Seamless Navigation in Audio Files

Chris J. Wellekens

Department of Multimedia Communications
Institut Eurécom, France
christian.wellekens@eurecom.fr

Résumé

New audio services require editing tools for audio files. Indexing is a solution for fast access to specific information which could be speaker identity, location of speaker intervention on the file, topic identification. Good editing tools for text files have been available for many years and a solution for seamless navigation in an audio file could be the recognition of the content of the file to be edited (speech to text) but this requires in general, large vocabulary speaker independent recognizers giving acceptable results only for cooperative speakers restricting their speech to a domain for which a language model can be learned. Also even in that case, detection of musical chunks, intervention of a given speaker and segmentation in speakers remain interesting challenges. Mastering the complete indexing techniques will open the market for appealing consumer applications producing audio (but also video) programs on demand. Clearly access to multimedia databases and multimedia archives will be easier.

1. Introduction

For many years, text editors have offered a lot of very useful functionalities like "search", "cut and paste" that nowadays, no user could do without. Increase of storage capacities and improvement of networking capabilities together with compression algorithms have boosted the use of audio and video information which are the most natural interaction between human beings and their environment.

We are now facing the problem of fast access to the information contained in these multimedia files.

Audio files are probably more difficult to deal with from the signal processing point of view. Indeed, the human hearing system is so much sophisticated that we are able to understand speech and to recognize speaker's voice in very noisy environments. A model of human hearing is thus extremely difficult to build and to a wide extent, the hearing process is still poorly mastered. Video signals can be more easily processed f.i. in terms of segmentation but the difficulty lays in the more difficult semantic interpretation due to the extreme variability of a scene: different orientations, different scales, different lightings,.... up to different aspects of a same scene (a person wearing glasses or different dresses, bearded, ...) or meaning of sequence of actions.

Audio is simpler from that point of view since the variability is usually not intentional (except for impostors in speaker recognition) and can thus be accounted for with fast adaptation methods.

The development of powerful tools will allow uncountable new applications f.i. for professionals like journalists or physicians (query for images, videos and audio records in large databases) but also for the unexperted customer (visual and audio documentation, TV or audio on demand, automatic analysis of personal phone box). Audio appliance companies are preparing new products able to locally store a huge amount of TV shows. With efficient editing tools, these video and audio materials can be indexed (and so segmented) and video on demand can be built from parts of the stored data. To be user friendly, the whole reconstruction task should be automatic and it thus relies on the quality of content based multimedia indexing which still remains the weak point.

Archiving huge amounts of audio/video information is almost useless without efficient query tools. In audio, a query could be just a spoken request which could be recognized or even immediately compared with recorded information.

Video description of a query could be much more sophisticated since showing one picture of a scene is generally not specific enough to find an acceptable replay of information (a face picture does not allow recovery of all appearances of the person). The problem of archiving is particularly acute for national audio-visual heritage institutes like Institut National de l'Audiovisuel (INA) in France which has the mission to store the whole audio-visual french production (movies, TV shows, audio records,...).

In many cases, the use of audio information can help video indexing. Audio-visual indexing may benefit the reciprocal support of both media f.i. at the level of semantics but also at the lower level of audio-visual speech recognition where lipreading improves the speech recognition scores in noisy environment. The problem of fusion of multiple media informations is still an open problem: indeed, either the features belonging to different media can be combined to transform the recognizer in a single decision tool with hybrid data or decision scores of multiple decision tools can be combined in view of a final decision.

2. Segmentation

A very important problem is the detection of the nature of a signal or more specifically how to segment a signal in chunks having an homogeneous content versus a given criterion (same speaker, music, speech or noise,...). The recent development of the vocal dialing of GSM put the focus on the voice activity detector (VAD) since the limitation of bandwidth imposes to avoid transmission of noise or non-speech and to replace it at the receiver by a locally generated noise. The need of VAD was also already well known for speech enhancement using noise subtraction since the noise spectrum is to be evaluated in noise segments before being subtracted from subsequent spoken segments: the role of VAD is critical and difficult since by definition of this application, it should be efficient even at low SNR. Different techniques have been proposed using the

power of the signal and zero crossing detection. Using power can be implemented as a decision with hysteresis [**?**]: below a given threshold, the decision is "non-speech" and above another threshold the decision is "speech". The ambiguity between the two thresholds is alleviated by using a majority decision from the neighboring frames. In GSM, the few amount of available memory prevents storage of long speech segments required for threshold evaluations.

Another solution is to train a very simple HMM on speech and non-speech in a supervised way and then to recognize speech and non-speech frames. Training of non-speech is critical since speech models are quite characteristic (mainly for voiced segments) while the variability of non-speech is high and may degrade the behavior of the detector; adaptation of the non-speech model is thus recommended for almost each application. A very interesting approach has been proposed recently by Renevey [**?**] where the coherence of speech segments has been recognized as much higher than that of speech. Therefore entropy can be used as a criterion to discriminate between speech and non-speech (higher entropy).

Another application is the separation between music and speech. Here neural networks have been trained and used to discriminate between speech and music and even between different types of music (jazz, pop, classical music,...) [**?**]. Automatic indexing of broadcast news requires this kind of separation as a preprocessing since many jingles are associated with the spoken presentation. In case of background music, source separation would be required but the available information on the records is generally not enough to solve this problem.

We will analyse in a subsequent section the role of the segmentation of a multispeaker speech file in speakers and of speaker identification for indexing as well as for improvement of the recognition rates [**?**].

In a similar way, language detection can also be used to activate the adequate phoneme or word models in a multilingual recognition task and even activate an on-line translation tool which is almost the ultimate goal of speech processing.

3. The Use of Large Vocabulary Speaker Independent Recognizers (LVSIR)

The most straightforward way to solve the problem of audio indexing is to use a powerful speech recognizer and then to search with a text editor. However this is the source of many problems since the recognizer is never perfect mainly if it is large vocabulary and speaker independent.

An interesting application of LVSIR is the transcription of the voice mail in written text which is demonstrated at AT&T research. Real time is not reached but it does not matter in such an application where by definition, delays are expected between recording and reading. From the recognized text, information can be recovered by using standard text editors.

A lot of experiments have also been conducted on the database Broadcast News Hub4 in many labs in Europe, Japon and the USA [**?**, **?**, **?**, **?**, **?**]. Several European projects like THISL were devoted to indexing by recognition.

LVSIR are not designed to deliver information on speaker identity; on the contrary, the recognizer tries to get rid of all information able to discriminate between speakers. We will see in a subsequent section how speaker identification can be useful for indexing tasks.

4. Keyword Spotting

Keyword spotting has been one of the first specialized tools studied for audio information access.

4.1. Garbage models

In the first attempts HMM word models for the keywords were created and trained as well as a garbage model which would stand for all other words except for keywords. The detection score degraded very quickly with the number of possible keywords. All keywords had to be specified in advance. An improvement is to use a dynamic garbage model. The emission probabilities of this model are no longer computed from parametric distributions associated with the states but are for each input vector the average of the emission probabilities of the most probable keyword states but the two or three best ones. The model is said dynamic since it has no own parameters but borrows probabilities from keyword models. It is a representation of an event that, by construction, will not fit so well for keywords as keyword models.

4.2. Phonemic lattices

To be more flexible versus the number of keywords and instead of using an LVSIR at the level of words, acoustic-phonetic decoding can be used. All efforts should then be put on the quality of the phonetic decoding and on the generation of a lattice of N-best solutions [**?**]. The lattice is generated off-line once forever and stored as a companion file of the audio one. A specific word described by its phonetic transcription can be searched for in the lattice and some rules on contiguity and overlapping between phoneme location hypotheses are applied to cope with their inaccuracy. A major advantage of this approach is that no a priori list of keywords is required and any word can be searched for (of course like in text search, very short words are detected everywhere!). However, since there is a wide variability in word pronunciations, the canonic phonetic transcription found in a dictionary is not the only entry to be used. Generation of pronunciation variants are required to increase the recovering rate of a keyword. Different techniques have been tested for the generation of the phonemic lattice including frame labelling [**?**], HMM model [**?**] and the REMAP technique [**?**, **?**, **?**] where use is made of more discriminant emission probabilities generated with an hybrid HMM/ANN phoneme labeller. Tests have been performed on TIMIT but also on CNN sport news. Evaluation criteria are difficult to define [**?**]: indeed the nature of the database plays a very important role as well as the frequency of the searched keyword in that database. ROC's (Receiver Operating Curves) are traditionally used when we have to compare the behavior of a system in terms of false alarms and no-detections (typical for speaker identification). Indeed a system is defined in a plane of these two kinds of errors and the ROC is the locus generated by a threshold variation in the decision system. However here we should be independent on the frequency of the keywords in the text and the false alarm rate is replaced by the probability of false alarms per keyword and per hour. Another criterion in case of sorting by content (i.e. if we try to sort messages in categories according to the keywords) could be the accuracy which is defined as follows. Let us assume that the searched keyword appears N times in the database. For each sentence, the likelihood that it contains the keyword is computed. The keyword spotter sorts the sentences in decreasing order of likelihood. In an ideal system, the first N sentences contain the keyword and the accuracy is 1. Of course,

errors are possible and let us assume that the m-th occurrence of the word appears in the n-th sentence ($n > m$): the accuracy is then n/m for this occurrence. The average accuracy on all occurrences gives a good measure of the recalling capability of the system. It is independent on the size of the database if the frequency of the keyword remains constant along the database. This is no longer true in case of keyword detection: indeed, the larger the database the lower the accuracy. Another efficiency measure is the gain of time in sorting a database: it is defined as the ratio between the number of sentences the user must not listen to the number he should have listened to if he had no keyword detector at his disposal.

5. Speaker information

Speaker identification has been tackled in a tremendous number of approaches and the first subsection does not intend to analyse the respective merits of all of them: a book would not be enough. Only major directions will be outlined and put into relation with the specifications of applications. In a second subsection, the problem of segmentation in speakers of a speech record will be introduced.

5.1. Speaker identification

The goals of indexing are quite different from those of electronic commerce where the identity of a speaker is required to validate a transaction and is thus a speaker verification. Indeed, speaker models in e-commerce can be built in advance from a reasonable amount of data collected off-line. The user claims his identity which is checked against the model. Prompted sentences by the system can circumvent the problem of the password utterance fraudulously recorded. In that case, models of phonemes for each user are trained off-line and the amount of data for enrolment is quite high. Indexing can only use the file to be indexed but on the other hand, the problem of impostors is irrelevant. Several solutions for identification have been proposed:

- vector quantization where each speaker is represented by his codebook; identification relies on the cumulated quantization error [**?**],
- a generalization of the VQ is the ergodic model where transition probabilities control the jump from one centroid to another,
- use of specific HMM trained on the user data [**?**],
- trained neural networks for each speaker require a lot of data for the enrolment [**?**]
- comparison of covariance matrices of the enrolled data and test data with different measures of similarity [**?**]
- comparison with centroids in the vector space generated off-line (eigenvoices) [**?**, **?**, **?**].
- generalized log-likelihood ratio with a world model where each speaker model and a world model are gaussian mixture models (GMM). Training data are recorded in an enrolment phase [**?**].

All these applications require enrolment and the quality of speaker identification is estimated by ROC's but also by the amount of enrolment data.

5.2. Segmentation in speakers

The purpose of segmentation is to segment recorded audio sessions in homogeneous chunks each one containing a single speaker. In a subsequent operation, chunks are grouped to form a database for a given speaker. From these homogeneous (if segmentation and grouping are accurate) chunks, speaker models (GMM) can be trained. Using these models, the audio session can be segmented in speakers exactly as in continuous speech recognition where sentences are segmented in words by using word models and Viterbi alignment.

The first experiments are due to Gish [**?**] who worked on the automatic analysis of dialogs between plane pilots and air traffic controllers. Of course, the different SNR's made the communication quality strongly asymmetric so that equal chunks of the record could be rather easily be identified as "pilot" or "controller" with a resolution of the chunk length. It became clear that for more general applications, other techniques were required and a popular technique is the use of split sliding windows along the audio files. More specifically, Gaussian models of the recorded signal are estimated in two contiguous windows and also for the union of the two windows. The product of the likelihoods of the two windows is compared with the likelihood of the union of the two windows using a BIC (Bayesian Indexing Criterion) which penalizes the representation having globally less degrees of freedom [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. A generalized likelihood criterion is plotted and its maxima are detected under some contraints to avoid artefacts due to numerical errors and locally weak representation of the signal. Their maxima correspond to speaker turn candidates. In a second pass, the segmentation points are confirmed [**?**].

The next step is the grouping of chunks assumed to belong to the same speaker. Since the number of different speakers is unknown, a search in a grouping tree is made and a threshold depending on BIC again decides when to continue grouping (grouping to the leaves of the tree leads to a single speaker) [**?**].

When the grouping is completed, the segmentation of the speech data may be considered as new database from which we can train speaker models (f.i. GMM). With the trained models and a Viterbi a new segmentation can be obtained and iteratively, more accurate speech models can be built together with an improved segmentation. This is exactly the same method as for embedded training of word or phoneme models with a Viterbi algorithm which ends up with a segmentation as a by product.

6. Conclusions

Automatic analysis of audio files has been shown to be a very appealing domain generating many useful applications. Different technologies must converge to improve the content based information retrieval from basic tools like VAD to LVCSR able to transcribe a general speech file. This paper tried to list technologies that may contribute to the creation of audio search engine and to draw an overview of the underlying techniques. Our will was not to be exhaustive and this contribution is widely biased by the research activity of the speech group of Eurécom. The list of references is far from reflecting the efforts done in all multimedia laboratories concerned with this domain crucial for the development of information society.

7. References

[1] Ph. Gelin, C.J. Wellekens, Keyword Spotting for Video Soundtrack Indexing, *IEEE Conf. Acoustics, Speech and Signal Processing, ICASSP-1996* , Atlanta (USA).

[2] Ph. Gelin, C.J. Wellekens, Keyword Spotting Enhancement for Video Soundtrack Indexing,*ICSLP 96*, Philadel-

phia, October 96.

[3] Ph. Gelin, C.J. Wellekens, REMAP for video soundtrack indexing, *ICASSP 97*, Munchen, 1997.

[4] Ph. Gelin, C.J. Wellekens, Keyword Spotting for Multimedia Document Indexing, *SPIE 97*, Dallas, Nov 97.

[5] P. Delacourt, D. Kryze, C.J. Wellekens, Speaker-Based Segmentation for Audio Data Indexing, *ESCA-ETRW Workshop: Accessing Information in Spoken Audio*, Cambridge (UK), April 1999.

[6] P. Delacourt, C.J. Wellekens, Audio Data Indexing: Use of Second Order Statistics for Speaker Based Segmentation, *Proc. ICMCS 99*, Florence, June 1999.

[7] P. Nguyen, R. Kuhn, J-C. Junqua, N. Niedzielski, C.J. Wellekens, Voix propres: Une représentation compacte de locuteurs dans l'espace des modèles, *CORESA 99*, Sophia Antipolis, France.

[8] P. Delacourt, C.J. Wellekens, Segmentation en locuteurs d'un document audio, *CORESA 99*, Sophia Antipolis, France.

[9] P. Nguyen, C.J. Wellekens, J.-C. Junqua, Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environmemt, *Eurospeech 1999*, Budapest, Hungary.

[10] P. Delacourt, C.J. Wellekens, "A first step into speaker-based indexing", *1st European Workshop on Content-based Multimedia Indexing (CBMI'99)*, Toulouse, France, October 25-27, 1999.

[11] P. Delacourt, J.F. Bonastre, C. Fredouille, S. Meignier, T. Merlin, C.J. Wellekens, "Différentes Stratégies pour le Suivi du Locuteur" , *RFIA2000 : Reconnaissance des Formes et Intelligence Artificielle*, Paris, 01-03 Février 2000.

[12] P. Delacourt, J.F. Bonastre, C. Fredouille, T. Merlin, C.J. Wellekens, "A Speaker Tracking System Based on Speaker Turn Detection for Nist Evaluation", *ICASSP-2000*, Istanbul, Turkey, 05-09 juin 2000.

[13] P. Delacourt, C.J Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing", Speech Communication, vol. 32, 2000.

[14] P. Delacourt, C.J. Wellekens, "Regroupement par le locuteur de messages vocaux", *Coresa 2000*, Poitiers, 19-20 octobre 2000.

[15] P. Nguyen, R. Kuhn, J.-C. Junqua, N. Niedzielski, C.J. Wellekens, "A compact representation of speakers in the model space", *Annales de Télécommunications*, nov. 2000.

[16] H. Gish, " Robust discrimination in automatic speaker identification", *ICASSP 1990*, pp. 289-292, 1990.

[17] P. Renevey, Doctoral thesis, EPFL, 2001.

[18] H. Bourlard,Y. Konig, N. Morgan, "REMAP: Recursive Estimation and Maximization of A Posteriori Probabilities. Application to Transition-Based Connectionist Speech Recognition", Internal report of ICSI, TR-94-064, August 1995.

[19] F. Bimbot, Y. Magrin-Chagnolleau, L. Mathan, "Second order statistical measures for text independent speaker identification", Speech Communication, vol. 17, nos 1-2, pp.177-192, Aug. 1995.

[20] S. Chen, J.F. Gales, P. Gopalakrishnan, R. Gopinath, H. Printz, D. Kanevsky, P. Olsen, L. Polymenakos, "IBM's LVCSR system for transcription of broadcast news used in the 1997 HUB4 English evaluation", *DARPA Speech Recognition Workshop*, 1998.

[21] F. Kubala, H. Jin, L. Nguyen, R. Schwartz, J. Makhoul, "The 1996 BBN Byblos Hub-4 transcription system", *DARPA Speech Recognition Workshop*, 1997.

[22] P. Woodland, J.F. Gales, D. Pye, S. Young, "The development of the 1996 HTK Broadcast News transcription system", *DARPA Speech Recognition Workshop*, 1997.

[23] M. Harris, X. Aubert, R. Haeb-Umbach, P. Bayerlein, "A study of broadcast news audio stream segmentation and segment clustering", *Eurospeech 1999*.

[24] P. Delacourt, "Indexation de données audio: segmentation et regroupement par locuteurs", Thèse doctorale, Institut Eurécom, 2000.

[25] P. Gelin, "Détection de mots clés dans un flux de parole: application à l'indexation de documents multimedia", Thèse doctorale, Institut Eurécom, 1997.

[26] C. Montacié, M.-J. Caraty, "Sound Channel Video Indexing", *Eurospeech 1997*, pp. 2359-2362, 1997.

[27] T. Matsui, S. Furui, "Comparison of text independent speaker recognition method using VQ-distortion and discrete/continuous HMMs", *ICASSP*, 1992

[28] D.A. Reynolds, "Speaker identification and verification using Gaussian mixtures models", Speech Communication, vol. 17, pp. 91-108, 1995.

[29] J-L. Gauvain, L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 HUB4-E transcrption system", *DARPA Broadcast News Workshop*, 1999.

[30] Y. Bennani, P. Gallinari, "Connectionist methods for speaker verification (Tutorial)", *ESCA Workshop on Speaker Recognition, Identification and Verification*, Martigny, 1994.