Retrieval of Speech Documents using Information Retrieval Techniques. Ph.D. Thesis, Department of Computer Science, Faculty of Applied Science, RMIT University, Melbourne, Victoria, Australia.

Ng, K. and V. W. Zue (1997) Subword unit representations for spoken document retrieval. In *Proceedings of the European Conference on Speech Communications and Technology, EUROSPEECH*, Rhodes, Greece, 1607 – 1610.

Ng, K. (2000) Subword-based Approaches for Spoken Document Retrieval. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

NGSW (2005) AODL Grant Info
http://www.ngsw.org/ Site visited 21/2/2005.

NPR (2000) Current Online | Streaming audio how-to
http://www.current.org/stream/stream020npr.html Site visited 14/7/2005.

NPR Archives (2004) NPR: Archives
http://www.npr.org/archives/index?loc=hometext.html Site visited 21/2/2005.

Nuance (2005) Nuance - The Speech Recognition Leader
http://www.nuance.com/ Site visited 21/2/2005.

Palmer, D. D. (1997) *Wittgenstein for Beginners*. New York, New York, USA: Writers and Readers Publishing.

PBMDP (2005) PBCore Metadata: Welcome
http://www.utah.edu/cpbmetadata/background/AtAGlanceFAQs.html Site visited 14/7/2005.

Philips Speech Processing (2005) Philips Speech Recognition Systems - Home
http://www.speechrecognition.philips.com/ Site visited 21/2/2005.

PHP FAQ (2005) PHP: FAQ: Frequently Asked Questions - Manual
http://uk2.php.net/FAQ.php Site visited 21/2/2005.

Powell, J. and V. Howell (1996) *Derrida for Beginners*. New York, New York, USA: Writers and Readers Publishing.

Prechelt, L. and R. Typke (2001) An interface for melody input. In *ACM Transactions on Computer-Human Interaction,* Vol. 8, Issue 2, 133-149.

Quinn, E. (2000) SpeechBot: The First Internet Site for Content-Based Indexing of Streaming Spoken Audio. Technical Whitepaper, Compaq Computer Corporation, Cambridge, Massachusetts, USA.

Rough ‘n’ Ready (2004) What We Do: Speech & Language Processing: Rough'n'Ready[tm]
http://www.bbn.com/speech/roughnready.html Site visited 9/9/2004.

SALT (2005) Speech Application Language Tags (SALT) Forum – Home

http://www.saltforum.org/ Site visited 18/7/2005.

SALT 1.0 (2002) Speech Application Language Tags (SALT) 1.0 Specification
http://www.saltforum.org/saltforum/downloads/SALT1.0.pdf Site visited 18/7/2005.

SALT FAQ (2005) Speech Application Language Tags (SALT) Forum – FAQ
http://www.saltforum.org/faq/faq.asp Site visited 18/7/2005.

Salton, G. and M. J. McGill (1983) *Introduction to Modern Information Retrieval*. New York, New York, USA: McGraw-Hill.

ScanSoft (2004) ScanSoft - The leading supplier of speech and imaging solutions
http://www.scansoft.com/ Site visited 21/2/2005.

SCANMail (2003) AT&T Labs Research – SCANmail
http://www.research.att.com/projects/SCANmail/ Site visited 21/2/2005.

Schäuble, P. and M. Weschler (1995) First Experiences with a System for Content Based Retrieval of Information from Speech Recordings. In *Working Notes of The International Joint Conference on Artificial Intelligence* (*IJCAI) Workshop*: *Intelligent multimedia information retrieval*, M. T. Maybury (Chair), Montreal, Canada, 59 – 69.

Shazam (2005) Shazam Entertainment Ltd - music recognition over your mobile phone
http://www.shazamentertainment.com/home.shtml Site visited 21/2/2005.

Smeaton, A. F., M. Morony, G. Quinn and R. Scaife (1998) Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries* (ECDL), C. Nikolaou and C. Stephanidis (Eds.) , Crete, 429 – 442, Heraklion.

Sonoda, T., T. Ikenaga, K. Shimzu and Y. Muraoka (2003) A Melody Retrieval System on Parallelized Computers. In *Entertainment Computing*, Naktsu and Hoshino (eds.), Boston, Massachusetts, USA: Kluwer Academic Publishers.

Spärck Jones, K. and P. Willett (Ed.) (1997) *Readings in Information Retrieval*. San Francisco, California, USA: Morgan Kaufmann Publishers.

Spärck Jones, K. P. Jourlin, S.E.Johnson and P.C.Woodland (2001) The Cambridge Multimedia Document Retrieval (MDR) Project: Summary of experiments. Technical report 517.

SpeechFind (2003) SpeechFind: Search the Speech from Last Century
http://speechfind.colorado.edu Site visited 9/9/2004.

Sphinx Resources (2005) CMUSphinx: The Carnegie Mellon Sphinx Project
http://cmusphinx.sourceforge.net/html/system.php Site visited 20/3/2005.

SSML (2004) Speech Synthesis Markup Language (SSML) Version 1.0
http://www.w3.org/TR/speech-synthesis/ Site visited 20/3/2005.

Takeshita, A., I. Takafumi and T. Kazuo (1997) Topic-based Multimedia Structuring In *Intelligent Multimedia Information Retrieval*, M. T. Maybury (Ed.), Menlo Park, California, USA: AAAI Press/MIT Press, 259 – 277.

Themefinder (2001) Themefinder
http://themefinder.org/ Site visited 20/3/2005.

THISL (1998) The THISL Spoken Document Retrieval Project
http://www.dcs.shef.ac.uk/research/groups/spandh/projects/thisl/overview-oct98/ Site visited 20/3/2005.

TIMIT (2004) LDC Catalog
http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1 Site visited 26/9/2004.

Tuerk, A., S.E. Johnson, P. Jourlin, K Spärck Jones and P.C. Woodland (2000) The Cambridge University Multimedia Document Retrieval Demo System. In *Proceedings of ACM SIGDIR Conference*, Athens, Georgia, USA, 394.

TREC (2004) Text REtrieval Conference (TREC) Home Page
http://trec.nist.gov Site visited 20/3/2005.

van Rijsbergen, C. J. (1979) *Information Retrieval*. 2nd edition. London, England, United Kingdom: Buttersworth.
http://www.dcs.gla.ac.uk/Keith/Preface.html Site visited 20/3/2005.

Van Thong, JM., P.J. Moreno, B. Logan, B. Fidler, K. Maffey and M. Moores (2001) SPEECHBOT: An Experimental Speech-Based Search Engine for Multimedia Content in the Web. Technical Report Series, Cambridge Research Laboratory, Compaq Computer Corporation, Cambridge, Massachusetts, USA.

VBWG (2005) W3C Voice Browser Activity
http://www.w3.org/Voice/ Site visited 20/3/2005.

Video Mail (1997) Video Mail Retrieval Using Voice
http://mi.eng.cam.ac.uk/research/Projects/vmr/vmr.html Site visited 20/3/2005.

Virage (2004) Virage, Inc.
http://www.virage.com/ Site visited 20/3/2005.

VoiceXML (2004) Voice Extensible Markup Language (VoiceXML) Version 2.0
http://www.w3.org/TR/voicexml20/ Site visited 20/3/2005.

Voorhees, E. M. and D. Harman (2001) Overview of the Ninth Text Retrieval Conference (TREC-9). In *Proceedings of the Ninth Text REtrieval Conference* (TREC-9), E. M. Voorhees and D. K. Harman (Eds.), Gaithersburg, Maryland, USA, 1 – 13.

W3C (2005) World Wide Web Consortium
http://www.w3.org/ Site visited 20/3/2005.

Wall L., T. Christiansen, R. Schwartz and S. Potter (1996) *Programming Perl, Second Edition*. Sebastopol, California, USA: O'Reilly and Associates, Inc..

Warnke, V., S. Harbeck, E. Nöth and H. Niemann (1997) Topic Spotting using Subword Units. In *Proceedings of 9 Aachener Kolloqium "Signaltheorie", Bildund Sprachsignale*, Aachen, Germany, 287 – 291.

Wechsler, M. (1998) Spoken Document Retrieval based on Phoneme Recognition, Doctor of Technical Sciences Dissertation, Swiss Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland.

X+V (2004) XHTML+Voice Profile 1.2
http://www.voicexml.org/specs/multimodal/x+v/12/ Site visited 20/3/2005.

Xmode (2004) ScanSoft - X|mode Multimodal System
http://www.scansoft.com/xmode/ Site visited 9/9/2004.

Yan Y., X. Wu, J. Schalkwyk and R. Cole Development of CSLU LVCSR: The 1997 DARPA HUB4 Evaluation System. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop* (BNTUW-98), Lansdowne, Virginia, USA. http://www.nist.gov/speech/publications/darpa98/ Site visited 20/3/2005.

Young S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland (2002) *The HTK Book*. Cambridge, England, United Kingdom: Cambridge University Engineering Department.

Zhou, B. and J. H. L. Hansen (2002) SpeechFind: an Experimental On-Line Spoken Document Retrieval System for Historical Audio Archives. In *Proceedings of ICSLP-2002: International Conference on Spoken Language Processing*, Vol. 3, 1969 – 1972, Denver, Colorado, USA.

Appendix A: ScanSoft Speech Technologies

SpeechWorks network speech solutions (Network Speech, 2004)

OpenSpeech – An ASR solution that is optimized for VoiceXML.

SpeechPearl Telephony – Is referred to on Scansoft's website as a supported legacy product and is ASR software designed for integration into voice processing platforms and developed for a wide range of telephony speech applications, from high-density digit recognition up to several million words vocabularies. SpeechPearl consists of a set of modules for application development and system integration.

SpeechWorks 6.5SE – Is referred to on Scansoft's website as a supported legacy product and is a software product for building network-based ASR services.

SpeakFreely - Is referred to on Scansoft's website as a supported legacy product and has speech recognition capabilities based on statistical models of spoken language and provides one form of natural language capabilities.

SpeechWorks family of embedded solutions (Embedded Speech, 2004)

SpeechWorks VoCon 3200 – A high accuracy and large vocabulary speech recognition engine.

SpeechWorks VoCon SF - A smaller footprint engine for platforms with constrained memory and CPU resources. Appropriate for embedded hardware and software applications, including those in automotive, telematics, consumer electronics and mobile communications

Speech Works ASR-1600 – A speech recognition engine designed with the games industry in mind for creating hands-free and multimodal interfaces.

Dragon MediaIndexer - creates an XML speech index of spoken content using ASR while simultaneously creating a streamable, encoded version of the content in real time (MediaIndexer, 2004).

Dragon NaturallySpeaking - offers home and small office users powerful speech recognition features to maximize productivity. NaturallySpeaking can launch programs, create documents and reports, and manage the desktop by voice. Also includes TTS (NaturallySpeaking, 2005).

ScanSoft Audio Mining Development System which includes an SDK and other tools for developers (AudioMining, 2005).

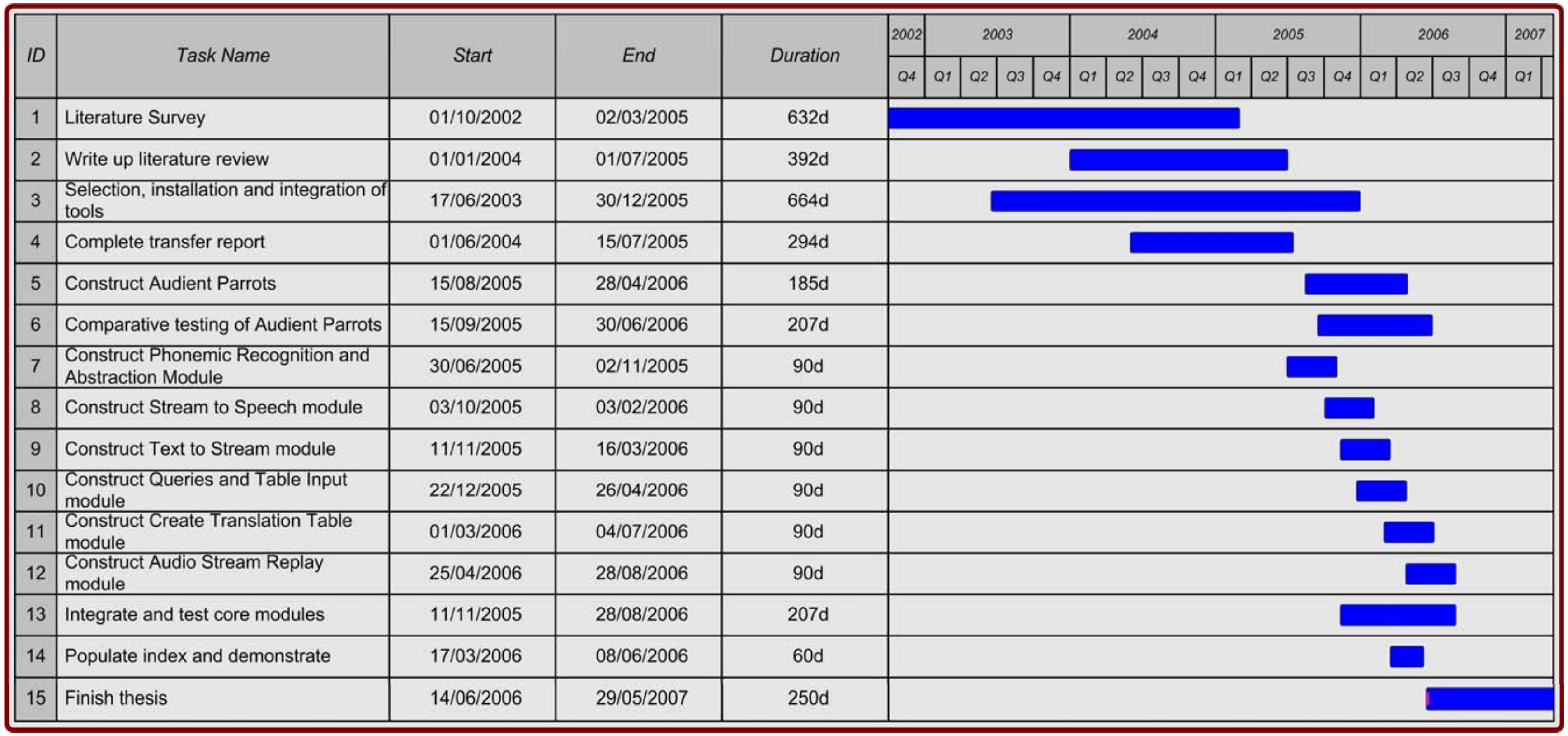
ViaVoice – Is similar to Dragon NaturallySpeaking in its applications. **X|Mode Multimodal System** – combines ASR and Text to Speech (TTS) technologies with mobile Internet and multimedia technology to enable rapid development and deployment of multimodal applications, combining voice, visual and audio interfaces on a single mobile device and a single session (Xmode, 2004)

Appendix B: M-PEG 7 Example

```
    <?xml version="1.0" encoding="iso-8859-1" ?>

- <!--  TU Berlin Spoken Content Demonstrator v1.0: http://www.nue.tu-
    berlin.de/forschung/projekte/mpeg7/

  -->

- <Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
    xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">

- <DescriptionUnit xsi:type="SpokenContentLatticeType" id="latFirst">

- <Header xsi:type="SpokenContentHeaderType">

- <PhoneLexicon id="GlobalLexicon" numOfOriginalEntries="43" phoneticAlphabet="other">

  <Token>...</Token>

  <Token>Q</Token>

  <Token>p</Token>

  …{Complete phone lexicon shortened for example – Ted Leath}

  <Token>6</Token>

  <Token>e</Token>

  </PhoneLexicon>

- <ConfusionInfo id="GlobalConfusionInfo" numOfDimensions="43">

  <Insertion>2522 465 185 87 1109 273 337 116 196 159 351 85 37 175 55 96 213 439 1079 48 120 142
    257 279 268 96 156 72 24 136 241 148 254 111 76 44 26 114 51 28 762 312 1</Insertion>

  <Deletion>1744 5663 114 771 5217 1656 339 713 402 1085 1423 417 81 587 222 195 698 832 3358 133
    1114 779 1201 712 891 311 634 133 39 1007 897 314 1096 287 329 153 45 385 129 42 3780 2721
    5</Deletion>

  <Substitution dim="43 43">7196 10 28 0 218 2 68 5 44 5 21 1 2 24 3 29 5 75 157 7 7 4 11 9 9 2 6 1 1 6
    7 9 39 8 2 1 3 4 3 1 47 42 0 69 5731 20 75 89 396 39 106 70 154 53 195 15 7 92 10 129 193 182 9 114
    182 9 14 14 16 8 6 3 30 39 8 45 14 8 3 2 17 12 2 58 41 0 10 17 880 23 43 15 18

…{Complete confusion matrix shortened for example – Ted Leath}

2  5 111 7 89 9 34 17 92 5 18 9 23 3 3 36 7 65 8 46 146 5 94 5 89 433 578 161 40 40 21 19 103 92 585
    156 27 24 18 175 119 17 172 2866 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 0 0 0 0 0 2 3 0 0 0 0
    0 0 0 0 0 0 0 35</Substitution>

  </ConfusionInfo>

- <SpeakerInfo id="SpeakerX" phoneLexiconRef="#GlobalLexicon" provenance="unknown">

  <SpokenLanguage>de</SpokenLanguage>

  </SpeakerInfo>
  </Header>

- <Block num="0" audio="unknown" defaultSpeakerInfoRef="#SpeakerX">

- <MediaTime>

  <MediaTimePoint>2003-11-10T00:00:00</MediaTimePoint>

  </MediaTime>

- <Node num="0" timeOffset="0">

  <PhoneLink nodeOffset="1" probability="1.000000e+000" acousticScore="-2.757310e+003" phone="0" />

  </Node>

- <Node num="1" timeOffset="41">

  <PhoneLink nodeOffset="1" probability="1.652989e-001" acousticScore="-8.868900e+002" phone="4" />

  </Node>

- <Node num="2" timeOffset="52">

  <PhoneLink nodeOffset="1" probability="3.370868e-002" acousticScore="-5.289200e+002" phone="9" />

  </Node>

- <Node num="3" timeOffset="59">

  <PhoneLink nodeOffset="1" probability="7.427358e-002" acousticScore="-1.356090e+003" phone="37"
    />

  </Node>

- …{Complete phone lattice shortened for example – Ted Leath}

- <Node num="30" timeOffset="304">

  <PhoneLink nodeOffset="1" probability="1.261858e-001" acousticScore="-1.430400e+002" phone="23"
    />

  </Node>

- <Node num="31" timeOffset="306">

  <PhoneLink nodeOffset="1" probability="1.703330e-001" acousticScore="-1.380200e+002" phone="41"
    />

  </Node>

- <Node num="32" timeOffset="308">

  <PhoneLink nodeOffset="1" probability="1.849971e-002" acousticScore="-2.353340e+003" phone="0" />

  </Node>

  <Node num="33" timeOffset="349" />
```

```
</Block>

</DescriptionUnit>

</Mpeg7>
```

Appendix C: Project schedule



ID	Task Name	Start	End	Duration
1	Literature Survey	01/10/2002	02/03/2005	632d
2	Write up literature review	01/01/2004	01/07/2005	392d
3	Selection, installation and integration of tools	17/06/2003	30/12/2005	664d
4	Complete transfer report	01/06/2004	15/07/2005	294d
5	Construct Audient Parrots	15/08/2005	28/04/2006	185d
6	Comparative testing of Audient Parrots	15/09/2005	30/06/2006	207d
7	Construct Phonemic Recognition and Abstraction Module	30/06/2005	02/11/2005	90d
8	Construct Stream to Speech module	03/10/2005	03/02/2006	90d
9	Construct Text to Stream module	11/11/2005	16/03/2006	90d
10	Construct Queries and Table Input module	22/12/2005	26/04/2006	90d
11	Construct Create Translation Table module	01/03/2006	04/07/2006	90d
12	Construct Audio Stream Replay module	25/04/2006	28/08/2006	90d
13	Integrate and test core modules	11/11/2005	28/08/2006	207d
14	Populate index and demonstrate	17/03/2006	08/06/2006	60d
15	Finish thesis	14/06/2006	29/05/2007	250d