Overview of the 2015 Workshop on Speech, Language and Audio in Multimedia

Guillaume Gravier, Gareth J. F. Jones, Martha Larson, Roeland Ordelman

Overview of the 2015 Workshop on Speech, Language and Audio in Multimedia

Guillaume Gravier
CNRS
IRISA & Inria Rennes
guig@irisa.fr

Gareth J.F. Jones
Dublin City University
School of Computing
gareth.jones@computing.dcu.ie

Martha Larson
Delft Univ. of Technology
Multimedia Computing Group
m.a.larson@tudelft.nl

Roeland Ordelman
Univ. of Twente & Netherlands
Institute for Sound and Vision
roeland.ordelman@utwente.nl

ABSTRACT

The Workshop on Speech, Language and Audio in Multimedia (SLAM) positions itself at at the crossroad of multiple scientific fields—music and audio processing, speech processing, natural language processing and multimedia—to discuss and stimulate research results, projects, datasets and benchmarks initiatives where audio, speech and language are applied to multimedia data. While the first two editions were collocated with major speech events, SLAM'15 is deeply rooted in the multimedia community, opening up to computer vision and multimodal fusion. To this end, the workshop emphasizes video hyperlinking as an showcase where computer vision meets speech and language. Such techniques provide a powerful illustration of how multimedia technologies incorporating speech, language and audio can make multimedia content collections better accessible, and thereby more useful, to users.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval—*Content Analysis and Indexing*

Keywords

Multimedia, Audio, Speech, Language

1. INTRODUCTION

The workshop on Speech, Language and Audio in Multimedia (SLAM) aims at bringing together researchers working in the fields of speech, language and audio processing with application to the analysis of, indexing of and access to multimedia data. In the context of SLAM, we define multimedia data as content captured, and often also edited, to create a message intended for humans, often combining multiple modalities. Typical data containing audio or language and fitting this definition are broadcast videos, video lectures, music and clips, or social media data. Note that in most cases, speech, language and audio are important carriers of semantics in multimedia content. In particular, language is of utmost importance for understanding the nature of the message. Yet these sources of information are often overlooked, or remain poorly exploited by researchers working to develop new multimedia technologies.

SLAM is by nature interdisciplinary, existing at the intersection of multiple scientific communities, including the research area of multimedia. Apart from traditional workshop objectives of sharing scientific results and visions, one of the goal of SLAM is to establish and make visible a cross-domain scientific community. To this end, SLAM gathers players from the fields of speech and audio processing, of natural language processing and of multimedia to share recent research results, discuss ongoing and future projects, explore potential areas for interdisciplinary collaboration or sharing or ideas, and develop new benchmarking initiatives of mutual interest to multimedia and language researchers.

SLAM'15 is the third edition of the workshop. The workshop series was initiated in 2013 by the Intl. Speech Communication Association (ISCA) Special Interest Group on Speech and Language in Multimedia[1], with support from the IEEE Special Interest Group on Audio and Speech Processing in Multimedia[2]. Previous editions [1, 6] were collocated with Interspeech, the premier international conference on speech communication yearly organized by ISCA. However, the long-term goal is to establish SLAM as a regular workshop, collocating in alternance with major multimedia conferences and major speech and language conferences, as a bridge between these domains.

Organizing SLAM'15 as an ACM Multimedia event is thus in logical continuation from the preceding editions. Previous efforts to highlight the importance of SLAM-related topics at ACM Multimedia include the workshop on Searching Spontaneously Conversational Speech [5] and the workshop on Audio and Multimedia Methods for Large-Scale Video Analysis [4]. SLAM transcends these efforts because it emphasizes language and audio in addition to speech, goes beyond the application of search and the large-scale issues, and

[1] http://slim-sig.irisa.fr
[2] http://www.computer.org/portal/web/tcmc/SIGASP

MM'15 October 26-30, 2015, Brisbane, Australia

also puts an explicit focus on work in which speech, language or audio is *one* or multiple modalities.

In order to emphasize its connection with the multimedia community, SLAM'15 features a special focus on video hyperlinking, a multimedia task that was recently introduced in international benchmarks, where computer vision and image processing meets speech, language and audio. Details on the video hyperlinking session are given in Sec. 4.

2. WORKSHOP SCOPE AND GOALS

Multimedia data are available in enormous volumes in 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. A large number of those sources include audio, speech and language, which are considered as key modalities for structuring and understanding, e.g., speaker turns, topic segmentation, entities and keywords. Analyzing audio, speech and language in multimedia data raises specific challenges that arise both from the nature of the data (e.g., semantics, variety, variability, multimodality) and from the typical use-cases considered in multimedia applications. Typical challenges specific to the nature of multimedia data include robustness in face of high variability in quality, efficiency to handle very large amounts of data, semantics shared across modalities, or potentially high error rates in transcription affecting spoken language processing. An important set of challenges also arises from the application scenarios in which users make use of multimedia. User needs evolve rapidly, often during the experience of interacting with multimedia data, and users have a difficult time expressing their own needs concretely, or formulating them explicitly in a way that can be directly exploited by a system. All of these factors may also make it difficult to create the large and representative data sets needed in order to develop solutions to multimedia challenges.

Worldwide, several national and international research projects are focusing on audio analysis of multimedia data. Similarly, various benchmark initiatives have devoted effort to offering tasks related to multimodal multimedia challenges (e.g., TRECVid, CLEF, MediaEval). The aim of the workshop is to support these efforts by drawing attention to their importance, and to the variety of interdisciplinary solutions that are needed in order to address current challenges in multimedia.

3. TOPICS

The workshop embraces a broad range of contributions including research work, project descriptions, evaluation initiatives, demonstrations and applications. The common denominator is that each emphasizes the added value of speech and/or language and/or audio for multimedia technology.

This year the workshop includes papers that address a wide variety of topics. Examples include: the exploitation of the structure or music, the challenge processing of spoken clinical reports, in a way that combines speech recognition and human checks, the use of multimodal data in training systems that are capable of detecting spoken words, and finally, how audio information can be integrated into an overall visual object retrieval system. A large portion of the workshop is devoted to the showcase of the task of video hyperlinking, which we turn to next.

4. VIDEO HYPERLINKING CASE STUDY

A key feature of SLAM'15 is the focus on video hyperlinking, with a dedicated session and round-table. Video hyperlinking is the task of organizing a collection of videos with (hyper)links from segments of interest, also know as *anchors*, to relevant segments within the collection. The video hyperlinking task was recently introduced in the MediaEval international benchmarking initiative [3, 2], moving to TRECVid in 2015.

Interestingly in the context of SLAM, accurate video hyperlinking approaches require the integration of multiple modalities, in particular image and language, where language sources are typically automatic transcripts, subtitles or program synopsis. The multimodal nature of the video hyperlinking task makes it an emblematic case study where the speech and language modalities are perfectly complemented by audio and vision. The workshop program features a number of contributions where audio and natural language processing are used for video hyperlinking, possibly in conjunction with images. A panel discussion focused on discussing the past, present and future of hyperlinking will conclude the workshop. The panel will aim at an understanding of which approaches are most promising and how they can be evaluated. The goal is to shape research directions at the crossroad of the scientific communities involved in SLAM and nurturing future implementations of video hyperlinking benchmarks.

5. ACKNOWLEDGEMENTS

We would like to thank the authors, presenters, and reviewers, whose efforts made the workshop possible.

6. REFERENCES

[1] F. Bechet and G. Gravier, editors. *Proceedings of the 1st IEEE/ISCA International Workshop on Speech, Language and Audio in Multimedia*, 2013.

[2] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2014. In *Working Notes Proc. of the MediaEval Workshop*, 2014.

[3] M. Eskevich, G. J. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. Van de Walle, P. Galuscakova, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking. In *Proc. ACM Intl. Conf. on Multimedia Retrieval*, pages 287–294, 2013.

[4] G. Friedland, D. Ellis, and F. Metze, editors. *Proceedings of the 2012 ACM International Workshop on Audio and Multimedia Methods for Large-scale Video Analysis*, 2012.

[5] M. Larson, R. Ordelman, F. de Jong, J. Köhler and W. Kraaij, editors. *Proceedings of the 3rd Workshop on Searching Spontaneous Conversational Speech*, 2009.

[6] T.-P. Tan, G. Gravier, and K. M. Siti, editors. *Proceedings of the 2nd IEEE/ISCA International Workshop on Speech, Language and Audio in Multimedia*, 2014.