Prior Abstracts: CCRMA DSP Seminar (MUS423)

Center for Computer Research in Music and Acoustics (CCRMA)*
Department of Music, Stanford University
Stanford, California 94305

Abstract

The CCRMA DSP Seminar[1] is a seminar for graduate students and visiting scholars pursuing music and audio applications of signal processing. This is a collection of prior abstracts for past quarters. The purpose is to provide a broadened view of CCRMA research in recent years. For a more complete summary, see the research section of the CCRMA Overview.[2]

*http://ccrma.stanford.edu/
[1]http://ccrma.stanford.edu/~jos/mus423/
[2]http://ccrma.stanford.edu/CCRMA/Overview/

Spring Quarter 2003-2004

Analysis and Resynthesis of Quasi-Harmonic Sounds: An Iterative Filterbank Approach

Harvey Thornburg and Randal Leistikow <{harv23,randal} at ccrma> (EE)

We employ a hybrid state-space sinusoidal model for general use in analysis-synthesis based audio transformations. This model combines the advantages of a source-filter model with the time-frequency character of the sinusoidal model. We specialize the parameter identification task to a class of "quasi-harmonic" sounds. The latter represent a variety of acoustic sources in which multiple, closely spaced modes cluster about principal harmonics loosely following a harmonic structure.

To estimate the sinusoidal parameters, an iterative filterbank splits the signal into subbands, one per principal harmonic. Each filter is optimally designed by a linear programming approach to be concave in the passband, monotonic in transition regions, and to specifically null out sinusoids in other subband regions. Within each subband, instantaneous frequency estimates averaged over all modes in the subband region are used to update the filter's center frequency for the next iteration. In this way, the filterbank progressively adapts to the specific inharmonicity structure of the source recording.

We demonstrate analysis-synthesis applications including standard time and pitch scaling, as well as new effect types exploiting the "source-filter" aspect.

Presentation Program

1. Harvey will begin by discussing the state-space sinusoidal model and Kalman-based analysis-synthesis engine, focusing on applications and rules for parameter tuning. (40 min.).
2. Randal will discuss the design of the filterbank responsible for extracting sinusoidal dynamics parameters for any quasi-harmonic sound model (40 min.).
3. Harvey will introduce, very briefly, a hybrid MATLAB/C code package useful in general linear-programming based FIR filter design. (20 min.)

Bayesian Identification of Closely-Spaced Chords from Single-Frame STFT Peaks

Harvey Thornburg and Randal Leistikow <{harv23,randal} at ccrma> (EE)

Identifying chords and related musical attributes from digital audio has proven a long-standing problem spanning many decades of research. A robust identification may facilitate automatic transcription, semantic indexing, polyphonic source separation and other emerging applications. To this end, we develop a Bayesian inference engine operating on single-frame STFT peaks. Peak likelihoods conditional on pitch component information are evaluated by an MCMC approach accounting for overlapping harmonics as well as undetected/spurious peaks, thus facilitating operation in noisy environments at very low computational cost. Our inference engine evaluates posterior probabilities of musical attributes such as root, chroma (including inversion), octave and tuning, given STFT peak frequency and amplitude observations. The resultant posteriors become highly concentrated around the correct attributes, as demonstrated using 227 ms piano recordings with -14 dB additive white Gaussian noise.

Psychoacoustic Issues in Audio Watermarking

Yi-Wen Liu <jacobliu at stanford> (EE)

Outline

- Overview of watermarking: terminologies and goals
- Applications of watermarking, and competing technologies:
  - Copyright protection
  - Copy protection (different!)
  - Piracy tracking
  - Broadcast monitoring
  - Tamper Detection
- Methods in audio watermarking
  - Spread spectrum
  - Time domain processing
  - Spectral manipulation using phase insensity
- The role of attackers
  - Categories
  - Game of watermarking
  - Counter measures and counter counter measures
- The Liu-Smith method (2003)
  - Watermarking based on parametric modeling
  - the Fisher information bound on data hiding rates
  - as an assymetric game
  - Implementation using the sine + noise + transient model
  - Advantages, disadvantages, applicability and usefulness
- Psychoacoustics:
  - How well can we hear a frequency shift?
  - How well can we hear a frequency shift in a noisy environment?
  - How well can we hear frequency shifts superposed?
- Sound demonstration

Sinusoidal Parameter Estimation based on Quadratic Interpolation of FFT Magnitude Peaks

Mototsugu Abe <abemoto at ccrma> (Visiting Scholar, Sony Corporation)

Among various approximate maximum likelihood (ML) estimators, quadratic interpolation of magnitude peaks in a Fast Fourier Transform (QIFFT) has been widely used due to its simplicity and accuracy, which is sufficient for most audio purposes. Indeed, it works nearly as well as a ML estimator for a single, time-invariant sinusoid (or well resolved multiple, quasi time-invariant sinusoids) when a large FFT zero-padding factor is used.

In practice, however, we are shackled with various restrictions, such as 1) we may wish to minimize zero padding for computational efficiency, 2) we may need to deal with interference from nearby signal components, and 3) we may have to consider time-varying sinusoids. All of these requirements raise the error bias and restrict the available range of the estimator.

In this talk, we theoretically predict and numerically confirm the errors associated with various parameter choices such as window types, lengths, and zero-padding factors, and provide precise criteria for designing the estimator. Also, as expansions of the QIFFT framework, we discuss simple bias correction functions and a method for estimation of AM and FM rates.

Piano Dispersion Filter Design

Julius Smith <jos at ccrma> (CCRMA)

This informal seminar presentation is devoted to the problem of designing digital filters to implement dispersion in a waveguide string model.

Winter Quarter 2003-2004

Acoustics of Percussion Instruments

Thomas Rossing <rossing at physics.niu.edu> (CCRMA Visiting Scholar
Professor of Physics, Northern Illinois University)

Most percussion instruments are impulsively excited and vibrate in a rather complex way, until their oscillations die out. At low to medium amplitudes, their vibrations can be conveniently described in terms of normal modes of vibration, but at large amplitude they may show distinctly nonlinear or chaotic behavior. This talk will describe some of the experimental techniques that are used to observe both linear and nonlinear vibrations in some percussion instruments.

Recent developments in musical sound synthesis based on a physical model

Julius Smith <jos at ccrma> (CCRMA)

This talk will begin with a brief review of general finite difference methods for discrete-time simulation of acoustic systems. Next, these methods are related to faster computational forms normally associated with digital signal processing techniques. In particular, the "digital waveguide" and "wave digital" paradigms are briefly reviewed and compared. More specialized methods will be discussed, which may be applied to the simulation of nonlinearities, distributed scattering, and model-order reduction.

Discrete-Time Peak and Shelf Filter Design for Analog Modeling

Jonathan Abel and David P. Berners <{abel,dpberner} at uaudio.com> (Universal Audio, Inc.)

A method for the design of discrete-time second-order peak and shelf filters is developed which allows the response of analog peak and shelf filters to be approximated in the digital domain. For filters whose features approach the Nyquist limit, the proposed method provides a closer approximation to the analog response than direct application of the bilinear transform. A peak filter and three types of resonant shelf filter are discussed, and design examples are presented.

A Filter Design Method Using Second-Order Peaking and Shelving Sections

Jonathan Abel and David P. Berners <{abel,dpberner} at uaudio.com> (Universal Audio, Inc.)

A method for designing audio filters is developed based on the observation that second-order peaking and shelving filters can be made nearly self-similar on a log magnitude scale with respect to peak and shelf gain changes. By cascading such second-order sections, filters are formed which may be fit to dB magnitude characteristics via linear least-squares techniques. A graphic equalizer interpolating prescribed band gains is presented, along with a filter minimizing the Bark-weighted mean square dB difference between modeled and desired transfer function magnitudes. It is noted that using second-order sections parametrized by transition frequency and gain provides a natural mechanism for slewing and interpolation between tabulated designs.

A method of automatic recognition of structural boundaries in recorded musical signals

Unjung Nam <unjung at ccrma> (Music)

Machine recognition of musical features in an audio signal is an important and desirable tool both for identification and retrieval of a given musical recording and for comparisons of multiple music recordings. A robust machine recognition system is critical for successful multimedia database management and useful for interactive music systems. Furthermore, it suggests new research opportunities in music theory, analysis and musicology.

Music is structured at a variety of levels ranging from the quasi-periodicity of frequency to the macro levels of large-scale musical forms. Intermediate levels of structure include structures such as motives and phrases. These structures constitute the salient perceptual units that listeners use to comparatively assess music in terms of the degree of similarity either within a given piece or between pieces. While human listeners are facile at distinguishing recurrence and contrast in music the same task has proven elusive in machine listening paradigms.

In this talk a methodology is proposed that attempts to derive salient and hierarchical musical structures from a raw audio signal by accessing the degree of novelty and redundancy throughout a musical signal.

From Spectral Pitch Estimation to Automatic Polyphonic Transcription: Recent Results

Harvey Thornburg and Randal Leistikow <{harv23,randal} at ccrma> (EE)

In this talk, we review our maximum likelihood spectral pitch estimator robust to undetected harmonics, spurious/interference peaks, skewed peak estimates, and unknown inherent deviations from ideal or other assumed inharmonicity structure. Inherent to our approach is a hidden discrete-valued descriptor variable identifying spurious/undetected peaks. The likelihood evaluation, requiring a computationally unwieldy summation over all descriptor states, is successfully approximated by a MCMC traversal chiefly amongst high-probability states. For historical interest, we compare our method to the classic maximum likelihood estimator of Julius Goldstein. Our method seems to compare favorably, and becomes crucially important in the case of low-frequency interference peaks where Goldstein's method suddenly breaks down. Furthermore, we exploit peak amplitude as well as frequency information, utilizing timbral domain knowledge where this may exist.

Recently, we have begun to extend these developments, presently pursuing a method for Bayesian spectral chord recognition, with our likelihood evaluation as the main computational engine. Octave, key, chord and tuning are abstracted, and may jointly be identified given only the peaks from a STFT frame. Time permitting, we may discuss the development and preliminary results, as well as outline a scheme for integrating framewise correspondences. The latter results in nothing more sophisticated than an HMM, where exact inference is possible (apart from the aforementioned MCMC steps in likelihood evaluation), and it is at least clear where to insert "domain knowledge" (from music theory) in the hierarchy. The latter provides an extremely simple and hopefully robust method for handling automatic transcription tasks, at least those involving polyphonic recordings of a single instrument.

Dynamic Range Compression based on Models of Time-Varying Loudness

Ryan Cassidy <ryanc at ieee.org> (EE)

We incorporate facts from recently proposed time-varying loudness models in the tuning of a contemporary level detection mechanism for the dynamic range compression of audio signals. An analysis of the time-varying and steady-state behaviour of such a mechanism is provided, thus revealing its shortcomings in light of well-known facts about steady-state loudness. To remedy this problem, the design of an equal-loudness filter is presented. The performance and implementation requirements of this filter are compared to the well-known A- and C-weighting schemes used in many sound level meters. Next the aforementioned tuning of the level detector to match recently proposed time-varying loudness models is presented.

Fall Quarter 2003-2004

Fundamentals of musical acoustics and mechanics of musical instruments: An overview

Antoine Chaigne <achaigne at ccrma> (CCRMA Visiting Scholar, Professor, ENSTA and Ecole polytechnique, France)

This overview will present most of the topics which will be developed in the upcoming lectures of the seminar.

One central question, when attempting to synthesize musical sounds from physical models, is to know the degree of refinement of the model needed. Using examples from the families of strings and percussive instruments, this presentation will focus on the necessary features that a physical model of musical instruments must contain in order to produce interesting sounds. Stringed and percussive instruments are characterized by the vibrations of at least some of their constitutive parts: in a guitar, for example, it is mandatory to take the vibrations of strings and soundboard into account. These vibrations are governed by partial differential equations complemented by initial and boundary conditions. Another crucial point is damping. Instrument makers know very well that changing the material of a top-plate has a pronounced effect on the resulting sound. Most of these effects are due to the rate of attenuation of the vibrations in the material itself. In xylophones and vibraphones, the presence of a tubular resonator under the vibrating bar modifies the sound dramatically. In this case, a good model must include the radiation of the bar, the excitation of the tube by the radiated wave and the resulting radiation of the tube. Finally, in all stringed instruments, and also in drums and timpani, one cannot reproduce the sound of the instrument accurately without considering the reaction of the radiated wave on the body itself.

Elements of numerical analysis with application to sound synthesis

Antoine Chaigne <achaigne at ccrma> (CCRMA Visiting Scholar, Professor, ENSTA and Ecole polytechnique, France)

This presentation is essentially a tutorial on basic concepts related to the use and properties of numerical analysis methods in the context of sound synthesis. The main results are illustrated by simple examples, such as piano strings, xylophone bars, and so on.

The physical behavior of musical instruments is described by continuous equations. These equations include derivatives and integrals vs. space and time. In order to solve these equations it is necessary to approximate the model by discrete formulations. Several strategies are possible: finite differences, finite elements, modal truncation, and others. In each case, fundamental questions arise: which spatial/time step should be selected? What are the consequences of these choices? How to ensure the stability of the model? Is it possible to predict the accuracy of a given method? The lecture will focus on the basic problems of stability and numerical dispersion, in the case of simple finite difference modeling of string and bar equations. Analogies and differences with respect to other methods will be discussed. It will be shown, in particular, that the well-known Courant-Friedrich-Levy (or CFL) condition for a second-order explicit finite difference scheme is equivalent to the Shannon (or Nyquist) sampling theorem in signal processing. Fruitful discussion on the comparison between waveguide modeling and finite difference modeling will certainly be of great interest.

Audio watermarking via sinusoidal modeling

Yi-Wen Liu <jacobliu at ccrma> (CCRMA EE Ph.D. student)

Audio watermarking refers to hiding bit-streams in sounds. Among existing applications such as copyright protection and broadcast monitoring, transparency and robustness are always required. Here, transparency means that watermarks should introduce no audible distortion, and robustness means that watermarks can not be erased without degrading the perceptual quality of sounds.

Many existing audio watermarking methods utilize the masking phenomena of human hearing. Fourier-like transform coefficients of a signal are amplitude-modulated to carry information. The modulation is made small enough to introduce distortion only below the masking threshold. Our method, instead, utilizes the frequency just noticeable difference (JND). The idea is to introduce small frequency modulation within JND, and then design watermark encoding and decoding procedures robust to selected types of signal manipulation, such as MP3 compression.

The presentation outline goes as follows:

- Brief introduction to watermarking: definition and requirements
- Sinusoidal modeling for watermarking
  - Analysis at encoder: sine + noise + transient
  - Synthesis at encoder: frequency quantization
  - Analysis at decoder: frequency re-quantization
- Listening tests
- Updates, discussions, and future directions

Automatic music identification

Avery Wang <avery at shazamteam.com> (Shazam Ltd)

We have developed and commercially deployed a flexible audio search engine. The algorithm is noise and distortion resistant, computationally efficient, and massively scalable, capable of quickly identifying a short segment of music captured through a cellphone microphone in the presence of foreground voices and other dominant noise, and through voice CODEC compression, out of a database of over a million tracks. The algorithm uses a combinatorially hashed time-frequency constellation analysis of the audio, yielding unusual properties such as transparency, in which multiple tracks mixed together may each be identified. Furthermore, for applications such as radio monitoring, search times on the order of a few milliseconds per query are attained, even on a massive music database.

Numerical experiments for piano and xylophone

Antoine Chaigne <achaigne at ccrma> (CCRMA Visiting Scholar, Professor, ENSTA and Ecole polytechnique, France)

One interest of developing physical models of musical instruments is to make numerical experiments. In practice, this means that the model is used for investigating, through simulations, the relevance of the constitutive parts of the instruments on the produced sounds. This strategy will be illustrated in this lecture by numerous examples related to both the piano and xylophone. For the piano, the model yields interesting results related to the propagation of waves along the string, the hammer striking position, the hammer/string mass ratio and the pulse waveform of the impact force. Future possible developments for this instrument will be presented. For the xylophone, the model shows the influence of mallet and bar properties on the initial transient of the sound, the relevance of the undercut profile and the function of the resonator. Extensions to marimbas and vibraphones will also be presented and discussed.

Application of Loudness/Pitch/Timbre decomposition operators to auditory scene analysis

Mototsugu Abe <moto at ccrma> (CCRMA Visiting Scholar
Sony Corporation, Japan)

In this presentation, I will present my former work at the University of Tokyo, which was completed in the mid 1990s and was a part of my doctoral thesis research.

The first half of the presentation is on "Loudness/Pitch/Timbre(*) Decomposition Operators." In this research, we constructed a set of operators as general audio signal processing tools in the time-frequency domain.

More concretely, we focus on the instantaneous change of audio in the Wavelet domain. The change is decomposed into three orthogonal components, and a method is given for projecting the change onto these components.

The second half is on an application of the operators to the problem of computational auditory scene analysis(*). In a (monaural) multi-stream sound(*), when frequency components of the sound change together in amplitude and frequency, they are grouped together as one auditory stream. Since the above operators provide us with a method for quantifying the instantaneous changes in amplitude and frequency, we utilize them for the initial stage. Then the estimated amplitude and frequency changes of the components are used to construct a probabilistic space in which peaks correspond to streams. The probability distribution may be updated with new data to follow streams through time.

Notes:

(*) Though "loudness", "pitch" and "timbre" are terms corresponding to human perception, they are used in our context to mean "amplitude change", "frequency change" and "the other types of change", respectively. (At the time, my adviser wanted to somehow relate this work to human perception, but we haven't done that yet.)

(*) The term "auditory scene analysis" is the title of a book written by Albert S. Bregman in 1990, in which he summarizes psychophysical characteristics of human perception in grouping/separating sounds which occur simultaneously or sequentially. In the 1990s, many researchers/engineers tried to simulate/implement them as a computational model on a computer.

(*) The term "multi-stream sound" is almost the same as "multi-source sound". However, "stream" corresponds to human perception regardless of how many sources actually exist, whereas "source" corresponds to an actual physical sound source.

Bayesian Two-Source Modeling for Separation of N Sources from Stereo Signals

Aaron Master <asmaster at ccrma> (CCRMA EE Ph.D. student)

We consider an enhancement to the DUET sound source separation system of Rickard and others, which allowed for the separation of N localized sparse sources given stereo mixture signals. Specifically, we expand the system and the related delay and scale subtraction scoring (DASSS) of Master to consider cases when two sources, rather than one, are active at the same point in STFT time-frequency space. We begin with a review of the DUET system and its sparsity and independence assumptions. We then consider how the DUET system and DASSS respond when faced with two active sources, and use this information in a Bayesian context to score the probability that two particular sources are active. We conclude with a musical example illustrating the benefit of our approach.

Polyphonic Instrument Identification Using Independent Subspace Analysis

Pamornpol ('Tak") Jinachitra <pj97 at stanford> (CCRMA EE Ph.D. student)

In this talk, I will present an ongoing work of a system which tries to identify the musical instruments played in a polyphonic mixture. The features used in classification are derived from the Independent Subspace Analysis (ISA) which somewhat decomposes each source, and the mixture, into its statistically "independent" components. Without re-grouping or actually separating the sources, these features can be used as fingerprints of each instrument, assuming the decomposed fingerprints are robust enough to the mixing process. The accuracy test on two-tonal instrument mixes (possibly with percussion) will be presented along with some tests on real songs. Interesting demos of the original ISA for source separation and auditory grouping will be given as an introduction.

Materials and sound properties. The role of damping.

Antoine Chaigne <achaigne at ccrma> (CCRMA Visiting Scholar, Professor, ENSTA and Ecole polytechnique, France)

In the daily life, we often knock on a table or on a wall in order to get information on their internal structure and materials. From the acoustic signal, our brain is able to recognize whether the objects we are seeing are made or glass, metal of wood, for example.

Therefore, if we want to simulate virtual structure-borne sound sources properly, the synthesis program must include specific features that allow recognition and discrimination among materials without ambiguity.

In this presentation, particular attention will be paid to the role of damping, whose role is crucial in the recognition of materials. It will be shown that accurate modeling of three major causes of damping (thermoelasticity, viscoelasticity and radiation) yields very realistic simulations of a large variety of materials (metal, wood, glass, carbon fibers,...). The presentation will be illustrated by comparison between measured and simulated sounds radiated by impacted plates.

References: `http://www.ensta.fr/~chaigne/articles/`
(papers `chaigne_plate.pdf` and `lambourg_plate.pdf`) http://www.ensta.fr/ chaigne/articles/

The sounds of gongs and cymbals

Antoine Chaigne <achaigne at ccrma> (CCRMA Visiting Scholar, Professor, ENSTA and Ecole polytechnique, France)

Gongs and cymbals belong to that category of instruments for which no physical model is available yet. Therefore, there is a real challenge for developing such models and implementing them in synthesizers. In order to do so, it is necessary to understand the main properties of these instruments and, in particular, the fact that the sound changes drastically depending on the strength of the impact.

The presentation will be mainly focused on experimental results that show, in particular, that the vibrations of gongs and cymbals are highly nonlinear, the magnitude of the transverse motion being often larger than the thickness of the structure. Accurate signal analysis shows that the frequencies which can be seen on sound spectra for strong impact are governed by very simple arithmetic rules between the eigenfrequencies of the structure. These frequencies are called "combination resonances". For very strong impact, specific signal processing tools must be applied in order to show that the signals are chaotic and governed by only a small number of degrees of freedom. In this case, the spectra do not exhibit isolated peaks but rather a "deterministic noise" which should be possible to reproduce with only a small number of nonlinear oscillators.

References: `http://www.ensta.fr/~chaigne/articles/Percussive_instruments/` (papers `ismagong_V1.pdf` and `ShellGONGFA02.pdf`)

Spring Quarter 2002-2003

Estimating performance parameters from the acoustic waveform of the violin

Arvindh Krishnaswamy <arvindh at ccrma> (EE)

Friday, April 4, 3:15 PM, CCRMA Library

In this friday's seminar I wish to talk about the following:

- My ongoing project on inferring control inputs to a musical instrument based on discriminant analysis. Please see

  `http://ccrma/~arvindh/pprs/icme03.pdf`

  for the paper that is going to be presented at IEEE ICME 03. It has an abstract and references which may be useful. I will present the ideas from this basic paper as well as discuss the numerous areas for future improvement which I have in mind.
- I will also give an overall 'research update' and talk about what research I plan to pursue on a long-term basis. Comments and suggestions will be appreciated. I am interested basically in an Analysis -¿ Feature extraction -¿ "Resynthesis from scratch" system for processing music/audio signals using musical models and acoustic signal models.
- An update on my latest thoughts on South Indian music will be given.

All of the above topics may be stretched, compressed, truncated or even changed depending on time and, most importantly, my level of preparation. ;-)

Comb Filter Modulation Effects

David Lowenfels <dfl> (CCRMA)

In this week's CCRMA DSP Seminar (Friday, 3:15pm, Ballroom), MA/MST student David Lowenfels will talk about his recent work in Virtual Analog synthesis. All are welcome. – Julius

ABSTRACT:

This is an invitation to come hear about my research into Virtual Analog synthesis. The BLIT model of Stilson and Smith has been extended with a single comb filter, allowing for effects such as waveform morphing, PWM, chorus, detune/unison, and hard sync. Forays into recursive FM (actually PM) have drastically reduced the computational complexity, superseding BLIT. I will show working demos and discuss my paper in progress.

David Lowenfels (http://ccrma/~dfl/)

Physical Modeling Synthesis of the Recorder

Hiroko Terasawa <hiroko at ccrma> (CCRMA)

Hi,

In the DSP seminar tomorrow, I will be presenting about physical modeling synthesis of recorder sound, which I worked on last year, and which I will present in ASA Nashville meeting next week. It will be a short presentation (3:30 - 3:50). Please drop in if interested. Your feedback will be greatly appreciated!

Thank you, - Hiroko

— ABSTRACT —

A time-domain simulation of the soprano baroque recorder based on the digital waveguide model (DWM) and an air reed model is introduced. The air reed model is developed upon the negative acoustic displacement model (NADM), which was proposed for the organ flue-pipe simulation [Adachi, Proc. of ISMA 1997, pp. 251–260], based on the semiempirical model by Fletcher [Fletcher and Rossing, The Physics of Musical Instruments, 2nd ed. (Springer, Berlin, 2000)]. Two models are proposed to couple DWM and NADM. The jet amplification coefficient is remodeled for the application of NADM for the recorder, regarding the recent experimental reports [Yoshikawa and Arimoto, Proc. of ISMA 2001, pp. 309–312]. The simulation results are presented in terms of the mode transient characteristics and the spectral characteristics of the synthesized sounds. They indicate that the NADM is not sufficient to describe the realistic mode transient of the recorder, while the synthesized sounds maintained almost resemble timbre to the recorder sounds.

Today, after Hiroko presents her upcoming ASA talk, Eric Lindemann will talk about his generalized sampling synthesis based on reconstructing phrase segments. Then after that, Arvindh Krishnaswamy will talk about his recent experience with nonlinear string synthesis algorithms. – Julius

AUDIO Quality Analysis with PEAQ

Opticom Inc.

This Friday (5/2/03) at 3:15pm in the CCRMA Ballroom, a representative from Opticom will present software for the objective measurement of sound quality. This is a long standing problem in audio signal processing, and one which few have addressed very thoroughly. All are welcome – Julius

Abstract

Audio quality is one of the key issues for a range of modern electronic equipment. Whether it is during the development of the equipment or while operating it, one will always be faced with the problem to determine and optimize the quality. New evaluation techniques based on modeling human perception have been devised to tackle this issue. This presentation intends to give a comprehensive overview on the latest standards for state of the art audio quality testing.

Up until now, one of the best ways to measure the sound quality of modern audio equipment has been to implement elaborate listening tests with experienced test subjects, so-called "expert listeners". The traditional measurement technologies have proven insufficient at providing any detailed information on the quality of modern digitally based circuit designs. Due to the increased use of such designs, in digital broadcasting, professional and consumer electronics, the storing of audio signals, as well as in the field of multi-media computing and Internet, it became necessary to develop an innovative, uniform method of measuring with which it is possible to objectively measure the quality of audio. After thorough verification, a model was recommended by the ITU-R as a measure for the perceived audio quality ("PEAQ") under recommendation BS.1387. The PEAQ measurement method is applicable to most types of audio signal processing equipment, both digital and analog.

Outline for the Presentation:

- Who is OPTICOM?
- Introduction to the problem
- Subjective Audio Quality Testing
- Objective Audio Quality Testing
- Principles of Psycho-Acoustics
- What is "PEAQ"?
- OPERA - A Comprehensive Solution

Winter Quarter 2002-2003

Prior Identification for Harmonic-Comb Piano Model

Harvey Thornburg <harv23 at ccrma> (EE)

Friday, Jan. 17 3:15 PM, CCRMA Library

In a previous seminar, we presented a "physically-informed" Bayesian sinusoidal model for reproduction of one vibrational mode of a single piano note. This model, based on solutions to the PDE of the physical model of Bensa et al. (2002) comprises a bank of coordinated exponentially decaying sinusoidal oscillators which exist in a roughly harmonic relationship (hence the name "harmonic comb"), but with adjustments for inharmonicity and frequency-dependent losses. Tolerances for parameter variation, deviations from model structure, and recording noise are introduced in the sense purely *functional* dependences are replaced by *statistical* dependences, i.e. conditional probability distributions.

Our statistical model is driven by the purely physical parameters: fundamental frequency, inharmonicity factor, and the frequency-dependent loss parameters, as well as the initial amplitudes and phases for each oscillator. Furthermore, in the interests of modeling multi-stage decay behavior, we must introduce *multiple* vibrational modes; each with its own set of physical parameters, initializations, etc. A single, exponentially decaying envelope for each partial does not suffice.

To identify so many input parameters from a such a complex statistical model, using nothing but the time-domain audio signal as "observation", is a daunting task. Nevertheless, we already possess a good deal of information about the input parameters and their relationships. The importance of the Bayesian methodology is that it allows prior information and domain knowledge to be introduced in the form of a prior distribution over the unknown parameters.

In this talk, we address the prior identification over all input parameters across multiple vibrational modes, using a precomputed database of partials' fundamental frequency, decay rate, and initial amplitude information from Bensa et al. (2002). We add to the set of "physical parameters", two input parameters modeling the sequence of initial amplitudes (consequential of the hammer strike). For a given piano note, mdeling of initial amplitude is linear on a log scale w.r.t log frequency, roughly justified by Hundley et al (1978).

Our initial attempt concerns the prior identification on a note-by-note basis.To this end, we explore two hypotheses concerning the interaction of multiple vibrational modes. One hypothesis, called the "discrete"approach, implied by numerous papers in the acoustics literature (e.g. Weinreich (1977) and Nakamura (1989), among others), implies the existence of discrete "mode classes", (i.e. "fast" and "slow" modes, the fast modes exhibit higher initial amplitudes etc.) Another hypothesis treats each mode as i.i.d, but attempts to discover correlations within the deviations, i.e. unusually fast modes exhibit unusually high amplitudes.

With the discrete mode-class approach, there is the difficulty of sorting the precomputed data according to mode class. A certain separability criterion is optimized; the exact optimization requires an exhaustive search. Nevertheless, we develop two methods: an approximate "Viterbi" algorithm, and a random search with simulated annealing. Both approximate algorithms demonstrate near-optimal performance on "known" sorting problems of similar size and dimension, though the "Viterbi" algorithm is considerably faster.

With the continuous approach, the main technical issue is the (suitably constrained) covariance identification. The direct maximum-likelihood approach has (at this time) proven intractable in closed-form. It is also unknown whether multiple local maxima exist for any configuration of the problem. As such, we pursue an iterative optimization based on the EM algorithm, which is guaranteed to converge to a local maximum of the likelihood function. An initialization trick in practice avoids the possible problem of converging to the wrong local maximum, although a theoretical guarantee has not yet been found. Nevertheless, we attain convergence to an acceptable result in very few (3-5) iterations.

Our results show that the data fails to support the discrete mode-class approach, despite ample support in the theoretical acoustics literature. The culprit, excessive variability in the range of partials' initial amplitudes and decay rates, induces significant overlap among the classes. On the other hand, the continuous approach finds success: the expected correlation between high initial amplitude and fast decay rate is established. As such, we extend the continuous approach to cover prior identification for the entire piano range. We introduce polynomial fits for the means of all input parameters as well as the covariance parameters. Fortunately, it seems unnecessary to pursue a "piecewise" fit as previously thought, which responds to the differing numbers of strings.

A Power-Normalized Wave Digital Piano Hammer

Julius Smith <jos at ccrma> (CCRMA)

This Friday at 3:15, in the CCRMA Library, I will present the recent ASA-02 paper by Stefan Bilbao, Julien Bensa, Richard Kronland-Martinet, and myself.

ABSTRACT

For sound synthesis purposes, the vibration of a piano string may be simply modeled using bidirectional delay lines or digital waveguides which transport traveling wave-like signals in both directions. Such a digital wave-type formulation, in addition to yielding a particularly computationally efficient simulation routine, also possesses other important advantages. In particular, it is possible to couple the delay lines to a nonlinear exciting mechanism (the hammer) without compromising stability; in fact, if the hammer and string are lossless, their digital counterparts will be exactly lossless as well. The key to this good property (which can be carried over to other nonlinear elements in musical systems) is that all operations are framed in terms of the passive scattering of discrete signals in the network, the sum-of-squares of which serves as a discrete-time Lyapunov function for the system as a whole. Simulations are presented.

Modeling vocal-tract influence in reed wind instruments

Gary Scavone <gary at ccrma> (CCRMA)

In this week's DSP seminar, I'll present some early explorations regarding the influence of upstream vocal tract resonances in reed wind instrument performance and modeling. First, I'll provide a demonstration of vocal tract effects with my saxophone and review previous research in this area. Then I'll discuss a digital waveguide simulation that I've made and demo its behavior. Finally, I'll open the floor to feedback and suggestions.

Recent Work in Self-Similarity Analysis of Audio

Jonathan Foote <foote at fxpal.com> (FX Palo Alto Laboratory, Inc.)

Matthew Cooper and Jonathan Foote
FX Palo Alto Laboratory, Inc.

We will present some new approaches and recent results from our work in the analysis of audio. Starting with a similarity matrix consisting of all pairwise measurement of spectral similarity, we review how the matrix can be used for segmentation and rhythmic analysis of music. We present more recent work on how this analysis can be used for retrieving music by rhythmic similarity, and how the matrix can be factorized to locate repeating segments, such as verses and choruses in popular music.

Autumn Quarter 2002-2003

Copyright protection of synthesized objects: a geometric interpretation

Yi-Wen Liu <jacobliu at stanford> (EE)

Synthesized multimedia objects are emerging everywhere now. One can talk on the phone to a virtual representative that speaks a synthesized tongue, drink soda of synthesized taste, such as *Coke*, or even fall in love with *Simone*, a synthesized character. It becomes urgent to protect such objects as intelectual properties, for the synthesis of them often involves a lot of computation power and human hours. This talk presents a mathematical definition of the term *synthesis*, and aims to provide a framework for the design of robust watermarking algorithms to, hopefully, protect copyrights to synthesized objects. In particular, this talk proposes a strategy of watermarking before synthesis, and presents a geometric method to evaluate its robustness to random attacks.

- **PART I: INTRODUCTION**
  - Synthesized multimedia objects: creation and distribution
  - A mathematical definition of *synthesis*
  - A mathematical formulation of the watermarking game
  - A review of the duality between audio watermarking and audio coding Copyright protection by watermarking before synthesis (WBS)
- **PART II: ANALYSIS**
  - A simple case study – protecting single-tone cell phone ringers
  - Problem formulation – frequency estimation under Gaussian white attacks
  - Solving the problem by designing a robust frequency estimator
  - The Marti Gras beads(MGB) geometric interpretation
  - Calculating the Cramér-Rao bounds (CRB)
  - Generalization to K dimensional manifold MGB interpretation
  - Calculation of variance of parameter estimation error based on K-manifold MGB interpretation
  - Comparing the MGB variance and CRB
- **PART III: POSSIBLE FUTURE DIRECTIONS**

  More case studies – e.g., non-stationary sinusoids, synthesized speech,
  - new music by physical modeling
  - Customizing WBS according to various types of anticipated attacks
  - Vector quantization on manifolds
  - Adaptive learning on manifolds

Doppler Simulation and the Leslie

Julius Smith <jos at ccrma> (CCRMA)

with Stefania Serafin (Music), and Jonathan Abel and David Berners (Universal Audio)

An efficient algorithm for simulating the Doppler effect using interpolating and de-interpolating delay lines is described. The Doppler simulator is used to simulate a rotating horn to achieve the *Leslie* effect. Measurements of a horn from a real Leslie are used to calibrate angle-dependent digital filters which simulate the changing, angle-dependent, frequency response of the rotating horn.

The *Doppler effect* causes the pitch of a sound source to appear to rise or fall due to motion of the source and/or listener relative to each other. The Doppler effect has been used to enhance the realism of simulated moving sound sources for compositional purposes, and it is an important component of the "Leslie effect."

The *Leslie* is a popular audio processor used with electronic organs and other instruments. It employs a rotating horn and rotating speaker port to "choralize" the sound. Since the horn rotates within a cabinet, the listener hears multiple reflections at different Doppler shifts, giving a kind of *chorus effect*. Additionally, the Leslie amplifier distorts at high volumes, producing a pleasing "growl" highly prized by keyboard players.

In this presentation, an efficient algorithm for digital simulation of the Doppler effect is presented, and the algorithm is applied to the problem of rotating-horn simulation for the Leslie effect.

Gaussian Magic

Julius Smith <jos at ccrma> (CCRMA)

At the CCRMA DSP Seminar Friday, 10/11/02, at 3:15-4:05 in the CCRMA Library, I will present a tutorial review of the amazing Gaussian (a.k.a. "bell curve" or "normal curve"). In particular, the interesting and elegant case of $\exp(-pt^2)$ for p complex will be treated (a Gaussian-windowed chirp), and applications to sinusoidal modeling of audio will be mentioned. In addition, I will review other classic results as time permits, such as the maximum entropy property, closure under multiplication and convolution, properties of moments, appearances in physics, the central limit theorem, and so on. – jos

More Gaussian magic!

Aaron Master <asmaster at ccrma> (EE)

At the CCRMA DSP Seminar this Friday, 3:15-4:05 in the CCRMA Library (or Ballroom... we'll see) I will present even more on sinusoidal / chirp modeling. This session will be very much related to Julius's presentation last week and will include:

- A review of my summer research in chirp modeling
- A quick summary of Yi-Wen's conclusions on chirp modeling
- Applications of chirp parameter estimation techniques by Yi-Wen, Julius, and myself
- An answer to the oft-raised question from the last seminar: "What happens when we have affine amplitude modulation of the chirp?"
- "Competitive" results of the different algorithms under various SNR conditions, phase offsets, and amplitude modulation. (Let me preview the answer: Gaussians are amazing).

Hope to see everyone there! —Aaron

3D Sound Project

Stephanie Salaun <salauns at ccrma> (CCRMA Visiting Researcher)

Hi everybody!

On Friday, October 25th, at 3h15 in the CCRMA Library, I am presenting part of my 3D sound project (it's not quite finished). Here is the abstract for those who are interested:

In every day life sounds are all around us in the horizontal and vertical planes. One of the methods for reproducing virtual sound is to binaurally record a sound and play it back through headphones or through loudspeakers. When using headphones the sound is played back with only a small transmission loss, so the signal is almost reproduced correctly at each ear.

However playing back the sound through loudspeakers creates crosstalk between them, an unwanted effect that may be removed by signal processing (crosstalk cancellation algorithm).

When trying to produce 3D sound through loudspeakers, we have to consider the loudspeakers 'setup, the crosstalk cancellation algorithm, the room, and other factors. This method will reproduce the 3D sound accurately at only one head position, known as the "sweet spot".

My system has four loudspeakers above the listener splitting the frequency into two frequency bands. The crosstalk cancellation is achieved using adaptive filtering.

Currently, I have having some difficulties with the implementation of my algorithms and I hope to get some feedback during the seminar.

Stephanie

Constrained EM Estimates for Harmonic Source Separation

Pamornpol Jinachitra ("Tak") <pj97 at stanford> (EE)

A constrained iterative method for harmonic source parameter estimation is proposed based on an EM algorithm by Feder&Weinstein(1988) with an intent for harmonic source separation. The algorithm is attractive in a guarantee of convergence to local minimum and the ability to deal with each partial separately. The problem of coinciding partials and interference among them in general is mitigated by the constraints on the "weak" partials on the stronger ones of the same harmonic source. A useful scheme to determine the weakness of a partial is proposed. The constrained iteration is shown to give highly accurate estimates of the sinusoidal parameters which results in a good source separation in most cases of highly overlapping spectra. However, the encountered problem of amplitude estimate inaccuracy of coinciding partials has not been addressed.

Sinusoidal Peak Estimators

Julius Smith <jos at ccrma> (CCRMA)

I will present some Matlab simulation results comparing several peak-frequency estimators, specifically Kay's method, Lank's method, and our old friend, the quadratically interpolated FFT dB magnitude peak. The estimators are compared to each other and to the Cramer-Rao lower bound.

Gaussian Chirp Rate Estimation

Aaron Master <asmaster at ccrma> (EE)

I will show the mathematical equivalence between a modified version of the Smith estimator and the Marques+Almeida estimator for frequency chirp parameter in a Gaussian-windowed signal. I will also explain the iterative aspect of the Marques and Almeida estimator, as well as their magnitude and phase estimation technique. Results comparing one iteration of the Smith and M+A methods will be presented. The various impacts of parabolic curve fitting, single point second order differencing, and averaged second order differencing will be considered for each method.

Watermarking Parametric Representations for Synthetic Audio: a Mid-Term Research Update

Yi-Wen Liu <jacobliu at stanford> (EE)

I will propose a system to watermark parametric representations for synthetic audio. The system combines quantization index modulation at the encoder and maximum likelihood parameter estimation at the decoder. To guarantee error-free data hiding under expected types of attacks, knowledge of Fisher information and Cramer-Rao bounds is applied to the system design. Experiments show that, merely by quantizing the frequency of sinusoidal tones, one can achieve 50b/s of data hiding that is robust to perceptually shaped additive attacks such as an MP3 compression.

Winter Quarter 2001-2002

Pitch Detection for Automatic Transcription

Ryan Hinton <rhinton at stanford> (EE)

The goal is to produce a general transcription tool with output appropriate for music notation software. The tool should accept a CD-quality WAV file of music based on the Western tempered scale. The source of the music should not be constrained: it should be able to process pitches from instruments, human voices, electronic synthesizers, or from any other source. Furthermore, the tool should accept reasonably many voices. The tool should record to a file the pitches in the music along with their start and stop times. One possible output file format is a MIDI file: several popular notation packages notate music from MIDI files. Other output file formats may be more appropriate depending upon file format complexity and eventual tool feature set.

Future algorithms may include determining which key the music is in; how much the music is off from the true key (i.e. a violin tuned a little off from a true "A"); whether a particular voice is sharp or flat; and the tempo and meter of the music. The tool could even attempt to identify particular instruments.

The tool may also be enhanced to accept different sample rates; different scales (i.e., American vs. international basis frequency and non-Western music); and different input and output file formats.

I will put together a summary of the goal and the challenges I know of along with algorithms I have considered.

Excitations models in self-sustained oscillators with a focus on the bowed string

Stefania Serafin <serafin at ccrma> (Music)

Talk 1:

The goal of this talk is to analyze different kinds of excitation models with a focus on friction-driven self-sustained oscillators like the bowed string. Different kinds of models have been proposed to reproduce the action of a bow exciting a violin string. Nowadays the availability of fast processors allows implemention of highly refined models in real-time. In this talk, I will present the evolution of friction models from the very simple one proposed by Coulomb to the latest discoveries on bowed string modeling.

Talk 2: Randal Leistikow will present his latest results constructing a "piano filter bank" which sets a passband around each piano key and rejects all other keys.

Applications of Probabilistic Monitoring to the Stabilization of the Joint Segmenter/Rhythm Tracker

Harvey Thornburg <harv23 at ccrma> (EE)

Harvey Thornburg
CCRMA DSP Seminar
Friday, Feb. 8 2002

An unfortunate property of the joint segmenter/rhythm tracker, where signal and event processing are decoupled, is that bursts of segmentation errors (misses, false alarms) lead to adjustments in the rhythm tracker's belief state which reinforce future segmentation errors of the same type. Though the joint engine is still locally stable, too large a disruption (involving many missing segments) leads to a of global instability which manifests in tempo period doubling. An interesting point for discussion is whether this instability is an inherent property of the decoupling or whether some structural adjustments might eliminate it. The former appears most likely to be true.

A variety of simple approaches can mitigate this instability, such as tuning the rhythm tracker to respond more slowly to tempo variations, or artificially increasing the entropy of the switching transition matrix. However, these measures, being unneccessarily conservative, tend to bias the overall estimation. A better approach is to exploit additional, independent information such that the tempo usually exists in a certain range (say, 50-200 bpm). Since this information is actually "true", we should not in theory experience adverse effects from bias.

In this talk, we will formalize the concept of "independent information", and show how it leads to a "feedback control" interpretation in the context of segmentation/rhythm tracking, thus providing a natural guard against instability. This approach can also be referred to as "probabilistic monitoring". Additionally, we discuss substantial pitfalls induced by computationally necessary approximations. So far, results give only mixed success in the application to joint segmentation/rhythm tracking, and some improvements are desired.

Perceptually similar orthogonal sounds and applications to multichannel acoustic echo cancellation

Yi-Wen Liu <jacobliu at stanford> (EE)

With the availability of increased communication bandwidths in recent years, people have become interested in full-duplexed, multichannel sounds for telepresence services because of the potentials for providing much better hearing experiences. Nevertheless, in any full-duplex connection of audio network, the problem of acoustic echoes arises due to the coupling between loudspeaker(s) and microphone(s) placed in the same room.

Moreover, it is known that the cancellation of multichannel acoustic echoes is a mathematically ill-conditioned problem. What happens is that, due to the high correlation between signals in multiple channels, an adaptive echo canceler tends to converge to a degenerate solution and fails to find the true coupling paths.

Although several types of algorithms for decorrelating the channels have been proposed to regularize the problem in the context of speech teleconferencing, these algorithms are all developed based on the criterion that any sound other than the speech signals should not be heard. However, it is not necessarily so in applications such as video games and performing arts where back-ground sound effects and background music are common and desired practices.

In this talk, it is presented to utilize background sounds for multichannel echo canceling.

Methods are developed to generate arbitrarily many orthogonal and perceptually similar sounds from a mono source, and the sounds are fed into a multichannel echo canceler for the canceler to better identify the echo paths. 2-channel and 5-channel simulations will be demonstrated.

Design of MLS Measuring System for Objective Room Acoustical Parameters

Stephanie Salaun <salauns at ccrma> (CCRMA)

A room can be defined by different acoustical parameters (decay curve, reverberation time, early decay time, clarity..). In order to measure these parameters, different techniques can be used like periodic pulse testing, time-delay spectrometry (TDS) and maximum-length sequences (MLS).

This project deals with a system measuring room acoustical parameters using the Maximum-Length Sequence (MLS) method. This system is based on both a DSP and a PC where the DSP generates the MLS and extracts the room impulse response. The PC calculates the parameters and displays them in a graphical user interface (GUI) which also controls the measurement.

Introduction to wavelet filter banks

Julius Smith <jos at ccrma> (CCRMA)

Subject: DSP Seminar, CCRMA Library, Friday 3:15

This week, Harvey Thornburg will present on Bayesian transient estimation in the first half, and I will present an introduction to wavelet filter banks in the second half.

Peak-Adaptive Phase Vocoder

Aaron Master <asmaster at ccrma> (EE)

In this talk, based on my ICASSP-02 presentation this year, I will describe a new method for refined estimation of amplitude and frequency trajectories in sinusoidal modeling. The method consists of performing a phase-vocoder style modeling of individual peaks in the short-time Fourier transform (STFT). When a peak corresponds to a true sinusoid for the duration of the analysis window, the instantaneous amplitude and frequency reduce to constants. More generally, however, narrow-band amplitude and frequency modulations can be accurately recovered over the duration of the analysis frame. The refined estimates can replace the usual piecewise-linear amplitude and frequency trajectories in sinusoidal modeling, allowing for more accurate modeling of musical attacks and pulsed waveforms such as voice. Instantaneous phase may be optionally tracked as well. In summary, the peak-adaptive phase vocoder can be viewed as a phase vocoder having analysis channels which are constructed adaptively about time-varying peaks in the STFT.

Autumn Quarter 2001-2002

Polyphonic Music Transcription

Clint Martin <cmartin_121 at yahoo.com> (EE)

In the DSP seminar on Friday, October 5, I will be discussing polyphonic music transcription. A music-transcription system takes a recorded piece of music, tries to accurately determine what was played, and outputs some sort of high-level musical notation; the output could be a MIDI file or sheet music. Part of the presentation will briefly cover different approaches people are taking to the polyphonic music-transcription problem and will demonstrate the results obtained by two recent music-transcription software packages. The majority of the presentation will be spent discussing my research on the topic, which includes an original signal segmentation scheme and a note-finding algorithm, based on the research of Andranick Tanguiane, that takes the partials present in the music and tries to determine the notes that were played. The theory and performance of the note-finding algorithm will be discussed in depth, and sound files will be played that demonstrate the system's strengths and weaknesses.

Bayesian Segmentation and Rhythm Tracking: Part I: Constructing and Identifying Probabilistic Models of Rhythm

Harvey Thornburg <harv23 at ccrma> (EE)

DSP Seminar
Harvey Thornburg
Friday, October 19 2001

Segmentation of polyphonic music is quite difficult, especially in the context of automatic transcription, where one task is to find the change times which correspond explicitly to note onsets. Thanks to rhythmic structure, the pattern of onsets is highly regular. The pattern of onsets admits a probabilistic structure (parameterized joint distribution over all onsets) known as the *rhythm model*. One goal is to exploit this "regularity" to improve the overall segmentation performance and better adapt segmentation to the problem of onset detection, via Bayesian segmentation: the rhythm model serves as a machine for generating priors about future/past onset locations used subsequently in segmentation. In a practical situation the rhythm model is incompletely specified; even basic attributes such as tempo, meter, and so forth must be identified as well: any stream of onsets serves as data with which to learn and identify parameters in the rhythm model. Hence, a fundamental task in automatic transcription is the joint segmentation and rhythm modeling.

In this talk (first in a two-part series) I will focus on rhythm modeling, and the identification of rhythm models from onset data. The sequel will focus on Bayesian segmentation and the issues of interfacing segmentation and rhythm modeling such that both operate in tandem on an audio signal. Since the focus is on modeling "events", this talk may be less "DSP-centric" and more "machine learning" than the usual seminar, but these aspects are very important to a problem which is fundamental to signal processing (where there exists any kind of regular structure to the evolution of a signal model's parameters)

To begin, I introduce the rhythm model's representation. This has two components: I first review the "metronome" model of Cemgil et al. for a constant-interval sequence with unknown tempo. Next, I introduce simple Markovian model for the sequence of note intervals and metric positions. The meter can be obtained as a function of the parameters (transition probabilities) of this model. The key challenge is the obtainment of a common sparse representation encompassing a variety of common meters. Finally, to conclude the representation part, I show how everything may be integrated in the framework of a linear switching state space model, where the metronome state captures the evolution of tempo/event position, and the switching state captures the evolution of rhythmic interval/metric position. The issue of linearization will also be discussed here.

The remainder of the talk focuses on inference and learning in switching state-space models. The inference step involves tracking "state" quantities like tempo, the expected

time location of a change, the current interval, metric position and so on. One obtains the conditional distribution of the states at all event locations given all observed onset times (smoothing) or just past and present observations (filtering) Since any particular event location is a function of the metronome state, the inference step supplies the appropriate prior distributions to use in segmentation.

Exact inference in switching state-space models is intractable; as we will see, the representation of the conditional distribution involves an exponentially growing mixture of Gaussians. (This is by contrast to a hidden Markov model, in which the only hidden state is discrete: we could have an exponentially growing mixture of scalar probabilities, but we don't incur this as a cost because a mixture of scalars is scalar.) In this talk I compare two approximate methods: Viterbi inference and a direct junction-tree based approach with distributional approximations. The junction-tree approach is based on a general method for organizing inference steps in probablistic networks called "variable elimination" or the "junction tree algorithm".

The learning step, by contrast, involves identifying parameters specifying "dynamic" quantities which describe the evolution of the state (in this case, the Markov transition probabilities, because everything else is fixed). Parameters are chosen to maximize the likelihood of the observations; the resultant optimization is solved via the EM algorithm. The likelihood computation is handled by the inference step, which highlights the need for accurate local approximations across groups of consecutive switching states (and thus the need to develop the junction tree algorithm.)

Making teleconferencing a better experience—stereophonic solutions

Yi-Wen Liu <jacobliu at stanford> (EE)

This talk is going to be a summary of my work on a stereophonic teleconferencing project at the department of broadband communication services of AT&T Research Labs during the past summer. We believed that although there have been a lot of efforts recently on improving video, the audio quality of commerially available, mono-sound teleconferencing systems is not very good. To build up a sound system devoted to teleconferences, in which many people often need to be able to argue at the same time, we claim that a stereo-sound, full-duplex system is a must.

It turns out that the cancellation of the echoes that arise due to the coupling between multiple loudspeakers and microphones is a mathematically ill-conditioned and hence still unsolved problem. Perceptually, these echoes "enlarge" the conference room and may degrade the speech clarity seriously. We propose a new idea to digitally watermark sounds from different loudspeakers so that the coupling paths from multiple loudspeakers to each of the microphones can be distinguished and the stereophonic echo cancelling problem is reduced to parallel monophonic echo cancelling problems.

A preliminary example of echo cancellation using digital watermarks will be demonstrated and the performance will be evaluated.

Bayesian Segmentation and Rhythm Tracking: Part I: Constructing and Identifying Probabilistic Models of Rhythm

Harvey Thornburg <harv23 at ccrma> (EE)

DSP Seminar
Harvey Thornburg
Friday, October 19 2001

This seminar will begin as an informal continuation of my previous seminar on the use and identification of probabilistic models for rhythm. I will conclude this part by giving more examples and proofs concerning the junction tree algorithm, as well as focusing on the interpretation of the resultant recursions as a bank of extended Kalman filters which expands during a switching state update and contracts upon the distributional approximation. What's really important for understanding the recursions is that they can be verified directly, using conditional independence relations. The fact the recursions "work" doesn't rest on the abstract graph-theoretical concepts of the junction tree algorithm, which makes the whole process a bit easier to understand. The junction tree algorithm can be thought of as a template for organizing the computations one needs to develop these recursions from scratch. Had it been around in the 1960's we would have immediately discovered all smoothing approaches to the Kalman filter, as well as all useful algorithms for HMM's and multilayer HMM's!

I will also show results of tracking and EM in the cases of accelerando, decelerando etc, because I didn't get to this point either, due to difficulties with the EM and with the use of noninformative priors (singular Fisher's information matrices) which have now been resolved.

Next I will begin the dual set of topics: what to do when you have a prior distribution over a segment time and wish to perform segmentation on the actual signal. I will introduce the basic problem, then discuss the various representations of distributions over segment times. Finally, I hope to complete coverage of the online (real-time) sequential test as well as show some interesting results and specializations from this test. However, I could certainly postpone discussion of these issues if there's anything else someone in the group wants to present.

What remains to be covered later, in a third seminar, are strategies for integration, such that we can really do joint segmentation and rhythm modeling using the actual signal. As well, we can discuss Bayesian offline segmentation methods. A byproduct of the Bayesian offline methods is we now have a useful nonparametric method for representing the posterior probability of change anywhere in the signal, based on "information in local windows". The original purpose was to obtain a general, "empirical Bayes" prior for offline segmentation based on quasi-perceptual criteria. This "changeogram" method I think relates to some of the issues in Clint's seminar.

If anyone is interested, Chapter 2 of the report covers all of these topics except the integration.

Also, please feel free to ask questions during any time of the presentation.
–Harvey

Two upcoming ASA talks on woodwind modeling

Gary Scavone <gary at ccrma> (CCRMA)

Tonehole radiation directivity measurements
Gary P. Scavone and Matti Karjalainen

Abstract:
Measurements have been conducted in an anechoic chamber for a comparison to current acoustic theory with regard to radiation directivity from a cylindrical pipe with toneholes. Several difficulties arise in measurements of this sort, including (1) the generation of sufficient driving signal strength at the pipe input for pickup by an external microphone; (2) external source-to-pickup isolation; (3) measurement contamination due to nonlinear driver distortion. Time-delay spectrometry using an exponentially swept sine signal was employed to determine impulse responses at points external to the experimental air column. This technique is effective in clearly isolating nonlinear artifacts from the desired linear system response along the time axis, thus allowing the use of a strong driving signal without fear of nonlinear distortion. The experimental air column was positioned through a wall conduit into the anechoic chamber such that the driver and pipe input were located outside the chamber while the open pipe end and tone holes were inside the chamber, effectively isolating the source from the pickup. Measured results are compared to both transmission-line, frequency–domain simulations, as well as time–domain digital waveguide calculations.

Time-domain synthesis of conical bore instruments
Gary P. Scavone

Abstract:
A series of approaches are presented for discrete time–domain synthesis of conical bore instrument sounds. This study uses a simple clarinet-like system, involving a memoryless nonlinear "reed" function and a distributed cylindrical air column model, as a point of departure for subsequent development. The generation of steady, self-sustained oscillations in such a system is complicated by properties inherent to truncated conical waveguides. Alternative methods include a structure equivalent to Benade's "cylindrical saxophone," with two separate parametrization schemes. Finally, a more general "virtual" model is presented that is capable of synthesizing both half-wave and quarter-wave resonators. This structure provides a rich variety of possible sounds and offers an interesting perspective on conical waveguides.

Spring Quarter 2000-2001

3D Positional Audio

Rachel Wilkinson <rwilkinson at ausim3d.com> (Ausim)

The first DSP Seminar this quarter will be this Friday, April 6th at 3:15 at the CCRMA Ballroom.

When we hear a sound, how do we tell where it's coming from? We often instinctively turn to pinpoint the origin of a new sound, but how do we know which way to turn? Moreover, why is the capability of determining source direction diminished or lost when we listen to audio over headphones? What's missing from the headphone signal?

The answers lie in how our brains process the sound waves caught by our ears, to determine where sounds we are hearing originate. Studies of these processes have revealed certain distinct features or cues that the brain has learned to use to localize sounds. By understanding and intelligently recreating these cues, technologists can now synthesize surprisingly realistic virtual aural environments.

Physical and empirical modeling are technologies that can be used as bases for simulating natural sounds. With a good mathematical model of the physical or acoustic characteristics of the sound source, the environment through which it propagates, and the listener, flat sound sources can be altered to sound as if they had actually propagated through and interacted with a physical environment.

This presentation will explore new research into ways of creating realistic 3D sound simulations. We will discuss processing and delivery strategies as well as psychoacoustic factors. Ideas will be generated for future applications and areas of further research.

E. Rachel Wilkinson
Audio and Acoustics Applications Specialist
AuSIM, Inc.
http://www.audiosimulation.com

Introduction to Multirate, Polyphase, and Wavelet Filter Banks

Julius Smith <jos at ccrma> (CCRMA)

This two-hour talk contains tutorial introductions to

- multirate digital systems,
- perfect reconstruction filter banks,
- quadrature mirror filters,
- polyphase filter bank analysis,
- multi-input, multi-output allpass filters,
- paraunitary filter banks, and
- wavelet filter banks (particularly the dyadic wavelet filter bank).

Additionally, it is demonstrated how polyphase filterbank analysis naturally converts the "filter bank summation" (FBS) representation of the STFT to an "overlap add" (OLA) representation, while simultaneously showing that the STFT implements a perfect reconstruction filter bank (usually oversampled).

This talk is a synthesis of Music 420 lecture overheads with contributions from past teaching assistants Scott Levine and Harvey Thornburg. They can be perused (on campus) at http://ccrma/~jos/JFB/.

Linear prediction analysis of voice under the presence of sinusoidal interference

Aaron Hipple, Yi-Wen Liu, and Kyungsuk Pyun <hipple—jacobliu—kspyun at stanford> (EE)

We are interested in tackling the single channel sound source separation problem of voice and non-voice signals. An interesting task would be to separate singing from instrumental accompaniment, pianos or guitars for example. In that case, it is crucial to make estimation of the glottal source of the voice part in the presence of interfering sinusoids.

The focus of our ongoing research is to study the linear prediction analysis of voice and try to come of with new methods to separate voice and non-voice from a single channel mixture.

Particularly, we've worked on an adaptive linear prediction (LP) analysis framework that is based on the LMS algorithm. The adaptive algorithm is causal, and has the potential of following the statistics of the voice more closely. However, the estimation of the LP coefficients is fluctuating around the optimal solution due to the nature of the LMS algorithm.

The outline of the talk is as follows,

- Introduction to the sound source separation problem: can it ever be done? Does it need to be done? What are the benefits of solving it?
- Literature review of the methods: Computational Auditory scene analysis (CASA) versus blind source separation (BSS) and Indepedent component analysis(ICA)
- brief review of linear prediction analysis of speech/voice
- Experimental results of LP analysis of voice under the presence of sinusoidal interference
- Possible direction for sound source separation: statistical methods and heuristic methods

The Bayesian Approach to Segmentation and Rhythm Tracking

Harvey Thornburg <harv23 at ccrma> (EE)

Segmentation refers to the detection and estimation of abrupt change points. Rhythm tracking refers ultimately to the identification of tempo and meter, but in this context it refers more generally to the identification of any higher-level "structure" or "pattern" amongst change points. Therefore the name "rhythm tracking", despite being catchy, is somewhat over-optimistic as to the goals of this project. Nevertheless, it serves until a better one can be found.

A naive approach to rhythm tracking is to use the partial output of a segmentation algorithm (say, the list of change *times*) to feed a pattern recognition algorithm which learns the structure of this event stream. What "learning" means depends on the how the data is processed. In the online setting, which is easiest to develop, learning means to predict, i.e. to develop posterior distributions over future change times given those already observed.

There are a number of problems with the naive approach. First, it is open loop. We could, and should somehow "close the loop", i.e. use the tracker's output posterior distributions as priors for the segmentation. This requires a Bayesian framework for the segmentation itself, which will be the *specific* focus of this talk. There are really two motivations for developing Bayesian techniques:

1. For the specific problem of rhythm tracking, only a fraction of detected change times are likely to convey rhythmic information. As well as note onsets, there could be expressive timbral and/or pitch fluctuations, interfering signals, and so on, which either do not correspond to rhythmic information or do so less reliably than the onset information. Hence, it is desired that the underlying segmentation method learn to recognize specifically the changes corresponding to the desired structure, a phenomenon known as "stream filtering".
2. Independent of rhythm tracking, there is still the concern that we make efficient use of large data samples. When change times exhibit a regular pattern (i.e. anything that differs from Poisson) this means events in a certain location will give information about those far away. Since any practical (i.e. linear time) approximate change detection scheme uses only short windows with local information in the signal about a given change point, it is desired to communicate more global kinds of information, without resorting to large windows.

Second, structures based only on the change times do not seem to cope well with rests, subdivisions, and like phenomena. For instance, it is unlikely that a given list of interarrival times will simply repeat, so immediately we must cope with the problem of variable length cycles. Hence, lists of groups of rhythmic intervals is not a good intermediate representation. We have to be a bit more clever. A proposed solution, easily adapted to general frameworks for change detection, is to propagate information

about the "strength" of each change as well as the change time. Formally, we define a change point as the pair Tk,Ek, where Tk is the time of a possible change, and Ek is an indicator that the change actually occurred. Right now the "strength" is defined as the posterior likelihood P(Ek—Tk, Ek-1, Tk-1, Ek+1, Tk+1, Ek-2, Tk-2, Ek+2, Tk+2, ...). This definition probably needs refinement. This likelihood may be computed using familiar methods (i.e. subspace tests). Intuitively, P(Ek—Tk, ...) relates to information about dynamics, and also encodes the indications of "rests" (P(Ek—Tk, ...) ¡¡ 1). In this representation, encoding of temporal information becomes very easy (the "metronome" model), and all the crucial information gets encoded in the patterns of strengths.

During the talk, I will focus mostly on the Bayesian framework for *sequential* change detection when the model after change is unknown. Here the appropriate prior concerns the *location* of the next change point, as well as the prior probability the change has occured before time zero (I find one can just set this to zero, but it's needed to calculate expected cost). Our stopping rule minimizes an expected cost function involving the event of a false alarm, the total number of miss-alarms, and the number of samples observed. Minimizing the latter also minimizes detection delay. Practically, we must also know something apriori concerning the distribution of models after change. There seem to be two promising approaches: the marginalized approach where we average over the entire space of candidate models (we must have a prior distribution for these models), and the multimodel approach where we sample the alternative model space to get some "representative" set. The way I do this using the fewest models possible is to sample at the "poles" of significance shells. I have found multimodel sampling approaches to work better in practice, but a thorough analysis is still being worked out.

It seems the "strength" prior cannot be integrated here, but we can use it in subsequent "backtracking" using the Bayesian subspace test. That part needs more refinement and discussion.

Linear Prediction of Voice in the presence of Sinusoidal Interference

Yi-Wen Liu <jacobliu at stanford> (EE)

This presentation is about the ongoing research by Aaron Hipple, Kyunsuk Pyun, and Yi-Wen Liu for Stanford's EE373 adaptive signal processing series instructed by professor B. Widrow of the EE department.

We are interested in the sound source separation problem of voice and non-voice signals from single microphone mixtures. An interesting task would be to separate singing from instrumental accompaniment such as guitar or piano sounds.

Particularly, we've worked on an adaptive LP framework that is based on the Widrow's LMS algorithm. The adaptive algorithm is causal, and has the potential of following the statistics of the voice more closely. However, the estimation of the LP coefficients is fluctuating around the optimal solution due to the nature of the LMS algorithm.

The outline of the talk is as follows:

- Motivations: commercial ones and academic ones
- what is voice? What is that "voiceness" which makes a voice sound sound like voice?

- brief review of the adaptive linear combiner/predictor
- brief review of linear prediction analysis of speech/voice
- Comparison between LMS-based adaptive LP and windowing-based framewise LP

- Toward sound source separation: a LP front-end plus an ICA backend; simple experiments and results

The advantages of wavelets to FFT for audio analysis and how to evaluate octave-band filter bank and improve it

Kyungsuk Pyun <kspyun at stanford> (EE)

This is about 60-75 minutes talk, followed by Harvey Thornburg's 30-45 minutes talk.

As an extension to my CCRMA Open house lecture last Thursday on "A fast and efficient octave-band filter bank for audio analysis", I would like to talk about "the advantages of wavelet approach for audio analysis, compared to FFT based one and how to evaluate the performance of octave-band filter bank and improve it"

This talk will have 3 sub-parts;

1. The advantages of wavelets over FFT for audio signal processing

In image processing, wavelets recently dominated FFT so that JPEG2000 already chose wavelets as a stanford. But in audio processing, this has not happened yet. I would like to address this issue and discuss the advantage of wavelets over FFT and possibility of using it in audio signal processing, especially when a fast signal processing is needed in lowpower wireless device like portable MP3 player. Then I will talk about other researchers' work using wavelet method.

Another important question is which wavelet to use for audio signal. I would like to discuss why Gabor wavelet has the definite advantage for an audio signal.

2. How to measure the performance of the filter bank

To get a feature vector, filter bank output need to go through amplitude compression and DCT(dicrete cosine transform) to decorrelate the output. After front-end processing, popular distortion measure like L2 norm or LDA(linear disriminant analysis) as was used in Dr. Yoon Kim's thesis on NLP ceptrum need to be used. This measure will be compared to one from FFT based filter bank.

3. How to improve the performance

On my octave-band filter bank, the binary tree is expanded only to the low frequency band. If the energy of music signal is concentrated mainly on the high frequency band(i.e. signal from an instrument like piccolo), then the binary tree need to be expanded to the high frequency band.

One way to do this is to expand the tree in full and prune it adaptively depending on the input frequency characteristics. Dr. Naoki Saito proposed LDB(linear discriminant bases) method to do this on his Yale Ph.D. thesis on "Local feature extraction and its application using a library of bases" on Dec. 1994. I wil explain how the proposed algorithm works. Basically after full expansion of the tree, K principal subband components are selected, where K¡M (M is the number of subbands when the tree is full grown). The core of the algorithm is to use an additive measure like minimum entropy to prune the tree, which makes the algorithm runs O(n*logn). This is the same order of speed as the FFT algorithm. This approach reuses the same prototype lowpass and highpass filter for an efficiency.

Thank you and I hope to see you all,
kyungsuk(peter)

Automatic music style classification: towards the detection of perceptually similar music

Unjung Nam <unjung at ccrma> (Music)

This is a preliminary study on music classification system that tries to cluster two pieces of music with similar musical content and classifies them into two different genres according to their timbral information.

I will present an overview on the works in the field of audio/music information retrieval, a general procedure of building music classification system and discuss an experiment I conducted on my system.

Bandlimited Interpolation and Virtual Analog Synthesis

Julius Smith <jos at ccrma> (CCRMA)

At tomorrow's DSP seminar (3:15, Thursday, CCRMA Ballroom), I will present a tutorial introduction to bandlimited interpolation and, time permitting, virtual analog synthesis. The lecture overheads can be perused online at

http://ccrma/~jos/Interpolation/

and

http://ccrma/~jos/VirtualAnalog/.

In search of golden HRTFs

Klaus A. J. Riederer <Klaus.Riederer at hut.fi> (HUT)

(This talk was ultimately moved to the Hearing Seminar.)

For commercial applications head-related transfer functions (HRTFs) are typically greatly simplified for more real-time performance with less signal processing power. However, by definition HRTF is a measured response from a point in the free-field space to a point in the ear canal. Hence, it contains all the monaural and binaural cues needed for spatial hearing (for the particular sound incident and person). The essence is to understand in what extent are these auditory cues (perceptually) important, although matters underlying are complex. Author's wide-scaled analyses of carefully measured HRTFs show excellent system quality, high HRTF repeatability and effects of various factors to HRTFs, including ear plug type, experimenter's experience/carefulness, hairstyle and headpieces. The vast inter-person variance makes it difficult to find robust common trends or groups in HRTFs. A mathematical index of deviance helps to segregate atypical responses, hinting to possible artifacts and/or deviating anatomical structures. This measure also relates to quantitative analysis towards a generic HRTF model allowing a deeper understanding of (auditory) spatial hearing. Unfortunately, HRTFs are not sufficient to understand the true multisensory spatial hearing. Its multisensory interactions with vision, motion, bone conduction, tactile and vestibular sensing with their cognitive impacts need also to be considered in order to reveal the true enigma of spatial hearing. Elaborate multimodal perceptual experiments are thus being devised. Brain mechanisms are investigated by magnetoencephalography (MEG) measurements based on individual HRTFs. Results show differences in binaural neural processing between azimuth and elevation angles of 3-D sounds. Related publications available at www.lce.hut.fi/~kar/publications.html

Klaus A. J. Riederer
Helsinki University of Technology
Laboratory of Computational Engineering, Cognitive Science, and Technology
Laboratory of Acoustics and Audio Signal Processing
P.O. Box 9400, FIN-02015 HUT, Finland
E-mail: , URL: http://www.lce.hut.fi/~kar

Winter Quarter 2000-2001

Parameter Engine for the Synthesis Tool Kit (STK)

Bryan Cook <bacook at ccrma> (EE)

The first DSP Seminar of this quarter will be this Thursday, Jan 11th at 3:15 at the CCRMA Ballroom.

I will be giving an overview and demo of the software infrastructure I've been developing for the Synthesis Toolkit (STK) which I have dubbed the Parameter Engine.

Music synthesis algorithms, physical modeling algorithms in particular, are controlled by a large parameter space. Managing and exploring these parameters can often be tricky.

The Parameter Engine tries to address these problems by addressing these goals:

1. Easy registration of synthesis parameters
2. Easy mapping of external controls to internal parameters
3. Easy mapping of parameter ranges
4. Easy parameter editing, loading and saving.

In addition to the Parameter Engine demo, I will also discuss interfacing STK to LabVIEW and give an overview of UIUC's Vanilla Sound Server (VSS) platform which already has STK integrated.

-Bryan Cook
bacook at ccrma

An efficient octave-band wavelet filter bank for voice analysis

Kyungsuk Pyun <kspyun at stanford> (EE)

For speech recognition systems, information lost at the front-end stage is not recoverable, which makes front-end processing crucial in the whole system. In this talk, a new speech front-end feature extraction system using a 12 bandpass filter bank will be proposed to be used for low-power small computers.

The need to perform signal processing at extremely low power, such as wearable computers, motivates the study of FIR filters having few taps with small integer coefficient. For example, digital watches do not have floating point operations.

The proposed speech front-end wavelet filter bank consists of 12 bandpass filters covering a frequency range from 1Hz to 4KHz. There are two sets of filter banks. The first one is regular wavelet expansion Octave band filter banks. The second set of filters was implemented using same prototype filters except downsampled by 2/3 at the very beginning of tree structure. Anti-aliasing filters were used before processing. Because of this downsampling factor of 3, sampling rate is 16128Hz, which is divisible by 3 and close to 16KHz but not exactly equal. The signal is stored in a one dimensional buffer before processing, just enough to be processed by lowpass and highpass filters, which makes both filters symmetric noncausal filters. Index of buffer is divided as follows. At first every other samples are taken. This is 2 downsampled version. After eliminating this samples, again every other samples are taken among remainders. The process continues in this way. One way to implement bandpass filter can be implemented sinc * sinusoid. During upsampling of bandpass filter, frequency of sine goes down and width of sinc function increase. In frequency domain this is equivalent to the fact that both the center frequency of bandpass filter goes down and the bandwidth goes down. That is, filter becomes closer to lowpass filter with smaller bandwidth. What's unique in proposed approach is the organization of the C code, namely one loop going through the samples, performing just a few assignment statements per sample. Our algorithm run on the order of N. Also the filter coefficients are chosen so that only fast computer arithmetics like addition and shifting are used. Multiresolution approach of wavelet transform gives the fixed resolution for each band. In this approach it is about 6 samples/cycle at maximum sensitivity of each filter. In this talk, frequency responses of 12 filters will be shown and filter bank output for chirp signal, several speech signals.

Segmentation of Audio Based on Stationarity Measures

Harvey Thornburg <harv23 at ccrma> (EE)

Time: Thursday 1/18 3:15 pm

This DSP seminar I will give a research update. As the focus last time was on transient modeling, this time will concern segmentation. Time permitting, I will present several topics loosely connected to the problem of *maximum likelihood estimation* of segment locations.

The first possible topic is to outline the steps in an exact, rigorous justification of the segment-wise least squares criterion as an asymptotic maximum likelihood estimator (MLE) for the Gaussian AR model, in the case where the innovations variance is unknown and not assumed either to be constant. Though the results are what one would expect using vague, intuitive arguments about initial conditions "washing out", it is nice at some point to have a proof. I follow a presentation indicated in Kay's "recursive MLE" paper, but, instead of leaving the "dirty work" up to a frustrating, hundreds of page presentation in Grenander and Szego, I derive and use the Gohberg-Semencul formula for the inverse of a finite Toeplitz matrix followed by a straightforward use of Gray's theory of asymptotically equivalent matrices. Besides clarifying the presentation, the "Gohberg-Semencul" approach seems to indicate systematic estimation biases for finite data, and suggests how to correct them.

The next topic is to formulate the idea of "subspace tests", used to test whether a parameter lies in a linear subspace, based on the MLE. Unlike conventional order selection criteria, it is possible to specify a confidence level (probability of false alarm) for these tests, which becomes useful in adjusting the "sensitivity" of the segmentation. There is a straightforward application to change detection, which to my knowledge appears for the first time. A "nested subspace test" is presented that recursively computes N subspace tests in an N-dimensional parameter space using only N multiplies and divides, involving no explicit matrix inversions. All of this leads to a new segmentation algorithm, which is much simpler and more accurate than those previously presented. I will discuss this and show results on speech.

Finally, I will review the $O(N^2)$ exact dynamic programming solution for the MLE of segment locations, then discuss something I am currently working on which is how to adapt Blake's "weak string" approach to get an approximate solution in much less time. The exact MLE is far too slow to use in practice, so this investigation is important. Basically, the MLE objective is relaxed by permitting, but heavily penalizing, parameter variations within a segment. The resultant cost objective may be transformed, by a change of variable, into a sum of quasiconvex functions. If the quasiconvex functions were convex the overall objective would be convex and then we could use gradient descent. The idea, which Blake calls GNC ("gradated nonconvexity") is to design related convex functions and morph them over time to the original functions, meanwhile running gradient descent. The discussion concerns exactly how the GNC can be adapted for the MLE cost objective at hand.

In addition, tomorrow night I will give a concise overview of the entire transient modeling project at the Motorola/AES meeting. I hope to see everyone at both.

Thanks,
–Harvey

Multimodel Coding

Julius Smith <jos at ccrma> (CCRMA)

At tomorrow's DSP seminar (3:15, Thursday, CCRMA Ballroom), I will present overheads from my talk last week at the Institute for Mathematics and its Applications (IMA). This was a workshop consisting primarily of image compression researchers, so my remarks were aimed in that direction.

Abstract: Musical Signal Models for Audio Rendering

This talk will summarize several lines of research going on in the field of music/audio signal processing that are applicable to audio compression and data reduction. While the techniques were motivated originally by the desire for realistic "virtual musical instruments" (including the human voice), the resulting rendering models may be efficiently transmitted to a receiver as a "specialized decoder" in software form which is then "played" by a very sparse data stream. In most cases, there is also a straightforward tradeoff between rendering quality and computational expense at the receiver. Since all models are built from well behaved audio signal processing components, the distortion at low complexity levels tends to be of a high level character, sounding more like a different instrument or performance than a distorted waveform.

Natural Voice Synthesis

Vicky Lu <vickylu at ccrma> (EE)

In previous quarters, I have developed an singing synthesis model for the non-nasal voiced sound. The associated analysis/re-synthesis procedure has shown that this model can generate naturally sounding voices with different voice qualities ranged from the press, normal to breathy mode. However, in addition to the analysis/resynthesis ability, a good synthesizer should provide easy controls on the model parameters. By exploring the correlation between the model parameters, we could reduce the complexity on the control interface.

To generate sustained voiced sound via the synthesis model proposed, we need to specify pitch contour, glottal excitation strength (Ee) contour, glottal shape (Rd) contour and vocal tract filter. For the breathy mode, we also need to specify model parameters for pitch synchronous Gaussian amplitude modulated noise. These parameters are often correlated to each other.

In the non-breathy phonation mode, it turns out that Rd and Ee are highly correlated. We can induce Rd parameter from Ee, hence, we only need to specify Ee contour which is more intuitive. In the breathy phonation mode, high-passed NHR (noise to harmonics ratio) is introduced to represent the aspiration strength. NHR is also highly correlated to Rd. Therefore, NHR is used to control the degree of breathiness in the breathy mode.

For sustained voiced sound synthesis, the dynamic fluctuations of the model parameters are the key factors for generating human-like sound. In this talk, I will also talk about the methods to describe the fluctuations for the purpose of synthesis.

Pitch Tracking Applied to South Indian Classical Music

Arvindh Krishnaswamy <arvindh at stanford> (EE)

I will first give an introduction to South Indian classical music, also known as "Carnatic Music." I will briefly cover fundamental concepts like 'Ragas,' 'Talas' and 'Gamakas' and how they are used in practice.

I will also talk about pitch intervals used in Carnatic music and some experiments and measurements I did myself with my violin and electronic tambura. I will present my pitch tracking scheme, and explain some differences with traditional methods.

I will also comment on my ideas on how to approach the Swara/Gamaka pattern identification problem, and also how to use the concept of Ragas to classify pieces of music.

Pattern recognition approaches to invert a violin physical model

Stefania Serafin <serafin at ccrma> (Music)

In this talk I will introduce different techniques to estimate the input parameters for a bowed string physical model based on pattern recognition. The aim is to invert the model, which means to obtain the correct input parameters corresponding to the right hand of the performer, i.e., the bow force, bow velocity and bow position from the output of a synthetic model and from recordings made on real instruments.

The ultimate goal is to be able to extract meaningful parameters from recordings made on real instruments and map those parameters to synthetic models, in order to obtain expressive synthesis.

All the approaches consist of assigning an a-priori probability to each set of data, and calculating the likelihood of each target spectrum using a set of training data for different values of bow velocity, bow position and bow force.

I will present the current results of this ongoing research made in collaboration with Julius Smith and Harvey Thornburg.

Abstracts: Autumn Quarter 2000-2001

Adaptive Additive Synthesis for Nonstationary Sounds

Dr. Axel Röbel <roebel at ccrma> (Berlin Technical University)

I will present a new approach to additive synthesis which is intended to model transient and nonstationary sounds by means of superposition of partials with segment-wise linear amplitude and frequency evolution. Due to the nonstationarity of the partials, STFT analysis is not applicable (there exists actual research on estimating slope parameters from STFT but I do not rely on this), but the parameters are optimized by an adaptive approach. The higher flexibility of the model requires regularization to favor reasonable (desired) solutions.

The talk will have the following sections:

- Description of the problem, the model, and partial tracking
- Optimization objective
- Analytical investigation of the tracking performance in case of simple partials and block wise analysis
- Regularization terms that are used to prevent non unique and undesired solutions
- Sound examples
- Description of planned work at ccrma:
  - physical interpretation of the results obtained so far
  - improve frequency resolution without decreasing parameter flexibility
  - regularization for physical meaningful results
  - handling of non-physical partials

Seminar overheads are available on the Web at

http://ccrma/~roebel/addsyn/

Time-variant modal decomposition and uses in transient modeling

Harvey Thornburg <harv23 at ccrma> (EE)

My talk focuses on a "piecewise-slowly-nonstationary" parametric model for tracking multiple sinusoids. The ultimate goal of the model is an analysis-synthesis method for transient sounds, where the assumption of local stationarity (so important to frequency-domain methods) fails to hold.

In the piecewise model, we assume sinusoidal parameters vary smoothly except on a finite set of points where abrupt changes occur. After we segment according to boundaries of abrupt change, we constrain AR parameters to vary according to a basis where the number of functions indicates the degree of smoothness. Then order selection criteria may be applied, to choose both the number of components and their degree of smoothness. To extract sinusoidal parameters from this "AR" representation, we use Kamen's time-variant cascade decomposition, then show exactly how this can be used for the modal decomposition. The modal decomposition is followed by a process of state estimation for amplitude/phase information.

However, the modal decomposition does not guarantee that "smooth" or "slow" AR parameter variations result in this same behavior for the sinusoids. Fortunately, I can show in the case of convergent AR parameters, corresponding sinusoidal parameters will have limit cycles which trace out a convex path, thus enabling a feedback stabilization of Kamen's recursions which yields at least an asymptotically smooth trajectory for the sinusoids. However, the stabilization fails in the case of real-valued modes.

In the talk, I review the entire system (both segmentation and modeling tasks), then focus on the specifics of time-variant mode extraction as it relates to the problem of smoothness.

A statistical pattern recognition approach to speech endpoints detection

Yi-Wen Liu <jacobliu at stanford> (EE)

ABSTRACT: The talk will summarize what I've done during the past summer as an intern at a startup company called VerbalTek, which is devoted to speech recognition applications on smaller platforms such as cellphones and PDAs. The problem we were facing was the noise-robustness issue of speech recognition, and it turned out that accuracy of speech endpoints detection (EPD) was what really mattered. Therefore, we developed a new EPD algorithm that is based on statistical pattern recognition. The key idea of our algorithm is to apply "N-weatherperson theory" on speech/non-speech classification of each frame. The performance of this algorithm was shown to be better than conventional heuristic methods under various types of noisy environments.

OUTLINE

Motivation: noise-robustness of EPD

Method and Theory

- Features selection
- Modeling the feature statistics
- Application of "N-weatherperson theory" to speech/non-speech classification
- EPD based on classification

Experiments and results

- EPD under realistic noisy environment
- Recognition rates at various levels of SNR

Music DSP in the Ivy League

Perry Cook <prc at cs.princeton.edu> (Princeton CS/Music)

Perry Cook slipped silently away from Stanford CCRMA in December of 1995. Since then he's reappeared occasionally like a bad penny, but mostly he's been roaming the woods of Central New Jersey, walking the same paths that were trod by Albert Einstein, John Von Neumann, Alvin Turing, Brook Shields, and countless lyme-tick infested deer. This talk will be a shocking expose on what Perry has been doing these last five years. Much of it will even have to do with DSP. Topics to include: Performance interfaces for computer music, analysis and synthesis of sound effects, Music Information Retrieval tools and systems, and Banded Waveguides (for stiff structures).

Turbulence noise modeling for the breathy singing voices

Vicky Lu <vickylu at ccrma> (CCRMA/EE)

The focus of my research is to improve the naturalness of the vowel tone quality via glottal excitation modeling. Based on the source-filter type synthesis model, I have proposed to use the LF-model for the glottal wave shape in conjunction with pitch-synchronous, amplitude-modulated Gaussian noise, which adds the flow-induced turbulence component to the glottal excitation. The turbulence noise model is especially important for perceiving breathy sound quality. Recently, I have been seeking for a better physically self-explained turbulence model.

In this talk, I will first give a review on the acoustic theory of the turbulence noise for human voice. I will also give a summary of the existing turbulence models in the speech literature. Based on these studies, a series pressure source is adopted to model the turbulence source. I will derive the pressure source from the estimated glottal excitation and vocal tract shape parameters, which are obtained via SUMT algorithm in my previous study. The manigtude and the spectrum of pressure source are then estimated afterwards. Since the area of the glottis is varying while the turbulence is generated, the turbulence noise is non-stationary, hence, the spectrum of the turbulence noise is time-varying. There are two alternatives to model the spectrum of the pressure source. One is to use a fix spectrum shaping filter which is the average spectrum. The other alternative is to use a time-varying ERB model. I will describe these two models in this talk.

Abstracts: Spring Quarter 1999-2000

Detection and Modeling of Transients in Analysis-Synthesis Frameworks

Harvey Thornburg <harv23 at ccrma> (EE)

My talk for the CCRMA DSP seminar (4/7 at 3:15 pm) consists of an update for my research on the detection and modeling of transients in analysis-synthesis frameworks. By contrast to some previous approaches, the focus is on parametric models and information-theoretic detection criteria. There are essentially six problem areas:

1. Develop model structure(s) for extensions to rapid decay and fast local variations
2. Detect abrupt changes, estimate locations.
3. Identify appropriate model structures for each region
4. Specify family of transformations
5. Stitch regions back together (in resynthesis)
6. Residual processing (nonparametric, allow artifacts)

For this talk, my focus is #(2), the joint detection and temporal estmation of abrupt change points (segment boundaries). I will discuss the current state of online and offline methods, then detail a new solution which gives an efficient computational structure for integrating features of both methods.

–Harvey

Glottal Excitation Modeling for the Singing Voice

Vicky Lu <vickylu at ccrma> (CCRMA/EE)

Naturalness of the sound quality is essential for the singing synthesis. Since 95% in singing is voiced sound, the focus of this study is to improve the naturalness of the vowel tone quality via the glottal excitation modeling.

In addition to the abilities of flexible pitch and volume control, the desired excitation model is expected to be capable of changing the voice quality so that the voice quality can be modified from laryngealized (pressed) to normal to breathy phonation.

To trade off between the complexity of the modeling and the analysis procedure to acquire the model parameters, we propose to use the source-filter type synthesis model, based on a simplified human voice production system. The source-filter model decomposes the human voice production system into three linear systems: glottal source, vocal tract and radiation. The radiation is simplified as a differencing filter. The vocal tract filter is assumed all-poled for non-nasal sound. The glottal source and the radiation are then combined as the derivative glottal wave. We shall call it as the glottal excitation.

The effort is then to estimate the vocal tract filter parameter and glottal excitation to mimic the desired singing vowels. The de-convolution of the vocal tract filter and glottal excitation was developed via the convex optimization technique. Through this de-convolution, one could obtain the vocal tract filter parameters and the glottal excitation waveform.

The next step is to build the glottal excitation synthesis model after the vocal tract filter has been found. Since both the wave-shape of the glottal excitation and the aspiration noise are important factors to change the breathiness of the sound quality, the glottal excitation is considered as two parts: one is the smoothed quasi-periodic derivative glottal wave and the other one is the glottis noise (turbulence noise). These two components are separated via wavelet decomposition of the glottal excitation waveform from de-convolution. The coarse wave-shape of the smoothed derivative wave is intended to be modeled via the LF model. The noise part is then roughly modeled as a pitch synchronous amplitude modulated Gaussian noise with larger power around the glottal closure instants. Due to the model mismatch and the source-tract interaction, a model for the residual fine structure of the smoothed derivative glottal wave becomes necessary. (This part is still under the survey. Inputs are very welcomed.)

In this talk, the de-convolution and glottal excitation modeling results for both synthetic data and real baritone recordings will be shown. I will focus on the synthetic data simulation results and discuss the impact of aspiration noise, GCI detection error and source-filter.

Wave Digital Filters (WDFs) Applied to Physical Modeling Problems

Stefan Bilbao <bilbao at ccrma> (EE)

Hi everyone,

I'll be presenting material this week on wave digital filters (WDFs) applied to physical modelling problems. There will be two installments:

Tuesday, 3:15, CCRMA Ballroom: Quick recap of wave digital filtering principles (i.e. classical network theoretic principles, wave variables, scattering junctions, the bilinear transform etc.), then the beginning of an overview of the generalization of WDFs to the multi-D case, where they can be applied to the numerical integration of systems of partial differential equations.

Friday, 3:15, CCRMA Ballroom: Sequel of the overview, then applications of the technique towards simulating the vibration of stiff beams, plates and cylindrical shells, as well as some matlab simulations.

Stefan

—

Adaptive Additive Synthesis for Nonstationary Sounds

Dr. Axel Röbel <roebel at ccrma> (Berlin Technical University)

I will present a new approach to additive synthesis which is intended to model transient and nonstationary sounds by means of superposition of partials with segment-wise linear amplitude and frequency evolution. Due to the nonstationarity of the partials, STFT analysis is not applicable (there exists actual research on estimating slope parameters from STFT but I do not rely on this), but the parameters are optimized by an adaptive approach. The higher flexibility of the model requires regularization to favor reasonable (desired) solutions.

The talk will have the following sections:

- Description of the problem, the model, and partial tracking
- Optimization objective
- Analytical investigation of the tracking performance in case of simple partials and block wise analysis
- Regularization terms that are used to prevent non unique and undesired solutions
- Sound examples
- Description of planned work at ccrma:
  - physical interpretation of the results obtained so far
  - improve frequency resolution without decreasing parameter flexibility
  - regularization for physical meaningful results
  - handling of non-physical partials

Seminar overheads are available on the Web at

http://ccrma/~roebel/addsyn/

Impact of String Stiffness on Virtual Bowed Strings

Stefania Serafin <serafin at ccrma> (Music)

Recent work in the field of bowed-string synthesis has produced a real-time instrument which, despite its simplicity, is able to reproduce most of the phenomena that appear in real instruments. Current research consists of improving this model, including refinements made possible by the improvement of hardware technology and the development of efficient digital signal processing algorithms.

In particular, I will focus on a technique used to model string *stiffness*, whose main effect is to disperse the sharp corners that characterize the ideal Helmholtz motion.

Another improvement concerns modeling the high frequency violin body resonances using a 3D waveguide mesh. The playability of all the different features of the model will be examined.

Application of Wavelet and other orthonormal transforms to a Perceptual Audio Codec with Switching Windows

Yi-Wen Liu <jacobliu at stanford> (EE)

Perceptual audio codecs using long windows produce an artifact called "pre-echo" when coding at 64kbps or lower bitrates. To eliminate pre-echoes, we need to switch to shorter windows whenever a sudden increase in signal intensity is detected. However, there are some difficulties to do perceptual coding when we switch to shorter windows:

- at lower bit rates ($\leq$64kbps), the bit reservoir is not big enough even to code the side information
- due to duality of Fourier transform, frequency resolution is low when we use short windows. Hence, it doesn't fully make sense to apply frequency domain masking.

I will present a hybrid audio coder that does perceptual coding at long windows and wavelet coding at short windows. The coder aims to provide better sound quality than a plain perceptual coder does at 64kbps. Following is the outline of my talk:

- brief review about the following keywords: perceptual audio coding, window switching, wavelet decomposition, etc.
- pre-echoes and motivation of window-switching
- attack detection algorithm
- the criterions for perfect reconstruction
- wavelet approximation as a problem of minimizing the L2 norm of the error
- experimental results: comparison of the error energy using different wavelet basis
- discussion about future work

Abstracts: Spring Quarter 1998-1999

Speech Feature based on Bark Frequency Warping — the Non-Uniform Linear Prediction (NLP) Cepstrum

Yoon Kim <yoonie at ccrma> (CCRMA/EE)

For this week's CCRMA DSP seminar:

I'll be presenting some recent results on the NLP technique for speech/speaker recognition. For those of you who missed my talk in the winter quarter, NLP cepstrum is a speech feature based on Bark frequency warping and multiple all-pole (LP) modeling. It has shown to effectively suppress speaker-dependent information while preserving the linguistic component of speech segments.

Also, I'll be talking about solving spectral estimation / filter design problems using convex optimization. The above promises to be a good basis for achieving speaker normalization, i.e. a process of eliminating speaker-dependent characteristics of speech uttered by multiple speakers.

Hope to see you !

Yoon

Time: 3:15pm Thursday, Apr 8
Place: CCRMA Ballroom

Joint Estimation of Glottal Source and Vocal Tract Filter from Speech Signal and some Fundamental Frequency Detection Methods

Vicky Lu <vickylu at ccrma> (CCRMA/EE)

For this week's seminar, I will start my talk at 3:15 and Prof. Risset will give his talk at 4:30.

I will give a update for joint estimation of glottal source parameters and vocal tract filter from speech pressure signal that I talked about last quarter. I will give some sound results that reconstructed from the estimated parameters. Finally, I will have a summary on what I leanrned from these simulations for voice synthesis.

Hope to see u there and give me suggestions!

Thanks!!

Vicky

Multiscale Modeling of Audio Textures

Harvey Thornburg <harv23 at ccrma> (EE)

Today I will present a joint project with Caroline Traube, concerning multiscale modeling of audio textures. The ultimate goal is to obtain a generalization of the 1/f processes rich enough to handle highly complex, nonstationary sounds such as rainfall, crackling fire, and outputs of turbulent/chaotic systems. An important aspect of such a generalization is that models be easily adapted to the hypothesized correlation structure of real signals, such that new classes of transformations may be developed. The model should tell us exactly *to what extent* and *how* this scale-correlation structure exists.

For today, the pertinent framework is that of compression. Based on the work of Albert Beneviste, jointly with Michele Basseville, Alan Willsky and many other researchers, we have developed a compression algorithm designed to measure and capture the scale-correlations structure of 1-D signals. The algorithm is analogous to LPC for time series, using the more general Levinson-Schur (or direct-PARCOR) decomposition since it is impractical to set up Wiener-Hopf equations. I tried (I think successfully) to redevelop the method for time series, which I will present carefully as part of the background.

Two important developments from the previously published work, which are *vital* in the transition from theory to implementation are 1) direct elimination of redundancy structure from multiscale Levinson recursions and 2) a correction for finite data length. Up to half the residuals get lost when assuming infinite data. This fact is independent of the data size. 2) is especially important because it makes the technique implementable on real signals, and it allows for the kind of windowing one might use to adapt to changing scale-correlation structure in real signals.

Time permitting I will present a brief history of the field.

Hope to see everyone there, and apologies for the lateness of the notice.

–Harvey

Impact of Torsion Waves and Friction Characteristics on the Playability of Virtual Bowed Strings

Stefania Serafin and Julius Smith <serafin at ccrmaand jos at ccrma> (CCRMA)

Tomorrow's DSP seminar will be presented by Julius and me. This week we will talk about physical modeling of bowed string instruments. In particular, we focus on their "playability", examining in which zones of a multidimensional parameter spaces "good tone" is produced. We focus on the influence of the torsional waves and on the shape of the friction curve. The aim is to analyze which elements of bowed string instruments are fundamental for bowed string synthesizers, and which can be neglected, to reduce computational cost. An application that will be shown is an implementation in Max/MSP of a bowed string model that can be controlled in real time using a graphical tablet. The playability region obtained is mapped to the parameters given by the tablet, to make easier to "play" the model.

Hope to see you there.

Stefania

Time: 3:15pm Thursday, May 6.
Place: CCRMA Ballroom

How to Calculate Constant Q Profiles and the Derivation of Toroidal Models of Inter-Key Relations

Hendrik Purwins <purwins at ccrma> (Berlin Technical University)

This talk will be a more technical follow up to the overview at the CCRMA Colloquium 2 weeks ago. I will show three different derivations of toroidal models of inter-key relations (ToMIR):

- geometric explanation,
- emergence in the Self Organizing Feature Map (SOM, Kohonen 82) trained by Shepard cadences previously processed by an auditory model,
- emergence in a SOM trained by averaged constant Q (cq-)profiles of Chopin's preludes recorded by Cortot (1933/34).

In method (3) the cq-profiles are 12-dimensional vectors, each component referring to a pitch class. They can be employed to represent keys. Cq-profiles are calculated with the constant Q filter bank (Brown & Puckette 92). This filter bank gives equal resolution for all regions in the logarithmic frequency domain. The cq-profiles are also used for key recognition, and for investigating pitch use in Bach, Chopin, and Alkan.

IN MORE DETAIL:

The constant Q transform is employed for deriving the ToMIR in (iii), and for key analysis. The algorithm is efficiently implemented by calculating the kernel of the constant Q filters in the frequency domain in advance and exploiting the sparsity of that kernel. The 12-dimensional cq-profile is the concentrated spectral information supplied by the constant Q transform. Each component of the profile indicates the strength of one pitch class. A cq-profile is calculated by summing up all values (bins) of the constant Q transform that belong to the same pitch class. Cq-profiles have the following advantages: (a) The 12 components correspond to values in probe tone ratings. In a psychological probe tone experiment a quantitative description of a key is derived from rating the different pitch classes within the tonal context of that key (Krumhansl & Kessler 82). (b) Calculation is possible in real-time. (c) Stability is obtained with respect to sound quality. (d) Cq-profiles are transposable.

A cq-hierarchy is a sequence of 24 cq-profiles, each representing a different major or minor key. Cq-hierarchies enable the study of pitch use, and automatic key recognition. We derive cq-hierarchies from cq-profiles of cadential chord progressions played on the piano or averaged cq-profiles of Prelude [& Fugue] cycles in all keys by Bach, Chopin, Alkan, and Scriabin. We investigate to what degree cq-hierarchies depend on (1) musical interpretation, (2) the recorded instrument (piano or harpsichord), and (3) the piece of music. Results given by different interpretations by Cortot and Pogorelich (Chopin op. 28), and Gould and Feinberg (Bach 'Well-Tempered Clavier', Book I) and different instruments indicate a negligible influence of (1) and (2) on cq-hierarchies. Piano cq-profiles can be directly compared to the sum of a probe tone rating and its transposition

at the fifth, weighted by the strength of the third partial of piano tones. Cq-hierarchies display small but significant differences between Bach (WTC I + II) and romantic music (Chopin op. 28, Alkan op. 31) in the use of perfect fifths, major sixths and major sevenths in major, and fourths, major sixths, and major and minor sevenths in minor.

In order to apply cq-hierarchies to key recognition, a special distance measure is introduced: the 'fuzzy distance' is the modified Euclidean distance, which weights each component according to its inverse variance. In the calculation of key distances we account for the fact that components of the cq-profile, which represent the key are consistently stable from one piece to another. The strength of the minor sixths and the major and minor sevenths in minor varies according to the use of natural, melodic, and harmonic minor scales. The fuzzy distance between keys emphasizes stable values in the cq-profiles, whereas varied components (e.g. minor sixths, major and minor sevenths in minor) are de-emphasized. There are no errors in key assignment for all pieces of Bach's WTC I.

Real-Time Chord Recognition of Musical Sound: A System Using Common Lisp Music

Takuya Fujishima <fujishim at ccrma> (CCRMA)

I describe a realtime software system which recognizes musical chords from the input signal of musical sounds. I designed an algorithm and implemented it in Common Lisp Music/Realtime environment. The system runs on Silicon Graphics O2 and, partly on linux. Experiments verified that the system succeeded in correctly identifying chords even on orchestral sounds. Details follow.

Traditionally, musical chord recognition is approached as a combination of polyphonic transcription to identify the individual notes and following symbolic inference to determine the chord. This approach often suffers from recognition errors at the first stage. They result from various kinds of noise and overlaps of harmonic components of individual notes in the spectrum, and are difficult to avoid.

My chord detection algorithm also has two stages, but each of them differs from the traditional one. At the first stage, it does not identify the individual notes. Instead, it generates a "pitch class profile (PCP)." PCP is an intensity set of the twelve chromatic pitch classes. It automatically unifies various dispositions of a single chord class. At the second stage, my algorithm does not do symbolic, rule-based inference. Instead, it does numerical processing on PCP to find the most likely root and chord type. Here I have tried two numerical methods for comparison. I have also introduced a couple of heuristics to the system so as to improve the accuracy.

I have implemented my algorithm using Common Lisp Music Realtime extension (CLM/RT). CLM is a flexible set of COMMON-LISP-based synthesis and signal processing tools. CLM/RT is a realtime extension of CLM. CLM/RT compiles the LISP code to binary executables to realize the realtime sound processing and the graphical user interface. Using CLM/RT, my system takes in the audio signal from either a microphone or a sound file, does chord recognition, and displays the result in a realtime manner. The system runs on Silicon Graphics O2.

I used various sound sources including electronic keyboards and recordings on CDs. As for CDs, I chose tonal harmonic music from the classical and popular repertoires. In the experiment, I input the sound to the system through a line-in, or a soundfile. The system displayed the recognition results. Then I compared them with the chord names which I put to the music materials in advance to evaluate the accuracy of the results.

The results of my experiments showed that my system could recognize triadic harmonic events, and to some extent more complex chords such as sevenths and ninths, at the signal level. The system put correct chord names even to complicated orchestral sounds. Each numerical method tried seemed to have its advantage and disadvantage, in terms of computational cost and accuracy. The heuristics introduced did improve the accuracy in total.

A Boundary Element Model for the Cylindrical Acoustic Tube

Professor Shyh-Kang Jeng <skjeng at ccrma> (Taiwan National University)

For next week's DSP seminar from 3:15 PM, I will give a talk about the propagation and scattering of acoustic wave in a cylindrical tube, with an emphasis on the higher-order modes. I will show two or three short movies for the propagation of fundamental and higher-order modes. I will also show the (colored) field distribution of the waveguide modes. The excitation and the equivalent filters of the higher modes will be addressed. The conventional concept of impedance discontinuity will be also shown as an approximation. Some recent results of the BEM analysis of horn-like transitions between two cylindrical tubes will be given. One of which verifies the calculations by David Berners using a simpler, but less general approach. The possible applications of these higher-order mode analyses to improve the digital waveguide model will be also proposed.

Please come and give me suggestions.

Sincerely,

Shyh-Kang Jeng

A Robust Constant Q Spectrum for Polyphonic Pitch Tracking

Professor Benjamin Blankertz <blanker at uni-muenster.de> (Institute for Mathematical Logic University of Muenster, Germany)

In this talk a method of improving the accuracy of frequency determination using the phase vocoder technique is presented. This gives the possibility to calculate a reliable constant Q spectrum even at a high time resolution, which is e.g. required for polyphonic pitch tracking. Employing the constant Q spectrum reduces such tasks to a pattern recognition problem.

The derivation of an explicit form of the error function of the phase vocoder in terms of frequency and phase of the sinusoidal components of the input signal allows the calculation of parameters for a 3-term Blackman window that minimize the expected error in the vicinity of a given frequency. This allows a robost and precise frequency determination.

A Hybrid Waveguide Model of the Transverse Flute

Mark Bartsch <bartscma at flyernet.udayton.edu> (Visitor, University of Dayton / EE)

Recent years have seen substantial improvements in the modeling and synthesis of jet-reed instruments such as flutes and organ pipes. Developments in the modeling of woodwinds using so- called digital waveguides have allowed the efficient acoustical modeling of the main body and toneholes of the instrument. Further, a new and more complete model of the jet's behavior in the presence of the edge and the resonator has been formulated for recorderlike instruments in [M. P. Verge, A. Hirschberg, and R. Causse, J. Acoust. Soc. Am.101, 2925–2939 (1997)]. A new simulation model of the transverse flute is presented which combines the contributions of digital waveguide modeling and the new source model. This new model is defined almost entirely by physical parameters (such as dimensions of the instrument) rather than by the arbitrary adjustment parameters often employed. The model is evaluated by comparing the effects of certain parameters on the model's operation with their effects on the performance of an actual flute. The model is further evaluated for its tuning characteristics by comparing its frequencies of oscillation with the sounding frequencies of a simple flute with matched physical parameters. [Work supported by the University of Dayton Honors Program.]

Abstracts: Winter Quarter 1998-1999

Joint Estimation of Glottal Source and Vocal Tract Filter from Speech Signal and some Fundamental Frequency Detection Methods

Vicky Lu <vickylu at ccrma> (CCRMA/EE)

For this week's seminar, I will talk about joint estimation of glottal source and vocal tract filter from speech signal and some fundamental frequency detection methods.

The goal is to build a voice model which is flexible enough for voice modification and interpolation between different states, and yet simple enough such that we can estimate the model parameters to reproduce a known speech signal. To ease the estimation, the source filter type voice model is chosen. At this talk, I will review the source filter type speech model and some methods to estimate glottal source and vocal tract filter from speech signal. A primitive model and its associated estimation procedure (not complete yet ...) will be introduced.

Fundamental frequency estimation is often necessary for pitch synchronize joint estimation of glottal source and vocal tract filter. The fundamental frequency is defined as a glottal cycle here. I will talk about 2 frequency domain methods (cepstrum based method and Harmonic product spectrum) and 2 time domain methods (AMDF and SRFD).

See u there !

vicky

Speech Feature based on Bark Frequency Warping — the Non-Uniform Linear Prediction (NLP) Cepstrum

Yoon Kim <yoonie at ccrma> (CCRMA/EE)

In statistically based speech recognition systems, choosing a feature that captures the essential linguistic properties of speech while suppressing other acoustic details is crucial. This could be more appreciated by the fact that the performance of the recognition system is bounded by the amount of linguistically-relevant information extracted from the raw speech waveform. Information lost at the feature extraction stage can never be recovered during the recognition process.

Some researchers have tried to convey the perceptual importance in such features by warping the spectrum to resemble the auditory spectrum. One example is the Perceptual Linear Prediction (PLP) method proposed by Hermansky, where a perceptually motivated filterbank is used to warp the spectrum, followed by scaling and compression of the spectrum. While the PLP provides a good representation of the speech waveform, its ability to model peaks of the speech spectrum – formants – depends on the fine structure of the FFT spectrum. This for instance could hinder the process of modeling formants of female speech through filterbank analysis, since there are fewer harmonic peaks under a formant region than in the male case. Also, various processing schemes used require memory, table-lookup procedure and/or interpolation, which might be computationally inefficient.

I will talk about a new method of obtaining parameters from speech that is based on frequency warping of the vocal-tract spectrum. The warping is achieved by using the Bark Bilinear Transform (BBT) on a uniform frequency grid to generate a grid that incorporates the non-uniform resolution properties of the human ear. Experimental results showing the superior performance of the NLP cepstrum with respect to conventional speech features will be presented.

Transaural Stereo and HRTF Modeling Review

Julius Smith <jos at ccrma> (CCRMA)

I will present some interesting reading I did recently on the subject of 3D sound and "transaural stereo". Consider the following questions:

- What is transaural stereo exactly? Does it really work? How well?
- What is the Cooper-Bauck "shuffle structure" for cross-talk cancellation? How does it improve on the original Schroeder-Atal structure?
- What is "diffuse field equalization" exactly?
- How long is an HRTF in time, and how much can it be shortened?
- How are HRTF filters designed these days?
- Is it worthwhile to implement FIR HRTFs using the FFT, or is time-domain convolution faster?
- What are some of the main patents in 3D auditory display and transaural stereo?

If you don't know the answers to any of the above questions, you might be interested in coming!

Julius

A Perceptually Based Audio Signal Model with Application to Scalable Compression

Tony Verma <verma at furthur.stanford> (EE)

This week i'll be giving a first run of my orals talk. As such comments, suggestions, and questions will be greatly appreciated.

A Perceptually Based Audio Signal Model with Application to Scalable Compression

Abstract (which is a bit long...)

Audio delivery in network environments such as the Internet where bandwidth is not guaranteed, packet loss is common and where users connect to the network at various data rates demands scalable compression techniques. Scalability allows each user to receive the best possible audio quality given the current network condition. In addition, because the separation principle for source and channel coding does not apply to lossy packet networks, an audio source coding technique that explicitly considers channel characteristics is desirable. These goals can be achieved by using a higher level description for audio than the actual waveform. This talk will focus on a method for extracting meaningful parameters from general digital audio signals that takes into account the way humans perceive sound; moreover, application of this parametric model to scalable audio compression will be discussed.

The model consists of three major components: sines, transients and noise. These underlying signal components are found during the analysis stage of the model. Quantizing and compressing the resulting model parameters allows for efficient storage and transmission of the original audio signal. The talk will cover enhancements made to current sine models. These enhancements allow explicit perceptual information to be included in the sinusoidal model. In addition, a novel transient modeling technique will be covered.

The three part model provides an efficient, flexible and perceptually accurate representation for audio signals. It therefore is appropriate for scalable compression over lossy packet networks. The efficiency of the model ensures high compression ratios. Flexibility simultaneously allows scalability and robustness to channel characteristics such as packet losses because subsets of model parameters represent the original signal with varying degrees of fidelity. Perceptual accuracy ensures that parameter subsets reasonably represent the original signal while the complete parameter set represents the original exactly in a perceptual sense.

Most current techniques for audio compression (e.g., MPEG audio layer 3 and AAC, Real Audio's G2, etc.), use a subband decomposition in conjunction with psychoacoustic models to compress the actual audio waveform itself. No model of the signal is assumed. These compression techniques have been very successful for targeted fixed bit rates; however, they cannot be scaled in large steps without severe loss in quality. This is evident in the case of Real Audio where a database will store many versions of an audio signal at various bitrates (e.g., 92Kbps, 64Kbps, 32Kbps, 20Kbps, and

16Kbps) and quality. Because using an underlying model allows meaningful subsets of parameters to describe the original signal, one compressed bitstream (e.g., 96Kbps) can be stored. Embedded within this bitstream are lower bitrate versions (e.g., 64Kbps, 32Kbps, 20Kbps, and 16Kbps) that can be easily extracted. Sound demos of the audio compression scheme will be played.

Hope to see you there!

-Tony

Musical instrument mechanics

Jim Woodhouse <jw12 at eng.cam.ac.uk> (University of Cambridge)

An informal overview of interesting mechanical vibration problems associated with musical instruments. The talk will range over tuned percussion, the musical saw, and the design of the violin body.

Bowed-string modeling and simulation

Jim Woodhouse <jw12 at eng.cam.ac.uk> (University of Cambridge)

Following on the previous colloquium, this talk will present a more detailed investigation of modeling the physical action of a bowed string, including the effects of string torsional properties and bending stiffness and the finite width of the ribbon of bow hair. Recent work will also be described in which the physics of rosin has been investigated. In terms of detailed physics, although not necessarily in terms of effectiveness for musical simulation, the model for rosin friction which has always been used in the past will be shown to be quite wrong.

Autumn Quarter 1998-1999

1. **09/24:** Jyri Huopeniemi (Visiting Researcher): HRTF filter design

Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design

(to be presented at the AES 105th Convention, San Francisco, Sept. 26-30,1998)

Jyri Huopaniemi
Nokia Research Center
Helsinki, Finland
jyri.huopaniemi at research.nokia.com
(work carried out while the author was a visiting researcher at CCRMA, Stanford University)

In this presentation, the problem of modeling head-related transfer functions (HRTFs) is addressed. Traditionally, HRTFs are approximated in real-time applications using minimum-phase reconstruction and various digital filter design techniques, yielding FIR or IIR structures. In this work, binaural auditory modeling has been applied to HRTF filter design analysis, and design methods have been compared from the auditory perception point of view. This paper presents applicable perceptually valid smoothing and filter design techniques and discusses listening test results for localization and timbre degradation using individualized HRTFs.

2. **10/08:** Scott Levine (EE) thesis defense preview (real thing 10/16)

Audio Representations for Data Compression and Compressed Domain Processing
(practice run for Scott's PhD/EE thesis defense)

Scott Levine <scottl at ccrma>
CCRMA/EE

In the world of digital audio processing, one usually has the choice of performing modifications on the raw audio signal, or data compressing the audio signal. But, performing modifications on a data compressed audio signal has proved difficult in the past. This thesis provides a new representation of audio signals that allows for both very low bit rate audio data compression and high quality compressed domain processing and modifications. In this context, processing possibilities are time-scale and pitch-scale modifications.

This new audio representation segments the audio into separate sinusoidal, transients, and noise signals. During determined attack transients regions, the audio is modeled by well established transform coding techniques. During the remaining non-transient regions of the input, the audio is modeled by a mixture of multiresolution sinusoidal modeling and noise modeling. Careful phase locking techniques at the time boundaries between the sines and transients allow for seamless transitions between representations. By separating the audio into three individual representations, each can be efficiently and perceptually quantized. Demos of audio compression to 32 kbps (22:1) and time-scale modification between 50% to 200% will be played.

3. **10/15**

Acoustic Stability of a Cylinder with Conical Cap

Julius Smith <jos at ccrma> (CCRMA)

This talk presents a proof of stability for a cylindrical acoustic tube terminated with a conical cap. The reason this takes some work is that, as is well known, the reflection and transmission transfer functions at the cylinder-cone junction are all *unstable* one-pole filters. Another

strange manifestation of this acoustic configuration is that the reflection transfer functions for pressure waves at the junction converge to -1 (the transmittances converge to zero), while from basic physical reasoning the reflectance of the conical cap as a whole must converge to +1. It is shown that this "paradox" is associated with *two* pole-zero cancellations at dc in the conical cap reflectance, as seen from the cylinder.

4. **10/22**

Estimation of model parameters for the two-mass glottal model

Vicky Lu <vickylu at ccrma> (CCRMA/EE)

The objective of this work is to build a better glottal source model for the singing synthesizer. To obtain a higher quality of voiced sound, glottal source modeling is very promising. In the speech literature, a two-mass model has been used for quite some time. One advantage of the two-mass model is that it considers the source-tract interaction automatically, and this is important for the high-pitched voice. Moreover, the model is not too complicated for computation. Hence, a two-mass model is chosen for my singing synthesizer.

However, choosing proper system parameters of the two mass model for different pitches is not trivial. This is the problem I addressed during my stay at NTT in Japan (Sep. 2nd to Oct. 13th). I assume that I can obtain the glottal flow data from the measurements of oral flow. Using this glottal flow data, the system states (displacements and velocities of the two masses, damping of the masses, and relative coupling stiffness) are then obtained using the Extended Kalman Filter (EKF).

In my talk, I will introduce the two-mass model I used. I will also show some preliminary simulation results and discuss the estimation formulation. A videotape introducing the experimental procedure will be shown.

5. **10/29**

Investigation of Perceptually Relevant Features for Speech Processing

Yoon Kim <yoonie at ccrma> (CCRMA/EE)

Choosing an optimal feature set for speech is essential in analyzing, recognizing and synthesizing speech. Also, in the problem of speaker normalization, the choice of parameters for the analysis model becomes crucial. Traditionally, researchers in the speech recognition field have used the weighted cepstrum as the feature for describing the acoustic aspects of speech. Cepstrum is often derived from LPC coeffcients by a simple transformation formula.

Some researchers have tried to convey the perceptual importance in such features by warping the spectrum to resemble the auditory spectrum. They found that the formant structure of the speech spectrum could be captured well using significantly less number of parameters than the normal LPC analysis. Also, Hermansky showed that the warped spectrum agrees well with the results in psychoacoustics, such as the effective second formant, and the 3.5 bark spectral-peak integration theory.

Smith and Abel introduced methods to derive the optimal conformal mapping for warping the spectrum to resemble the bark spectrum. (Bark Biliear Transform). The transformation itself is an all-pass type transfer function, thus mapping the unit circle onto itself. For a rational transfer function, the bark bilinear transform (BBT) preserves the order of the polynomials in the numerator and the denominator. Thus, we can obtain a warped cepstrum by taking the BBT of the LPC predictor polynomial, then converting it into the cepstrum (LPC-BBT Cepstrum).

In this talk which serves as a report of research in progress, I'll compare the performance in terms of vowel separability between the LPC-BBT Cepstrum and conventional methods. I'll also address the issue of computational complexity in obtaining these features. Finally, I will talk about future plans and applicable problems.

6. **11/05:** Peter Lindener (VR): TI 326x00 DSP applications

7. **11/12**

Sampling Rate Estimation Techniques in the Discretization of Nonlinear Systems

Harvey Thornburg <harv23 at ccrma> (CCRMA/EE)

Nonlinearities are prevalent in acoustic and circuit modeling. Since nonlinear elements tend to expand the bandwidth of their input, direct implementation in a digital system results in aliasing. A multirate system can be used to reduce aliasing at additional computational cost. The problem then becomes one of choosing an appropriate sampling rate at which to run the nonlinear element.

To this end, one must first be able to estimate the output spectrum given certain "worst-case" statistics of the input, (i.e. power,bandwidth). Many useful nonlinearities are memoryless; they are often hard (nonanalytic), or defined by an implicit relation. A technique based on iterated convolutions and Taylor series is extensible to general Volterra systems; however, hard elements cannot be dealt with and much complexity is required for the implicit cases. I will present a new technique for the memoryless case based on decomposition into a set of FM operators, and show how the frequency-domain computations are simplified when brought into the "amplitude domain". This technique is valid for any bounded nonlinear map.

Areas of application include soft saturation (sigmoid) elements and quantization of narrowband signals. I will show a completely general method for variably soft sigmoid construction which supersedes methods previously discussed (such as p-circle methods), and present sampling rate estimations for different measures of softness over a range of overdrive settings.

Classical theories model quantization effects as additive broadband noise. If this "noise" is truly uncorrelated with itself and the input signal, aliasing of the quantization error will be imperceptible. I will attempt to show, using spectral estimation techniques, for what bit depths the classical assumptions break down, and to what degree quantization and aliasing couple in the digitization of a narrowband signal.

8. **11/19:** Craig Sapp (CCRMA): Optical Gesture Recognition

9. **12/03**

Sinusoidal and Transient Modeling Using Frame-Based Perceptually Weighted Matching Pursuits

Tony Verma <verma at furthur.stanford> (CIS/EE)

The talk will focus on a well known subject at CCRMA: sinusoidal modeling. I'll first talk about a method of sinusoidal modeling that explicitly takes into account the psychoacoustics of human hearing using extensions to an analysis-by-synthesis technique known as matching pursuits. An efficient transient model that uses sinusoidal modeling in a rotated space will then be discussed. Current noise models fit well with the sine and transient models forming a flexible compact model for many audio signals. The three part model coherently captures salient features of an audio signal which is important for meaningful modifications such as time-scaling and pitch-shifting, exmples of which will be played.