Session 5aMUe

Musical Acoustics and Signal Processing in Acoustics: Signal Representations and Models of Musical Sounds II

James W. Beauchamp, Cochair
University of Illinois, 1002 Eliot Drive, Urbana, IL 61801, USA

Bertrand David, Cochair
Télécom Paristech (ENST) / TSI - CNRS LTCI, 46, rue Barrault, Paris Cedex 13, 75634, France

Contributed Paper

11:00

5aMUe1. Early use of the Scott-Koenig phonautograph for documenting performance. George Brock-Nannestad (Patent Tactics, Resedavej 40, DK-2820 Gentofte, Denmark, pattac@image.dk), Jean-Marc Fontaine (Université UPMC - Ministère de la Culture - CNRS - IJRA - LAM, 11, rue de Lourmel, F-75015 Paris, France, jmfontai@ccr.jussieu.fr)

Acoustics in the 1850s combined listening, observation and notation. This was "real-time", catching any phenomenon as it appeared. If it was repeatable, one could prepare for it. Continuous data rather than observation data enabled a very different analysis from observation and notebooks. Édouard-Léon Scott's invention of the phonautograph enabled this. A surface moved below a stylus that was vibrated by sound in air. Originally a blackened glass plate, the scientific instrument maker Rudolph Koenig contributed by devising a very long axis representing time, because now blackened paper was wrapped around a cylinder on a threaded shaft. A two-dimensional representation of the individual vibrations was obtained. Scott both deposited a sealed letter with the Paris Academy of Sciences in January, 1857 and filed a patent application in April, 1857. Later he deposited a further sealed letter and in 1859 he filed an application for patent of addition. Following the thoughts expressed and documented in his manuscripts here analyzed in context and comparing with Koenig's production it is feasible to see how they were dependent on each other, although they had different purposes in mind. The paper concentrates on Scott's interests in performance, and a number of original tracings are discussed.

Invited Papers

11:20

5aMUe2. Sparse representations of audio: from source separation to wavefield compressed sensing? Remi Gribonval (INRIA, IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France, remi.gribonval@irisa.fr)

Sparse signal representations, which are at the heart of today's coding standards (JPEG, MPEG, MP3), are known to have had a substantial impact in signal compression. Their principle is to represent high-dimensional data by a combination of a few elementary building blocks, called atoms, chosen from a large collection called a dictionary. Over the last five years, theoretical advances in sparse representations have highlighted their potential to impact all fundamental areas of signal processing. We will discuss some current and emerging applications of sparse models in musical sound processing including: signal acquisition (Compressed Sensing - sampling wave fields at a dramatically reduced rate) and signal manipulation (e.g., source separation and enhancement for digital remastering). We will conclude by discussing the new algorithmic and modeling challenges raised by these approaches.

11:40

5aMUe3. Towards a hierarchically sparse model for audio signals. Laurent Daudet (UPMC Univ Paris 06, LAM / IJLRA, 11 rue de Lourmel, 75015 Paris, France, daudet@lam.jussieu.fr)

In this paper we discuss a major issue that arises when building sparse atomic models for music signals: in which space (/ dictionary) should we represent the signals? Having extremely redundant dictionaries is very informative for the locally most important atoms, but is irrelevant for the detail atoms that represent residual signals. Also, what are the advantages of dictionaries composed of learned atoms (which brings the issue of representativeness) compared to generic dictionaries? We here advocate for non-fixed dictionaries, with a multi-layered hierarchical decomposition: the first layer roughly describes the signal, in an extremely redundant, signal-tailored, structured dictionary. This layer is also very sparse, most of the information being carried by the atoms' parameters, similarly to parametric representations. Subsequent layers give more and more details, increasing the data amount while reducing overcompleteness and structural model constraints. These techniques have obvious applications for audio coding, but are also useful for scalable music data mining. This research is supported by the French ANR (DESAM project).

12:00

5aMUe4. Bayesian inference in hierarchical non-negative matrix factorisation models of musical sounds. Ali Taylan Cemgil (University of Cambridge, Trumpington Street, CB2 1PZ Cambridge, UK, atc27@cam.ac.uk), Tuomas Virtanen (Tampere University of Technology, Inst. of Signal Processing, FI-33101 Tampere, Finland, tuomas.virtanen@tut.fi)

There has been a surge of interest to efficient audio and music modeling using tools from statistical machine learning. One such technique, that has been particularly successful, is non-negative matrix factorisation (NMF). However, a detailed theoretical understanding of this success is missing, as well as links to other modeling strategies such as sinusoidal or harmonic models. To fill this gap,

we describe NMF in a statistical framework, as a hierarchical generative model consisting of an observation and a prior component. We show that particular choices lead to standard NMF algorithms as special cases, where parameter estimation is carried out via maximum likelihood. Starting from this view, we develop extensions that facilitate more realistic acoustic modeling (such as spectral smoothness or harmonicity of natural sounds) and alternative inference techniques via Gibbs sampling and variational Bayes, which allow us to do principled comparisons between alternative models via Bayesian model selection. Our novel construction, where we make use of Markov chains of Gamma random variables, retains conjugacy and enables us to develop models that fit better to real data while retaining attractive features of standard NMF such as fast convergence and easy implementation. We illustrate our approach on polyphonic pitch estimation.

12:20

5aMUe5. On sinusoidal modeling of nonstationary signals. Axel Roebel (IRCAM, 1, pl. Igor-Stravinsky, 75004 Paris, France, roebel@ircam.fr)

In this presentation we are going to give an overview over a number of techniques that have been developed in our group to improve the modeling of musical (nonstationary) signals using sinusoidal models. One of the basic problems with sinusoidal models is the fact that the underlying theory is derived assuming stationary sinusoids, while in the real world all sinusoidal components are non stationary. The two topics that will be covered main are, first a technique that allows to distinguish between nonstationary sinusoidal components, noise and transients, and second a new technique for parameter estimation of nonstationary sinusoids using frequency domain demodulation.

12:40-1:40 Lunch Break

Invited Papers

1:40

5aMUe6. Advances in the tracking of partials for the sinusoidal modeling of musical sounds. Sylvain Marchand (LaBRI - CNRS, University of Bordeaux 1, 351 cours de la Liberation, F-33405 Talence, France, sylvain.marchand@labri.fr)

Whereas sinusoidal modeling is widely used for sounds, polynomial models are still used for the model parameters, which can hardly handle modulations (vibrato and tremolo) present in musical sounds. Moreover, the partial tracking algorithms are often designed under stationarity assumptions. Advances in partial tracking may come out of the modeling of the partials themselves. We consider their parameters (frequencies and amplitudes) as predictable and slow time varying: First, the future of any partial can be determined from its past evolutions; second, no audible frequency should appear in the spectral content of these evolutions, otherwise it would question the perceptive consistency of the model. We then choose to handle nonstationary sinusoidal modeling by a deterministic approach based on linear prediction of the partial evolutions and partial discrimination based on the spectral properties of these evolutions. The underlying model for each partial is now a sum of sinusoids, thus leading to a two-level sinusoidal modeling, well suited for musical sounds, where modulations are important. The enhanced partial tracking algorithm also handles the case of crossing partials, without the need for any probabilistic approach. Better modeling the deterministic part of polyphonic sounds leads to enhanced source separation and time scaling algorithms.

2:00

5aMUe7. Adaptive subspace methods for high resolution analysis of music signals. Roland Badeau (Télécom Paristech (ENST) /TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, roland.badeau@enst.fr), Bertrand David (Télécom Paristech (ENST)/TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, bertrand.david@enst.fr)

In the field of music signals analysis, the tonal part of a broad variety of sounds is often represented as a sum of slowly varying sinusoids. The Fourier transform remains a prominent tool for estimating the parameters of this model, due to its robustness and to the existence of fast algorithms. Its main drawback relies in its spectral resolution, bounded by the length of the analysis window. Subspace-based high resolution (HR) methods are conversely not constrained by this limit, since they rely on the particular geometrical structure of the signal model. Nevertheless, they have been seldom used in audio signal processing, mainly due to their high computational cost. Based on recent advances in the field of subspace tracking, enhanced adaptive algorithms for HR analysis have thus been developed, leading to a high resolution time-frequency representation of the signal, called HR-ogram. The application of these algorithms to music signals, made difficult by the high dynamics and the presence of colored noise, has required the tuning of well-adapted preprocessing techniques. The whole tool is now mature, and allows a high quality separation of the tonal part of various musical sounds. This research is supported by the French ANR under contract ANR-06-JCJC-0027-01 (DESAM).

2:20

5aMUe8. Towards an adaptive subspace-based representation of musical spectral content. Bertrand David (Télécom Paristech (ENST)/TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, bertrand.david@enst.fr), Roland Badeau (Télécom Paristech (ENST)/TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, roland.badeau@enst.fr)

This study presents an algorithm based on an adaptive framework model for musical sound signals assumed to be composed of slowly varying frequency components surrounded by additive noise. These components appear as contours in a time-frequency representation, as for instance, a spectrogram. To extract these contours, an often used solution is to estimate the parameters (amplitude, frequency, and phase) of each component at each frame and then to link them from one to the next with the help of a distance measure or an HMM. Conversely, our method attempts to *update* the estimated values from one time instant to the next. It relies on principal subspace tracking (with respect to time) together with gradient descent to individually update each of the component parameters. Finally, each extracted contour, which represents the frequency and amplitude variation of a single component, is available for subsequent

processing. Applications are demonstrated in the fields of harmonic plus noise decomposition and analysis/transformation/synthesis. This research is supported by the french Institut Telecom, TAMTAM project.

2:40

5aMUe9. Damped sinusoids and subspace based approach for lossy audio coding. Olivier Derrien (Université de Toulon, Av Georges Pompidou, BP 56, 83162 La Valette du Var, France, olivier.derrien@univ-tln.fr), Gaël Richard (Télécom Paristech (ENST) /TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, gael.richard@enst.fr), Roland Badeau (Télécom Paristech (ENST)/TSI - CNRS LTCI, 46, rue Barrault, 75634 Paris Cedex 13, France, roland.badeau@enst.fr)

The new subspace-based techniques recently introduced appear to be well adapted for the parameters estimation of a damped sinusoids + noise signal model. These high-resolution (HR) methods have a better frequency resolution than the Fourier analysis, but they are rarely used in audio coding. Although HR methods would be suitable for parametric coding at low bitrates, we show that they are also efficient for high-bitrate coding where state-of the art codecs are usually transform-based. Our coding scheme first includes a 8-band PQMF filter-bank decomposition. Then, each subband signal is segmented in onsets and a maximum-order HR analysis is performed on each segment with the ESPRIT algorithm. For each component of the model, frequency, damping, amplitude and phase are quantized. The residual signal is not coded. This codec is compared to a MDCT framework, where transform and quantization are the same as in a MPEG-AAC but without inclusion of perceptual modelling and entropy coding. Preliminary objective and subjective tests show the potential of this approach which requires, on mostly tonal signals, significantly less bits than the traditionnal MDCT method for a given quality.

3:00

5aMUe10. Auditory model based analysis of polyphonic music. Anssi Klapuri (Tampere University of Technology, Korkeakoulunkatu 1, 33720 Tampere, Finland, anssi.klapuri@tut.fi)

This study is about the use of an auditory model to extract multiple pitches from polyphonic music signals. One goal was to identify the conditions where pitch analysis using an auditory model is advantageous over more conventional time or frequency domain approaches. It is shown that these conditions include especially the processing of band-limited signals or signals where important parts of the audible spectrum are corrupted by band-limited interference. An efficient implementation strategy is described which reduces the computational complexity of the auditory model roughly by factor 10. Further prospects of bandwise processing and redundant signal representations in general are discussed.

3:20

5aMUe11. On perceptual distortion measures and parametric modeling. Mads G. Christensen (Aalborg University, Niels Jernes Vej 12 A, DK-9220 Aalborg, Denmark, mgc@es.aau.dk)

Over the past two decades, there has been much interest in incorporating human sound perception in signal processing algorithms, often in the form of perceptual distortion measure or an approximation thereof. An example of this is MDCT-based audio coding where a good quality can be achieved at very low bit-rates by taking masking effects into account. More recently, the same principles have been applied to parametric modeling and coding of audio signals. We discuss the inherent tradeoffs in choosing a perceptual distortion measure and a parametric model and the pros and cons of various ways of implementing such perceptual distortion measures is discussed. An important question that we seek to answer is whether perception should be taken into account in the estimation of model parameters or this should be done in a separate step.

Contributed Paper

3:40

5aMUe12. Efficient coding of a xylophone sound using spikogram nonredundent coding. Rolf Bader (University of Hamburg, Institute of Musicology, Neue Rabenstr. 13, 20354 Hamburg, Germany, R_Bader@t-online.de)

Sensory systems try to use the incoming data most efficiently. Studies of Lewicki et al. lately showed, that a representation of a spike train just representing the sound and not having any redundancy is how the auditory system of the cat represents incoming sounds. To compare this theory with musical instruments, a xylophone sound was analyzed in terms of a spikogram. Here, gammatones of a 64 channel filterbank with different attack and decay values are superposed in a way to reconstruct the original sound. Only those gammatones were used which are needed to result in the sound used as input and so no redundancy is present in the analysis of the sound. It was found, that the most reasonable fit of the gammatone shape with the empirical data indeed made the representation most efficient and so the xylophone sound is most easily represented by the auditory system. Although more analysis are needed here, musical instruments could show up to be built in a way to fit a most efficient coding by listeners and so fulfill a middle-of-the-road rule of not too much and not too few information so that listeners are interested in but not overtaxed by the sounds.

4:00-4:20 Break

4:20

5aMUe13. A frequency shifting model of pitch. Paris Smaragdis (Adobe Systems Inc., 275 Grove St., Newton, MA 02466, USA, paris@media.mit.edu)

We present a model useful for tracking of melodies of sounds that can have arbitrary harmonic structure (including inharmonic instruments and noise sources). This model is based on a spectral shift assumption which is capable of tracking melodic movements of an instrument regardless of the irregularity of its spectrum. This technique can be used to simultaneously estimate the spectral character of the instrument to be analyzed in addition to its melody. It is capable of dealing with multiple instances of the same instrument, thereby recognizing chords as well as notes, and can also extract multiple melodies in audio signals composed out of many instruments.

4:40

5aMUe14. Modeling vocal sounds in polyphonic musical audio signals. Masataka Goto (National Institute of Advanced Industrial Science and Technology (AIST), IT, 1-1-1 Umezono, Tsukuba, 305-8568 Ibaraki, Japan, m.goto@aist.go.jp), Hiromasa Fujihara (National Institute of Advanced Industrial Science and Technology (AIST), IT, 1-1-1 Umezono, Tsukuba, 305-8568 Ibaraki, Japan, h.fujihara@aist.go.jp)

This paper describes our research aimed at modeling vocal sounds (singing voices) in available music recordings and its applications to singer identification, singer similarity, and lyrics synchronization. Our predominant-F0 estimation method, PreFEst, can obtain the melody line by modeling the input sound mixture as a weighted mixture of harmonic-structure tone models (probability density functions) of all possible F0s and estimating their weights and the tone models by MAP estimation. Since the PreFEst was designed for general melodies, we extended it to specialize in vocal melodies by using vocal timbre models --- vocal and nonvocal GMMs. Those GMMs are trained beforehand and used to evaluate the vocal probability. The GMMs are also used to identify vocal regions, but its strategy should depend on applications. For singer identification and singer-similarity calculation, since the purpose is to model singer's identity by training each singer's vocal GMM, certainly reliable vocal regions should be identified even if most true regions were missed. On the other hand, for lyrics synchronization, since the purpose is to align each phoneme to the estimated vocal melody, vocal regions should be identified without missing any true regions. We achieved this by biasing log likelihoods provided by vocal and non-vocal GMMs.

5:00

5aMUe15. Music and speech signal processing using harmonic-temporal clustering. Jonathan Le Roux (University of Tokyo, Sagayama/Ono Laboratory, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan, leroux@hil.t.u-tokyo.ac.jp), Hirokazu Kameoka (NTT Communication Science Laboratories, NTT Corporation, 3-1 Morinosato wakamiya, 243-0198 Atsugi, Kanagawa, Japan, kameoka@eye.brl.ntt.co.jp), Nobutaka Ono (University of Tokyo, Sagayama/Ono Laboratory, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan, onono@hil.t.u-tokyo.ac.jp), Alain De Cheveigne (CNRS, Universite Paris 5, Ecole Normale Superieure, 29 rue d'Ulm, 75230 Paris, France, alain.de.cheveigne@ens.fr), Shigeki Sagayama (University of Tokyo, Sagayama/Ono Laboratory, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan, sagayama@hil.t.u-tokyo.ac.jp)

We present here the principle of the recently introduced harmonic-temporal clustering (HTC) framework and its applications in both music and speech signal processing. HTC relies on a precise parametric description of the harmonic parts of the power spectrum through constrained Gaussian mixture models. The model parameters of all the elements of the acoustical scene are estimated jointly by an unsupervised 2D time-frequency clustering of the observed power density. HTC is effective for multi-pitch analysis of music signals and F0 estimation of single and multiple speaker speech signals in various noisy environments. It also enables to perform extra processing of monaural music and speech signals, such as isolation or cancellation of a particular part, noise reduction and source separation.

5:20

5aMUe16. Timbre transposition based on time-varying spectral analysis of continuous monophonic audio and precomputed spectral libraries. James W. Beauchamp (University of Illinois, 1002 Eliot Drive, Urbana, IL 61801, USA, jwbeauch@uiuc.edu), Mert Bay (University of Illinois, 212 W. Healey, Apt. 303, Champaign, IL 61820, USA, mertbay@uiuc.com)

A sinusoidal model for solo musical sounds consisting of time-varying harmonic amplitudes and frequencies allows for convenient temporal and spectral modifications. With a harmonic model, analysis frames can be grouped by fundamental frequency (F0) and then clustered in terms of their harmonic spectra. The resulting cluster centroid spectra are used as spectral libraries. When continuous audio monophonic passages are analyzed in the form of harmonic components, F0 vs. time data are used to guide the extraction of parameters from the sound in order to find appropriate library spectra for resynthesis. Two methods for finding appropriate spectra are: (1) best rms match with the incoming spectra and (2) best spectral centroid match. These give similar results, but centroid matching yields smoother spectra over time. Timbre transposition is performed by using a library that belongs to another instrument. We have found that when the target instrument has a unique timbral quality based on its spectrum, the synthesis sounds mostly like that instrument. However, if the target instrument's spectral characteristic is not sufficiently differentiated from the source, the source timbral quality may dominate, probably due to its temporal behavior being transmitted. Results will be demonstrated by audio examples.

5:40

5aMUe17. Comparison of the sound of a grand and an upright piano using wavelets. Grigorios Plitsis (Greece, grigoriosplitsis@merseymail.com)

Wavelet analysis is useful for extracting patterns and thus analyzing signals. Although the Fourier analysis can reveal different features of a signal, it is less appropriate for describing transient phenomena and sudden sound changes. As a result, Fourier-based music reconstruction cannot exactly imitate the physical sound, as it is not possible to know at the same time a specific frequency as well as the time of occurrence of this frequency. Wavelet analysis is capable of highlighting different attributes of a signal. Different types of wavelets are thus used in the present study in order to compare the sound produced by a grand piano with that produced by an upright piano.

6:00

5aMUe18. Modeling of piano sounds using FEM simulation of soundboard vibration. Luis I. Ortiz-Berenguer (Universidad Politecnica de Madrid, Ctra.Valencia km7, 28031 Madrid, Spain, lortiz@diac.upm.es), Francisco J. Casajus-Quiros (Universidad Politecnica de Madrid, Ctra.Valencia km7, 28031 Madrid, Spain, javier@gaps.ssr.upm.es),David Ibanez-Cuenca (Universidad Politecnica de Madrid, Ctra.Valencia km7, 28031 Madrid, Spain, dibanez@alumnos.euitt.upm.es)

Pattern-matching methods for polyphonic transcription of piano sounds require a set of patterns that can be obtained by modeling the piano-sound spectra. The modeling should take into account not only the string stiffness but also the effect of the soundboard impedance on the string vibration. Studies on that effect corresponding to a wide range of impedance values have previously been carried out by the authors. However, actual impedance values for real pianos must be used in the model. Although the impedance value of a few grand-pianos have been measured by the authors, these results are not significative enough to create a model. Thus, a FEM simulation of soundboard vibration is proposed to obtain nearly-actual impedance values. The simulation considers several cases of vibrating plates from the simplest rectangular one and increasing the similarity to real piano soundboards. The quality of the simulation is verified comparing the obtained results with either recognized theoretical results for the simplest cases or measured values for the more complex ones. The complexity of the simulated soundboard is limited to the case that produces only slight variations in the modeled spectrum. [This work has been supported by Spanish National Project TEC2006-13067-C03-01/TCM.]

FRIDAY MORNING, 4 JULY 2008 ROOM 202/203, 11:00 A.M. TO 12:00 NOON

Session 5aMUf

Musical Acoustics: Plucked and Struck Idiophones II

Thomas D. Rossing, Cochair
Stanford University, CCRMA, Department of Music, Stanford, CA 94305, 26464 Taaffe Rd, Los Altos Hills, CA 94022, USA

Charles Besnainou, Cochair
Institut Jean le Rond d'Alembert, Laboratoire d'Acoustique Musicale, 11, rue de Lourmel, Paris, 75015, France

11:00

5aMUf1. Characterizing the sound of an African thumb piano (kalimba). David M.f. Chapman (Scientific Consultant, 8 Lakeview Avenue, Dartmouth, NS B3A 3S7, Canada, dave.chapman@ns.sympatico.ca)

The kalimba is an African percussion instrument whose notes are generated by vibrating metal tines of various lengths attached to sound board or sound box. Unlike a vibrating string or organ pipe, the overtone structure is anharmonic, that is, the overtone frequencies (which determine sound quality) are not simple integer multiples of the fundamental frequency (which determines pitch). The ratio of the first overtone to the fundamental is in the range 5.3-5.9, depending on the tine geometry. The vibrating tines are modeled as a clamped-supported-free beam and the observed overtone structure is shown to be in accordance with this model. Audio examples will be provided.

11:20

5aMUf2. The jew's harp, experimental study and modeling. Charles Besnainou (Institut Jean le Rond d'Alembert, Laboratoire d'Acoustique Musicale, 11, rue de Lourmel, 75015 Paris, France, chbesnai@ccr.jussieu.fr), Joel Frelat (Institut Jean le Rond d'Alembert, Lab. d'Acoustique Musicale, 11, rue de Lourmel, 75015 Paris, France, frelat@lmm.jussieu.fr),Adrien Mamou-Mani(Institut Jean le Rond d'Alembert, Lab. d'Acoustique Musicale, 11, rue de Lourmel, 75015 Paris, France, mamou-mani@lmm.jussieu.fr)

Under its archaic aspects jew's harp is a musical instrument highly subtle. Indeed, a metal blade (or wooden) attached to a rigid frame put into vibration by the musician, and coupled to the buccal resonator allow nice tune. The skilt of the jew's harp focuses on the conformations of this cavity whose function is to select the right components of the vibration to be amplify. In our study, we have modelled a playing technique which involves blowing during the blade vibrates. In the lake of breath, the spectre of sound produced by the blade is odd, i.e., it includes at first approximation odd components (n+1) multiple the fundamental. Whereas when the musician adds breath the spectrum turns into a spectrum containing all components of basic integer multiples (n). This work takes place in the context of studies of vibrating systems under prestress and loaded. In that case the load, and the prestress are generated by the musician breath by bending the blade On the other hand, experimental studies are compare with the model results.

11:40

5aMUf3. Sound of the HANG. Thomas D. Rossing (Stanford University, CCRMA, Department of Music, Stanford, CA 94305, 26464 Taaffe Rd, Los Altos Hills, CA 94022, USA, rossing@ccrma.stanford.edu), David