GRIDNEWS: A DISTRIBUTED AUTOMATIC GREEK BROADCAST TRANSCRIPTION SYSTEM

D. Dimitriadis[1], A. Metallinou[2], I. Konstantinou[1], G. Goumas[1], P. Maragos[1] and N. Koziris[1]

[1] School of ECE, National Technical Univers. of Athens, Zografou, Athens 15773, Greece
[2] School of ECE, Univers. of Southern California, Los Angeles, CA 90089-2560, USA

Email:[ddim, maragos]@cs.ntua.gr, [ikons,goumas,nkoziris]@cslab.ece.ntua.gr, metallin@usc.edu

ABSTRACT

In this paper, a distributed system storing and retrieving Broadcast News data recorded from the Greek television is presented. These multimodal data are processed in a grid computational environment interconnecting distributed data storage and processing subsystems. The innovative element of this system is the implementation of the signal processing algorithms in this grid environment, offering additional flexibility and computational power. Among the developed signal processing modules are: the Segmentor, cutting up the original videos into shorter ones, the Classifier, recognizing whether these short videos contain speech or not, the Greek large-vocabulary speech Recognizer, transcribing speech into written text, and finally the text Search engine and the video Retriever. All the processed data are stored and retrieved in geographically distributed storage elements. A user-friendly, web-based interface is developed, facilitating the transparent import and storage of new multimodal data, their off-line processing and finally, their search and retrieval.

Index Terms— Computer architecture, distributed database systems, multimedia systems, speech recognition, user interface

1. INTRODUCTION

The wide use of digital systems and personal computers has led to the creation of an extensive amount of multimodal information. The management of such increasing amount of data poses novel technological and research challenges. In this context, the need to encompass more computing power and storage capabilities for automatic processing of such data has increased drastically. While the technology on information retrieval of text data is quite mature due to the widespread of World-Wide-Web, this is not yet true for audio and video signals. This is due to the fact that the nature of this kind of data makes information retrieval, i.e. automatic speech recognition in this case, a significantly harder task compared to text searching.

The "GridNews" system, as shown in Fig.1, was developed in the context of an interdisciplinary research project in the area of multimodal processing and indexing. This project was focused on the implementation of a distributed platform performing keyword spotting on Broadcast Data (BN) from television and radio news programmes. This platform is based on grid technologies designed and implemented for the needs of the project. The grid computing approach was followed due to the computational and data intensive tasks. The primary objective of the GridNews system is to perform *Audio Diarization* on multimodal data and return short video files containing the selected keywords. The "diarization" term describes the processes of segmenting and classifying the audio streams [1] and finally, transcribing them in written text. The proposed system functionality can be considered analogous to that of the "YouTube" system but greatly differentiated in the automatic video segmentation and annotation processes that only the GridNews system offers. Its implementation combines keyword extraction from video files using ASR algorithms with efficient and scalable distributed storage media content and indexing of the extracted metadata.

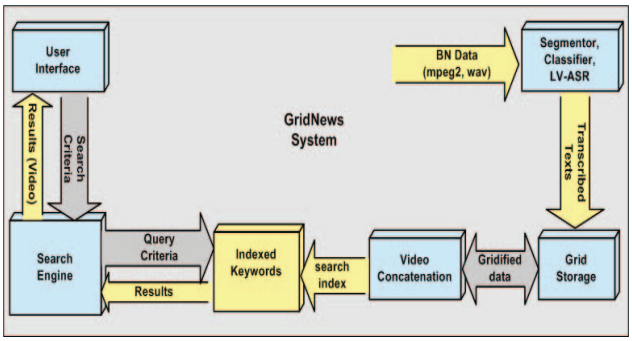

Fig. 1. GridNews System Overview

Audio streams from BN contain sounds stemming from various sources, such as speech from multiple speakers and music, all corrupted by different types and levels of noise. The number of the different classes into which the input audio can be segmented and classified depends mainly on the input quality of the audio-streams, the amount of available data and the desired level of detail in the audio classification process. Herein, the audio diarization module categorizes the input signal into two complementary classes, i.e. speech and non-speech. The first class contains speech in the presence of various levels of noise and speech from single or multiple speakers of both genders. On the other hand, the second class contains noise, silence or music. This system precedes the ASR module, therefore it is important that the non-speech segments are correctly recognized and excluded from the speech recognition process. Audio diarization, also, includes the problem of event detection summarized in speaker changes or transitions between speech and non-speech regions. Metric-based approaches have been proposed in literature performing well in this event detection task. Such approaches measure the distance between two neighboring windows of the input signal and detect an event if this distance exceeds a certain threshold

This work was supported in part by the Greek research program 'ΕΗΓ10-GridNews', the grants ΠΕΝΕΔ − 2003 $E\Delta - 865, 866$ and 554 [co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)], and the EC Network of Excellence 'MUSCLE'. The work was performed when A.Metallinou was with NTUA.

[2]. Various distance measures can be used e.g. the Bayesian Information Criterion (BIC) described in [3], the Weighted Mean Distance criterion (WMD) or the T^2 criterion proposed in [4].

Finally, concerning the task of *Large Vocabulary Speech Recognition* (LVASR), two principal approaches are usually followed. The first one is the phone-based approach, producing either a phone transcription or some kind of phone lattice. The other one is the word-based approach yielding a word-level transcription that could be directly used for the word spotting task. Herein, we have followed the first approach since a lightweight recognizer was considered adequate despite the adverse conditions present in most BN data, e.g., speaker variability and recording conditions, that affect greatly the ASR performance.

2. SIGNAL SEGMENTATION MODULE

The audio-based event detection module is implemented using a metric approach, i.e. a combination of *Bayesian Information Criterion* (BIC), T^2 and *Weighted Mean Distance* (WMD) criteria [3, 4].

BIC criterion considers two consecutive windows of the audio stream and models their probability distributions using gaussians. Their distance is estimated and in case of exceeding a certain threshold, an event is considered at the mid-point between these two windows. Otherwise no event is detected and the algorithm continues running. However, BIC is not computationally efficient and requires long windows to reliably estimate the data-driven probability distributions. On the contrary, the T^2 and WMD criteria are employed when not enough data are available or smaller window sizes are considered. The implemented algorithm sequentially scans the audio stream looking for event points between consecutive windows. While no such points are detected the window sizes gradually increase until a new event is detected and the window lengths are reinitialized. The criterion to be applied depends solely on the window length and as it increases, first the WMD, then T^2 and finally the BIC criterion are sequentially used. The desired detection is based on a 2-passages algorithm, combining 2 different kinds of features, the MFCC for the 1^{st} pass and the Fractal Coefficient [5] for the 2^{nd} one. The 2 passages procedure decreases greatly the false alarm detection rates. In general, MFCC-based segmentation yields large falsely-detected events, but the 2^{nd}-pass eliminates most of them.

Under the scope of this project, different audio-based features have been investigated and their performance has been assessed, before finally employing the MFCCs and the Fractal Coefficient. In general, the MFCCs perform adequately for the needs of the segmentation, classification and ASR tasks of GridNews. However, for the segmentation task, we have concluded that the MFCC features should be augmented with one-dimensional features like the zero-crossing rates (ZCR), pitch, the spectral flux [6, 7, 8] or the fractal dimension (Fractal Coefficient) of the signal [5, 9], improving the overall performance. These features are selected because they combine computational efficiency and good performance. All features are estimated over $40\ msec$ frames with no overlap. Finally, the Segmentor examines whether the transitions between speech and non-speech sections last at least $2\ sec$, otherwise it ignores any such potential event. So, a small pause or a short noise between speakers will not be considered as a non-speech region.

3. SIGNAL CLASSIFICATION MODULE

The Classifier utilizes statistical modeling methods to label the audio segments, extracted by the segmentation module, either as speech or non-speech classes. For the classification task, a single Hidden Markov Model (HMM) with two states is trained, one for the speech and another one for the non-speech class. The state corresponding to the speech class is modeled by 8 gaussian mixtures and the non-speech class is modeled by 4 gaussian mixtures. This HMM used for the speech/non-speech classification task is trained using only the audio modality of the collected BN data and their corresponding transcriptions. The provided hand-labeled, rich transcriptions include information about the speaker identity, the speaker changes and the level of background noise for the speech regions. On the other hand, the non-speech regions are tagged as silence, music or noise. The features employed are 39 MFCCs, i.e. $C_0 + 12$ cepstral coefficients and their 1^{st} and 2^{nd} time-derivatives. The frame lengths equal to $40\ msec$ with frame overlap of $10\ msec$. The HTK Toolkit is used for both the feature extraction and training processes.

As mentioned above, the Classifier takes as input the segments estimated by the audio-stream segmentation module. We assume that each one of these segments belongs only to one of the two possible audio classes, i.e. all segments are considered homogenous, since the segmentation module detects transitions between speech and non-speech regions and segments the original audio-stream accordingly. The HMM classifies each frame of the audio segment to the most probable state. Finally, a majority-rules criterion is applied over all labeled frames to classify the entire segment in either the speech or non-speech class.

4. LARGE-VOCABULARY SPEECH RECOGNITION MODULE -LVASR

The implemented *Greek Large-Vocabulary Speech Recognition* system (Greek LV-ASR) is based on the statistical Hidden Markov Models (HMM) framework, i.e. the HTK platform, considering all of its restrictions. For LV-ASR tasks, both the Acoustic Model (AM) and Language Model (LM) are trained using a Greek BN database, i.e. the GridNews BN Corpus. This database contains multimodal BN data suitable for the AM and written text for the LM.

The AM consists of triphone-based continuous density left-right HMMs. The Greek dictionary has 33 phonemes, creating 4489 triphones (as appeared in the training Corpus). Therefore, 4489 triphone-based 3-state HMMs are modeled with 12 gaussian mixtures per state. The extracted features are 39 MFCC coefficients plus their Cepstral Mean Subtraction (CMS), estimated over $30msec$ frames with $20msec$ overlap. The HMMs are trained over 15h of spontaneous-speech and BN data.

The LM is the probability distribution over all possible words, providing the information of how frequently a word sequence can appear. The most commonly used language models are the *N-gram LM models*, where the value N is the order of the LM. For $N = 2$, the LM is called *Bigram* and every word probability depends only on the immediately preceding one. The word probabilities are estimated over a text Corpus. Herein, the Corpus consists of 100 million words, provided by the *Athens News Agency* (ANA), the national news agency of Greece for the needs of the project. Herein, we have employed a Bigram LM because of the HTK Toolkit (version 3.3) restrictions. The vocabulary of the system consists of the $50k$ most frequent Greek words found in the training data. Finally, the perplexity of the trained LM is 80.59.

5. GRID SYSTEM DESIGN

The grid platform designed for the purposes of the GridNews system consists of three subsystems. The *Distributed Execution platform*

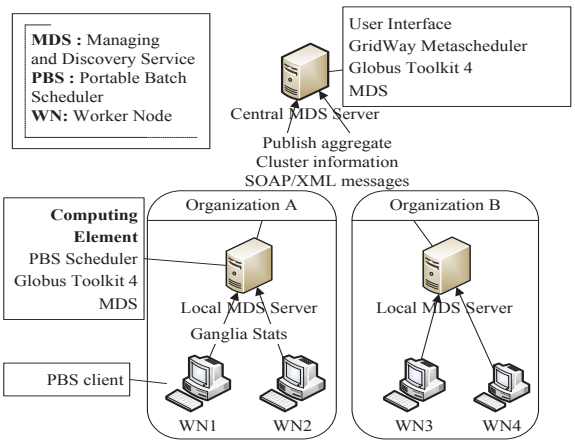

Fig. 2. Execution architecture

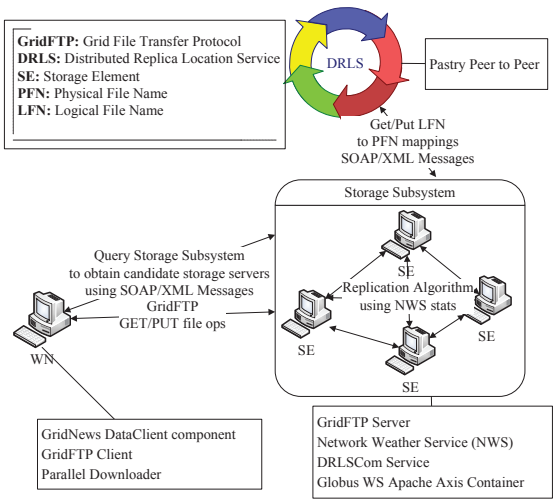

Fig. 3. Storage architecture

(C-QUEUES) is responsible for the scheduling and dispatching of computationally intensive tasks to the candidate worker nodes according to their availability and current load. The *Distributed Storage platform* (G-STORE) accounts for storing, indexing and retrieving the large amount of segmented videos exported by the multimodal signal processing tools. Finally, the *Keyword Indexing platform* (R-MERGE, SEARCH) deals with the indexing of the extracted keywords and the file segments in which they were encountered, to support efficient keyword search functionality.

Distributed Execution Platform

Grid computing offers support for seamless integration of computational resources through Virtual Organizations (VOs) [10]. In the scope of GridNews, this is performed by a two level scheduling. Each organization is equipped with a local job scheduler (OpenPBS). Local schedulers are registered in a higher level meta-scheduler (GridWay) and publish aggregated cluster usage information. Jobs are submitted to the User-Interface machine, where GridWay is installed. The meta-scheduler dispatches jobs to the local clusters according to the collected usage information. The aggregation of cluster usage information is performed by a hierarchical grid directory service named MDS. The architecture of the execution platform of GridNews (C-QUEUES) is depicted in Figure 2.

Distributed Storage Platform

The basic operations of the storage subsystem are the file transfer (upload/download) and the Distributed Replica Location Service (DRLS). DRLS is a distributed Peer to Peer index containing locations of replicated file instances through LFN (Logical Filename) to PFN (Physical Filename) mappings as described in [11]. The storage subsystem offers two basic primitives: PUT and GET file.

Upon a PUT file client request, the system locates the most suitable storage servers using a replica selection algorithm to replicate the client file. Storage servers periodically publish usage information using the Network Weather Service tool (NWS). The replica selection algorithm keeps a sorted list of less loaded servers according to a number of weighted metrics such as CPU usage, available bandwidth and free disk space. According to the replication factor decided, top k servers are selected from the server list to perform the storage and replication of the user file.

Upon a GET file client request, the system contacts DRLS to find available replicas of the requested file. When the replica list is retrieved, a client side parallel download is issued. The parallel download mechanism is based on the BitTorrent protocol: The client issues k parallel connections with the replica servers. Each connection downloads a chunk of the requested file. The client keeps issuing connections with the fastest replica servers until all the requested file is downloaded. BitTorrent in the context of Grid computing has been studied in [12]. The architecture of the Storage layer is presented in Figure 3.

Keyword Indexing platform

The automatic keyword extraction tool exports a text file with the recognized keywords and the time segment in which they were founded. An application that imports the extracted information in a relational database has been developed to optimize the search procedure. Processed videos are stored in a specific database table, i.e. “videos”. During video processing, the ASR module extracts the time intervals that contain an entire phrase (e.g, a sentence). The initial BN file is chopped using a number of time markers. These time markers are stored in an other table “video-parts”, with a key that is a foreign key to the table of videos, (one to many relationship). Each time segment can contain one or more keywords. Keywords are stored in the table “keywords”, along with their key, that is a foreign key to the table “video parts” (one to many relationship).

With the aforementioned design, the user can perform keyword search through the GridNews web portal. The portal software executes a query to the table keywords. The video parts that contain the requested keyword are returned to the user in a google-like url list. When the user clicks on a result, the portal software locates and serves the requested video to the portal user.

System Implementation

For the development of GridNews, we used *Globus v4.0* [13] as grid middleware. Globus is the most widely-used program for the creation of grid infrastructures. Globus does not constitute a completed platform of software, it only provides the essential components that were used by the GridNews system. What is more, Globus follows the Service Oriented Architecture (SOA) through the use of Web Services Resource Framework (WSRF). GridNews storage and execution platforms export their functionality through Grid WSRF compliant Services, and as of this, they can easily be

extended/integrated with other SOA applications. The functionality of GridNews services is exported as a WSDL thus facilitating the implementation of platform independent client software.

System Deployment
The hardware of GridNews platform is constituted by many heterogeneous elements, leading us to the use of virtualization. Recent studies of virtualization techniques in Grid environments [14] have shown that they offer a very useful layer of abstraction. The use of virtual computers may add a small administrative cost during installation, but through the offered flexibility the platform management is greatly simplified: GridNews storage and execution nodes can easily scale to a large number with little administrative effort.

GridNews system was deployed in a number of virtual machines using the Xen virtualization platform. More specific, we have used two virtual machines to act as computing elements/heads of local clusters, four virtual machines as worker nodes, three virtual machines as storage nodes, and a single virtual machine that consisted the portal site and the user interface node. These virtual machines were deployed in a single physical machine, a 4X Dual Core AMD Opteron(tm) Processor 875 2.2GHz with 16Gb of RAM.

6. RESULTS

The overall performance of the GridNews system concerning the keyword spotting in BN data is evaluated for each one of the individual modules. This modular evaluation scheme offers further insight making their deficiencies easier to detect since all modules are of different nature. In this context, the module evaluations are:
Successful Event Detection Rates: BN data are dominated by long segments labeled as "speech". However, large sections containing noise, music or silence co-exist. As described, the event detector was focused in locating the transitions between sections of speech and non-speech. Therefore, BN data containing different kinds of transition have been used for the evaluation process. All transitions have been manually annotated. An event detection is considered successful when the transition has been detected within a time interval of $0.5\ sec$, therefore, the detection precision is $0.5\ sec$. The system performance is measured in terms of the *Success* rate is 89.74% and the *False Alarm* rate is 13.89%.
Correct Classification Rates: After segmenting the BN data into homogenous sections, these sections are categorized into two different classes using an HMM model. This HMM is trained with 8h of BN data and tested on 2h of non-overlapping data. The training and testing data have been manually tagged and transcribed. The classification decision is based on the majority of all frame decisions (majority-wins criterion). Finally, the GridNews performance for the classification is *Correct Classification Rate*$= 94.11\%$ at the frame-level.
Correct Recognition Rates: The Recognizer performance is measured in terms of correct word transcriptions of the speech segments, used thereafter for the keyword spotting task. Significant effort has been paid to ensure good performance and decoding speed, since they are inversely proportional due to the LM complexity and the dictionary size. The testing set consists of 2h of BN data. The performance of the Recognizer in terms of *Correct Recognition Rate* equals to 61.58% for mixed recording conditions.

The performance of the implemented system will be further enhanced when more sophisticated system modules will be introduced. In a future system implementation, the LM will be getting enriched automatically by importing contemporary news articles from selected internet sites. Further, more audio classes will be included in the Classifier module, enabling the training of targeted HMMs, e.g., male/female and noisy speech models.

7. ACKNOWLEDGMENT

We wish to gratefully acknowledge help from A. Zlatintsi, V. Anastasopoulou and the *Athens News Agency* (ANA), for collecting and manually annotating the GridNews BN multimodal Database.

8. REFERENCES

[1] S. E. Tranter and D. A. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, pp. 1557–1565, 2006.

[2] T. Kemp and M. Schmidt and M. Westphal and A. Waibel, "Strategies for Automatic Segmentation of Audio Data," *Proc. IEEE Int.Conf Acoust.,Speech, Signal Process.*, 2000.

[3] S. S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Technical Report, IBM Watson Research Center*, 1998.

[4] R. Huang and J. H. L. Hansen, "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, pp. 907–919, 2006.

[5] P. Maragos and A. Potamianos, "Fractal dimensions of speech sounds: Computation and application to automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 105, 1999.

[6] A. Samouelian and J. Robert-Ribes and M. Plumpe, "Speech, Silence, Music And Noise Classification Of Tv Broadcast Material," *Proc. IEEE Int. Conf. on Spoken Lang. Process.*, 1998.

[7] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997.

[8] L. Lu and H.-J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. Speech, Audio Process.*, vol. 10, pp. 504–516, 2002.

[9] P. Maragos and F.-K. Sun, "Measuring the Fractal Dimension of Signals: Morphological Covers and Iterative Optimization," *IEEE Trans. on Signal Processing*, vol. 41, pp. 108–121, 1993.

[10] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," *Int. J. High Perform. Comput. Appl.*, vol. 15, pp. 200–222, 2001.

[11] A. Chazapis, A. Zissimos, and N. Koziris, "A peer-to-peer replica management service for high-throughput grids," *Proc. Int. Conf. on Parallel Processing*, pp. 443–451, 2005.

[12] A. Zissimos, K. Doka, A. Chazapis, and N. Koziris, "Gridtorrent: Optimizing data transfers in the grid with collaborative sharing.," in *Proc. 11th Panhellenic Conference on Informatics (PCI2007)*, Patras, Greece, May 2007.

[13] I. Foster and C. Kesselman, "Globus: a metacomputing infrastructure toolkit," *Int. J. High Perform. Comput. Appl.*, vol. 11, no. 2, pp. 115, 1997.

[14] I. Foster, T. Freeman, K. Keahey, D. Scheftner, B. Sotomayor, and X. Zhang, "Virtual clusters for grid communities," *Cluster Computing and Grid (CCGRID)*, 2006.