Automatic Metadata Extraction from Spoken Content using Speech and Speaker Recognition Techniques

Héctor Delgado, Javier Serrano, Jordi Carrabina

Software-Hardware Prototypes and Solutions Lab, Autonomous University of Barcelona, Spain
hector.delgado@uab.cat, javier.serrano@uab.cat, jordi.carrabina@uab.cat

Abstract

Today information extraction plays a significant role in management of massive data quantities for different purposes. One of the open challenges in this field is the automatic extraction of information from audio streams. This paper describes a useful metadata extraction system which performs a powerful combination of speech and speaker recognition tasks. The system carries out the speech transcription through a Catalan language recognizer based on Hidden Markov (HMM) tied-state cross-word triphones acoustic models, Mel Frequency Cepstral Coding (MFCC) and N-gram language modeling. In addition, a speaker diarization is performed using HMM based segmentation and Perceptual Linear Prediction (PLP) feature extraction. Both speech-to-text transcription and speaker diarization can be utilized as annotation data for multimedia content. In order to make indexing and retrieval more flexible and efficient, the extracted metadata is stored using the MPEG-7 multimedia content description interface. The system has been successfully tested on the recordings of the plenary sessions of the Catalan Parliament.

Index Terms: Metadata extraction, Automatic speech recognition, Speaker diarization, GMM, HMM, MPEG-7

1. Introduction

Multimedia content volume has increased hugely since storage capacity has become almost unlimited, and information technologies has extended widely. Content management, indexing and retrieval are key challenges that are increasingly becoming more difficult, due to the massive data quantity. Since manual annotating of such massive content is not viable, it forces content to be annotated automatically through information extraction approaches.

Speech inside audio streams is a huge information source where data may be obtained from, in an automatic way. Speech technologies are able to extract different kinds of information from audio. Instead of manual annotation of audio information, speech technologies offer automatic extraction which is significantly efficient in terms of time and accuracy, particularly in huge amount of repositories. Spoken content automatic transcriptions of the audio/video streams are also suitable for automatic on-line or off-line subtitling, word spotting, or as support for either foreign or hearing impaired people. On the other hand, automatic speaker diarization allows direct access to the parts where particular speakers participate.

This paper describes a useful metadata automatic extraction system which focuses on the spoken content. The system performs a powerful combination of speech and speaker recognition tasks. The metadata extraction system carries out a speech transcription by means of a Catalan recognizer based on cross-word tied-state triphone HMMs, MFCC and N-gram language models for Catalan language. In addition, a speaker diarization is performed using HMM based segmentation and PLP feature extraction. For both transcription and speaker diarization, a comparison of performance has been performed with different configurations, in order to improve results. Once the information has been extracted, it is automatically stored using the MPEG-7 content description interface.

The system may be used in different application domains. Concretely, the system has been successfully applied on the recording of the plenary sessions of the Catalan Parliament. Parliament daily generates big amounts of video and audio files. Therefore, it is a domain where content management is a key task, and it could take advantage of the automatic metadata extraction system.

The paper is organized as follows. Section 2 refers to the problem of managing big amounts of audiovisual content and how spoken content can be exploited to extract metadata in order to obtain suitable annotation information. Section 3 describes in detail how the system works, the experiments carried out, the results obtained and a discussion. Finally, some conclusions are given in section 4.

2. Spoken content management and automatic metadata extraction

Multimedia content management is a key task in any big repositories of audiovisual content. Annotation metadata is indispensable for an effective management, but it cannot be obtained in a manual way since it is an inviable and very time-consuming task. For that reason, information extraction must be done in an automatic way. Metadata extraction could be performed at several levels (from video and audio), but one of the most important information source is the spoken content. Numerous approaches have been developed in automatic speech recognition and automatic speaker recognition. Thus it is a good idea to exploit these kind of systems to extract metadata.

Particularly, speech-to-text transcriptions and speaker diarizations are specially useful as annotation data for multimedia content. Therefore an automatic metadata extraction that acts over the spoken content and extract the speech transcription and performs the speaker diarization is proposed.

The speech transcriptions of the spoken content provides a wide range of possibilities. Usually ASR systems output contains information about each recognized word and temporal information about the occurrence. On the management side, indexing and retrieval can be carried out utilizing the spoken content transcription. On the user side, this data can be used to different purposes, like to perform keyword spotting, providing direct access to the desired words inside a given audiovisual

stream, or as an complementary resource to the audiovisual content (subtitles).

Regarding the speaker diarization, it has also applications for both managers and users. On one hand, speaker diarization may be utilized as annotation data to be applied on content indexing and retrieval purposes. On the other hand, users can take advantage of the diarizations when browsing through content, allowing direct access to the segments where particular speaker participates.

The extracted metadata must be stored in a convenient way in order to be utilized for indexing and retrieval purposes. Therefore, such metadata is stored using the MPEG-7 multimedia content description interface, centering on the spoken content description schemes. It allows to apply the great variety of indexing and retrieval techniques developed in previous research about spoken document retrieval over MPEG-7 descriptions. The fact of following MPEG-7 also raises content exchange and compatibility between different platforms.

The automatic metadata extraction from spoken content can be addressed to a great variety of domains, where it is necessary to manage great quantities of audiovisual content (broadcast news, meeting, etc). One target domain is the audiovisual material recorded at the Parliament. Everyday the plenary sessions are registered on audio and video. In Catalonia, such material from the Catalan Parliament is in the public domain and is made available to citizens on the Internet. It has 2 important consequences. First, a better accessibility to content would improve the user experience when browsing trough content. Secondly, there are public resources available to the research community to be used to develop speech recognizers (for instance, the transcriptions of the sessions can be used to train language models) and speaker diarization systems.

3. Experiments and results

This section describes the experiments in detail. Both speaker diarization and speech-to-text transcription have been performed using the Hidden Markov Model Toolkit (HTK) [1].

Once the transcription and the speaker diarization are obtained, the results are parsed and the MPEG-7 description is generated.

3.1. Speaker Diarization

The current subsection describes the speaker diarization system set-up in detail, particularly the training and test data, the feature extraction and the speaker modeling and configuration.

3.1.1. Training and test data

The training and test data are recordings from the Catalan Parliament. The original 16 bit, one-channel and 48 kHz audio has been down-sampled to 16 kHz. 4 hours of speech were used for training the models and 1 hour for testing.

3.1.2. Feature extraction configuration

Some techniques has been successfully applied on speaker recognition tasks, such as LPC, LPC-Cepstra, MFCC and PLP. In previous work carried out at CEPHIS about speaker diarization in broadcast news audio [2], PLP features have been empirically proven to be beneficial for this purpose. For that reason, PLP method has been chosen for the current task. Firstly, a speech signal processing is made for each type of feature. A 0.97 coefficient pre-emphasis filter is applied, and a 25 ms Hamming window that scrolls each 10 ms is used to obtain signal frames. Then, a feature vector of 12 PLP coefficients is obtained from each frame. Finally, the energy coefficient, delta- and delta-delta features (time derivatives) are added to the feature vectors. Further detailed explanation of the PLP technique can be found in [3].

3.1.3. Speaker Modeling

After a previous study of the Parliament audiovisual content, a categorization of the participating speakers must be carried out. Generally, one can divide the participants into these categories: the premier, the government ministers, the president of the Parliament and the representatives of the parliamentary groups.

In addition, there are sound events that are not properly speech, such as background noise, murmur or silence. It must be taken into account that politicians remain for at least 4 years, and usually 1/3 of the parlamentarians change. Consequently, it worth to develop models for each member of the parliament.

Having done the categorization, it is necessary to decide how to model the different categories in a useful way that facilitates direct access to each speaker participation. Some participants are especially relevant, such as the premier, the president of the Parliament and the government members. For these speakers it is highly useful to have their particular segments. On the other hand, some speaker like the representatives of the parliamentary groups could be merged into a general 'other speakers' category.

After this study, models will be developed for the following cases: single-model for the premier, single-model for the president of the Parliament, one single-model for each government minister, one 'shared' model for the representatives of the parliamentary groups and single-model for silence/background noise.

3.1.4. HMM configuration

An HMM for each case listed in subsection 3.1.3 is created in this stage. Each HMM has three states in a left-to-right topology. Only the central state has a GMM as emitting probability density function. Diagonal covariance matrices have been proved to be beneficial, thus they are used here. Particularly, there are three main reasons to use only diagonal covariance instead of full covariance matrices [4]. Firstly, the density modeling of an Mth order full covariance GMM can be achieved using a larger order diagonal covariance GMM. Furthermore, diagonal covariance matrices GMM are computationally more efficient than full covariance matrices GMM. Finally, empirically diagonal matrix GMM have been observed that outperform full matrix. Then, the single-gaussian model is split into 8, 16, 32 and 64 mixture gaussians. The model parameters are iteratively re-estimated using the implementation of the Baum-Welch algorithm in HTK (HERest tool).

Once the model set is obtained, the diarization is performed through the Viterbi algorithm (HVite tool). One general HMM is generated from the individual models creating a model loop.

3.2. Speech-to-text transcription

A description of the speech-to-text system is given next, focusing on the training and test data, acoustic modeling and language modeling.

3.2.1. Training and test data

The acoustic models have been trained using the SpeechCon Catalan speech corpus [5]. The corpus has spontaneous and read speech from 550 speakers, recorded with four microphones at different distances. Each utterance is stored in 4 independent (one per microphone) 16 bit, 16 kHz uncompressed audio files.

The test consists of 13 minutes of speech extracted from the recordings of the Catalan Parliament plenary sessions. The original 16 bit, one-channel and 48 kHz audio has been downsampled to 16 kHz. The nature of the training and test data is very different (clean speech versus noisy, non spontaneous speech). For that reason, an adaptation stage will be necessary in order to improve accuracy.

The audio files are then parametrized into a 39-dimensional feature vector consisting on 12 cepstral coefficients plus the 0th coefficient, deltas and delta-deltas.

3.2.2. Acoustic modeling

Firstly, a set of 40 HMMs (39 monophones plus 1 silence model) is obtained. The HMM consists of 3 output states with self-loops, in a left-to-right topology. This set is calculated by means of a flat initialization, and each model is re-estimated using the Baum-Welch algorithm. A short pause model is then created by cloning the central state of the silence model and adding a skipping transition.

The phone-level transcriptions are converted into cross-word transcripts. New triphone models are created by cloning the central state of its corresponding monophone. Every triphones with the same central monophone share their transition matrices. Then the model parameters are re-estimated again.

Since the triphone set do not cover the all the possible triphones in the language, they are synthesized and their states are tied to physical models states. It also contributes adding a more robust set of models. The tying process is carried out through decision tree clustering that uses linguistically motivated questions about a triphone's context. The resulting tied-state triphones are re-estimated.

Finally, the single gaussian models are split and re-estimated consecutively into 2, 4, 8, 16 and 32 gaussian components, obtaining the final set of tied-state triphone models.

In addition, a further improvement of the ML models is carried out. A set of discriminatively trained models is developed from the ML set using MMI, running 4 iterations of the EBW algorithm (HMMIRest tool).

3.2.3. Language modeling

The language model is a 64k word based 3-gram LM that has been developed using the transcriptions of the plenary sessions of the Catalan Parliament, consisting of around 24 million words. The completed 167000 word vocabulary was reduced to 64k for two reasons. Firstly, the majority of the words inside the original vocabulary have very few occurrences and they are considered as rare words that are hardly pronounced. Therefore suitable probabilities cannot be calculated for those uncommon words. Secondly, the decoder used for the experiments imposes a restriction on the vocabulary size of 64k words. Thus, the 64k most common words in the corpus were taken. The utilized training tool was also HTK (LBuild).

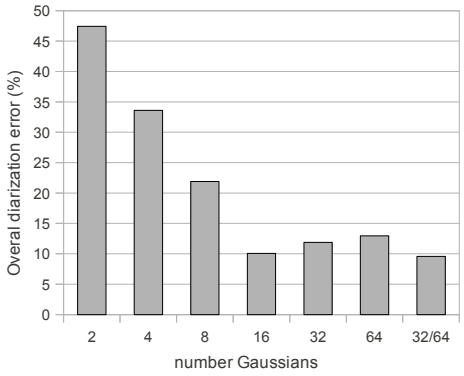

Figure 1: *Speaker diarization results.*

3.3. Results and discussion

3.3.1. Speaker diarization results

This section shows the results obtained from the experiments carried out to test the speaker diarization system. The speaker diarization results evaluation has been made following the Spring 2006 Rich Transcription Meeting Recognition Evaluation Plan (NIST) [6]. This plan proposes a "who spoke when" diarization scoring.

The total error percentage is called Overall Diarization Error (ODE). Fig. 1 shows the ODE obtained using 2, 4, 8, 16, 32 and 64 gaussians. Furthermore, a combined model of 32 gaussians for particular speakers and 64 gaussians for 'other speakers' and silence has been tested.

Analyzing the results, it can be observed that the ODE decreases significantly when incrementing the number of gaussians, up to 16 components. However, this trend changes when increasing the number of gaussians at this point. It has been observed that the models for known speakers perform better when using 64 gaussian components, whereas the 'other speaker model' works better with 32 components. Thus, a combined 32 and 64 gaussian model set has been evaluated, decreasing the ODE up to a 11.68% ODE. In any case, the combined model hardly outperform the 16 gaussian models (10.06% ODE), and it may not worth it due to the computational cost.

3.3.2. Speech-to-text results

After carrying out the speech-to-text experiments, the obtained results are shown below. Firstly, the Acoustic models have been tested on clean speech, recorded in office environment. Fig. 2 shows the Word Error Rate (WER) for the different trained models: maximum likelihood models (ML) and discriminatively trained models using MMI criterion. The same models have been evaluated performing MLLR speaker adaptation as well.

The same experiments have been done using real noisy speech from the Catalan Parlament, whose figures are depicted in Fig. 3. In this case, the adaptation carried out is not based on particular speaker, but on the general features of the speech in the Parliament.

It can be observed that there is a significant difference of performance when changing the characteristics of the audio. For clean audio recorded in office environment, that is quite similar to the features of the training corpus, WER decreases up to 20,21% and 16,14% using speaker adaptation and discrimi-

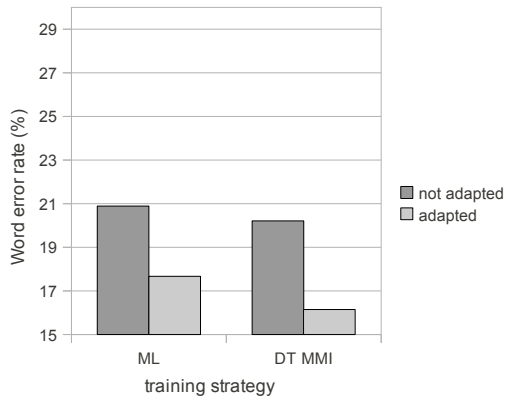

Figure 2: *Word error rate in clean speech experiments.*

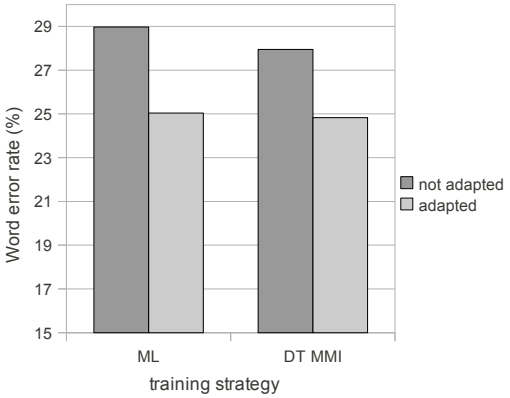

Figure 3: *Word error rate in parliament speech experiments.*

natively trained models. When using Parliament speech, the accuracy decreases to 24,3 %due to the environmental noise and overlapped speech, although an improvement of around a 3% is achieved with MLLR adaptation.

3.4. Metadata representation in MPEG-7

Once the transcription and the speaker diarization have been obtained, it is necessary to put them together to form the MPEG-7 description. For that purpose, a software have been developed in Java. The tool parses the files generated by HDecode (transcription) and HVite (speaker diarization), and generates a MPEG-7 description in XML, using the Spoken Content Description Scheme [7]. In relation to the speech-to-text transcription, the best word output is represented by means of the *SpokenContent* and *Lattice* descriptors.

4. Conclusions

We have proposed a system that agglutinates speech recognition and speaker diarization. The system have been successfully applied on the recording of the plenary sessions of the Catalan Parliament. The speech transcription is performed using a set of tied-state cross-word triphone models and a 3-gram LM. In relation to the speaker diarization, a categorization of the speakers that participate in the Parliament recordings was carried out and a set of HMM was developed using different number of gaussian components. Then, HMM based speaker diarization was done.

As far as the speech-to-text transcription is concerned, the results state the importance of the audio quality in automatic speech recognition. Noisy speech means an important degradation of accuracy. It can be managed by means of the application of robust ASR and noise reduction techniques. However, word error rates using clean audio decreases under 17%. This figure is supposed to get lower using a more robust LM trained with a better textual corpus. In addition, discriminatively trained models have improved error rates around 1%, but in some cases it could not worth it due to discriminative training is a very time-consuming process. In other works, discriminative training has resulted to be beneficial for large vocabulary tasks. Thus, more work is needed in order to get better error rates with discriminatively trained models. Discriminative training using MPE criterion has been proven to be useful for large vocabulary tasks. Therefore, it should be tested as future work. Finally, the use of speaker adaptations improves significantly the accuracy.

Focusing on the speaker diarization system, results reveal that PLP features are adequate to perform the speaker segmentation in our data set. This means that PLP features extract inter-speaker variability correctly. Regarding the GMM, an increment in the components number not always improve error rates. Particularly, increasing mixture components over 16 does not contribute with any improvement.

Finally, the extracted metadata was stored according to the MPEG-7 content description interface automatically through the Java parser. That results very useful because it allows the use of a great variety of content based indexing and retrieval techniques.

5. Acknowledgements

This work has been partly founded by the Spanish Ministerio de Ciencia y Innovacion, Project Reference TEC2008-03835/TEC. This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700

6. References

[1] S. Young and G. Evermann, *The HTK Book*. Cambridge University Engineering Department, March 2009.

[2] H. Delgado, "Segmentación de Vídeo mediante Reconocimiento de Locutores," Master's thesis, Universitat Autónoma de Barcelona, June 2009.

[3] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] Asunción Moreno, Albert Febrer and Lluis Márquez, "Generation of Language Resources for the Development of Speech Technologies in Catalan," in *Proceedings of the Language Resources and Evaluation Conference LREC 06*, Genova, Italy, 2006.

[6] NIST, *Spring 2006 (RT-06S) Rich Transcription Meeting*, 2006.

[7] "MPEG-7 Overview," Internet: http://www.chiariglione.org/mpeg/standards/mpeg-7/.