International Journal of Research in Advanced Engineering and Technology
ISSN: 2455-0876; Impact Factor: RJIF 5.44
Received: 10-09-2019; Accepted: 12-10-2019
www.newengineeringjournal.in
Volume 5; Issue 4; October 2019; Page No. 43-46

Spoken content metadata extraction using speech and speaker recognition approaches

Pardeep Sangwan
Department of ECE, Maharaja Surajmal Institute of Technology, New Delhi, India

Abstract
The collection of information today plays an important role in processing massive amounts of data for different purposes. Automatic extraction of information from audio streams is one of the open challenges in this area. The present work explains a method to extract metadata for performing combined tasks to identify the speakers utilizing 'Hidden Markov (HMM) tied-state crossword tri-phones acoustic models, Mel-Frequency Cepstral Coefficients (MFCC) and N-gram language modelling'. The device performs speech transcription through a Catalan language recognizer. In addition, a diarization of the speaker is performed using segmentation based on HMM and extraction of the feature 'Perceptual Linear Prediction (PLP)'. For multimedia content, voice-to-text conversion as well as speaker diarization may be utilized as descriptive information. The storage of metadata is done with the help of MPEG-7 to make indexing and retrieval more versatile and effective. The device was successfully tested on the recording of the Catalan Parliament's plenary sessions.

Keywords: metadata extraction, automatic speech recognition, speaker Diarization, GMM, HMM, MPEG-7

1. Introduction
Today, the amount of multimedia content is increasing dramatically as IT sector is growing rapidly and hence almost infinite storing capability is achieved. Due to the massive quantity of data, managing contents, its index and retrieving the data are key problems which becoming complexer. Because manual annotation of such vast data is not feasible, content is forced to be automatically annotated by approaches to information extraction. Speech inside audio streams is a huge source of information from which data can be automatically extracted. Speech technologies are capable of extracting various types of audio data. Automated audio/video stream transcriptions of spoken material are also ideal for automated subtitles formation, and assistance for international and/or hearing-impaired people. Moreover, automatic speaker diarization can be used to directly access the sections in which different speakers are active. This paper describes an automated extraction method for useful metadata that focuses on the spoken text. Using a Catalan recognizer based on crossword tied-state triphone HMMs, MFCC and N-gram dialect models for Catalan language, and the metadata extraction method performs a voice transcription. In addition, diarization of speakers is done using segmentation based on HMM and extraction of PLP functionality. A comparison of performance with different configurations has been performed for both transcription and speaker diarization performance improvement. After data collection has been collected, the MPEG-7 content definition interface must automatically save it. The device can be used in various fields of use. In particular, the system was utilized to record Catalan Parliament's plenary sessions producing a large quantity of video and audio files on a daily basis. It is therefore an area where content management is a key activity and the automated metadata extraction program will take advantage of it. The paper is structured in the following manner. Section 2 relates to the issue of managing large volumes of audio-visual material and how to use spoken content for extracting metadata for obtaining correct descriptive information. Section 3 explains working of system, tests performed, result achieved, and a summary. Lastly, in section 4, some conclusions are given.

2. Spoken Content Management and Automatic Metadata Extraction
Managing multimedia content is a crucial activity in any broad audio-visual content repositories. Annotation metadata is important for efficient management, but its manual collection is not possible as it is an inviable and time-consuming process. Therefore, the processing of data must be done automatically. Metadata extraction can be achieved at multiple levels. In automatic speech recognition and automatic speaker recognition, various approaches have been established. It is therefore a good idea to use such systems to collect metadata. Especially as annotation information for multimedia content, speech-to-text transcriptions and speaker diarizations are particularly useful. Thus, an automated extraction of metadata is introduced, which works on the spoken material and extracts the transcription of speech and performs the diarization of the speaker. The spoken content's speech transcriptions offer a wide range of possibilities. Usually the output of ASR systems contains information about each acknowledged word and the occurrence time information. Indexing and retrieval can be done on the management side using the transcription of spoken data. On the user side, this information can be used for various purposes, such as spotting keywords, providing direct access to the desired words within a given audio-visual channel, or complementing audio-visual content (subtitles). As for the diarization of speakers, it also has applications for managers and users alike. On the one side, the diarization of speakers can be used as annotation information to be used for indexing and retrieving content purposes. On the other hand, when browsing through content, users can take advantage of the diarizations, allowing direct access to the segments in

which a particular speaker is involved. To be used for indexing and retrieval purposes, the extracted metadata must be conveniently stored. This metadata is therefore stored using the multimedia content description system of MPEG-7, which focuses on the description of spoken communication schemes. This allows the wide range of indexing and retrieval techniques developed over MPEG-7 descriptions in previous research on spoken information retrieval to be applied. The audio-visual content recorded at the Parliament is one target area. The plenary sessions were captured on audio and video every day. These Catalan Parliament content is in the public domain in Catalonia and is made available on the Internet to people. It has two major consequences. Second, improving web usability will boost user experience when browsing content troughs. Second, the research community has public resources available to improve speech recognizers and speaker diarization systems.

3. Experimental Setup

Each section is a detailed description of the experiments. The Hidden Markov Template Toolkit (HTK) [1] was used to perform both speech diarization and speech-to-text transcription. After receiving the transcription and the diarization of the speaker, the results are analysed and the definition of MPEG-7 is produced.

3.1 Speaker Diarization

This subsection explains in detail the configuration of the sp eaker diarization system, particularly the training and test in formation, the extraction of the function and the modeling a nd configuration of the speaker.

3.1.1 Training and Test Data

This data is recorded from Catalan Parliament.

3.1.2 Extracting Features

Various methods, such as LPC, LPC-Cepstra, MFCC and PLP, have been successfully applied to speaker recognition tasks. PLP features have been empirically proven to be beneficial for this purpose in previous work at CEPHIS on speaker diarization in broadcast news audio [2]. PLP approach has therefore been chosen for the current task. First, for each form of function, a speech signal processing is performed. A pre-emphasis filter of 0.97 coefficient is applied, and a Hamming window of 25 ms is used to scroll every 10 ms to obtain signal frames. From each frame, a feature vector of 12 PLP coefficients is then obtained. Finally, the characteristic vectors are applied to the energy e quation, delta and deltadelta properties (time derivatives).
Thorough explanation can be found in [3] of the PLP methodology.

3.1.3 Speaker Modelling

A categorization of the participating speakers must be carried out after a previous study of the Parliament's audio-visual content. In general, the participants can be divided into these categories: The Prime Minister, the ministers of the government, the leader of the parliament and the members of the parliamentary parties. Furthermore, audio phenomena such as background noise, murmur, or silence are not properly spoken. It must be borne in mind that politicians stay for at least four years, and usually change 1/3 of parliamentarians. Therefore, it is worth designing templates for each parliament member. Participants which are very important, such as the Prime Minister, Parliament president and members of the government. It is highly useful for these speakers to have their unique parts. On the other hand, certain speakers such as legislative party members could be grouped into a general category of' other speakers.

3.1.4 Configuring HMM

At this stage, a HMM is created for all cases given 3.1.3. In left-to-right topology, each HMM has three states. Only the central state has a GMM function emitting density of likelihood. It has been shown that diagonal covariance matrices are useful, so they are used here. There are three main reasons, in particular, for using only diagonal covariance rather than total covariance matrices [4]. Second, with a larger order diagonal covariance GMM, the density simulation of a mth order maximum covariance GMM can be accomplished. In addition, GMM matrices with diagonal covariance are computationally more efficient than GMM matrices of maximum covariance. Finally, it was found that empirically diagonal matrix GMM exceeds the complete matrix. The single-gaussian model is then divided into 8, 16, 32 and 64 gaussian mixtures. Re-estimation of parameters is done iteratively utilizing “Baum Welch algorithm”. The diarization is performed via the Viterbi algorithm (HVite tool) once the template set is obtained. A general HMM is created by constructing a model loop from the individual models.

3.2 Speech-to-Text Transcription

Speech-to-text system is explained in the following subsection:

3.2.1. Training and test data

The acoustic models are equipped using the Catalan language corpus of Speech Con [5]. The corpus has 550 speakers spontaneous and read speech, captured at different distances with four microphones. Utterances are stored in 16-bit, 16 kHz uncompressed audio files in 4 separates (one per microphone). The test consists of 13 minutes of speech from the recordings of the plenary sessions of the Catalan Parliament. The original 16 bit, one-channel and 48 kHz audio was sampled down to 16 kHz. There are very different types of training and test data (clean speech versus noisy, non-spontaneous speech). For this purpose, to improve accuracy, an adaptation stage will be required. These recorded files are then parameterized into a 39 dimensional vector with 12 cepstral coefficients plus the 0^{th} coefficient, deltas and delta-deltas.

3.2.2. Acoustic Modelling

Next, a range of 40 HMMs is obtained (39 monophones plus 1 model of silence). The HMM consists of 3 self-looping output states in a topology from left to right. Each array is determined using a flat initialization and the Baum-Welch algorithm re-estimates each model. Then a short pause model is developed by cloning the silence model's central state and adding a transition skipping. Transcriptions at the phone stage are translated into transcripts of crosswords. New triphone models were developed by cloning their corresponding monophone's central state.
Each triphone shares their transition matrices with the same central monophone. Then the parameters of the model will

be re-estimated. Since the triphone array does not cover all of the language's possible triphones, they are synthesized and their state is related to state of physical models. It also helps to incorporate a more stable array of designs. The tying process is accomplished by clustering decision tree using linguistically motivated questions about the context of a triphone. Its re-estimates the resulting tied-state triphones. Finally, the single Gaussian models were subdivided and subsequently re-estimated into 2, 4, 8, 16 and 32 Gaussian elements.
In addition, the ML models will be further improved. Using MMI, a set of discriminatively trained models is built from t he ML array, running 4 EBW algorithm iterations (HMMIR est tool).

3.2.3. Language Modelling

The language template is a 64k word based3-gram LM developed using the transcriptions of the Catalan Parliament's plenary sessions, consisting of approximately 24 million words. For two factors, the completed 167000-word vocabulary has been reduced to 64k. Second, there are very few occurrences in most of the words inside the original vocabulary and they were regarded as uncommon, hardly pronounced words. Therefore, it is not possible to calculate appropriate probabilities for these uncommon terms. Second, a limit on the vocabulary size of 64k words is enforced by the decoder used for experiments. Therefore, the most common 64k words are taken in the corpus.

4. Result

4.1. Speaker Diarization

The assessment of the outcomes of the diarization of speakers preceded the Rich Transcription Meeting Recognition Assessment Program (NIST) of Spring 2006 [6]. This program suggests a diarization ranking "who spoke when." Total Diarization Error (ODE) is the overall error factor. Fig. 1 Displays the ODE of 2, 4, 8, 16, 32 and 64 Gaussians. In contrast, a composite version of 32 Gaussians was tested for common speakers and 64 Gaussians was tested for' other speakers' and silence. Analysing the results, it can be found that when increasing the number of gaussians up to 16 elements, the ODE decreases significantly. This pattern, however, shifts as the number of Gaussians at this stage increases. Models for known speakers have been found to perform better when using 64 gaussian components, while the' other speaker version ' works better with 32 components. A composite array of 32 and 64 Gaussian models were tested, increasing the ODE to an ODE of 11.68 percent. In any case, the combined model barely outperforms the 16 Gaussian models (10.06 percent ODE), and due to the computational expense, it may not be worth it.

4.2. Speech-to-Text Results

The results obtained will be shown below after performing the speech-to-text experiments. The Acoustic models were first tested on clean voice, captured in the office environment. Fig. 2 Displays the Word Error Rate (WER) for the various trained models: maximum probability models (ML) and discriminatively trained models using the MMI criterion. Even performing MLLR speaker adaptation, the same models were evaluated. The same experiments were made using the Catalan Parliament's actual noisy voice, the figures of which are represented in Fig. 3. The adaptation carried out in this case is not based on a single speaker, but on the general features of Parliament's voice. It can be found that when adjusting the characteristics of the sound, there is a major difference in performance. For clean audio recorded in the office environment, word error rate declines by up to 20.21 & 16.14 % utilizing speech adaptation and discriminatively trained models. When using Parliament voice, due to environmental noise and overlapping speech, the accuracy decreases to 24.3 percent, although with MLLR adaptation an increase of about 3 percent is achieved.

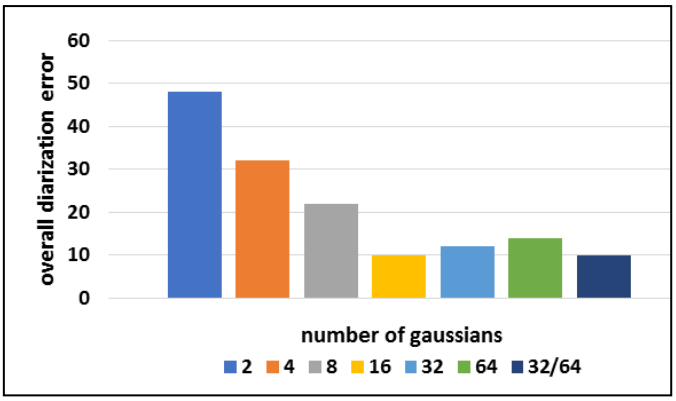

Fig 1: Speaker Diarization results

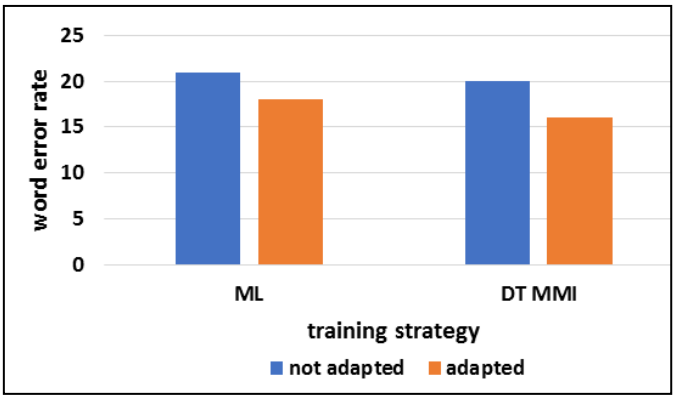

Fig 2: Word error rate in clean speech experiments

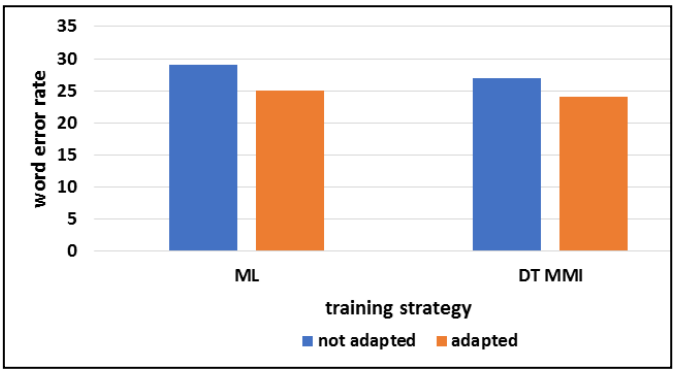

Fig 3: Word error rate in parliament speech experiments

5. Conclusion

We also suggested a scheme that incorporates the identification of speech and the diarization of speakers. The device was successfully applied to record the Catalan Parliament's plenary sessions. A set of tied-state cross-word triphone models and a3-gram LM are used to perform the speech transcription. With regard to the diarization of

speakers, a categorization of speakers participating in the recordings of the Parliament was performed and a collection of HMM was created using different number of Gaussian components. Then the diarization of the HMM-based speaker was completed. With regard to speech-to-text translation, the results indicate the significance of audio quality in the recognition of automatic expression. Noisy speech means the reliability is greatly reduced. Use comprehensive ASR and noise reduction methods, it can be controlled. Word error rates using clean audio, however, are dropping below 17%. With a more reliable LM equipped with a stronger textual corpus, this figure should be lowered. However, discriminatively trained models have increased error rates around 1 percent, but in some cases, discriminative learning is a very time-consuming operation, which could not be worth it. Discriminative learning has been useful for broad vocabulary tasks in other plays. Therefore, more work is needed with discriminatively trained models to achieve better error rates. It has been shown that discriminative learning using MPE criterion is useful for broad vocabulary tasks. This should therefore be tested as future work. Furthermore, the use of speaker modifications significantly improves the reliability. Focusing on the speaker diarization process, the results show that the PLP features in our data set are appropriate to perform speaker segmentation. It means PLP software correctly extracting the variance of the interspeaker.

As far as GMM is concerned, an increase in the number of components does not always raise error rates. Increasing components of the mixture over 16 in particular does not lead to any change. Finally, the extracted metadata was automatically stored via the Java parser in accordance with the MPEG-7 content description interface. That results are very useful because it allows a wide variety of indexing and retrieval techniques based on content to be used.

6. References

1. He Y, Sainath TN, Prabhavalkar R, McGraw I, Alvarez R, Zhao D, *et al*. Streaming end-to-end speech recognition for mobile devices, arXiv preprint arXiv:1811.06621, 2018.
2. Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, Bates S, *et al*. Indatacenter performance analysis of a tensor processing unit, in ACM/IEEE Annual International Symposium on Computer Architecture (ISCA). IEEE, 2017, 1-12.
3. Sim KC, Narayanan A, Bagby T, Sainath TN, Bacchiani M. Improving the efficiency of forward-backward algorithm using batched computation in tensorflow, in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, 258-264.
4. Bagby T, Rao K. Efficient implementation of recurrent neural network transducer in tensorflow, in IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.
5. Soltau H, Liao H, Sak H. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition, in Interspeech. ISCA, 2017.
6. Virpioja S, Smit P, Gronroos SA, Kurimo M. Morfessor 2.0: Python implementation and extensions for morfessor baseline, Aalto University, Tech. Rep, 2013.