Fusion of Multimodal Information in Music Content Analysis*

Slim Essid and Gaël Richard

Institut Télécom, Télécom ParisTech, CNRS-LTCI
37 rue Dareau, 75014 Paris, France
Slim.Essid@telecom-paristech.fr, Gael.Richard@telecom-paristech.fr

Abstract

Music is often processed through its acoustic realization. This is restrictive in the sense that music is clearly a highly multimodal concept where various types of heterogeneous information can be associated to a given piece of music (a musical score, musicians' gestures, lyrics, user-generated metadata, etc.). This has recently led researchers to apprehend music through its various facets, giving rise to *multimodal music analysis* studies. This article gives a synthetic overview of methods that have been successfully employed in multimodal signal analysis. In particular, their use in music content processing is discussed in more details through five case studies that highlight different multimodal integration techniques. The case studies include an example of cross-modal correlation for music video analysis, an audiovisual drum transcription system, a description of the concept of informed source separation, a discussion of multimodal dance-scene analysis, and an example of user-interactive music analysis. In the light of these case studies, some perspectives of multimodality in music processing are finally suggested.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Multimodal music processing, music signals indexing and transcription, information fusion, audio, video

Digital Object Identifier 10.4230/DFU.Vol3.11041.37

1 Introduction

While the most natural way to perceive music is through its acoustic rendering, it is clear that it is a highly multimodal concept that can be sensed in a variety of ways: music is materialized in the head of a composer, or a trained musician reading a musical-score; it is translated into sound and motion in a performer's gestures or a dancer's movements and steps; it becomes visual art when it is illustrated by disc cover designs or transformed into an audiovisual production; not to mention its textual dimension that encapsulates not only the lyrics (in sung music) and editorial metadata, but also social web content such as user-tags, reviews, ratings, etc.

Consequently, treating music only through its acoustic realization appears to be quite restrictive, which has led researchers in the general field of music content analysis to apprehend it through its various facets, giving rise to *multimodal music analysis* studies. To our knowledge the earliest contributions along this line dealt with two modalities, that is the audio and score modalities, in order to perform music-to-score matching [10, 66]. The more complex visual modality has not been exploited in music analysis until the late 90s [60],

* This work was partially supported by the European commission with the 3Dlife project

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 37–52

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

in contrast to the speech processing domain where audiovisual speech recognition systems have been imagined in the 80s [59]. Not surprisingly, the earliest works on audiovisual music were dedicated to the analysis of piano music [60], [62], probably due to the possibility to segment the keyboard keys and track the musician's fingers positions on the keyboard more easily than with other instruments.

Since then, our field of interest has seen a variety of multimodal studies spanning a wide range of techniques and applications, an overview of which is proposed in this article. We will first provide a synthetic view of methods that have been successfully employed in multimodal research works in general and discuss their use for music processing. Subsequently, we will discuss a selection of case studies we have contributed to, and highlight the related future research directions that seem promising to us.

2 Multimodal Techniques

Multimodal processing techniques, in general, fall into one of two categories of a binary taxonomy: *early integration techniques* as opposed to *late integration techniques*.[1] The former refers to the process whereby a system directly exploits the "raw" low-level features used to describe each data stream, without any further transformations other than basic postprocessing (typically denoising, normalisation, resampling, etc.). By contrast, the latter is employed to indicate that the joint exploitation of the modalities is performed at a *decision-level*, typically by combining the outputs of intermediate monomodal classifiers. This distinction will be useful to understand the differences between the techniques presented hereafter. Another interesting distinction is the following: the effort of characterizing the "relationships" between the different modalities reflecting the content being analyzed is referred to as *cross-modal processing*, while the problem of efficiently combining the information conveyed by the different modalities (to perform a more thorough analysis of the content) is called *multimodal fusion*. Below we further describe the previous paradigms and discuss their exploitation in the field of music processing.

2.1 Cross-Modal Processing

The relationships between the modalities considered can be expressed in several different ways.

In the first place, when dealing with modalities having a temporal dimension (typically audio signals, video signals, or musical scores), it might be required to temporally align the different data streams in case they are not initially synchronized. In fact, achieving this synchronization may be one's ultimate goal: for instance when dealing with the audio and score modalities, this task is often referred to as *music-to-score alignment* (or *music-to-score synchronization*) [50, 41]. Since the latter is already well covered in other articles of this book, we will here assume that the data streams considered are temporally aligned.

Assuming synchronized features, many proposals have been made to measure a form of dependency between two heterogeneous data streams, part of which remain under-exploited in the music information retrieval community, despite their potential. For the sake of clarity, we make the assumption (without loss of generality) that two streams of data are considered: an audio stream and a video stream. Though the methods presented in the following have

[1] It is worth mentioning that hybrid approaches exist too.

been mainly applied to those two particular modalities, they can be used with any other parallel data streams whose dependency is to be characterized.

A number of techniques have been suggested to map the observed audio and visual feature vectors to a low dimensional space where a *measure of "dependency"* between them can be computed. Let us assume the n observed audio feature vectors $x_a \in \mathbb{R}^{D_a}$ are assembled column-wise in a $(n \times D_a)$-matrix X_a, and the corresponding visual feature vectors[2] $x_v \in \mathbb{R}^{D_v}$ are assembled column-wise in a $(n \times D_v)$-matrix X_v. The methods we describe here aim to find two mappings f_a and f_v (that reduce the dimensions of the audio and visual feature vectors), such that a dependency measure $S_{av}(f_a(X_a), f_v(X_v))$ is maximized. Various approaches can be described using this same formalism. Darrel *et. al.* choose the mutual information [8] as a dependency measure and seek single-layer perceptrons f_a and f_v projecting the audiovisual feature vectors to a 2-dimensional space. Other more popular approaches, for which closed-form solutions can be found, use linear mappings to project the feature streams:

- Canonical Correlation Analysis (CCA), first introduced by Hotelling [33], aims at finding pairs of unit-norm vectors t_a and t_v such that

$$(t_a, t_v) = \arg \max_{(t_a,t_v)\in\mathbb{R}^{D_a}\times\mathbb{R}^{D_v}} \operatorname{corr}(t_a^t X_a, t_v^t X_v)\,. \tag{1}$$

- An alternative to the previous (expected to be more robust than CCA) is Co-Inertia Analysis (CoIA). It consists in maximizing the covariance between the projected audio and visual features:

$$(t_a, t_v) = \arg \max_{(t_a,t_v)\in\mathbb{R}^{D_a}\times\mathbb{R}^{D_v}} \operatorname{cov}(t_a^t X_a, t_v^t X_v)\,. \tag{2}$$

- Yet another configuration known as Cross-modal Factor Analysis (CFA), and found to be more robust than CCA in [45], seeks two matrices T_a and T_v, such that

$$(T_a, T_v) = \arg \max_{(T_a,T_v)} (1 - ||T_a X_a - T_v X_v||_F^2) = \arg \min_{(T_a,T_v)} ||T_a X_a - T_v X_v||_F^2; \tag{3}$$

  with $T_a T_a^t = I$ and $T_v T_v^t = I$. $||X||_F$ denotes the Frobenius norm of matrix X.

Note that the previous techniques can be kernelized to study non-linear coupling between the modalities considered (see for instance [44, 31]).

The interested reader is referred to [33, 31, 45] for further details on these techniques, and to [25] for a comparative study. Examples of applications in the field of music content processing are mentioned in Section 2.4.

2.2 Feature-Level Fusion

Feature-level fusion is the (early integration) process of combining different types of features from different modalities into a common feature representation.

The most basic audiovisual feature fusion approach consists in concatenating the audio and visual feature vectors, x_a and x_v, to form a global feature vector $x_{av} = [x_a, x_v]$. However,

[2] The underlying assumption is that the (synchronized) audio and visual features are extracted at the same rate, which is often obtained by downsampling the audio features or upsampling the video features, or by using temporal integration techniques [40].

the dimensionality of the resulting representation is often too high, leading researchers to resort to dimensionality reduction methods.

A common approach is to use *feature transformation* techniques such as Principal Component Analysis (PCA) [5], Independent Component Analysis (ICA) [63], or Linear Discriminant Analysis [5]. An interesting alternative, is *feature selection* [30] which aims to select only useful descriptors for a given task and discard the others. Indeed, when applied to the feature vectors x_{av}, feature selection can be considered as a feature fusion technique whereby the output will hopefully retain the "best of x_a and x_v" *i.e.* a subset of the most relevant audiovisual features (with respect to the selection criterion).

Nevertheless, the two previous approaches can be considered as limited owing to the different physical nature of the audio-visual features to be combined. In particular, the features do not necessarily live in the same metric space, and are not necessarily extracted from the same temporal segments. Consequently, there has been a number of proposals attempting to address these limitations. One possible approach consists in building separate kernels for different features, before determining new optimal kernels (as convex combinations of the individual ones) in order to use them for classification [70]. Another possible approach of note is the construction of joint audiovisual representations, envisaged as *audiovisual atoms* in [38], and *audiovisual grouplets* in [39], both exploiting audiovisual correlations. The joint audiovisual representation may in particular be built using one of the audiovisual subspace methods described in Section 2.1 (see [45] for an example).

2.3 Decision-level fusion

Late fusion or the idea of combining intermediate monomodal decisions[3] in order to achieve a more accurate multimodal characterization of a content has been explored extensively, under various configurations.

Numerous works rely on *majority voting* procedures whereby final global decisions are made based on a weighted sum of individual voters, each typically corresponding to a decision taken on a particular modality. The weights are often chosen using either heuristics or trial-and-error procedures (see for example [46]). This idea can be better formalized using a Bayesian framework, which allows for taking into account the uncertainty about each classifier's decisions, as done in [36]. Also, solutions to deal with the potential imprecision of some modalities have been proposed using the *Dempster-Shafer* theory [19]. Another widely used strategy consists in using the monomodal classifiers outputs as features, on the basis of which a new classifier, that is expected to optimally perform the desired multimodal fusion, is learned [68].

The previous approaches do not account for the dynamic properties of the media streams considered, nor do they allow for encoding prior knowledge about the dependency structure in the data, in particular the temporal and/or cross-modal dependencies. To this end, sophisticated dynamic classifiers have been utilized, ranging from variants of (multi-stream) Hidden Markov Models (HMM) [28, 52, 43, 1], through more general Dynamic Bayesian Networks (DBN) [6, 27], to even more general graphical models such as Conditional Random Fields (CRF) [41, 3].

[3] These decisions are generally output by previously trained classifiers.

Table 1 Case studies presented. The "Modalities" are the ones taken into account in the corresponding case study; "Cross-modal" indicates whether the method presented performs cross-modal analysis; "Fusion" indicates whether it exploits multimodal fusion and "Section" is where in this chapter the case study is presented.

Case studies	Modalities	Cross-modal	Fusion	Section
Audiovisual correlation in music videos	audio, video	•	○	3.1.1
Audiovisual drum transcription	audio, video	○	•	3.1.2
Music in motion: analyzing dance scenes	audio, motion, depth, video, choreographies	•	•	3.2
Interactive music analysis	audio, human	•	•	3.3
Informed source separation	audio, score, human	•	○	3.4

2.4 Discussion

Many of the techniques mentioned above have been exploited in multimodal music content analysis research. Cross-modal analysis seems to be particularly popular within this domain. For instance, CCA has been used both for studying correlations between sounds and human motion or gestures [54, 55], and correlations between music and words, in view of creating a musically meaningful vocabulary [65]. Also, heuristic rules for the association of higher-level descriptors extracted from different modalities have been employed [4, 21]. In fact, it seems that approaches relying on heuristic rules are mainstream, be it for specific content analysis tasks, such as music video summarization [71], or more general classification problems (see for example [46] where the output of audio and visual classifiers are heuristically combined).

We believe there is a great potential in exploiting the more sophisticated cross-modal techniques and dynamic statistical models previously mentioned to be able to better express one's prior knowledge on the data structure (features dependency, temporal synchronisation, multi-scale effects, higher-level cross-modal concept relationships, etc.) and fully exploit the valuable information that is encoded in it. This of course entails a formalisation effort which is expected to be rewarding both in terms of performance and generalization insofar as the purpose of using common architectures for different applications can be pursued. Modeling the ambiguity and imprecision of intermediate (mono-modal) decisions thanks to the Dempster-Shafer theory of evidence is another interesting idea that is believed to hold much promise.

3 Case Studies

We now present particular multimodal music applications that we have treated in the past few years illustrating the techniques introduced in Section 2. Table 1 gives an overview of these case studies indicating the modalities considered and the class of techniques employed.

3.1 Audiovisual Music

3.1.1 Audiovisual Correlation in Music Videos

The first case study is dedicated to a specific aspect of multimodal signal analysis and aims at exploiting the correlation between the audio and visual modalities in music videos [21].

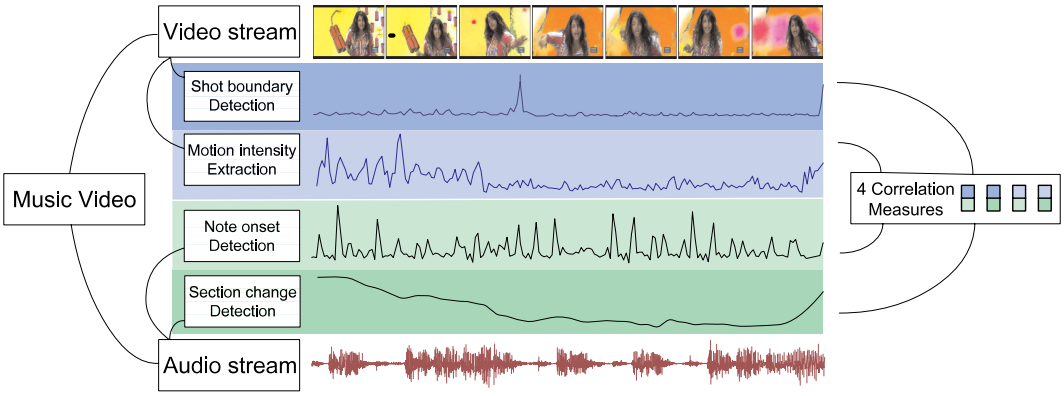

Figure 1 Overview of the audio-visual content structuring system (from [21]).

In the case of music videos, a large palette of semantic relationships between the audio and video streams may be used by the artists at the production stage. For example, mainstream music videos show dancers or performers, but some videos have a narrative content based on higher-level features of the song (such as structure or mood) while others explore new forms of visual metaphors [26, 42, 53].

In this case study (further described in [21]) high-level structures of the audio and video streams are separately extracted in order to measure the correlations between these structures. The objective in such an approach is to characterize the synchrony of significant events and changes in the music and the accompanying images.

It is clear that a large number of salient events can be defined both for audio and visual streams. In music signals, note or chord changes are obviously important events. Thus, an efficient mid-level temporal structuring of a music piece can be achieved by detecting the onsets of such events which coarsely capture the rhythmic properties of the music (many onset detection methods exist and the interested reader may consult the tutorial given in [2]).

In parallel, the events of interest to be extracted from the video include rapid movements such as dance steps, movements of musicians or any action sequence (similarly many approaches exist and such events can be for example detected using motion activity detectors [37]).

At a higher level, a music piece can be temporally segmented in sections, characterized by distinct dynamic, tonal or timbral properties and corresponding to the musical structure of the piece, *i.e.* choruses, verses, fill-ins, etc. Such segments can be either obtained by identifying large blocks in a self-similarity matrix computed on the signal (see for example [58, 7] in the framework of automatic summarization) or by exploiting novelty detection methods which allow for determining boundaries between homogeneous temporal segments [21].

For the video part, the higher level description is obtained by means of a segmentation into shots. In fact, shot changes events are semantically important in the sense that they may be correlated with the rhythm or section changes in the music.

These four segmentation processes produce detection functions (represented in Figure 1) ideally exhibiting peaks whenever an event or section change is detected. The detection functions can be thresholded to obtain the temporal location of salient events and segment boundaries, or directly considered to measure correlations.

The experiments reported in [21] have shown that the correlation between *note onset*

(music) and *shot changes (video)* is particularly appropriate for cross-media authoring or cross-media retrieval applications (e.g. audio retrieval from video or vice-versa video retrieval from audio). In the latter case, it obviously depends on the genre of the music videos. For instance, for narrative videos, where the music video has a strong narrative content and chronology, the proposed mid-level correlations are not adequate since they cannot capture such high level semantic links. Understanding music lyrics, music emotion from audio and video, represent some of the very attractive current and future lines of research in this domain.

This case study is thus an illustration of an exclusively cross-modal application, where multimodal fusion *per se* is not employed, in the sense that one is only interested in detecting the synchrony between the audio and visual streams and not in interpreting or automatically annotating the individual streams. Note that such a matching of the audio and video content at a structural level opens the path for numerous applications, ranging from temporal re-synchronization of mismatched audio and video streams to audio-driven video editing, or soundtrack retrieval by video query.

3.1.2 Audiovisual Drum Transcription

Drum transcription in polyphonic music is a particularly interesting case study for multimodal music analysis. Indeed, for many musical instruments (brass and woods in particular) a small visible movement of the musician's body or fingers may induce a large variation of the produced sound. On the contrary, the nature of the drum kit (e.g. consisting of several drum elements which are physically located at rather different locations) implies that a rather specific movement is needed from the drummer to hit each of the drum elements. It is then expected that multimodality is of great benefit for automatic drum transcription.

Even though a number of studies exist for drum solo transcription (see [18]) or for monomodal (audio-only) drum transcription of polyphonic music signals [24], [57], there has been only a few studies exploiting multimodality. A number of multimodal experiments were conducted by S. Dahl showing the relationship between body movements and emotions in marimba performances or the correlation between video features and musical accents in drumming [9].

In [22], a multimodal system for drum transcription is described exploiting both the video and audio modalities. In this work several early-fusion and late-fusion techniques were evaluated on drum-solos and it was shown that feature-level fusion by simple concatenation of audio and video features can achieve significant improvements compared to either of the monomodal transcription systems. However, with such simple integration schemes, it does not seem obvious that the strength of each modality is well exploited. In this initial system, there is indeed no intent to understand the semantics of the images or to extract higher-level features.

A different strategy is followed in [49] where the video modality is used as a detection process. More precisely (see Figure 2), the video sequence is first analyzed to detect the position of each drum element (drums and cymbals) in the scene, and more specifically the part of the instrument hit by the drum sticks. A geometric criterion is used to detect the drum tops (which are of circular shape). Then, a simple motion intensity feature coupled with foreground object segmentation is used to detect drum strokes on each of the detected drum tops. The transcription is obtained by identifying which drum instrument corresponds to each detected drum top. In parallel, the audio transcription system can also be used, as an additional source of information, to unequivocally assign each detected region to the corresponding drum instrument. Finally, once a video transcription is obtained, it can

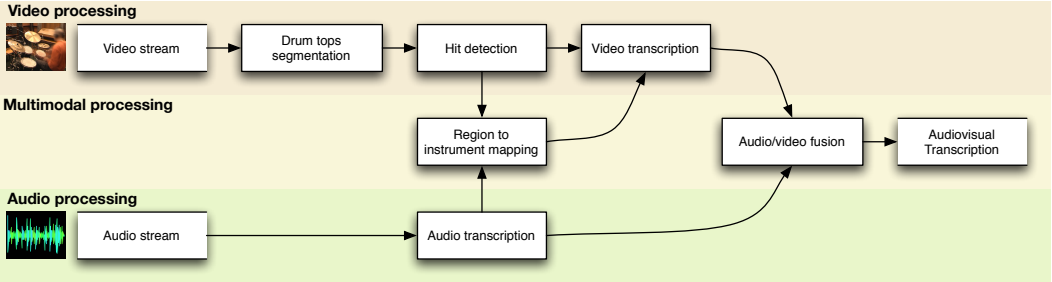

Figure 2 Overview of the audio/video analysis drum transcription system (from [49]).

be fused with an audio transcription or other video transcriptions obtained from different cameras.

This multimodal system outperformed both the monomodal systems and the system based on the traditional early and late fusion methods (the evaluation was performed on the Audiovisual ENST-Drums database [23]). One of the interesting lessons that can be learned from this work is that exploiting high-level information obtained from one modality to drive (or at least help) the processing of the other modality can be a better strategy than merely relying on direct feature-level or decision-level fusion.

3.2 Music in Motion: Analyzing Dance Scenes

Dancing is another manifestation of the multimodal nature of music. Indeed, it can be considered as a form of motion-rendering of music by dancers. For most dance styles, the analysis of a dancer's movements cannot be abstracted from the related music, as the steps and movements of the choreography are expected to be responses to particular musical events, an observation that has been successfully exploited in [61, 11].

We here describe a new multimodal dance dataset that is particularly challenging in terms of open research issues, namely the *3DLife dance dataset*[4] [14].

The dataset consists of multimodal recordings of Salsa dancers, captured at different sites with different pieces of equipment, as illustrated in Figure 3. This includes:

- synchronized 16-channel audio capture of dancers' step sounds, voice and music;
- synchronized 5-camera video capture of the dancers from multiple viewpoints covering whole body, plus 4 non-synchronized additional video captures;
- inertial (accelerometer + gyroscope + magnometer) sensor data captured from multiple sensors on the dancers' bodys;
- depth maps for dancers' performances captured using a Microsoft Kinect;
- original music excerpts;
- different types of ground-truth annotations, for instance, annotations of the music in terms of beats, annotations of the choreographies with step time codes relative to the music and ratings of the dancers' performances (by the Salsa teacher).

Over 20 dancers have been captured, each performing 2 to 5 solo Salsa choreographies among a set of 5 pre-defined ones. The dancers have been instructed to execute these choreographies respecting the same musical timing, *i.e.* all are expected to synchronize

[4] http://perso.telecom-paristech.fr/~essid/3dlife-gc-11/

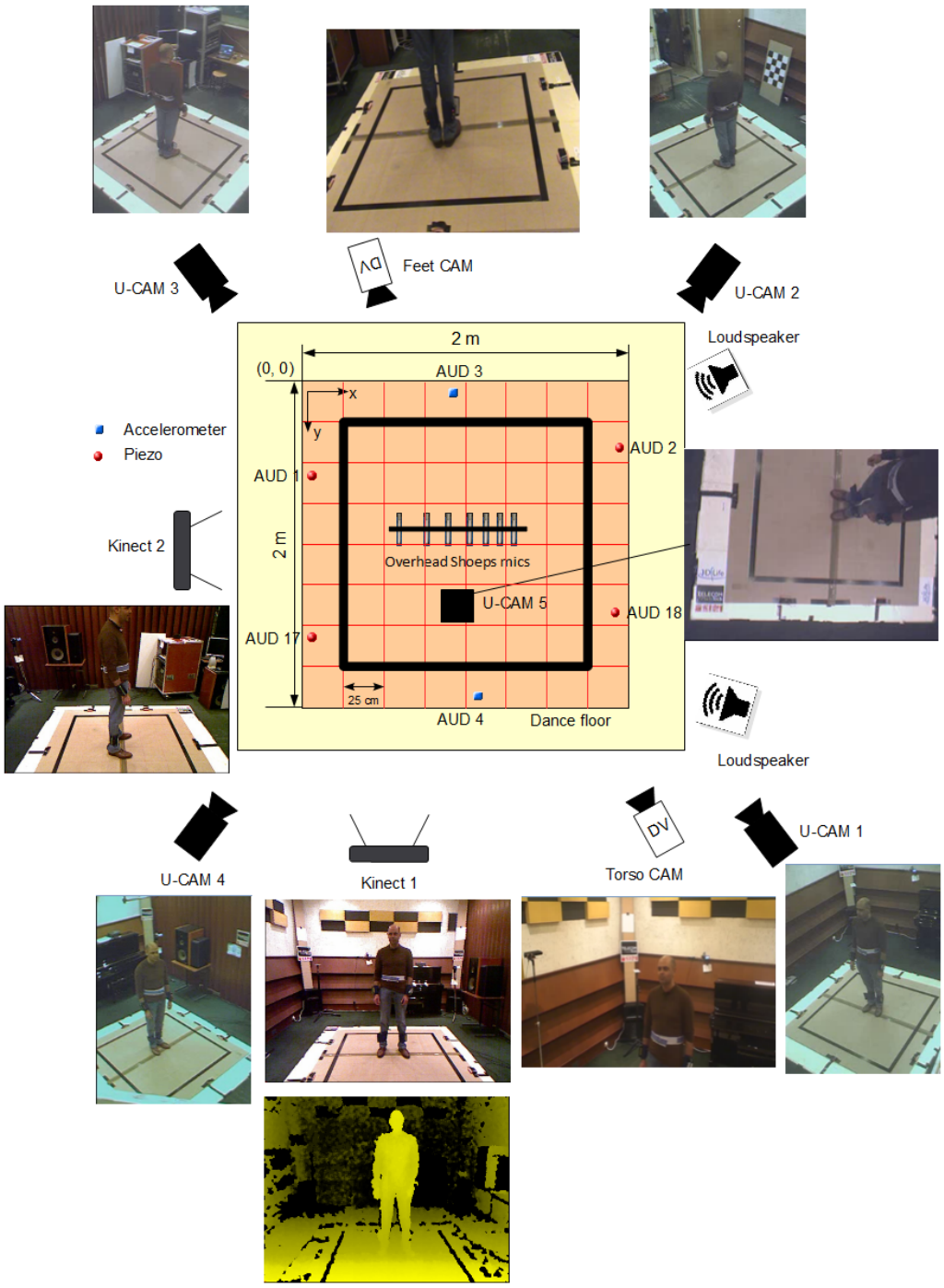

Figure 3 Recording setup at Telecom ParisTech studio.

steps/movements to particular music beats. Salsa music was chosen for this data corpus as it is a music genre that is centered at dance expression, with highly structured, yet not straightforward rhythmic patterns.

The dancers' degree of mastering of Salsa is variable. In particular there are two reference dancers which are considered as the dance teachers whose performances are viewed as the ideal templates to be followed by the other "student-dancers". In fact, this dataset has been designed in view of a broad application scenario that is an online virtual environment for dance teaching (see [14] for more details).

A number of exciting research questions are raised by such a scenario, many of which are intimately connected to multimodal music content analysis issues, in particular:

- multimodal dance performance analysis, including dance step/movement tracking and recognition;
- dance performance rating, which may involve the alignment of a dance-student performance against the teacher's performance for comparison, and/or the analysis of the student's "sense of rhythm" by assessing his/her movements timing with respect to musical timing;
- musical rhythm analysis using the analysis of the timing of a (reliable) dancer's movements;
- automatic dance synthesis for virtual agents.

Some of these tasks have already been approached. For instance encouraging results have been obtained for automatic dance performance rating [13], though more sophisticated approaches are needed towards a more accurate evaluation of a performance that would allow for highlighting a dancer's mistakes across the duration of a choreography.

3.3 User-Interactive Music Analysis

The analysis of some forms of music which cannot be represented by musical scores, in particular *electro-acoustic* music [48], cannot be envisaged without taking into account the viewpoint of a human analyst, for instance a musicologist. This is owing to the highly subjective nature of such an analysis that is linked to high-level cultural and cognitive processes.

Hence, interactive schemes have been considered for the development of electro-acoustic music analysis systems [29]. This scenario is considered as a particularly challenging multimodal scenario in which the music takes two forms: on the one hand, an audio recording (and possibly its visual waveform or spectrogram representation), and on the other hand the analyst's mental perception of the recording. Here the goal is to reach a representation of the recording that matches, insofar as it is feasible, its representation in the mind of the musicologist, in a reasonable period of time. Such a representation often takes a graphical form in which visual objects are chosen by the analyst to represent sound objects. The interested reader is referred to [35] for examples.

In his work, Gulluni has focused on electro-acoustic music pieces that can be represented as the superposition of *sound objects*. Using *relevance-feedback* and *active learning techniques* (see [29] for more details on these techniques), satisfactory performance has been obtained at transcribing such a content into sound objects [29].

Many exciting extensions could be addressed in the continuation of this work. Notably, the user could be equipped with more advanced interfaces, such as EEG[5] headsets, in his/her

[5] ElectroEncephaloGraphy: "the recording of electrical activity along the scalp", see [67] for more details.

interaction with the computer analysis system (which has been so far limited to keyboard and mouse feedback), thus allowing it to take into account their cerebral feedback while listening to the music. Even more general physiological recordings could be employed with the aim to characterize the user's emotional responses to the content, for example ECG, blood pressure, sweat activity, etc.

3.4 Partially-Informed Source Separation

The gradual shift of the general domain of music signal processing from the analysis of isolated notes or monophonic signals to the more challenging and more realistic case of polyphonic music explains the increase of interest for source separation paradigms. Indeed, one of the popular means to deal with polyphony is to first split the signals into individual sources (or components) that can then be individually processed as monophonic signals [51, Section V]. Even if the source separation is, in many situations, not explicit (and may only provide a mid-level representation on which subsequent processing would be easier), it remains a very challenging task for common music recordings (e.g. mono or at best stereo recordings of complex polyphonic signals).

However, performances of source separation systems can be significantly improved by incorporating some prior information about the sources and the mixing process. In unsupervised source separation, this information can be given in form of a specific source model (as for example the source/filter model used in [12] for singing voice separation). But in some cases, one may have access to a richer information that describes the content. This additional information can be provided by a user [64] or by a more or less accurate transcription of the music signal (see for example [32], [16] for score-informed transcription systems). In [64], the goal is to separate the singing voice from the polyphonic recording using some information provided by a user. To that aim, the user mimics the desired source by simply singing or humming the main melody. The source separation is then performed using both the original polyphonic music signal and the user provided input. Since the user's signal is simpler to process (no polyphony) and carries many audible similarities with the original signal in both frequency and temporal behaviors, it greatly helps the source separation.

In some cases, one may have access to a more or less accurate transcription of the polyphonic music by means for example of a MIDI score. The usefulness of this MIDI score (possibly obtained on the Web) depends on its quality or in other words on its accuracy to represent the original recording content.

In real case scenarios, it is usually important to first align the score to the audio recordings (see for example [41, 17, 34, 50]). Then, once aligned, the score is used to guide the source separation. For example, the score is used in [69] for obtaining improved spatial information about the sources in a stereo source separation problem. In other works, the aligned MIDI score is used as priors in the probabilistic model (such as Probabilistic Latent Component Analysis in [20] and [32]). The MIDI score can also be used to define harmonic filters which are built from the fundamental frequencies of each active notes [15]. It is also possible using score informed source separation to focus on specific parts of the music. For example in [16], an automated approach is proposed for the decomposition of a monaural piano recording into sound sources corresponding to the left and the right hands.

The different examples discussed above all exploit another source of information, other than the original audio signal. In all cases, this leads to significant improvements in separation quality.

However, it seems reasonable to assume that the strategy followed in these studies can be extrapolated to a much wider set of information sources including for example the lyrics

of the song, the gender of the singer or possibly his or her emotional state. The availability of cover versions, of some of the separated sources as in recent informed source separation methods ([47],[56]) or of user tags for appropriate source models selection also appear to be extremely valuable sources of information.

4 Conclusion

Signal processing for music analysis is a vibrant and rapidly evolving field of research. The richness and complexity of the music content call for methods that take into account music-specific characteristics including concepts such as pitch, harmony, rhythm, and instrumentation. Nevertheless, a growing trend in music analysis is to tackle the problem in a more global manner and to exploit, whenever possible, the multimodal or multi-faceted aspects of music. In this paper, we have proposed a short synthetic view of some methods that have been successfully used in multimodal signal processing. We have also briefly discussed five case studies as recent examples of successful exploitation of multimodality in music processing. In the light of these case studies, it seems clear that multimodality in music processing is very promising. Although many important challenges in this field are ahead of us, we would like to highlight three main directions for future work:

- **Towards extended multimodality**: Most current studies focus on a limited number of modalities (audio and video, audio and score, audio and tags, ...). Since music is by nature truly multidimensional there is a great interest to incorporate multiple information sources or modality for music analysis tasks (including source separation), such as for example song lyrics, singer/performer's motion and emotional state, user tags, physiological signals (EEG[6], ECG[7], ...), etc.
- **Towards extended cross-modality**: There are no particular reasons why cross-modality should be expressed through simple linear couplings. There is thus a clear perspective to extend the current approaches to non-linear coupling between modalities using for example "kernelized correlations".
- **Towards extended user interaction**: In most studies, the user is not directly involved in the music analysis stage. It seems however important to strengthen the involvement of users by further developing the concept of relevance feedback or active learning which should allow for designing better human-aware multimodal music systems.

5 Acknowledgments

This article is largely based on the works of several students mostly from Telecom ParisTech. We warmly thank Olivier Gillet, Sébastien Gulluni, Kevin Mc Guinness, Romain Hennequin, Cyril Joder and Antoine Liutkus. Part of this work was also conducted with the support from the European Commission with the 3Dlife Network of Excellence [8].

References

1 E. Argones Rua, H. Bredin, C. Garcia Mateo, G. Chollet, and D. Gonzalez Jimenez. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications*, 12(3):271–284, May 2008.

[6] ElectroEncephaloGram
[7] ElectroCardioGram
[8] http://www.3dlife-noe.eu/3DLife/

2 J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.

3 K. Bousmalis and L. Morency. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 746–752, March 2011.

4 C. Chen, M. Weng, S. Jeng, and Y. Chuang. Emotion-based music visualization using photos. *Advances in Multimedia Modeling*, pages 358–368, 2008.

5 C. Chibelushi, J. Mason, and N. Deravi. Integrated person identification using voice and facial features. *IEE Colloquium on Image Processing for Security Applications (Digest No.: 1997/074*.

6 T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *16th IEEE International Conference on Pattern Recognition*, volume 3, pages 789–794, 2002.

7 M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

8 T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

9 S. Dahl. The playing of an accent - preliminary observations from temporal and kinematic ana. of percussionists. In *Journal of New Music Research*, volume 29(3), pages 225–234, 2000.

10 R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, 1984.

11 Y. Demir, E. Erzin, and Y. Yemez. Evaluation of audio features for audio-visual analysis of dance figures. In *European Signal Processing Conference (EUSIPCO)*, 2008.

12 J.-L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, October 2011.

13 S. Essid, D. Alexiadis, R. Tournemenne, M. Gowing, P. Kelly, D. Monhagan, P. Daras, A. Dremeau, and N. O'Connor. An advanced virtual dance performance evaluator. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2012.

14 S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, R. Tournemenne, N. E. O'Connor, P. Daras, and G. Richard. A multimodal dance corpus for research into real-time interaction between humans in online virtual environments. In *ICMI Workshop On Multimodal Corpora For Machine Learning*, Alicante, Spain, November 2011.

15 M. Every and J. Szymanski. A spectral-filtering approach to music signal separation. In *International Conference on Digital Audio Effects, DAFx'04*, Napoli, Italy, October 2004.

16 S. Ewert and M. Müller. Score-informed voice separation for piano recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, Miami, USA, 2011.

17 S. Ewert, M. Müller, and R. B. Dannenberg. Towards reliable partial music alignments using multiple synchronization strategies. In *Proceedings of the International Workshop on Adaptive Multimedia Retrieval (AMR)*, Madrid, Spain, September 2009.

18 D. FitzGerald and J. Paulus. Unpitched percussion transcription. *Signal Processing Methods for Music Transcription*, 2006.

19 S. Foucher, F. Lalibert, G. Boulianne, and L. Gagnon. A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

20 J. Ganseman, P. Scheunders, G. Mysore, and J. Abel. Evaluation of a score-informed source separation system. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, Utrecht, August 2010.

21 O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Trans. on Circuit and Systems for Video Technology*, March 2007.

22 O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

23 O. Gillet and G. Richard. Enst-drums: an extensive audio-visual database for drum signals processing. *Proceedings of the International Society for Music Information Retrieval Conference*, 2006.

24 O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, March 2008.

25 R. Goecke and J. Millar. Statistical analysis of the relationship between audio and video speech parameters for australian english. In *ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003*, pages 133–138, September 2003.

26 M. Gondry. *The Work of Director Michel Gondry*. Director's Series, Vol. 3, DVD, Palm Pictures, 2003.

27 J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes. Dbn based multi-stream models for audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

28 G. Gravier, G. Potamianos, and C. Neti. Asynchrony modeling for audio-visual speech recognition. In *Proceedings of the second international conference on Human Language Technology Research*, pages 1–6, San Diego, California, 2002. Morgan Kaufmann Publishers Inc.

29 S. Gulluni, O. Buisson, S. Essid, and G. Richard. An interactive system for electro-acoustic music analysis. In *International Conference of Music Information Retrieval*, 2011.

30 I. Guyon and A. Elisseeff. An introduction to feature and variable selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

31 D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

32 R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

33 H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3 - 4):321 – 377, 1936.

34 N. Hu, R. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 2003.

35 INA - GRM. Portraits polychromes. `http://www.inagrm.com/accueil/collections/portraits-polychromes`.

36 Y. Ivanov, T. Serre, and J. Bouvrie. Error weighted classifier combination for multi-modal human identification. Technical Report MIT-CSAIL-TR-2005-081, MIT, 2005.

37 S. Jeannin and A. Divakaran. Mpeg-7 visual motion descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11:720–724, 2001.

38 W. Jiang, C. Cotton, S. Chang, D. Ellis, and A. Loui. Short-term audiovisual atoms for generic video concept classification. In *Proceedings of the 17th ACM International Conference on Multimedia*, pages 5–14. ACM, 2009.

39 W. Jiang and A. Loui. Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 123–132, Scottsdale, USA, 2011.

40 C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):174–186, 2009.

41 C. Joder, S. Essid, and G. Richard. A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transaction on Audio, Speech and Language Processing*, 19(8):2385–2397, November 2011.

42 S. Jonze. *The Work of Director Spike Jonze*. Director's Series, Vol. 1, DVD, Palm Pictures, 2003.

43 E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. HMM based structuring of tennis videos using visual and audio cues. In *Proceedings of the 2003 International Conference on Multimedia and Expo*, pages 309–312, 2003.

44 P. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–378, 2000.

45 D. Li, N. Dimitrova, M. Li, and I. Sethi. Multimedia content processing through cross-modal association. In *ACM International Conference on Multimedia*, Berkeley, CA, USA, November 2003.

46 A. Lim, K. Nakamura, K. Nakadai, T. Ogata, and H. Okuno. Audio-visual musical instrument recognition. In *National Convention of Audio-Visual Information Processing Society*, March 2011.

47 A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, September 2011.

48 P. Manning. *Electronic and computer music*. Oxford University Press, Jan 2004.

49 K. McGuinness, O. Gillet, N. O'Connor, and G. Richard. Visual analysis for drum sequence transcription. In *European Signal Processing Conference (Eusipco)*, Poznan, Pologne, sep 2007.

50 M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

51 M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

52 A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audiovisual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2. IEEE, 2002.

53 V. Niches. *Extraordinary Music Videos*. DVD, EAF Music„ 2002.

54 K. Nymoen, B. Caramiaux, M. Kozak, and J. Torresen. Analyzing sound tracings: a multimodal approach to music information retrieval. In *First International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM)*, 2011.

55 H. Ohkushi, T. Ogawa, and M. Haseyama. Music recommendation according to human motion based on kernel CCA-based relationship. *EURASIP Journal on Advances in Signal Processing*, 2011(1):121, December 2011.

56 M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721 –1733, August 2011.

57 J. Paulus and A. Klapuri. Drum sound detection in polyphonic music with hidden markov models. *EURASIP J. Audio Speech Music Process.*, 2009:14:1–14:9, January 2009.

58 G. Peeters, A. L. Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

59 E. Petajan, B. Bischoff, D. Bodoff, and N. Brooke. An improved automatic lipreading system to enhance speech recognition. In *CHI '88 Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 19 – 25, 1988.

60 J. Saitoh, A. Kodate, and H. Tominaga. Integrated data processing between image and audio - musical instrument (piano) playing information processing. In *6th International Conference on Image Processing and its Applications*, pages 438–442 vol.1, July 1997.

61 T. Shiratori, A. Nakazawa, and K. Ikeuchi. Detecting dance motion structure through music analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

62 P. Smaragdis and M. Casey. Audio/visual independent components. In *International Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, 2003.

63 P. Smaragdis and C. M. Audio visual independent components. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 709–714, 2003.

64 P. Smaragdis and G. J. Mysore. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, pages 69–72, 2009.

65 D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet. Identifying words that are musically meaningful. In *International Conference of Music Information Retrieval*, pages 405–410, 2007.

66 B. Vercoe. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 199–200, 1984.

67 Wikipedia. Electroencephalography. http://en.wikipedia.org/wiki/Electroencephalography.

68 P. Wilkins, T. Adamek, D. Byrne, G. Jones, H. Lee, G. Keenan, K. Mcguinness, N. E. O'Connor, A. F. Smeaton, A. Amin, Z. Obrenovic, R. Benmokhtar, E. Galmar, B. Huet, S. Essid, R. Landais, F. Vallet, G. T. Papadopoulos, S. Vrochidis, V. Mezaris, I. Kompatsiaris, E. Spyrou, Y. Avrithis, R. Morzinger, P. Schallauer, W. Bailer, T. Piatrik, K. Chandramouli, E. Izquierdo, M. Haller, L. Goldmann, A. Samour, A. Cobet, T. Sikora, and P. Praks. K-space at TRECVid 2007. In *TREC Video Retrieval Evaluation: TRECVID 2007*, November 2007.

69 J. Woodruff, B. Pardo, and R. Dannenberg. Remixing stereo music with score-informed source separation. In *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, Victoria, October 2006.

70 Y. Wu, C.-Y. Lin, E. Chang, and J. Smith. Multimodal information fusion for video concept detection. In *International Conference on Image Processing*, pages 2391 – 2394, October 2004.

71 C. Xu, X. Shao, N. Maddage, and M. Kankanhalli. Automatic music video summarization based on audio-visual-text analysis and alignment. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 361–368. ACM, 2005.