Blending the physical and the virtual: multi-modal signal processing for music interfaces (tutorial proposal)

George Tzanetakis
Dept. of Computer Science
University of Victoria, Canada
gtzan@cs.uvic.ca

ABSTRACT

In the past few years there has been a signficant increase of interest in rich mult-modal user interfaces that go beyond the traditional mouse/keyboard/screen interaction. Such interfaces use a variety of sensing (and actuating) modalities to receive and present information to users. Unlike simple interfaces like a mouse or keyboard, they typically require techniques from signal processing and machine learning in order to extract and fuse high level information from noisy, high dimensional signals over time. The availability of affordable rich controllers with multiple modalities such as the Nintendo Wii and the Microsoft Kinect has opened many exciting new possibilities for the development of such interfaces. These rich interfaces pose many interesting signal processing challenges while offering fascinating possibilities for new research. The goal of this half day advanced tutorial is to cover signal processing and machine learning techniques that are needed for the design and development of such interfaces. New interfaces for musical expression offer great examples and early prototypes of rich, interactive multi-modal interfaces and will be used as case studies to illustrate the underlying concepts.

1. MOTIVATION

We are probably at the cusp of a new era in human computer interaction in which users will view the way we interacted with computers using keyboards and mice, as we view today punched cards and teletypes. The large diversity of affordable hardware in terms of sensors and actuators today, provides a fertile environment for creating rich multimodal interfaces. Frequently, their design and development requires interdisciplinary expertise combining concepts from digital signal processing (DSP) and human computer interaction (HCI).

There is an increasing interest in rich user interfaces that go beyond the traditional mouse/keyboard/screen interaction. In the past few years, their development has accelerated due to the wide availability of commodity sensors and actuators. For example, the Microsoft Kinect provides, at low cost, a structured light infrared depth camera, a regular color camera, and a microphone array. Moreover, smart phones contain a variety of additional sensors such as accelerometers that provide unique opportunities for control. Unlike traditional controllers such as a mouse that provide direct and simple sensor readings, these new rich interfaces provide high dimensional, noisy and complex sensor readings. Therefore sophisticated digital signal processing and machine learning techniques are required in order to develop effective human computer interactions using such intefaces. At the same time they offer fascinating possibilities of blending the physical and virtual world such as non-invasive full body control and augmented reality. Different algorithms and customizations are required for each specific application and possibly user so there is a large design space for novel research and contributions.

Multimedia research has always been an interdisciplinary research area and research in rich interactive multi-modal interfaces is even more so. Researchers especially coming from a computer science and human computer interaction background frequently don't have direct training in digital signal processing techniques. At the other end of the spectrum, researchers with a DSP background frequently don't have direct training in techniques from human-computer interaction especially related to the evaluation of user interfaces. On top of that familiarity with circuits and embedded computing is also frequently required. The goal of this tutorial is to provide an advanced overview of this topic with specific emphasis on digital signal processing and machine learning techniques although some of the other topics (HCI and physical computing) will also be touched. New interfaces of music will be used as particularly interesting and motivating case studies to illustrate the concepts.

Traditional musical instruments are some of the most fascinating artifacts created by human beings in history. The complexity and richness of control afforded by an acoustic musical instruments, such as a cello, to a professional musicians is impressive. Research in new interfaces for musical expression has explored how such complex and delicate control can be combined with the ability to interact with a computer. In some ways, research in this area has anticipated the development of augmented reality rich multimodal interfaces and provides good examples of rich sensory intearctions. Based on these considerations, new musical instruments will be used throughout this tutorial in order to present working case studies and examples for illustrating particular concepts. At the same time, the concepts covered

ACM Multimedia 2013, Barcelona, Spain

are general and can be applied to any type of rich multimodal human computer interaction. There is also a rich literature focusing on speech interaction (speech recognition, and synthesis) that can also be multi-modal (typically audio-visual). In this tutorial, we will only briefly touch on speech based interactions and only for ideas and concepts that are relevant to other types of interaction as it is a more specialized topic that is much more covered in existing literature and previous tutorials.

2. RELEVANCE AND TIMELINESS

Research in rich, interactive multi-modal interfaces is fundamentally part of multmedia research. It could even be considered as a new major paradigm in multimedia research which initially was focused on trasmission and representation of multimedia data, then became focused on analysis of multimedia data mostly in large collections, and now is more and more embracing interaction. Of course there is a large body of work that stretches back decades in this area but the last few years there have been some catalysts that have made research in this area more accessible and affordable. The first has been the availability of low cost sensing commercial sensing devices such as the W-ii and the Kinect. Even as recently as five years ago similar working on interfaces with similar sensing capabilities would have required specialized, custom-made hardware at significantly higher cost. Another catalyst has been the increased activity in physical and embedded computing with devices such as the Arduino and the lower cost of 3D printing. Finally the amazing growth and pervasive nature of smart phones means that a large number of computer users now have portable, ubiquitous computing devices with rich sensory capabilities. Finally this tutorial is particularly relevant to the increasing awareness of music and audio related research in the ACM Multimedia community as evidenced by the creation of a new track for submission in this area following two successful ACM MIRUM workshops the last two years.

To the best of my knowledge there have not been previous tutorials on the exact topic. Multi-modal human computer interaction has been covered but mostly in the context of speech and mobile search. Real-time algorithms have also been covered but focusing on computer vision and smart cameras. There have also been tutorial on music signal processing but not for the creation of real-time interactive systems. As a conclusion, even though different concepts covered by the tutorial have been presented in other contexts, the material is novel and timely.

3. TARGET AUDIENCE

The target audience is graduate students and researchers without a strong background in digital signal processing that are interested in designing and developing novel human-computer interactions using rich multi-modal interfaces. It also targets researchers interested in multi-modal interfaces that come from a computer science background and might not be familiar with the possibilities offered by modern digital signal processing techniques. It should also be of interest to researchers and graduate students with DSP experience. Even though many of the techniques used in the design and development of rich multi-modal interfaces, such as dynamic time warping (DTW) or Hidden Markov Models (HMM), are familiar to DSP practitioners, real-time interactive interfaces pose specific challenges such as real time implementation, causality, fault tolerance that are not as critical in other areas of DSP. Therefore, part of the tutorial will show how DSP can be adapted for this application context to perform tasks such as automatic calibration, gesture detection, and tracking. The design and development of rich multi-modal interfaces is inherently interdisciplinary and also involves knowledge that is typically not part of the standard electrical and computer engineering curriculum. For example, the evaluation of such interfaces is not as straighforward as evaluation is in other areas of DSP and requires understanding of concepts from human computer interaction. Because of the nature of this research, the literature tends to be scattered across many different communities which makes it harder to pursue for new comers to the field. This tutorial tries to cover the all the basic concepts needed to embark in this journey with particular emphasis on digital signal processing and machine learning and a specific focus on music interfaces and interaction.

4. OUTLINE

The tutorial is structured in 6 modules with length of either 15, 30 or 45 minutes for a total of 3 hours. The exact durations are flexible and will be determined by the interests of the participants. The following modules are proposed:

4.1 Motivation and Overview (15 minutes)

A historical overview of human computer interaction culminating with rich multi-modal interfaces with the specific emphasis on what motivated their evolution will be provided. In addition this module will provide a general overview of the field and discuss different ways of organizing the material in terms of both application areas and underlying concepts.

4.2 Sensors and Actuators (15 minutes)

In this module, we discuss the various types of sensors and actuators that are available today with an emphasis on their signal characteristics. Sensor examples include: accelerameters, force-sensing resistors, microphones and microphone arrays, color/depth cameras, and touch surfaces. Actuators include various types of motors, solenoids, and displays. We will also discuss specific hardware examples such as the Nintendo Wii controller, the Microsoft Kinect, and the Apple iPhone from a sensor/actuator perspective.

4.3 Signal Conditioning and Feature Extraction (45 minutes)

Unlike simple sensors such as an optical mouse, modern interfaces with multiple sensors typically require a layer of digital signal processing to provide useful data. Even though in some cases this functionality can provided by the software drivers of the device, in many cases custom signal conditioning and feature extaction algorithms need to be written. In this module we cover the basics of signal conditioning with topics such as denoising, dimensionality reduction, filtering, resampling, and dealing with missing samples with concrete examples using multi-modal sensor data. In addition, feature extraction is also covered. Topics covered include: various types of time-frequency analysis (short time fourier transform (STFT), wavelets), i basic image and video analysis (segmentation, optical flow, single and multiple object tracking), and modeling dynamics (periodicity detec-

tion, self-similarity matrices). The algorithms will be illustrated with specific examples using audio, video, and sensor data in the context of new interfaces for musical expression.

4.4 Dealing with uncertainty & time (45 minutes)

Once time series of feature vectors for the different sensing modalities have been computed, in many cases further processing is required for effective human interaction. A canonical example is gesture detection in which a particular pattern of sensor readings over time needs to be detected accurately despite noise due to both acquisition and user variability. In this module, techniques from machine learning and data mining such as supervised (classification and regression) and unsupervised (clustering) learning will be described from a multi-modal interface perspective. The algorithmic constraints of real time and causal implementation will also be covered. Interaction always takes place over time and therefore techniques for modeling uncertainty over time are needed. Techniques for time modeling such as Kalman Filters, Hidden Markov Models (HMM), nd Dynamic Time Wrapping (DTW) will also be described through specific examples from multi-modal interaction. Finally, various approaches to sensor fusion will be discussed.

4.5 Human-Computer Interaction (30 minutes)

|

The evaluation of user interfaces requires user involvement and concepts from the field of human computer interaction. Connections with the recent interest in quality of multimedia experience will aslo be made. In this part of the tutorial, general methodologies for design and evaluation of interfaces such as user centered and participatory design will be presented. A variety of methods for receiving feedback about an interface such as wizard of Oz prototyping, ethnography, and qualitative and guantitative analysis will also be surveyed. There are several statistical tests that are used, and in some cases misused, in human computer interaction research. Depending on time we will cover the definition and use of statistic tests such as t-test and ANOVA, pearson correlation, linear and non-linear regression. We will also cover the definition and use of statistic tests for ordinal data such as the Kruskal-Wallis test and Spearman correlation. The distinction between parametric and non-parametric tests will also be discussed. In all cases the concepts will be illustrated through specific examples from rich multi-modal user interfaces.

4.6 Integration, Implementation and Case Studies (30 minutes)

In this part of the tutorial, the integration of the various concepts that were covered into a working rich multi-modal interface is disussed. Rich multi-modal user interfaces need to operate in real-time and gracefully interact with users which necessitates more complex software architectures that what many researchers in digital signal processing are used to. We briefly touch upon some issues of software engineering such as unit testing, modular design and templates for user interaction such as the classic Model/View/Controller design pattern. A number of software environments, languages and packages, that can be used for the design and development of rich multi-modal interfaces, will also be briefly described through specific case studies. Examples include OpenCV, Open Frameworks, Marsyas, R, and Weka. The interfacing of sensors and actuators through micro-controllers, such as the Arduino, will also be discussed. In this part of the tutorial, specific case studies that utilize the concepts covered in the other parts from the field of new interfaces for musical expression will be presented. Depending on time the following systems will be presented: a surrogate sensor methodology for training non-invasive signal acquisition using direct sensors through classification and regression, teaching a virtual violinist to bow using machine learning to automatically determine sound quality, self-tuning and self calibrating music robotic instruments, extending the performance possibilities of a vibraphone using the Microsoft Kinect controller, and the Soundplane, multi-touch, pressure sensitive new music interface that provides highly expressive control possibilities. Although these user cases are drawn from the work of the presenter in order to have complete information about all the details of implementation and copyright clearance for all the figures they are representative of work in this area and cover a wide variety of different approaches.

5. MATERIAL DISTRIBUTED

The participants to the tutorial will receive digital (or hard) copies of the slides as well as an annotated bibliography of work relevant to the topic of approximately 100 citations. A website that will be created will also provide additional resources such as links to software and hardware (sensors, actuators), notes related to the topics covered, case studies, demonstration videos, and code examples.

6. PRESENTER BIO

George Tzanetakis is an Associate Professor in the Department of Computer Science with cross-listed appointments in ECE and Music at the University of Victoria, Canada. He is Canada Research Chair (Tier II) in the Computer Analysis and Audio and Music and received the Craigdaroch research award in artistic expression at the University of Victoria in 2012. In 2011 he was Visiting Faculty at Google Research. He received his PhD in Computer Science at Princeton University in 2002 and was a Post-Doctoral fellow at Carnegie Mellon University in 2002-2003. His research spans all stages of audio content analysis such as feature extraction,segmentation, classification with specific emphasis on music information retrieval. He is also the primary designer and developer of Marsyas an open source framework for audio processing with specific emphasis on music information retrieval applications. His pioneering work on musical genre classification received a IEEE signal processing society young author award and is frequently cited. He has given several tutorials in well known international conferences such as ICASSP, ACM Multimedia and ISMIR. More recently he has been exploring new interfaces for musical expression, music robotics, computational ethnomusicology, and computer-assisted music instrument tutoring. These interdisciplinary activities combine ideas from signal processing, perception, machine learning, sensors, actuators and human-computer interaction with the connecting theme of making computers better understand music to create more effective interactions with musicians and listeners. More details can be found http://www.cs.uvic.ca/ gtzan.