Audio and Multimedia Processing in Digital Libraries

Mounya Elhilali[1]

Institute for Systems Research, Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742, USA.

Abstract

In this era of rapid growth of digital libraries, it becomes very important to develop technologies necessary for audio and multimedia processing, as many of the archives and resources used nowadays are in the form of audio and speech recordings. This paper describes some of the requisite technologies for speech and audio processing. The techniques highlighted in this study involve audio coding, automatic indexing and classification, information retrieval, and fast sound playbacks. The main focus of this work is to provide a brief overview of the various techniques currently in use. A particular emphasis is given to the potential benefits that these existing audio processing tools can gain from the knowledge of the human auditory system.

1. Introduction

The new developments and advances of electronic archive and digital library technologies are and will be triggering more focus towards the importance of audio and multimedia resources. These later are certainly one of the valuable data objects that library systems need to incorporate with their collections. However, audio and multimedia data require special processing techniques specifically appropriate for them.

In this paper, we give an overview of the processing techniques relevant for audio processing. Section 2 reviews the distinctive features of sound data, and thus the type of technologies needed for audio processing. Some of the important audio collections currently available in digital library systems are presented in section 3. Then, section 4 addresses specific technologies in audio processing including audio coding, indexing, classification, information retrieval, and fast sound forwarding. In this study, we present the state of the art technologies available nowadays for audio processing. We also try to point out some of their reported or known limitations or shortfalls. And whenever appropriate, we mention some of the possible improvements or modifications of these techniques, based on certain facts about human audition and perception.

2. Audio Data & Digital Libraries

In addition to textual materials, digital libraries need to support other object models, such as a video, audio and imaging data. In this paper, we specifically address

[1] mounya@eng.umd.edu

the processing of sound data associated with different sources (sound recordings, audio components of video data…). The importance of audio archives cannot be denied; they provide insight into historical facts that could have been falsified in records; they also archive important resources of historical, social, cultural, economic and scientific value. Some examples of important audio corpora in digital library systems are cited in section 3.

In this paper, the data we are interested in include any sound recordings (speech or non-speech), audio component of video data, as well as multimedia data. In this later case, sound is usually incorporated with other effects (images, graphics, virtual reality objects). So, whether in the case of video or multimedia objects, the sound component has to be considered separately; as audio techniques are generally specific to sound objects –this issue is explained later in this section-. Nevertheless, audio processing usually goes hand with processing of other objects. For instance, the analysis of audio data plays a primary role in video parsing, and does usually complement the imaging information in providing a better classification of the audiovisual data.

At this level, the question is why do we need to address issues related to audio data separately from the other types of resources. Several reasons could be pointed out to respond to this question:

- ✓ *Voluminous data sizes:* It is a fact that one hour of compact-disk quality of digitized audio recording requires 635 megabytes of storage if not compressed. So, no matter how much advanced and cheap the data storage technologies become, it is necessary to address the need for compression and coding schemes, as most sound and multimedia resources require substantial amounts of storage space.
- ✓ *Networking issues:* The same one hour of uncompressed CD quality sound would also require network links that support rates of few hundreds megabits/sec. So, even with the use of compressed data, the capabilities and limitations of the underlying hardware are also crucial. This indeed is of major importance for audio and multimedia resources, which require the delivery of data in a timely and steady fashion. Thus, streaming protocols would be necessary in order to segment the data into uniform packets that could be reliably delivered throughout the system with no noticeable delays or audible distortions.
- ✓ *Sound compression quality assessment:* As the need for compression schemes has been addressed in the previous points, the question is the design of these schemes so as to maintain the quality of the sound recordings (in terms of intelligibility as well as quality). This quality assessment question has to be addressed with special consideration to the human perceptual capabilities, since the end user would be requiring listening to the sound part or video segment with no substantial perceivable degradation.
- ✓ *Indexation and information retrieval:* Another aspect of audio processing concerns analyzing the video or sound tracks to address tasks such as segmenting, indexing, cataloging and retrieving information. Techniques from various research areas have to be incorporated at this level; e.g. speech recognition (to identify the data content), natural language processing (to perform content-based indexing), and artificial intelligence (for information retrieval).
- ✓ *Fast sound browsing:* Finally, an important task that needs to be addressed is offering possibilities of fast navigation through audio recordings. These techniques help the

user find information of interest in the audio resources; in the same way one would skim through a text document or fast-forward a video segment.

Until recently, audio has not been valued as an archival source because of the difficulty of retrieving or processing information in large sound files. Technology challenges have to be taken in order to value audio recordings as one of the important resources of any digital library system.

3. Practical Experiences

In this section, we present practical implementations from widely known libraries; which have dedicated special efforts to incorporating valuable audio and video recordings to their digital library archives. These experiences are cited in this context as examples of applications of integrating audio corpora in digital libraries, and also as possible references for real case studies.

a. Library of Congress [15, 16]: The library of congress is known worldwide as the largest and most prestigious library collection in the world. Its archives comprise unique and valuable collections of various types, including books, articles, legislative papers, maps, photographs, sound recordings, video data …etc. As far as its audio resources are concerned, one of the library of congress largest collections are the American History records. These include political speeches (e.g. valuable audio recordings from the World War I era), and music sound tracks dating back to the late 19th and early 20th century (e.g. folk music recordings, Vaudeville and popular entertainment records from the period 1870-1920).

b. British Library [5]: The British library is the 250 years old national library of the United Kingdom. It provides one of the world finest collections of books, manuscripts, journals and other resources covering all areas of knowledge. Its digital library is an extended IT system with diverse functionalities, and access to great amounts of materials. Concerning its sound corpora -which is of interest for us in this study- the British Library has one of the world's largest sound archives. This sound collection is part of the National Sound Archive (NSA) opened since 1955. Its collection is a very extended one with over one million disks, in addition to tapes, video recordings …etc. As part of its most outstanding archives, the NSA has recently released an astonishing recording of Nelson Mandela's famous speech before his sentence to life imprisonment in 1964. This record brought back interest of many people in sound archives.

c. Informedia [2, 13, 23]*:* Informedia is a large project carried at Carnegie Mellon university to study the establishment and use of multimedia digital libraries. This project brings insights and innovations to various techniques used for automated video and audio indexing, navigation and retrieval. Technical details about this project can be found in [13].

4. Audio Processing Techniques

This section addresses with more details several issues raised in section 2. We specifically review the main techniques suggested or used for the different audio processing tasks needed in the framework of a digital library system. A detailed technical description of the various techniques is out of the scope of this study, but relevant references are pointed out whenever appropriate.

4.1 Coding of Audio Information

Coding audio data is a signal processing class of techniques referring to digitizing or compressing sound for efficient, secure storage and retransmission. Despite the rapidly growing advances of unlimited data storage capabilities, this processing stage is still necessary because of several facts about audio and multimedia data in general:

- ✓ Multimedia is extremely information rich,
- ✓ Files are very data intensive,
- ✓ Audio and video data are temporal data which need to be played out to users at constant rate,
- ✓ Digitized information contains redundancy, and in the case of multimedia data, defining these redundancies is a very subjective and fuzzy process, and thus very hard to achieve.

Up-to-date technologies in terms of audio compression or coding offer a wide variety of possibilities. They span a wide range going from waveform or *time-domain* coders, to frequency-based or *spectral* coders. These later exploit the non-uniform distribution of audio information across the spectral domain. More elaborate coders are also available on the market, and are known as *vocoders*. They are typically suited for speech signals, and are inspired by the knowledge of the speech production system (e.g. the human vocal tract). In addition, *hybrid* coding schemes are sometimes used by combining some of the above techniques.

The choice of a specific coding scheme depends on a tradeoff between data rate and signal quality. Also, special consideration should be given to the network topology over which the system is implemented. As a matter of fact, most digital library projects are based on local-area networks based on Ethernet connections, with sometimes dial-up links to the system. And due to the huge amount of multimedia information that has to be carried on such connections with an acceptable playing pace, streaming protocols need to be used in order to avoid any clicks or delays in playing the downloaded sound.

Therefore, the choice of coding schemes as well as network configurations is obviously application or system specific. They also depend on the nature of the audio data in the library's database. Different references [6, 7, 9, 12, 18, 21] in the literature address specific techniques in more details, depending on the distinctive features of the system developed.

However, and independently from any network configuration or system structure, there are certain issues that should be addressed when processing audio data. For instance, considering properties of the human auditory system can be very valuable in improving the performance of these techniques [11]. The premise of this approach is that effective audio processing can be realized by exploiting the perceptual tolerance of the human ear to certain acoustic deviations. In particular, some of these audition properties are: the fact that ear is relatively phase-insensitive, the presence of masking effects enabling strong signals to totally mask very low signals, the variability of ear sensitivity to different frequencies as a function of sound loudness …etc. Other advantages can also be drawn from the known redundancies in certain audio signals. For instance, speech signals generated by human speakers are subject to vocal tract limitations; e.g. the vocal tract shape and thus speech spectrum change relatively slowly compared to the sampling frequency.

In this section, we pursue further these potential advantages that can be gained from human perception, and auditory processing models. A test study is presented where certain audio signals are being compressed. The benefits drawn from incorporating auditory models are specifically addressed.

The particular compression test conducted here consists of compressing audio signals[2], using multi-rate processing and sub-band coding[3]. The signals are first decimated to rates where audible distortions are minimized; then the coding stage is performed using a multi-resolution filter-bank. A schematic of the stages of this analysis are shown in figure 1.

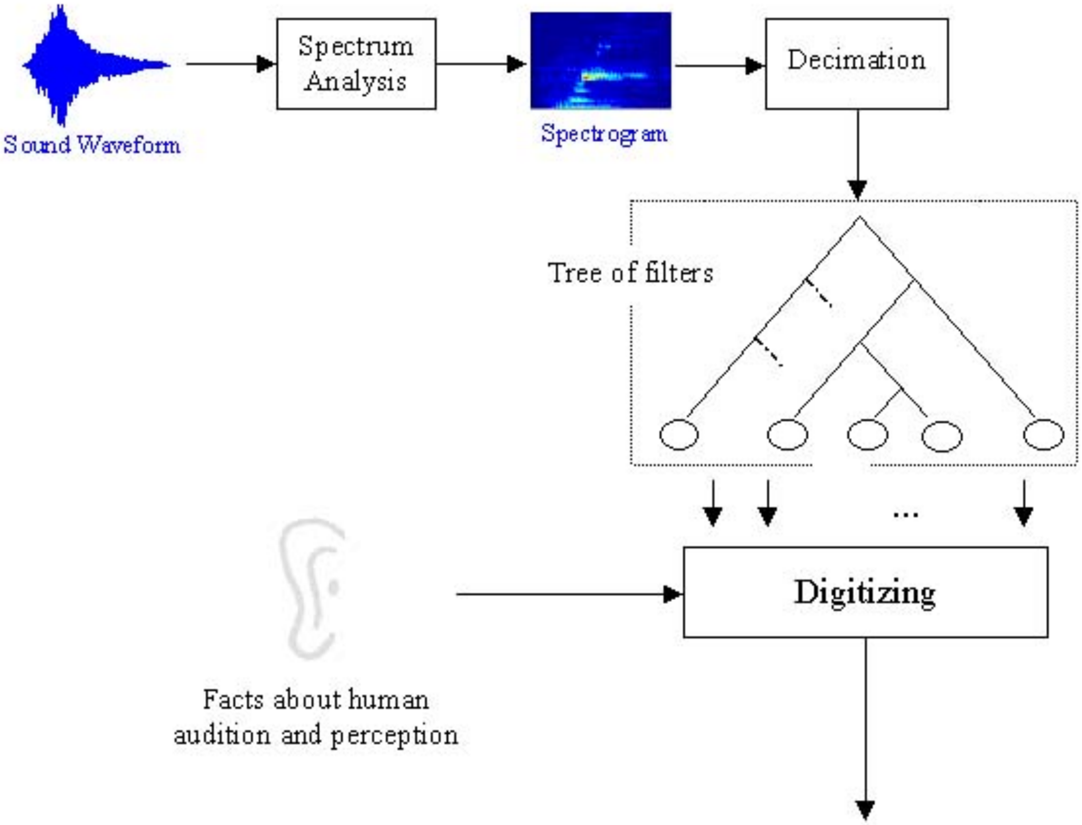

Fig1. *Stages of analysis for the music compression test.*
The model is based on multi-rate analysis and multi-resolution filter-banks.

[2] The audio signals used in this study are music clips originally sampled at 44.1 KHz (Compact-Disk quality). The choice of music clips for this compression task is due to the spectrum challenges of music data compared to other forms of audio signals. It is known that speech signals are focused at bandwidths up to 3 to 4Khz, while for music, ranges up to 7 - 20 KHz are usually preferred, and sometimes required [18]. The high frequency content of music data would certainly put limitations to the compression possibilities.

[3] This compression technique is a quite standard, and non-sophisticated one. But the aim of this study is not the compression scheme itself. Rather, we would like to be able to test the premises of the incorporation of human perception knowledge in improving coding techniques.

Moreover, in order to introduce further refinements to this scheme, i.e. achieve better compression ratios while maintaining an acceptable sound quality; we suggest the use of various human perception facts. Such factors include:

- ✓ Use of a frequency weighting function that varies according to sound loudness. Such perceptual function is given in figure 2.
- ✓ Benefiting from the masking mechanisms in the human ear, by allowing adjacent bands to have relatively high power, and thus masking the quantization errors. This is achieved by spreading the variances in signal sub-bands over adjacent bands.

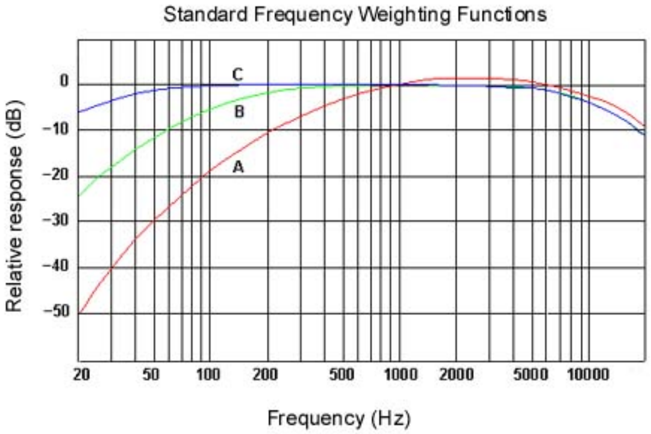

Fig 2. *The frequency weighting functions*[4].

Using these techniques in order to attain compression with minimally perceivable degradation, we managed to achieve rates as low as about 195 Kbps starting from audio signals at CD quality (i.e. about 1.4 Mbps). Incorporating the perceptual weightings did indeed help in improving the compression ratios, while disregarding these stages could only achieve rates of about 600 to 800 Kbps.

In order to assess the quality of the processed data, we perform a listening test with five panelist subjects. Each subject is given both the processed and non-processed sound, and is asked to score the processed signals based on the presence or absence of any audible distortions. These listening tests were used all along this study in order to decide about the cutoffs of the compression rates, in order to achieve minimum distortion compression. The subjects scores are used to compute a Mean Opinion Score (MOS), which is a classical subjective test used in assessing the performance of audio processing systems [8, 18].

4.2 Automatic indexing and Classification of audio materials

Unlike text documents where indexing techniques are based on digital text representations, indexing of audio is a much more challenging task since it is difficult to

[4] This study uses weighting A which is reported to provide the most accurate model of the human perception of loudness.

locate information in large audio materials. A digital library system would necessarily require a separate module for automatic indexing of audio objects. Automating this process is key in order to minimize human intervention, and thus the subjectivity that could be introduced by human classifiers. Different stages are necessary to achieve this task; and no standard algorithm can be presented, since the techniques are very specific to the types of audio objects available in the database. In this section, we present various aspects of some techniques that are considered standard among different applications. A schematic of the different stages is given in figure 3.

The first preprocessing stage that might be needed is segmentation. It is a primitive process needed to segment the data into uniform clips, thus allowing a more flexible classification –within the same file- of the different sounds, possibly different speakers …etc. Effectively, segmentation has to respect the 'natural' flow of the sound, so as to avoid introducing audible crude jumps when listening to the reconstructed signal. This is achieved by setting segment boundaries when an abrupt change in the sound features or content is detected. These features vary depending on the type of audio objects being processed, but they range from acoustic (physical sound behavior) to psycho-acoustic (related to human cognition) cues [14, 22].

The segmented sound is then analyzed for classification. The techniques used for this task are usually performed at two levels: sound type classification, and content-based classification. These methods operate at different levels of abstraction, and are usually complementary of each other. Sound classification refers to categorizing the audio segments into different classes (speech, music, natural/environmental sound, effects –car, bell-, applause, laughter…etc). The variability of the sound acoustic features is used to achieve this classification. Reference [26] suggests a Hidden Markov Model (HMM) based on sound time-frequency analysis for audio classification of environmental sounds.

The other aspect of the indexing process is a more challenging task: Content-based classification. In practice, two general paradigms are being used:

- ✓ A statistical approach [14, 17, 24] which uses features extraction, with an underlying statistical model. Based on a set of patterns for each class, boundary decisions are established in the feature space.
- ✓ A 'black box' approach using neural networks [14], which is basically a non-algorithmic method, and is more suitable for large problem spaces with high-dimensions and complex interactions between the different variables.

It is necessary to point out that both approaches need to be preceded by an analysis stage where feature vectors need to be extracted from the data. These features depend on the classification scheme, as well as the type of audio objects.

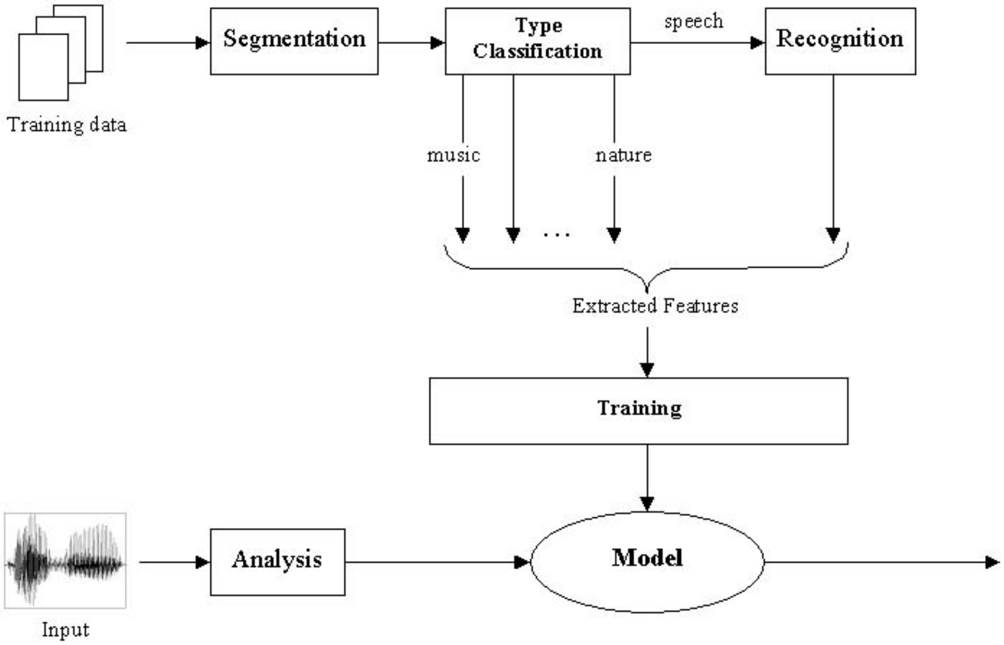

Fig 3. *Schematic of the indexation and classification stages for audio data.*

Finally, it is worth mentioning special issues that concern the class of speech data; which generally constitutes a large proportion of any audio database. Special analyses can be performed on speech allowing the system to offer more flexibility in retrieving information based on full-text searches. Such possibility would not be possible for non-speech data. The speech analysis should then include speech recognition modules, as well as speaker classification and segregation. The introduction of these techniques to the system would allow a more refined classification of the speech data; and thus better performance and flexibility of the retrieval system. Though subjects such as speech recognition and speaker and topic classification[5] are huge areas of research of their own, and various techniques to tackle them have been reported in the literature, we would like to address some of their challenges and difficulties described in some practical experiences. As reported in [2]:

> *Speech on a sound track may be indistinct, perhaps because of background noise or music. It may contain any word in the English language, proper nouns, slang, and foreign words. Even a human listener misses some words.*

These problems limit the performance of the speech recognizer. Even with the use of most state-of-the-art technologies, the Informedia project –presented in section 3- reports error rates ranging from 20 to 50% depending on the characteristics of the speech segment [2, 23]. These rates indicate the need for using additional techniques, along with the recognition algorithms in order to achieve better performance. Such techniques

[5] Topic classification refers to classifying data according to different topic addressed in the document. See [17] for more details.

include source separation algorithms which allow different simultaneous sounds to be separated, and thus reduces the effect of background interference with speech. Other techniques to consider would depend on the known features of the type of audio materials in the system (broadcast news, recordings of public speeches, sound tracks of video…etc).

4.3 Information Retrieval

The flexible indexation of the available audio data is certainly critical in providing users with an array of possibilities to query the sound archives. In the context of audio data, the 'traditional' keyword and full-text queries should be combined with other acoustic features to retrieve data from the system.

As it is certainly the case for normal text data, retrieving information from a large database is a technological challenge. Various techniques to address this issue have been suggested, and are being used (cluster analysis, probabilistic search, distance measures …etc). These same techniques are also appropriate for audio search. But, slight modifications should be considered to include retrieval searches based on non-text data in case of non-speech sound. In this case, the search techniques should be based on acoustic or perceptual features of the sound.

As for speech data, the information retrieval system should be able to use the indexation and recognition techniques to allow full-text searches. One possible concern is the fact that error rates from the recognizer have been reported to be quite high, as mentioned in section 4.2. Still, we should not expect the retrieval system to be highly affected by these errors because of the redundancies in language. Indeed, it is natural in speech that key words would be repeated several times in the same segment. This indicates that despite the misses in the speech recognizer, the probability of correctly picking at least one of the key words is still high.

4.4 Sound playback techniques

An important functionality that a digital library has to offer to its users is the possibility of quickly reviewing or skimming through the searched data. This task is particularly challenging for audio or speech recordings, as there is no natural way for humans to skim through sound information because of the transient nature of sound. As argued in [4, p. 4]: *"The ear cannot skim in the temporal domain the way the eyes can browse in the spatial domain."*

Aside from some signal processing techniques that could be applied (e.g. decimation); which have their own limitations[6], one should rather consider a more perceptually oriented approach where acoustic cues extracted from the data could be used.

In the case of speech for instance, one can exploit the known features and redundancies inherent in speech. Other than the acoustic features of the signal itself, other syntactic and semantic characteristics can also be used; and are indeed among the

[6] Decimation (i.e. reducing the sampling rate) causes shifts in the signal frequency components, and pitch which causes unnatural speech -Mickey Mouse like-

cues used by the human auditory system to understand speech, even when it is severely distorted by noise, or in the absence of entire words or even phrases.

In this framework, this study suggests the use of various techniques –based on human acoustics and perception- to achieve the fast-forwarding task.

a. Time Compression: The most widely used technique to compress sound is Time Scale Modification (TSM) or its variants [1, 10, 25]. It basically involves the adjustment of the speaking rate without affecting the pitch, or any other frequency components; i.e. the resulting signal sounds as if the same person is talking faster or slower with the same voice. Another possible method would be to benefit from the dichotic capabilities of the auditory system, where the human ear is able to integrate information from both ears [3] (see figure 4).

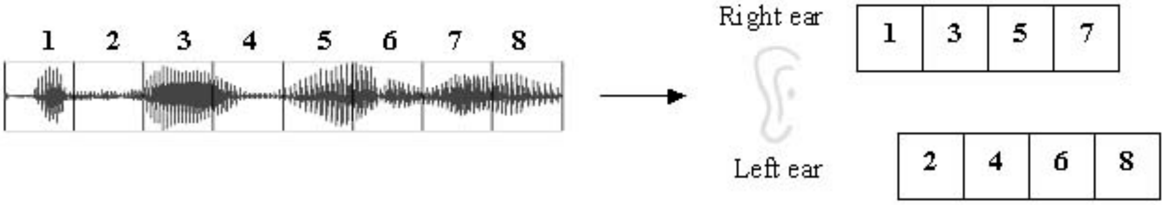

Fig 4. *Presentation of sound to the ears in a Dichotic system.*

This technique, referred to as dichotic time compressed speech (DTCS) was first introduced in the mid 60's by Scott [20]. Dichotic speech does not sound very natural at first, but human adaptation helps reduce the sensation of unusualness over time. However, one should also consider the need for special hardware in order to allow different sound segments to be played to each ear with a certain synchrony.

b. Silence Removal: Silence is usually an important part of audio signals. Still, silence segments could be removed for enabling fast forwarding of the data if they are sustained over extended time intervals. However, this technique should be performed very carefully so as not to confuse silence intervals with pauses between adjacent syllables, which are crucial in the speech recognition task. It is reported in [19] that:

> *Just as pauses are critical for the speaker in facilitating fluent and complex speech, so are they crucial for the listener in enabling him [or her] to understand and keep pace with the utterance.*

A typical silence removal algorithm is implemented by dividing the signal into frames. If the frame energy is consistently lower than a threshold for a long period of time, the corresponding frames can be removed. The choice of the threshold has to be very conservative, as certain consonant sounds have very weak energy; but should not be mistakenly removed.

c. Acoustic Segmentation: The aim of this technique is to segment the audio signal according to its most salient intervals (e.g. the beginning of new topics, sentences...etc). It allows a flexible and interactive browsing of the data, depending on the level of details preferred by the user. In this context, one can use various possible cues such as pauses (between different sentences or topics), speaker identification (to separate talkers in a conversation), prosody, and pitch variations (when emphasizing new ideas, introductions). These acoustic cues can ideally create a hierarchy of segments in the speech signal corresponding to main topics, subtopics …etc.

5. Conclusion

Based on this study, it becomes clear that incorporating audio corpora in digital library systems is certainly not an easy task. This paper gave an overview of the main techniques used in today's audio processing. And in order to consider the implementation of a real system, every topic addressed in this work would have to be carefully analyzed in a more technical study, where specific decisions about the algorithms and approaches to be used have to be made depending on the specificities of the application under consideration. Many references in this paper point out to the major techniques used nowadays, and could be used as basis for further technical details. Also, the special concerns and refinements (especially facts about human audition and perception) addressed in this study should be given some attention; as we showed that they can be important in improving the performance of any system, as well as learning from any shortfalls or challenges faced with in previous projects.

Acknowledgment

The work on music compression presented in section 4.1 was done in collaboration with Cliff Knoll, department of Electrical and Computer Engineering, University of Maryland at College Park.

References

1. *A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, S. Srinivasan, and G. Cohen. "Using audio time scale modification for video browsing," Proceedings of the 33rd Hawaii International Conference on Systems Sciences, 2000.*
2. *W. Y. Arms. Digital Libraries, in series: Digital Libraries and Electronic Publishing. Cambridge, Massachusetts: The MIT Press, 2000.*
3. *B. Arons. "Efficient listening with two ears: Dichotic time compression and spatialization," Proceedings of the Second International Conference on Auditory Display, Santa Fe, pp. 171-177, 1994.*
4. *B. Arons. "Speech-Skimmer: a system for interactively skimming recorded speech," ACM Transactions on Computer-Human Interaction, 4(1) pp. 3-38, 1997.*
5. *The British Library.* http://www.bl.uk.
6. *J. L. Chen, and B. S. Chen. "Model-based multirate representation of speech signals and its application to recovery of missing speech packets," IEEE Transactions on Speech and Audio Processing,* ***5****, pp. 220-231, 1997.*
7. *R. Cox, and P. Kroon. "Low bit-rate speech coders for multimedia communication," IEEE Communications Magazine, 34(12) pp. 34-41, 1996.*
8. *S. Dimolitsas, F. Corcoran, and C. Ravishankar. "Dependence of opinion scores on listening sets used in degradation category rating assessments," IEEE Transactions on Speech and Audio Processing, 3, pp. 421-424, 1995.*

9. *A. Gersho, and E. Raksoy. "Multimode and variable-rate coding of speech," in Speech Coding and Synthesis, W. Kleijn & K. Paliwal (eds). New York: Elsevier, pp. 256-288, 1995.*
10. *S. Ghaemmaghami, S. Sridharan, and V. Chandran. "Speech compaction using temporal decomposition," Electronics Letters, 34(24), 1998.*
11. *O. Ghitza. "Auditory models and human performance in tasks related to speech coding and speech recognition," IEEE Transactions on Speech and Audio Processing, 2(1), 1994.*
12. *B. Gold. "Digital speech networks," Proceedings of the IEEE, (65) pp. 1636-1658, 1977.*
13. *Informedia project. http://www.informedia.cs.cmu.edu.*
14. *M. Liu, and C. Wan. "A study on content-based classification and retrieval of audio database," the International Symposium on Database Engineering and Applications, pp. 339-345, 2001.*
15. *Library of Congress. http://www.loc.gov.*
16. *Library of Congress. A Digital Strategy for the Library of Congress, Committee on an Information Technology Strategy for the Library of Congress, Computer Science and Telecommunications Board, National Research Council. Washington DC: National Academy Press, 2000.*
17. *J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. "Speech and language technologies for audio indexing and retrieval," Proceedings of the IEEE, 88(8) pp. 1338-1353, August 2000.*
18. *D. O'Shaugnessy. Speech Communications: Human and Machine. IEEE press. Second edition, 2000.*
19. *S. S. Reich. "Significance of pauses for speech perception," Journal of Psycholinguistic Research, 9(4) pp. 379-389, 1980.*
20. *R. J. Scott. "Time adjustment in speech synthesis," Journal of the Acoustical Society of America, 41, pp. 60-65, 1967.*
21. *A. Spanias. "Speech coding: A tutorial review," Proceedings of the IEEE, 82, pp. 1541-1582, 1994.*
22. *G. Tzanetakis, and P. Cook. "Multi-feature audio segmentation for browsing and annotation," Proceedings of the IEEE workshop on Applications of Signal Processing to Audio and Acoustics, New York, pp 17-20, 1999.*
23. *H. D. Wactlar, M. G. Christel, Y. Gong, and A. Hauptmann. "Lessons learned from building a terabyte digital video library," Computer, pp.66-73, February 1999.*
24. *E. Wold, T. Blum, D. Keislar, and J. Wheaton. "Content-based classification, search, and retrieval of audio," IEEE Multimedia, 3(3) pp. 27-36, 1996.*
25. *P. H. W. Wong, and O. C. Au. "Fast browsing of speech material for digital library and distance learning," Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS'98, (3) pp. 615-618, 1998.*
26. *T. Zhang, and C. C. J. Kuo. "Hierarchical system for content-based audio classification and retrieval," Conference on Multimedia Storage and Archiving Systems III, SPIE, (3527) pp. 398-409, Boston, 1998.*