Diarization for the Annotation of Legal Videos

Ciro Gracia[*], Xavier Binefa[*], Emma Teodoro[°] , Núria Galera[°]
[*]*Universitat Pompeu Fabra, Barcelona, Spain*
[°]*UAB Institute of Law and Technology, Universitat Autònoma de Barcelona, Spain*

across these records. This tool is based on data recovering the legal structure of hearings. To grasp this structure automatically, we apply and compare different audio diarization algorithms to obtain the temporal boundaries of the speakers and their tracking across the hearing. Previous work on legal data will help us to apply diarization techniques into web services platforms (Ontomedia).

Keywords: Legal Multimedia, Diarization, Speech analysis, Gaussian Mixture Models

1. Introduction

For some years now, all civil hearings are tape-recorded in Spain. Under the Spanish Civil Procedure Act (1/2000), these videotapes become official records. All proceedings must incorporate them, and lawyers have to obtain them from the court, as they are distributed to the parties by the Court clerks.

Unfortunately, the content of the CDs is not generated to feed the existing legal large databases (public or private), and there are no tools in the market yet to help lawyers and judges with their daily management. They are forced to store and retrieve this multimedia information from the CDs they receive with any single case, one by one.

To make a prototype for video retrieval and classification, a detailed analysis of the legal structure of the hearings may provide information about which are the potential semantically rich patterns. Recognizing these patterns allows not only a more efficient navigation for query and referencing, but also a higher level interface that speeds up case analysis. In this paper we present the annotation and navigation tools, based on the recovering of the legal proceedings from the oral hearings. We also discuss some approaches to audio diarization as a first step to extract key metadata for efficient navigation. This paper follows up the analysis already performed elsewhere (Binefa et al. 2007, Casanovas et al. 2009). Research projects around Ontomedia allowed us keep researching on legal video segmentation and audio diarization. We will summarize some lessons learned in court hearing analyses, to apply automate methods to extract information from the Ontomedia interactive mediation processes.

Automated information extraction constitutes the challenge we face in this paper.

2. Defining semantically rich patterns in Legal Multimedia

As IDT researchers discovered during the acquisition knowledge process, Spanish civil oral hearings are not being recorded following documented guidelines. Written procedure acts, such as the Spanish LEC, provide a judicial program and agenda that are only partially followed in Court. Furthermore, all recordings do not exhibit the same properties, as on the whole they do not suggest that there is a formalized standard.

Technical implementation of recoding varies from court to court, and it is obvious that sound and video quality criteria should be optimized for storage. Usual quality configuration includes 8kHz mono sound and 320x288 pixels resolution on a wide field of view image encoded at 8 frames per second. Despite different technical quality configurations varying from court to court, usually civil hearing recordings share the camera point of view, and the common court room distribution. The camera position and orientation always ensures that neither the judge nor the rest of the court members appear in the camera field.

Inside this multimedia representation of the hearings, there are several semantically rich patterns. These patterns must provide enough context information to allow fast navigation and querying. Most of the valuable information comes from the structural elements of the hearings. Oral hearings are a moderated environment where certain procedures and rules govern the sequence evolution.

This feature can be exploited to define which existing patterns are useful to infer the legal structure of the hearing. Three major elements perform a powerful improvement for navigation capabilities of legal multimedia: (i) interventions, (ii) phase inference, and (iii) specific terminology spotting. Hearings are primarily characterized for being mostly a moderated environment where each participant has specific taking turns.

These spoken interventions provide a first structure level on the multimedia data. Interventions can be defined as a "someone is talking". In that sense, interventions provide: (i) a graphical representation which contains semantical meaning related to the speaker's specific roles; (ii) the representation of the oral hearing structure regarding the order and duration of the interventions. For example the cross-interrogation by lawyers to witnesses could be easily identified as the edition of a set of interrelated interventions.

2.1. KNOWLEDGE ACQUISITION

Stemming from the official videotapes, we worked on detailed transcriptions of 15 different civil oral hearings. We set up a focus group of 3 researchers and 3 professional lawyers and solicitors. Using standard eliciting techniques (namely, interviews, e-mail discussions and brain-storming), focusing on their legal professional knowledge (LPK), we obtained expressions and legal concepts actually used in court. We inserted these procedural and legal expressions within the different stages of the structured hearings. As a result of applying eliciting techniques, researchers and lawyers identified roughly 900 terms and legal practical expressions.

Our analysis of the audiovisual records shows that the actual development of civil hearings differs considerably from the provisions made by the law (i.e. we detected stages of the hearings that were not identified by de 1/2000 Act). The structure of the general proceedings contained in procedural statutes and regulations is not the same structure implemented in court. The procedural dynamics of hearings as they are performed follows patterns and rules shaped by practical constraints and routines.

To face these difficulties, we represented four types of oral hearings in workflows. Workflows are conceived as flexible and adapted structures able to contain (and interact with) other structures of practical knowledge (i.e. typologies, keywords lists, and content groups).

The basic units of workflows are procedural stages. Procedural stages may include: actions (red rombus) which imply to move forward or backwards to another stage; procedural legal concepts (green boxes); and content of legal concepts (orange boxes). Table I shows a typology of the civil hearings, and Figure 1 represents e.g. the development of hearings of precautionary and provisional measures (*Medidas Cautelares y Provisonales*)

Table I. Table of Spanish civil hearings

Preliminary hearing ***Audiencia previa***	Preliminary hearing *Audiencia previa*
Ordinary proceedings ***Juicio ordinario***	Ordinary proceedings *Juicio ordinario*
Verbal proceedings ***Juicio verbal***	Verbal proceedings *Juicio verbal*
	Proceedings regarding the capacity of persons (special proceedings) *Procesos sobre capacidad de las personas (procesos especiales)*
	Matrimonial proceedings (special proceedings) *Procesos matrimoniales y menores (procesos especiales)*
	Judicial division of states *División judicial de patrimonios*

	Exchange proceedings (special proceedings) *Procedimiento cambiario (procesos especiales)*
	Summon proceedings (special proceedings) *Procedimiento monitorio (procesos especiales)*
	Mortgage proceedings *Procedimientos hipotecarios*
	Extraordinary appeal for procedural infringements and cassation proceedings *Recurso extraordinario por infracción procesal y casación*
	Opposition to judicial determination of fees for undue fees *Impugnación tasación de costas por honorarios indebidos*
	Opposition to enforcement *Oposición a la ejecución*
Precautionary and provisional measures ***Medidas cautelares y provisionales***	Precautionary measures *Medidas cautelares*
	Provisional measures *Medidas provisionales*

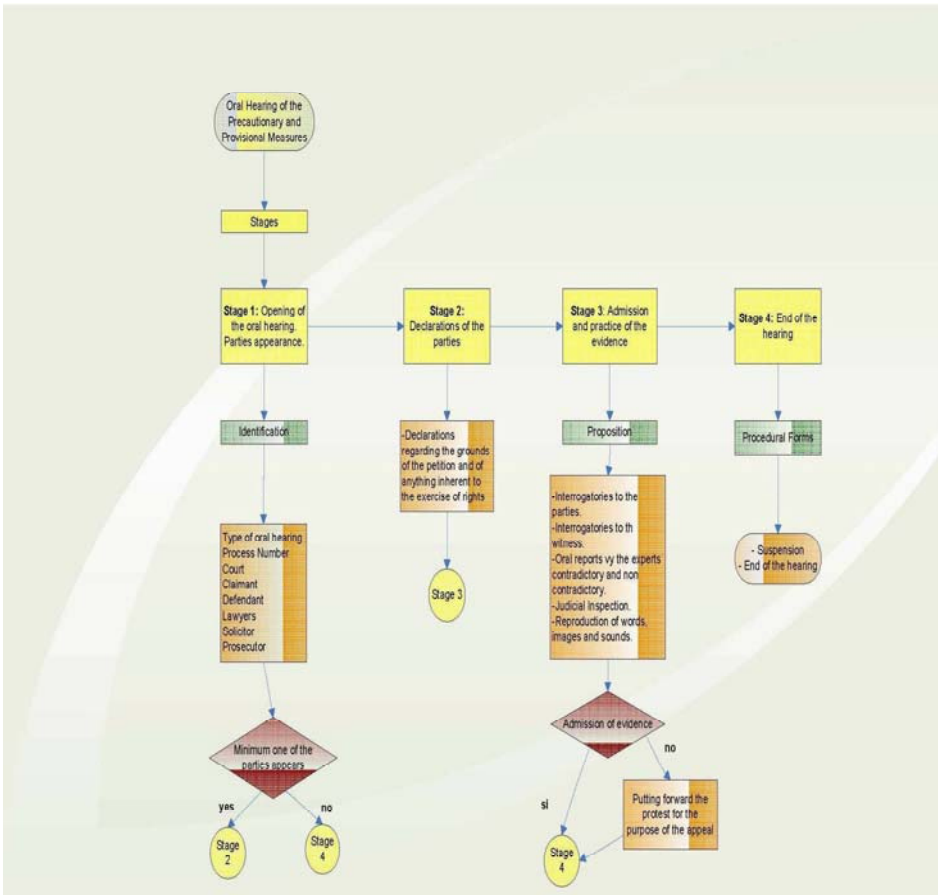

Figure 1. General structure of the Precautionary and Provisional hearings.

Stages can be used for group interventions into buckets. Each of them represents a specific part of the hearing procedure and provides a legal contextualization of the procedure development and its characteristics. For instance, in an overview of a hearing where stages are graphically represented and easily identifiable, the presence or absence of some stages (for example, practice of means of evidence) and their durations (number of interventions or minutes) provide information for navigation, hearing characteristics and analysis. The legal domain, as many other knowledge areas, includes specific terms which can be tracked on the speech signal.

Many elements recall on the acoustic signal as information source as far as video signal carries poor information. Some events can be easily recognized from video signal, e.g. elements that include movement or important changes in people's position. For example, when a witness enters the court room to testify or when a lawyer approaches to the judge's table to show documentation. However, the real challenge comes from their automatic extraction from the records.

2.2. INFORMATION EXTRACTION

The objective of this section is to discuss technical alternatives for extracting the three major elements described previously (interventions, phase inference, and specific terminology).

Detecting legal vocabulary from speech signal can be seen as a procedure referred in the signal processing literature as "word spotting" where an automatic speech recognition system (ASR) is trained to recognize a set of specific (small vocabulary) words (W.Kraaij, 1998). Acoustic template models can be also acquired by the navigation interface and searched exhaustively on the record producing a similarity function from which a set of plausible candidates can be suggested (Rabiner, 1993). More general approaches use general domain ASR systems for translate acoustic signal into text and further applying text processing techniques.

Stage inference requires a model of the legal casuistic obtained from empirical data which can take the form of a flow chart or automata. After a set of empirical observations, a set of Stages characteristics and the set of events that characterize the transitions between them are built, and a deterministic or stochastic model is stated ad hoc.

Other possibilities include the description of multimedia as a rich set of features and use automatic learning for the statistical differences between stage sequences. Similar techniques are often applied in human sequence evaluation, although the features defining stages should be very high level in order to obtain generalizable results (Cai, 2008; Peeters, 2003).

Interventions represent a speaker segmentation task which can be diarized. It will be described in depth in the next section.

2.3. THE PROTOTYPE SYSTEM

To explore the possibilities that improve navigation, we have developed a media exploration user interface that is heavily supported by a graphical visualization of speaker's indexing of the oral hearing. The user interface is based on a graphical representation of the interventions and the segmentation of the stages.

Figure 1. Media exploration graphical user interface. Horizontal segments represent speaker interventions. Vertical Column segments represent Phases presents in the media. Iconic representations of speakers help to filter interventions.

The interventions are represented as a set of horizontal segments, each one in a color that identifies the speaker owner. Stages are represented as columns with different sections, each one representing a filter operation on the displayed interventions. Different speakers can be identified by color or by an iconic representation that allows filter interventions by just displaying the speech segments of a given speaker. This graphical interface allows easy navigation and the basis for further annotation or documentation handling.

We have described it as an example of graphical exploitation of the patterns listed in the previous section. We have found that this kind of interfaces is appreciated by this specific group of users, so we have encouraged implementing speaker diarization as one of the basic elements for media annotation.

3. Diarization

The goal of diarization is to locate homogeneous regions within audio segments and consistently label them for speaker's identity, background noise, music, etc (Angueraphd, 2004). This labeled media finds applications in numerous speech processing tasks, such as speaker adapted speech recognition, speaker detection, speaker identification, and audio document retrieval.

In a legal multimedia context, this task can be seen as the process which detects speaker's turns and regroups those uttered by the same speaker. The scenario application does not involve music and other types of sources. Usually, signal processing literature often refers to this simplified version of diarization as speaker segmentation and typically subdivides the problem into a subset of problems and techniques depending on the prior knowledge of the environment. Most of the supervised approaches use previous knowledge on the number of the speakers, the infrastructure calibration and samples from speakers' voices to build models for later classification and selection. Blindly approaches do not assume any prior information and use metric approaches and clustering to estimate turn boundaries, the number of speakers and regroup speaker segments (Hyoung, 2005).

3.1. GENERAL OVERVIEW

To determine when each participant is speaking in a recording implies detecting boundaries between interventions, as well as identifying the speaker between these boundaries. Preliminary speech detection must be performed to segregate speech content from non speech content such as background noise in order to avoid further misdetections and distortions. The issues related to the segments of diarization are twofold: detecting the audio boundaries of the speakers and recognizing the segments belonging to the same speaker. From now on, we will refer to them as intervention segmentation and speaker identification. Detecting boundaries between speaker's interventions can be faced following two approaches which depend on the prior knowledge of the environment. On the one hand, metric-base methods measure the similarity between two neighboring speech regions and try to determine where an abrupt change in the speech characteristics is detected. This approach estimates boundaries where abrupt changes appear, usually pointing to speaker turns or a change point between speakers by finding homogeneous acoustic regions (Rodriguez, 2007). Given the set of segments, each one is assumed to belong to a sole speaker, so that the process to determine the number of speakers and which segments belong to the sole speaker must be faced without prior knowledge. The unsupervised approach classically faces this problem by hierarchical bottom-up clustering techniques, where a distance or dissimilarity criterium between each pair of segments is computed and the method puts together the two clusters which are closest (most similar), resulting in a new partition of the data. This process is computed in an iterative fashion until a certain stop criterium is reached (Hyoung, 2005). The number of estimated speakers corresponds to the resulting number of

clusters, and the segments belonging to each speaker are the different segments belonging to each cluster.

On the other hand, model-based segmentation relies on statistical models, e.g. Gaussian Mixture Models (GMM). These models are constructed for a fixed set of acoustic classes, such as anchor speaker, music, etc. from a training corpus. The incoming audio stream can be classified by maximum likelihood selection. Segment boundaries are assumed where a change in the acoustic class occurs (Kemp, 2000).

3.2. IMPLEMENTATION AND ALTERNATIVES

Diarization is a process usually applied to multimedia data as an offline process, but results can be easily improved adding specific control element to the recording environment. In legal multimedia, oral hearings usually involve a small and quasi fixed amount of roles: judge and court staff, lawyers of both sides and an unknown number of third parties which play a similar semantic role (witnesses). This relative small amount of elements and its variability allow to incorporate specific microphones and record each channel separately, making possible an easy online indexing just by activity detection on each channel.

Generally, diarization requires detecting activity as a preprocessing step, non speech noises add disturbance to posterior processes increasing the overall error, the number of participants in the media also increases error and processing time as mismatch in speaker identification tends to increase with the number of different speakers (population increases) (Reynolds, 1992; Reynolds, 1995). Having the possibility to spread different recording channels among different roles or participants allows increasing posterior diarization accuracy as well as an easy online indexing.

4. Speech Analysis

Speech signal contains patterns inside its temporal variability, and such patterns extend over a very wide range of time scales. This variability patterns are much clearer to analyze in frequency domain. Short time Fourier transform is a major tool for speech frequency domain analysis producing a representation commonly called spectrogram. One of the most successful approaches to sound signal analysis for segregation and recognition of acoustic events is the one that aims to analyze and reproduce the natural hearing process of human auditory system or also called human perception.

The study of human auditory system aims at determining the factors that allow humans to solve difficult situations where they are able to segment,

segregate, recognize, track and synthesize individual sources from a mixture of channels of sound streams (Hayking, 2005). Nowadays the signal processing techniques that extract human perception adapted features are the most used in speaker recognition/identification. Most of these features are based in Filter bank processing techniques as human auditory system behavior can be modeled by a set of critical band filter. Mel frequency cepstral coefficients (MFCC) are probably most common features extracted from speech signal for speech recognition and speaker recognition. Their physco-acoustic approach is based in the melodic scale of frequency which was proposed by Newman and Volkman in 1937. Melodic scale represents the perceptual scale of pitches judged by listeners to be equal in distance from one to another. Mapping between melodic scale and Hz is defined as:

$$Mel(hz) = 2592 * \lg_{10}(1 + \frac{hz}{700})$$

MFCC perform a band pass filtering on power spectrum where an optional number of band pass filters are placed equally into Melodic scale warping by this way Fourier frequency axis (Roch, ; Zhang, 2003; Stern,2006).

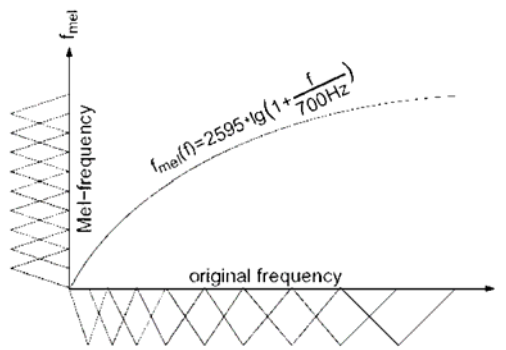

Figure 2. Frequency axis warping by melodic scale , filters equally spaced into Melodic scale results in logarithmic filters in Hertz frequency scale.

A common configuration can be found as consisting of 24 triangular shaped filters covering audible frequency range. The resulting coefficients are derived to Cepstral representation by taking the Discrete cosine Transform. Cepstral representation (Bogert, 1966; Childers, 1977; Oppenheim, 2004; Roch) refers to a representation where periodicity inside power spectra is analyzed, this analysis is closely related to homomorphic system theory developed by Oppenheim in middle sixties applied to the Source-filter model of speech production.

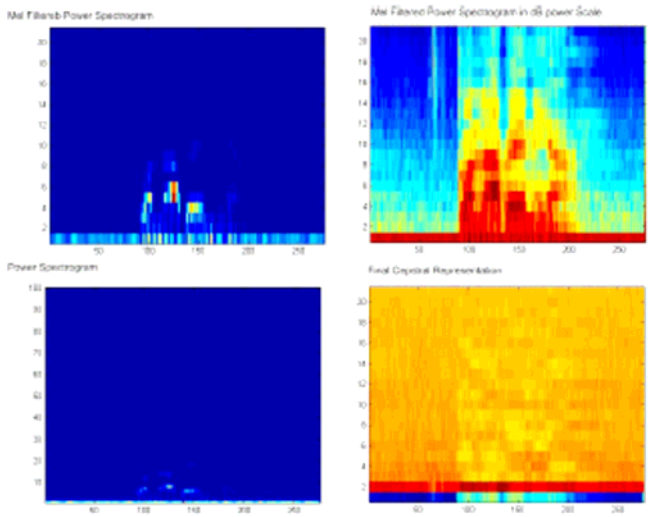

Fig. 3. MFCC Processing by Images

Posterior dimensionality reductions techniques can be addressed to keep most of the significant information while reducing noisy information. In our approach we have chosen framework for speech signal processing based in a 128-Melodic bank filtering and a posterior Principal component analysis in order to keep 10 principal dimensions to perform posterior training and classification.

5.1. SEGMENTATION

Segmentation has the objective of detecting speaker change point. Inside every speech region, a distance is computed between two adjacent given length windows. By sliding both windows on the whole audio stream, a distance curve is obtained. Peaks above a threshold in this curve are thus considered as speaker change points.

Most of the common distance metrics for change point detection are based on model Gaussian distribution for data on both adjacent windows. Methods try to compare both sides in terms of distribution overlapping, hypothesis testing or by model selection criterion. Classic metric functions are Kullback-Leiber, hostelling T2, Generalized likelihood ratio and Bayesian Information Criterion (Kemp, 2000). We have chosen Crossed BIC as an efficient implementation of BIC for segmentation, which has

been reported adequate for broadcast news diarization (Rodriguez, 2007; Anguera, 2004).

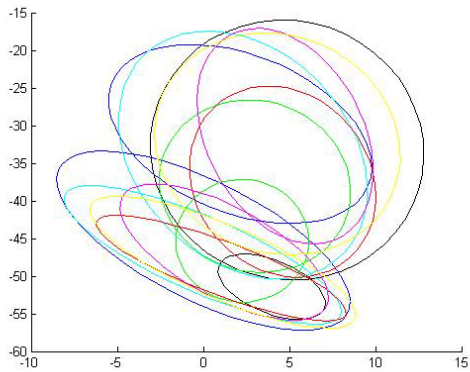

Figure 4. Sample Speakers Gaussian mixture modeling, it shows a high degree of overlapping between models.

Our Model Based segmentation is performed by using 15 seconds of speech samples from each participant in the media. A well known widely applied statistical model is used for approximate probability density for each participant acoustic features. This learning process in based on a Gaussian mixture modeling, followed by consecutive pruning stages to produce an acoustic modeling based only in each speaker's specific characteristics.

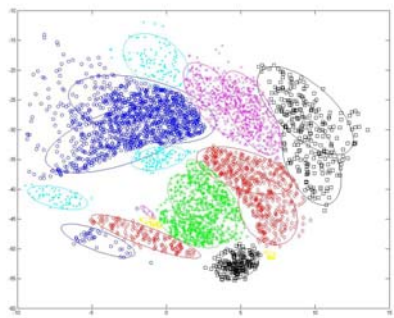

Figure 5. Sample Speakers Mixture Modeling after pruning

The selective use of feature vectors (Shrikanth, 2006) consists on reducing the overlap between speaker models. This normally implies discarding features that are common in speaker training data sets like some short silences or some turbulent phonemes. The influence of such common features to model overlap to model overlap is especially critical when the

training utterance length is limited. A method to reduce model overlap consists of generating the models and filtering the features of the next learning set, just leaving features that are correctly classified using maximum likelihood methods, and recomputing iteratively the model until the accuracy achieves the required level. Discarded features are keep to build a complementary model, during maximum likelihood classification feature vectors belonging to this model will be discarded. This approach pretends to make robust decisions by selecting only robustly assignable feature vectors.

4.2. HYBRID MODEL

Using a hybrid model, the diarization becomes a two step process: segmentation and identification. From the segmentation point of view, metric approaches are interesting as they are more efficient to compute and do not make prior assumptions of the acoustic classes. While model approaches are more precise, they do not generalize acoustic conditions absent in the models. Identifying which of the speech segments are owned by the same speaker can be seen as a speaker identification task, in which taking advantage of the prior build models, a maximum likelihood classification task is performed. Grouping same speaker with neither prior information nor assumptions typically require working with long utterances to ensure sufficient statistics (Viet-Bac Le, 2007). Most of them are related with covariance estimation of dissimilarity criterions and all the data must be available before grouping can start. However, the prior speaker model requirement is acceptable given that Speaker Identification techniques can provide a very high accuracy, allow online operation and do not require such long utterances.

4.3. EXPERIMENTS

We performed experiments evaluating the diarization error (DER) as the percentage of wrongly assigned time versus total time. Experiment corpus compromises files of legal media content. Inside these data files 7 speakers (3 females and 4 males) participate in 60 minutes of audio signal. Speakers change point detection obtained from the diarization are compared to manually annotated boundaries, considering a margin of half a second of tolerance. We show precision (PRC) and recall (RCL) measures coming from standard information retrieval techniques to evaluate the performance of both methods for boundary detection: hybrid and model. Recall takes into account the completeness of the retrieval and precision measures of the purity of the retrieval.

Method	Diarization Error
Metric segmentation	0.15
Model segmentation	0.18

Method	Boundary Precision	Boundary Recall
Metric segmentation	0.66	0.6
Model segmentation	0.36	0.53

Generated boundaries and labeling were incorporated to the navigation interface and were evaluated by final users on navigation task over tested recordings. The evaluation of diarization applied to the navigation interface resulted in helpful information to further improve the exploring and navigation capabilities of the final users on this media application.

5. Conclusions

In this paper we have presented an overview of legal multimedia documents obtained from civil oral hearings in Spain. We have also presented an analysis of its potential to boost legal professionals' capabilities on case management and hearing analysis. We defined which patterns inside the media can provide the best legal structure description and we presented an interface for its graphical presentation. We have introduced diarization as a powerful information source for navigation and a general overview of methods for its computation. We have also summarized some results from the hybrid approach to further develop the navigation interface.

Acknowledgements

This work has been developed within the framework of two different projects: (i) ONTOMEDIA: Platform of Web Services for Online Mediation, TSI-020501-2008); (ii) ONTOMEDIA: Semantic Web, Ontologies and ODR: Platform of Web Services for Online Mediation (CSO-2008-05536-SOCI). These projects lean on the results of FIT-350101-2006-26.

References

Anguera, X. (2003). *Robust Speaker Segmentation and Clustering for Meetings*, presented as PhD Thesis proposal in UPC (Barcelona, Spain).

Binefa, Xavier; Gracia, Ciro; Monton, Marius; Carrabina, Jordi; Montero, Carlos;Serrano, Javier; Blzquez, Mercedes; Benjamins, Richard v.; Teodoro, Emma;Poblet, Marta; Casanovas, Pompeu. *Developing ontologies for legal multimedia applications* (2007). In P. Casanovas, M.A. Biasiotti, E. Francesconi, M.T. Sagri (Eds.) "Proceedings of LOAIT 07 II Workshop on Legal Ontologies and Artificial Intelligence Techniques", pp. 87-102.

Bogert, B. P., Healy M.J.R., Tukey, J.W. (1984), *The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe-cracking*. In Rosenblatt, M. (Ed.), "Proceedings of the Symposium on Time Series Analysis", Wiley, New York, pp. 209-243.

Cai, R., Lu, L., Hanjalic, A. (2008). *Co-Clustering for auditory scene categorization,* IEEE Transactions on multimedia.

Casanovas, P., Binefa, X., Gracia, C., Montero, M., Carrabina, J., Serrano, J., Blzquez,M., Lpez-Cobo, J.M., Teodoro, E., Galera, N., Poblet, M. (2009), *The e-Sentencias pro-totype. Developing ontologies for legal multimedia applications in the Spanish Civil Courts*. In Casanovas, P.; Breuker, J.; Klein, M.; Francesconi, E. (eds) "Channeling the legal informational ood. Legal Ontologies and the Semantic Web", IOS Press, Amsterdam, pp. 199-219.

Childers, D.G., Skinner, D.P., Kemerait R.C. (1977), *The Cepstrum: A Guide to Processing*, "Proceedings of the IEEE", Vol. 65, No. 10, October, pp. 1428-1443.

Haykin, S., Chen, Z. (2005). *The Cocktail Party Problem Review*. Neural computation, Num. 17.

Hernando, J., Anguita, J. XBIC (2004). *Nueva medida para segmentacion del locutor hacia el indexado automatico de la señal de voz*. In "III Jornadas en Tecnologías del Habla".

Hyoung-Gook, K., Moreau, N., Sikora, T. (2006). *MPEG-7 Audio and Beyond Audio Content Indexing and Retrieval*. John Wiley and Sons.

Kemp, T., Schmidt, M., Westphal, M., Waibel, A. (2000). *Strategies for automatic segmentation of audio data*. Proc. Icassp

Kinnunen T. (2003), *Spectral Features for Automatic Text-Independent Speaker Recognition*, Joensuun Ylipisto.

Kraaij, W., van Gent, J., Ekkelenkamp, R., and van Leeuwen, D. (1998*), Phoneme based spoken document retrieval*. In "Proceedings of the fourteenth Twente Workshop on Language Technology TWLT-14", University of Twente, pp. 141-143.

Kwon , S, Narayanan, S. (2006), *Robust Speaker Identification based on selective use of feature vectors*, Pattern Recognition Letters, Num. 28 , pp. 85-89.

Kwon, S., Narayanan S., (2005, *Unsupervised Speaker Indexing Using Generic Models*. IEEE Transactions on Speech and Audio Processing, Vol. 13 No. 5 September.

Le, V.B., Mella, O., Fohr, D. (2007), *Speaker Diarization using Normalized Cross Likelihood Ratio*. In "10th International Conference on Speech Communication and Technology (Interspeech-Eurospeech07)".

Oppenheim, A.V., Schafe R.W. (2004), *From Frequency to Quefrency: A history of the Cepstrum*. Dsp History in IEE signal processing magazine

Peeters, G. and Rodet, X. (2003), *Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instrument databases*. "Proc.of the 6th Int. Conference on Digital Audio EFFects (DAFX-03)", London UK.

Rabiner, L.R., Juang, B.H. (1993), *Fundamentals of speech recognition*, Prentice Hall.

Reynolds A., Rose C (1995), *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Model,* IEEE Transactions on Speech and Audio Processing Vol. 3.

Reynolds, D.A. (1992), A *gaussian mixture modeling approach to text-independent speaker identification*, PhD thesis, Georgia Institute of Technology, August.

Roch, M. *Cepstral Processing Lectures*. San Diego State University.

Rodriguez, L.J (2007), *A Simple But Effective Approach to Speaker Traking in Broadcast News*, LNCS 4478, "Third Iberian Conference" IbPRIA, Vol.2, pp. 48-55.

Stern, R.M., (2006), *Signal processing for robust speech recognition*. Department of Electrical and Computer Engineering and School of Computer Science Carnegie Mellon University.

Yu Zhang, Y. (2003), Presentation. Seminar Speech Recognition - Mel-spectrum computation.