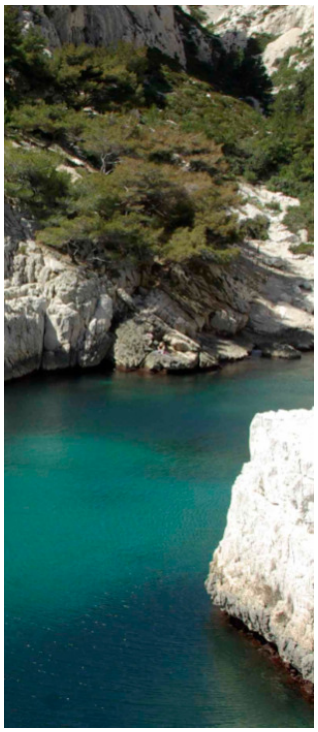
First Workshop on Speech, Language and Audio in Multimedia

InterSpeech satellite event | August 22/23, 2013 | Marseille (France)

Welcome



The first Workshop on Speech, Language and Audio in Multimedia (SLAM) aims at bringing together researchers working in speech, language and audio processing to analyze, index and access multimedia data.

Multimedia data are now available in very large amounts with a wide variety of formats and qualities, from professional content to user-generated ones:

- Lectures;
- Meetings;
- Interviews;
- Debates;
- Conversational broadcast;
- Podcasts;
- Social videos on the Web;
- etc.

Those data and the associated use scenarios raise specific challenges:

- Robustness facing the high variability in quality;
- Efficiency to handle very large amount of data;
- Semantics shared across modalities;
- Potentially high error rates in transcription;
- etc.

Worldwide, several national and international research projects are focusing on audio analysis of multimedia data. Various benchmark initiatives have been initiated such as

- TRECVID Multimedia Event Detection Track;
- MediaEval Benchmarking Initiative for Multimedia Evaluation;
- ETAPE Automatic Speech Processing Evaluation;

Latest news

Aug. 6th Proceedings are available, hosted by CEUR-WS.

Jun. 27th Workshop program is available.

Jun. 27th List of hotels close to the venue.

Jun. 23rd Workshop will take place in Pharo park.

Jun. 6th Registration website is now open.

About SLAM 2013

The workshop is organized in conjunction with Interspeech 2013 over 1.5 days, starting Thu. 22, 2013 at mid-day and ending Fri. 23, 2013 afternoon, right before the main conference. Marseille is conveniently connected by high-speed train to Lyon where the Interspeech conference will take place. The format of the workshop will include an invited talk, oral presentations of scientific work and a poster session for project and benchmark presentations.

The first SLAM workshop is jointly organized by the newly created ISCA SIG on Speech and Language in Multimedia and by the IEEE SIG on Audio and Speech Processing in Multimedia. This first edition is intended as the first of a series of workshop.

Questions regarding the workshop can be addressed to slam2013@easychair.org

Browse

- Call for papers
- Important dates
- Information to authors
- Information to participants

First Workshop on Speech, Language and Audio in Multimedia

InterSpeech satellite event | August 22/23, 2013 | Marseille (France)

People

General chairs

Guillaume Gravier
IRISA, Rennes, France

Frédéric Béchet
Aix Marseille Université, LIF-CNRS, Marseille, France

Scientific committee

Xavier Anguera
Telefónica

Frédéric Bechet
Aix Marseille Université, LIF-CNRS

Delphine Charlet
Orange Labs

Gerald Friedland
ICSI

Sadaoki Furui
Tokyo Institute of Technology

Guillaume Gravier
CNRS

Gareth Jones
Dublin City University

Martha Larson
TU Delft

Lin-Shan Lee
National Taiwan University

Georges Linarès
Université d'Avignon

Florian Metze
CMU

Organizing committee

Patrice Bellot
Aix Marseille Université, LSIS-CNRS

Hervé Bredin
LIMSI, CNRS

Benoît Favre
Aix Marseille Université, LIF-CNRS

Sylvain Meigner
LIUM, Université de Maine

Christian Raymond
IRISA, INSA Rennes

First Workshop on Speech, Language and Audio in Multimedia

InterSpeech satellite event | August 22/23, 2013 | Marseille (France)

Workshop Program

Schedule

Thursday August 22nd

- **14h00-14h15** Welcome & Introduction
- **14h15-16h00** Session 1 : Audio & Video event detection and segmentation
- **16h00-16h30** Coffe break
- **16h30-18h00** Session 2 : ASR in Multimedia documents
- **19h30-22h30** Banquet

Friday August 23rd

- **09h00-10h45** Session 3 : Multimedia person recognition
- **10h45-11h15** Coffee Break
- **11h15-12h15** Invited Talk : Sam Davies, BBC R&D
- **12h15-14h00** Lunch
- **14h00-15h45** Session 4 : Speaker & Speaker roles recognition
- **15h45-16h00** Coffee break
- **16h00-17h30** Session 5 : Multimedia applications and corpus
- **17h30-18h00** Discussion about SLAM & Closing

Keynote speaker

Sam Davies, BBC R&D

Sam Davies joined BBC R&D in 2007 working on a variety of projects including high frame rate television, object tracking in sporting events and image recognition. Since 2009 he has been working on the Multimedia Classification project, which has been identifying new techniques for metadata generation from audio and video content in the BBC archive. This work has resulted in prototypes which offer unique was to analyse the semantic and affective, or emotional, content of audio, visual and text documents.

In this talk we will present an overview of our work on the BBC's World Service Archive. This project uses automatic speech recognition and a novel technique for topic identification & disambiguation from noisy transcripts to enable automatic semantic tagging of programmes. Of course such processing across large archives is hard to scale so we'll present some of our work towards tackling this issue. We will also present our research on attempts to classify the entire BBC's archive of radio and television programmes broadcast since 1922. This project looks to classify programmes by creating new metadata from an analysis of programme content. This will primarily focus on the work we have done in identifying semantic content from audio (including music), along with a brief overview of our work on affective indexing. Here we will briefly introduce our multimodal work on identifying the mood or emotional component of a programme, focussing on mood identification from music, speech and non-speech sounds.

Program

Session 1 : Audio & Video event detection and segmentation

Hervé Bourlard, Marc Ferràs, Nikolaos Pappas, Andrei Popescu-Belis, Steve Renals, Fergus

Ms, Peter Bell, Sandy Ingram and Mael Guillemot

Processing and Linking Audio Events in Large Multimedia Archives: The EU inEvent Project

Abstract: In the inEvent EU project [1], we aim at structuring, re- trieving, and sharing large archives of networked, and dynam- ically changing, multimedia recordings, mainly consisting of meetings, video-conferences, and lectures. More specifically, we are developing an integrated system that performs audio- visual processing of multimedia recordings, and labels them in terms of interconnected “hyper-events” (a notion inspired from hyper-texts). Each hyper-event is composed of simpler facets, including audio-video recordings and metadata, which are then easier to search, retrieve and share. In the present paper, we mainly cover the audio processing aspects of the system, includ- ing speech recognition, speaker diarization and linking (across recordings), the use of these features for hyper-event indexing and recommendation, and the search portal. We present initial results for feature extraction from lecture recordings using the TED talks.

Benjamin Elizalde, Mirco Ravanelli and Gerald Friedland

Audio Concept Ranking for Video Event Detection on User-Generated Content

Abstract: Video event detection on user-generated content (UGC) aims to find videos that show an observable event such as a wedding ceremony or birthday party rather than an object, such as a wedding dress, or an audio concept such as music, speech or clapping. Different events are better described by different concepts. Therefore, proper audio concept classification enhances the search for acoustic cues in this challenge. However, audio concepts for training are typically chosen and annotated by humans and are not necessarily relevant to a specific event or the distinguishing factor for a particular event. A typical ad-hoc annotation process ignores the complex characteristics of UGC audio such as concept ambiguities, overlap concepts and concept duration. This paper presents a methodology to rank audio concepts based on relevance to the events and contribution to discriminability. A ranking measure guides an automatic or user-based selection of concepts in order to improve audio concept classification with the goal to improve video event detection. The ranking aids to determine and select the most relevant concepts for each event, discard meaningless concepts, combine ambiguous sounds to enhance a concept, thereby suggesting a focus for annotation and understanding of the UGC audio. Experiments show an improvement of the audio concepts mean classification accuracy as well as a better-defined diagonal in the confusion matrix. The selection of top 40 audio concepts using our methodology outperforms a best-accuracy-based selection by a relative 17.56% and a frame-frequency-based selection by 5.74%.

Diego Castán and Murat Akbacak

Segmental-GMM Approach based on Acoustic Concept Segmentation

Abstract: The amount of multimedia content is increasing day by day, and there is a need to have automatic retrieval systems with high accuracy. In addition, there is a demand for event detectors that go beyond the simple finding of objects but rather detect more abstract concepts, such as ``woodworking'' or a ``board trick.'' This article presents a novelty approach for event classification that enables searching by audio concepts from the analysis of the audio track. This approach deals with the acoustic concepts recognition (ACR) creating a trained segmentation instead a fixed segmentation as segmental-GMM approach with broad concepts. Proposed approach has been evaluated on NIST 2011 TRECVID MED development set, which consists of user-generated videos from the Internet, and has shown a EER of 40\%.

Diego Castán, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

Broadcast News Segmentation with Factor Analysis System

Abstract: This paper studies a novel audio segmentation-by-classification approach based on Factor Analysis (FA) with a channel compensation matrix for each class and scoring the fixed-length segments as the log-likelihood ratio between class/no-class. The system described here is designed to segment and classify the audio files coming from broadcast programs into five different classes: speech (SP), speech with noise (SN), speech with music (SM), music (MU) or others (OT). This task was proposed in the Albayzin 2010 evaluation campaign. The article presents a final system with no special features and no hierarchical structure. Finally, the system is compared with the winning system of the evaluation (the system use specific features with hierarchical structure) achieving a significant error reduction

in SP and SN. These classes represent 3/4 of the total amount of the data. Therefore, the FA segmentation system gets a reduction in the average segmentation error rate that is able to be used in a generic task.

Session 2 : ASR in Multimedia documents

Pierre Lanchantin, Peter Bell, Mark Gales, Thomas Hain, Xunying Liu, Yanhua Long, Jennifer Quinnell, Steve Renals, Oscar Saz, Matt Seigel, Pawel Swietojanski and Phil Woodland

Automatic Transcription of Multi-genre Media Archives

Abstract: This paper describes some recent results of our collaborative work on developing a speech recognition system for the automatic transcription or media archives from the British Broadcasting Corporation (BBC). The material includes a wide diversity of shows with their associated metadata. The latter are highly diverse in terms of completeness, reliability and accuracy. First, we investigate how to improve lightly supervised acoustic training, when timestamp information is inaccurate and when speech deviates significantly from the transcription, and how to perform evaluations when no reference transcripts are available. An automatic timestamp correction method as well as a word and segment level combination approaches between the lightly supervised transcripts and the original programme scripts are presented which yield improved metadata. Experimental results show that systems trained using the improved metadata consistently outperform those trained with only the original lightly supervised decoding hypotheses. Secondly, we show that the recognition task may benefit from systems trained on a combination of in-domain and out-of-domain data. Working with tandem HMMs, we describe Multi-level Adaptive Networks, a novel technique for incorporating information from out-of domain posterior features using deep neural network. We show that it provides a substantial reduction in WER over other systems including a PLP-based baseline, in-domain tandem features, and the best out-of-domain tandem features.

Christian Mohr, Christian Saam, Kevin Kilgour, Jonas Gehring, Sebastian Stüker, Alex Waibel

SLIGHTLY SUPERVISED ADAPTATION OF ACOUSTIC MODELS ON SUBTITLED BBC WEATHER FORECASTS

Abstract: In this paper we investigate the exploitation of loosely tran- scribed audio data, in the form of subtitles for weather fore- cast recordings, in order to adapt acoustic models for automati- cally transcribing those kinds of forecasts. We focus on dealing with inaccurate time stamps in the subtitles and the fact that they often deviate from the exact spoken word sequence in the forecasts. Further, different adaptation algorithms are compared when incrementally increasing the amount of adaptation mate- rial, e.g., by recording new forecasts on a daily basis.

Stefan Ziegler and Guillaume Gravier

A Framework for Integrating Heterogeneous Sporadic Knowledge Sources into Automatic Speech Recognition

Abstract: Heterogeneous knowledge sources that model speech only at certain time frames are difficult to incorporate into speech recognition, given standard multimodal fusion techniques. In this work, we present a new framework for the integration of this sporadic knowledge into standard HMM-based ASR. In a first step, each knowledge source is mapped onto a logarithmic score by using a sigmoid transfer function. Theses scores are then combined with the standard acoustic models by weighted linear combination. Speech recognition experiments with broad phonetic knowledge sources on a broadcast news transcription task show improved recognition results, given knowledge that provides complementary information for the ASR system.

Session 3 : Multimedia person recognition

Olivier Galibert and Juliette Kahn

The first official REPERE evaluation

Abstract: The REPERE Challenge aims to support research on people recognition in multimodal conditions. Following a 2012 dry-run, the first official evaluation of systems has been conducted at the beginning of 2013. To both help system development and assess the technology progress a specific corpus is developed. It current totals at 30 hours of video with multimodal annotations. The systems have to answer the following questions: Who is speaking? Who is present in the video? What names are cited? What names are displayed?

The challenge is to combine the various informations coming from the speech and the images.

Hervé Bredin, Johann Poignant, Guillaume Fortier, Makarand Tapaswi, Viet-Bac Le, Anindya Roy, Claude Barras, Sophie Rosset, Achintya Sarkar, Qian Yang, Hua Gao, Alexis Mignon, Jakob Verbeek, Laurent Besacier, Georges Quénot, Hazim Kemal Ekenel and Rainer Stiefelhagen
QCompere @ REPERE 2013
Abstract: We describe QCompere consortium submissions to the REPERE 2013 evaluation campaign. The REPERE challenge aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. First, four mono-modal components are introduced (one for each foregoing community) constituting the elementary building blocks of our various submissions. Then, depending on the target modality (speaker or face recognition) and on the task (supervised or unsupervised recognition), four different fusion techniques are introduced: they can be summarized as propagation-, classifier-, rule- or graph-based approaches. Finally, their performance is evaluated on REPERE 2013 test set and their advantages and limitations are discussed.

Benoit Favre, Géraldine Damnati, Frederic Bechet, Meriem Bendris, Delphine Charlet, Rémi Auguste, Stéphane Ayache, Benjamin Bigot, Alexandre Delteil, Richard Dufour, Corinne Fredouille, Georges Linarès, Jean Martinet, Gregory Senay and Pierre Tirilly
PERCOLI: a person identification system for the 2013 REPERE challenge
Abstract: The goal of the PERCOL project is to participate to the REPERE multimodal evaluation program by building a consortium combining different scientific fields (audio, text and video) in order to perform person recognition in video documents. The two main scientific challenges we are addressing are firstly multimodal fusion algorithms for automatic person recognition in video broadcast ; and secondly the improvement of information extraction from speech and images thanks to a combine decoding using both modalities to reduce decoding ambiguities.

Mohamed Hatmi, Christine Jacquin, Emmanuel Morin and Sylvain Meignier
Named Entity Recognition in Speech Transcripts following an Extended Taxonomy
Abstract: In this paper, we present a French named entity recognition (NER) system that was first developed as part of our participation in the ETAPE 2012 evaluation campaign and then extended to cover more entity types. The ETAPE 2012 evaluation campaign considers hierarchical and compositional taxonomy that makes the NER task more complex. We present a multi-level methodology based on conditional random fields (CRFs). Experiments were conducted using manually annotated training and evaluation corpora provided by the organizers of the campaign. The obtained results are presented and discussed.

Session 4 : Speaker & Speaker roles recognition

Benjamin Bigot, Corinne Fredouille and Delphine Charlet
Speaker Role Recognition on TV Broadcast Documents
Abstract: In this paper, we present the results obtained by a state-of-the-art system for Speaker Role Recognition (SRR) on the TV broadcast documents issued from the REPERE Multimedia Challenge. This SRR system is based on the assumption that cues about speaker roles may be extracted from a set of 36 low level features issued from the outputs of a Speaker Diarization process. Starting from manually annotated speaker segments, we first evaluate the performance of the SRR system, formerly evaluated on Broadcast radio recordings, on this heterogeneous set of TV shows. Consequently, we propose a new classification strategy, by observing how building show-dependent models improves SRR. The system is then applied on some speaker segmentation outputs issued from an automatic system, enabling us to investigate the influence of the errors introduced by this front-end process on Role Recognition. In these different contexts, the system is able to correctly classify 86.9% of speaker roles while being applied on manual speaker segmentations and 74.5% on automatic Speaker Diarization outputs.

Houman Ghaemmaghami, David Dean and Sridha Sridharan
Speaker Attribution of Australian Broadcast News Data

Abstract: Speaker attribution is the task of annotating a spoken audio archive based on speaker identities. This can be achieved using speaker diarization and speaker linking. In our previous work, we proposed an efficient attribution system, using complete-linkage clustering, for conducting attribution of large sets of two-speaker telephone data. In this paper, we build on our proposed approach to achieve a robust system, applicable to multiple recording domains. To do this, we first extend the diarization module of our system to accommodate multispeaker (>2) recordings. We achieve this through using a robust cross-likelihood ratio (CLR) threshold stopping criterion for clustering, as opposed to the original stopping criterion of two speakers used for telephone data. We evaluate this baseline diarization module across a dataset of Australian broadcast news recordings, showing a significant lack of diarization accuracy without previous knowledge of the true number of speakers within a recording. We thus propose applying an additional pass of complete-linkage clustering to the diarization module, demonstrating an absolute improvement of 20% in diarization error rate (DER). We then evaluate our proposed multi-domain attribution system across the broadcast news data, demonstrating achievable attribution error rates (AER) as low as 17%.

Carole Lailler, Grégor Dupuy, Mickael Rouvier and Sylvain Meignier
Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?
Abstract: Speaker identification is based on classification methods and acoustic models. Acoustic models are learned from audio data related to the speakers to be modeled. However, to record and annotate such data is time-consuming and labor-intensive. In this paper we propose to use data available on video-sharing websites like YouTube and Dailymotion to learn speaker-specific acoustic models. This process raises two questions: on the one hand, which are the speakers that can be identified through this kind of knowledge and, in the other hand, how to extract these data from such a noisy corpus that is the Web. Two approaches are considered to extract and annotate the data: the first is semi-supervised and requires a human annotator to control the process, the second is totally unsupervised. Speakers models created from the proposed approaches were experimented on the REPERE 2012 TV shows test corpus. The identification results have been analyzed in terms of speaker roles and fame, which is a subjective concept introduced to estimate the ease to model speakers.

Johann Poignant, Hervé Bredin, Laurent Besacier, Georges Quénot and Claude Barras
Towards a better integration of written names for unsupervised speakers identification in videos
Abstract: Existing methods for unsupervised identification of speakers in TV broadcast usually rely on the output of a speaker diarization module and try to name each cluster using names provided by another source of information: we call it "late naming". Hence, written names extracted from title blocks tend to lead to high precision identification, although they cannot correct errors made during the clustering step. In this paper, we extend our previous "late naming" approach in two ways: "integrated naming" and "early naming". While "late naming" relies on a speaker diarization module optimized for speaker diarization, "integrated naming" jointly optimize speaker diarization and name propagation in terms of identification errors. "Early naming" modifies the speaker diarization module by adding constraints preventing two clusters with different written names to be merged together. While "integrated naming" yields similar identification performance as "late naming" (with better precision), "early naming" improves over this baseline both in terms of identification error rate and stability of the clustering stopping criterion.

Session 5 : Multimedia applications and corpus

Abhigyan Singh, Martha Larson
Narrative-driven Multimedia Tagging and Retrieval: Investigating Design and Practice for Speech-based Mobile Applications
Abstract: This paper presents a design concept for speech-based mobile applications that is based on the use of a narrative storyline. Its main contribution is to introduce the idea of conceptualizing speech-based mobile multimedia tagging and retrieval applications as a story that develops via interaction of the user with characters representing elements of the system. The aim of this paper is to encourage and support the research community to further explore and develop this concept into mature systems that allow for the accumulation and access of

large quantities of speech-annotated images. We provide two resources intended to facilitate such work: First, we describe two applications, together referred as the ‘Verbals Mobile System’, that we have developed on the basis of this design concept, and implemented on Android platform 2.2 (API level 8) using Google's Speech Recognition service, Text-to-Speech Engine and Flickr API. The code for these applications has been made publically available to encourage further extension. Second, we distill our practical findings into a discussion of technology limitations and guidelines for the design of speech-based mobile applications, in an effort to support researchers to build on our work, while avoiding known pitfalls.

Larry Heck, Dilek Hakkani-Tur, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasarathy, Lisa Stifelman, Elizabeth Shriberg and Ashley Fidler
Multi-Modal Conversational Search and Browse
Abstract: In this paper, we create an open-domain conversational system by combining the power of internet browser interfaces with multi-modal inputs and data mined from web search and browser logs. The work focuses on two novel components: (1) dynamic contextual adaptation of speech recognition and understanding models using visual context, and (2) fusion of users' speech and gesture inputs to understand their intents and associated arguments. The system was evaluated in a living room setup with live test subjects on a real-time implementation of the multimodal dialog system. Users interacted with a television browser using gestures and speech. Gestures were captured by Microsoft Kinect skeleton tracking and speech was recorded by a Kinect microphone array. Results show a 16% error rate reduction (ERR) for contextual ASR adaptation to clickable web page content, and 7-10% ERR when using gestures with speech. Analysis of the results suggest a strategy for selection of multi-modal intent when users clearly and persistently indicate pointing intent (e.g., eye gaze), giving a 54.7% ERR over lexical features.

Korbinian Riedhammer, Martin Gropp, Tobias Bocklet, Florian Hönig, Elmar Nöth and Stefan Steidl
LMELECTURES: A MULTIMEDIA CORPUS OF ACADEMIC SPOKEN ENGLISH
Abstract: This paper describes the acquisition, transcription and annotation of a multi-media corpus of academic spoken English, the LMELectures. It consists of two lecture series that were read in the summer term 2009 at the computer science department of the University of Erlangen-Nuremberg, covering topics in pattern analysis, machine learning and interventional medical image processing. In total, about 40 hours of high-definition audio and video of a single speaker was acquired in a constant recording environment. In addition to the recordings, the presentation slides are available in machine readable (PDF) format. The manual annotations include a suggested segmentation into speech turns and a complete manual transcription that was done using BLITZSCRIBE2, a new tool for the rapid transcription. For one lecture series, the lecturer assigned key words to each recordings; one recording of that series was further annotated with a list of ranked key phrases by five human annotators each. The corpus is available for non-commercial purpose upon request.