2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

October 18-21, 2009
Mohonk Mountain House
New Paltz, NY, USA

IEEE

IEEE Signal Processing Society

IEEE Catalog Number: CFP09AUD
ISBN: 978-1-4244-3679-8
ISSN: 1931-1168

ORGANIZING COMMITTEE

General Co-Chair

Jacob Benesty
Université du Québec, INRS-ÉMT,
Canada
benesty@emt.inrs.ca

General Co-Chair

Tomas Gaensler
mh acoustics
USA
tfg@mhacoustics.com

Program Co-Chair

Jingdong Chen
WeVoice, Inc.
USA
jingdongchen@ieee.org

Program Co-Chair

Yiteng (Arden) Huang
WeVoice, Inc.
USA
arden_huang@ieee.org

Finance Chair

Michael S. Brandstein
Information Systems Technology Group
MIT Lincoln Lab, USA
msb@ll.mit.edu

Publications Chair

Eric Diethorn
mh acoustics,
USA
ejd@mhacoustics.com

Local Arrangements Chair

Heinz Teutsch
Avaya Labs Research
USA
teutsch@research.avayalabs.com

Publicity Chair

Sofiène Affes
Université du Québec, INRS-ÉMT,
Canada
affes@emt.inrs.ca

Far-East Liaison

Shoji Makino
University of Tsukuba,
Japan
maki@tara.tsukuba.ac.jp

GENERAL CHAIR'S INTRODUCTION

Dear WASPAA 2009 Attendee,

As general co-chairs of the organizing committee, we welcome you to the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)! The WASPAA 2009 committee has been organizing this workshop since the fall of 2007. Our hope is that you find this year's workshop program as interesting and inspiring as those in years past, and that you have opportunity to meet and socialize with friends and professional colleagues. As previous WASPAA attendants know, Mohonk Mountain House is a unique venue that provides an extraordinary environment for fruitful discussions, friendly arguments, and professional networking.

Just like any workshop or conference, WASPAA has a limited number of accepted papers. What distinguishes WASPAA from other meetings is that the number of attendees is also limited. Both WASPAA 2007 and 2009 reached the maximum number of attendees the venue can accommodate, which is about 145 participants. As happy as we are that WASPAA is a highly regarded professional meeting, it has been very hard for the committee to turn down a fairly large number of people interested in the workshop because of space limitations. That said, you as a presenter or attendee should enjoy the fact that you belong to the crowd that "made it" this year.

This year's program has been long in the making. After all reviewers' hard work, we could only accept 89 papers out of 181 submissions. Unfortunately, there was still a fairly large group of high-quality papers that had to be rejected. Accepted papers represent a broad spectrum of what is being researched and developed in the areas of music, speech, acoustics, and audio signal processing. The organizing committee has high expectations of many of the papers that will be presented during the course of the workshop.

WASPAA 2009 has the pleasure to present three keynote speakers from areas of music and audio processing, human speech processing recognition, and speech and audio coding. Dr. Barry Blesser, Blesser Associates, is a distinguished researcher and pioneer in the digital audio revolution. Prof. Jont Allen, from University of Illinois, Urbana-Champaign, has a deep knowledge and understanding of auditory modeling and processing. Last, but not least, we have Prof. Bastiaan Kleijn, from KTH, Stockholm, who has repeatedly surprised his research community with keen insights and new ideas in the field of audio coding. Please help us welcome these distinguished speakers who are expected to set the tone and raise the perception of this workshop to new heights.

Jacob Benesty and Tomas Gaensler
General Co-chairs WASPAA 2009

PROGRAM CONTENTS

WORKSHOP SCHEDULE — OVERVIEW

	Sunday, October 18	Monday, October 19	Tuesday, October 20	Wednesday, October 21
7:00 am 8:00 am		***Breakfast*** West Dining Room	***Breakfast*** West Dining Room	***Breakfast*** West Dining Room
8:00 am 8:45 am		***Keynote*** Barry Blesser Conference House	***Keynote*** Jont B. Allen Conference House	***Keynote*** Bastiaan Kleijn Conference House
8:45 am 10:25 am		***Lecture Session LM1*** Music Perception, Analysis & Synthesis Conference House	***Lecture Session LT1*** Signal Enhancement, Dereverberation, and Echo Control Conference House	***Lecture Session LW*** Speech and Audio Coding Conference House
10:25 am 10:40 am		***Coffee Break*** East Dining Room	***Coffee Break*** East Dining Room	***Coffee Break*** East Dining Room
10:40 am 12:20 pm		***Poster Session PM*** Music Processing and Source Separation East Dining Room	***Poster Session PT*** Speech Enhancement, Multichannel Audio Processing & Hearing Aids East Dining Room	***Poster Session PW*** Spatial Sound Perception & Reproduction, Speech & Audio Coding East Dining Room
12:20 pm 4:00 pm		***Lunch/Afternoon Break*** West Dining Room	***Lunch/Afternoon Break*** West Dining Room	***Lunch/Closing*** West Dining Room
4:00 pm 6:00 pm	***Registration*** Hospitality Room	***Lecture Session LM2*** Microphone Array Source Separation Conference House	***Lecture Session LT2*** Spatial Sound Perception Analysis & Reproduction Conference House	
6:00 pm 8:00 pm	***Dinner and Opening*** West Dining Room	***Dinner*** West Dining Room	***Dinner*** West Dining Room	
8:00 pm 10:00 pm	***Reception and Cocktails*** Sponsored by mh acoustics West Dining Room	***Demo Session DM and Cocktails*** Sponsored by RIM-Blackberry West Dining Room	***Demo Session DT and Cocktails*** West Dining Room	

KEYNOTE ADDRESS — MONDAY 8:00 AM – 8:45 AM

Signal Processing Not Spoken Here: The Human Experience of Aural Space

Barry Blesser
Blesser Associates, USA

The human experience of sound and the signal processing of its parameters are often only weakly related, each with its own language and concepts. On the one hand, physical acoustics follows the laws of nature and signal processing follows the rules of mathematics. On the other hand, cultural acoustics depends on the life styles of individuals, who can flexibly select a cognitive strategy based on mood, personality, and situation. These differences illustrate the dichotomy between the phenomenology of sensory experience and the predictive formalism of the scientific method. Specifically, the concepts of causation, quantification, and consistency exist in the language of signal processing but only weakly exist in cultural acoustics.

During one period in our history, spatial acoustics were viewed as undesirable noise; during other periods, spatial acoustics were viewed as sensory texture that gave each space a signature personality. For the first few years after the Boston Symphony Hall was built, it was considered to be a musical disaster. It is now considered to be one of the worlds ten best musical spaces.

Even though the acoustical properties of a space can strongly influence the emotions and behavior of the inhabitants, that influence is not readily measurable in controlled laboratory environments. The Perceptual Uncertainty Principles states that strong phenomena in real life may not be observable in the laboratory, and conversely, consistent laboratory results may not be applicable to real life. The paradigm of every discipline is like a filter that allows some aspects of a phenomenon to be observed while hiding and distorting others. Sensory anthropologists observe human behavior in natural settings, while perceptual scientists observe constrained behavior in sterile environments with very simple stimuli.

Consider a large physical space like a concert hall or cathedral. The physical acoustics of such spaces are so complex that they can never be accurately described because thermal waves make them time-varying. Simulated spaces are only approximations because there are no computers that can process sonic signals in three dimensions at the required temporal and spatial bandwidths of the human auditory system. Creating spatial illusions is possible but such illusions are often not consistent among individuals and sound sources. Moreover, illusions take advantage of evolutionary artifacts of our brain substrates, but there is no formal method for discovering such artifacts. For example, in-head localization of a sound source when listening with stereophonic headphones is such an example of an artifact. The perceptual difference between the decay of a guitar string and the reverberation of a concert hall is also a similar artifact. Artificial reverberation is actually an aural illusion of a space.

Modern neurophysiology research has clearly demonstrated that our brain retains its plasticity forever. We are how we live. A music conductor has better localization ability than the average individual. Those with weak vision can form an image of a space by listening. Spectral artifacts of a signal processing algorithm may initially be inaudible but become very loud as a user spends more time listening. Musical enthusiasts who attend classical music in real concert halls may become very sensitive to the aural influence of seat selection, while those that listen by headphones to modern music may enjoy the fact that the simulated space are artistic contradictions that could never be built.

Biosketch

Barry Blesser received his S.B., S.M. and Ph. D. degrees in electrical engineering from M.I.T. in 1964, 1965, and 1969 respectively, specializing in communications. Following graduation, he spend 9 years as an Associate Professor of Electrical Engineering and Computer Science at M.I.T. where he was active in research on cognitive perception and signal processing. Since that time, Dr. Blesser has pursued a consulting career in a variety of professional disciplines, including medical, character recognition, control systems, digital audio, and technology management.

Dr. Blesser was one of the pioneers in the digital audio revolution. He was the inventor of the first commercial digital audio reverberation system, the EMT-250; he helped start Lexicon, and he has published numerous papers on many aspects of digital audio. He was the system architect of Orbans Audicy digital audio workstation. In 1980, Dr. Blesser was president of the AES and has been elected member of the board of Governors several time. For the last 25 years he has served as a reviewer on the editorial board of the Journal of the AES, and is currently Consulting Technical Editor for the journal. He was the recipient of the AES Silver, Bronze, and Governors medals and received the award for best published paper several times. He is a Fellow of the AES.

Recently, Dr. Blesser has expanded his professional activities to include technology management and risk engineering. He has held the position of Director of Engineering and Chief Technology Officer in several major audio companies. He was a founder of Pencept, a company that had specialized in recognizing hand print characters. Dr. Blesser has several patents on artificial reverberation. He help found a new start up company, 25-Seven, Inc, specializing in temporal processing of audio signals for the broadcast industry.

In 2007, MIT Press published his book, Spaces Speak, Are You Listening? Experiencing Aural Architecture, which considers the auditory experience of space from a wide range of perspectives. The Associates of College and Research Libraries award it the Outstanding Academic Title for 2007.

Lecture Session — LM1: Monday 8:45 am – 10:25 am

Music Perception, Analysis, and Synthesis

Session Chair:
Daniel P.W. Ellis
Columbia University, USA

LM1–1 *8:45 am – 9:05 am*

Automatic Gain and Fader Control For Live Mixing

Enrique Perez Gonzalez and Joshua Reiss, *Queen Mary University of London, UK*

A cross-adaptive mixing device has been developed for the purpose of optimizing the gain levels of a live audio mixture. The method aims to achieve optimal mixing levels by optimizing the ratios between the loudness of each individual input channel and the overall loudness contained in a stereo mix. In order to evaluate the amount of loudness of each channel in real-time, accumulative statistical measurements were performed. The system uses a cross-adaptive algorithm to map the loudness indicators to the channel gain values. The system has applications in automatic mixing of live music, live mixing of game audio, and studio recording post-production.

LM1–2 *9:05 am – 9:25 am*

Acoustic Coupling in Multi-Dimensional Finite Difference Schemes for Physically Modeled Voice Synthesis

Matt Speed, Damian Murphy, and David M. Howard, *University of York, UK*

Finite-difference time domain approximation of the wave equation has been shown to provide a good approximation to acoustic wave propagation in multiple dimensions. Two dimensional models are often assumed to be an adequate compromise between the poor geometrical representation afforded by one-dimensional models and the additional computational loading incurred by three-dimensional modeling. This paper demonstrates the validity of multi-dimensional finite-difference schemes for obtaining accurate frequency responses of uniform cylinders, then compares simulations of coupled cylindrical resonators carried out in different dimensionalities. It is found that two-dimensional models exhibit erroneous low-frequency resonant behaviour where large discontinuities in acoustic admittance are presented by the geometry under simulation.

LM1–3 *9:25 am – 9:45 am*

Chord Recognition Using Measures of Fit, Chord Templates and Filtering Methods

Laurent Oudre, Yves Grenier, and Cédric Févotte, *CNRS LTCI, France*

This paper presents an efficient method for chord transcription of music signals. A succession of chroma vectors is calculated from the signal in order to extract the musical content of the piece over time. We introduce a set of chord templates for several types of chords (major, minor, seventh,...): different chord models taking into account one or more harmonics of the notes of the chord are considered. In order to fit the chroma vectors to the chord templates, we analytically calculate a scale parameter. The detected chord over a frame is the one minimizing a measure of fit between a rescaled chroma vector and the chord templates. Several popular measures in the probability and signal processing field are considered for our task. In order to take into account the time-persistence, we perform a post-processing filtering over the recognition criteria which quickly smooths the results and corrects random errors. The system is evaluated on the 13 Beatles albums and compared to state-of-art. Results show that our method outperforms state-of-art but more importantly is significantly faster.

LM1–4 *9:45 am – 10:05 am*

Unifying Semantic and Content-Based Approaches for Retrieval of Environmental Sounds

Gordon Wichern, Harvey Thornburg, and Andreas Spanias, *Arizona State University, USA*

Creating a database of user-contributed recordings allows sounds to be linked not only by the semantic tags and labels applied to them, but also to other sounds with similar acoustic characteristics. Of paramount importance in navigating these databases are the problems of retrieving similar sounds using text or sound-based queries, and automatically annotating unlabeled sounds. We propose an integrated system, which can be used for text-based retrieval of unlabeled audio, content-based query-by-example, and automatic annotation. Our system builds an ontology where sounds are connected to each other based on a measure of perceptual similarity, while words and sounds are connected by optimizing link weights given user preference data. Results on a freely available database of environmental sounds contributed and labeled by non-expert users, demonstrate effective average precision scores for both the text-based retrieval and annotation tasks.

LM1–5 *10:05 am – 10:25 am*

A Novel Framework for Recognizing Phonemes of Singing Voice in Polyphonic Music

Hiromasa Fujihara and Masataka Goto, *National Institute of Advanced Science and Technology, Japan*, Hiroshi G. Okuno, *Kyoto University, Japan*

A novel method is described that can be used to recognize the phoneme of a singing voice (vocal) in polyphonic music. Though we focus on the voiced phoneme in this paper, this method is design to concurrently recognize other elements of a singing voice such as fundamental frequency and singer. Thus, this method is considered to be a new framework for recognizing a singing voice in polyphonic music. Our method stochastically models a mixture of a singing voice and other instrumental sounds without segregating the singing voice. It can also estimate a reliable spectral envelope by estimating it from many harmonic structures with various fundamental frequencies (F0s). The results of phoneme recognition experiments with 10 popular-music

songs by 6 singers showed that our method improves the recognition accuracy by 8.7 points and achieves a 20.0% decrease in error rate.

10:25 am – 10:40 am
Coffee Break (East Dining Room)

Poster Session — PM: Monday 10:40 am – 12:20 pm

Music Processing & Source Separation

Session Chair:
Eric Diethorn
mh acoustics, USA

PM–1 Domain Decomposition Method for the Digital Waveguide Mesh

Moonseok Kim and Gary P. Scavone, *McGill University, Canada*

The digital waveguide mesh (DWM) is a discrete-time numerical method for modeling the propagation of traveling waves in multidimensional mechanical and acoustic systems. Despite the fact that the DWM is not as computationally efficient as a 1-D digital waveguide, it is still widely used for sound synthesis of musical instruments and for acoustical modeling of rooms because of the simplicity of the implementation. However, large-scale realization of the digital waveguide mesh is not adequate for many simulations because of its relatively high and direction-dependent dispersion error. The influence of dispersion error can be reduced by using a denser mesh structure though with extra computational costs. This paper presents a method for efficiently interconnecting rectangular DWM sub-domains of different mesh density. The method requires only small overlapped buffer regions to be added for each sub-domain. This allows the selective use of higher and lower density grids in a single simulation based on spatially-dependent dispersion error criteria.

PM–2 Synthesis of Guitar by Digital Waveguides: Modeling the Plectrum in the Physical Interaction of the Player with the Instrument

François Germain, *École Polytechnique, France*, Gianpaolo Evangelista, *Linköping University, Sweden*

In this paper, we provide a model of the plectrum, or guitar pick, for use in physically inspired sound synthesis. The model draws from the mechanics of beams. The profile of the plectrum is computed in real time based on its interaction with the string, which depends on the movement impressed by the player and the equilibrium of dynamical forces. A condition for the release of the string is derived, which allows to drive the digital waveguide simulating the string to the proper state at release time. The acoustic results are excellent, as it can be verified in the sound examples provided.

PM–3 Fast Bayesian NMF Algorithms Enforcing Harmonicity and Temporal Continuity in Polyphonic Music Transcription

Nancy Bertin and Roland Badeau, *CNRS-LTCI, France*, Emmanuel Vincent, *IRISA-INRIA, France*

This article presents theoretical and experimental results about constrained non-negative matrix factorization (NMF) in a Bayesian framework, enforcing both spectral harmonicity and temporal continuity. We exhibit fast multiplicative update rules to perform the decomposition, which are then applied to perform polyphonic piano music transcription. This approach is shown to outperform other standard NMF-based transcription systems, providing a meaningful mid-level representation of the data.

PM–4 Computing Predominant Local Periodicity Information in Music Recordings

Peter Grosche and Meinard Müller, *Saarland University and MPI Informatik, Germany*

The periodic structure of musical events plays a crucial role in the perception of tempo as well as the sensation of note changes and onsets. In this paper, we introduce a novel function that reveals the predominant local periodicity (PLP) in music recordings. Here, our main idea is to estimate for each time position a periodicity kernel that best explains the local periodic nature of previously extracted note onset information and then to accumulate all these kernels to yield a single function. This function, which is also referred to as PLP curve, reveals musically meaningful periodicity information even for non-percussive music with soft and blurred note onsets. Such information is useful not only for stabilizing note onset detection but also for beat tracking and tempo estimation in the case of music with changing tempo.

PM–5 Acoustic Topic Model for Audio Information Retrieval

Samuel Kim and Shrikanth Narayanan, *University of Southern California, USA*, Shiva Sundaram, *TU Berlin, Germany*

A new algorithm for content-based audio information retrieval is introduced in this work. Assuming that there exist hidden acoustic topics and each audio clip is a mixture of those acoustic topics, we proposed a topic model that learns a probability distribution over a set of hidden topics of a given audio clip in an unsupervised manner. We use the Latent Dirichlet Allocation (LDA) method for the topic model, and introduce the notion of acoustic words for supporting modeling within this framework. In audio description classification tasks using Support Vector Machine (SVM) on the BBC database, the proposed acoustic topic model shows promising results by outperforming the Latent Perceptual Indexing (LPI) method in classifying onomatopoeia descriptions and semantic descriptions.

PM–6 Guided Harmonic Sinusoid Estimation in a Multi-Pitch Environment

Christine Smit and Daniel P.W. Ellis, *Columbia University, USA*

We describe an algorithm to accurately estimate the fundamental frequency of harmonic sinusoids in a mixed voice recording environment using an aligned electronic score as a guide. Taking the pitch tracking results on individual voices prior to mixing as ground truth, we are able estimate the pitch of individual voices in a 4-part piece to within 50 cents of the correct pitch more than 90% of the time.

PM–7 Improving MIDI-Audio Alignment with Acoustic Features

Johanna Devaney, *McGill University, Canada*, Michael I. Mandel and Daniel P.W. Ellis, *Columbia University, USA*

This paper describes a technique to improve the accuracy of dynamic time warping-based MIDI-audio alignment. The technique implements a hidden Markov model that uses aperiodicity and power estimates from the signal as observations and the results of a dynamic time warping alignment as a prior. In addition to improving the overall alignment, this technique also identifies the transient and steady-state sections of the note. This information is important for describing various aspects of a musical performance, including both pitch and rhythmic aspects.

PM–8 Note Detection with Dynamic Bayesian Networks as a Postanalysis Step for NMF-Based Multiple Pitch Estimation Techniques

Stanisław A. Raczyński, Nobutaka Ono, and Shigeki Sagayama, *University of Tokyo, Japan*

In this paper we present a method for detecting note events in the note activity matrix obtained with Nonnegative Matrix Factorization, currently the most common method for multipitch analysis. Postprocessing of this matrix is usually neglected by other authors, who use a simple thresholding, often paired with additional heuristics. We propose a theoretically-grounded probabilistic model and obtain very promising results due to the fact that it encodes basic musicological information. The biggest advantage of our approach is that it can be extended without much effort to include various types of general information about musical signals, such as principles of tonality and rhythm.

PM–9 Multi-Voice Polyphonic Music Transcription Using Eigeninstruments

Graham Grindlay and Daniel P.W. Ellis, *Columbia University, USA*

We present a model-based approach to separating and transcribing single-channel, multi-instrument polyphonic music in a semi-blind fashion. Our system extends the non-negative matrix factorization (NMF) algorithm to incorporate constraints on the basis vectors of the solution. In the context of music transcription, this allows us to encode prior knowledge about the space of possible instrument models as a parametric subspace we term “eigeninstruments”. We evaluate our algorithm on several synthetic (MIDI) recordings containing different instrument mixtures. Averaged over both sources, we achieved a frame-level accuracy of over 68% on an excerpt of Pachelbel’s Canon arranged for doublebass and piano and 72% on a mixture of overlapping melodies played by flute and violin.

PM–10 Polyphonic Music Transcription Employing Max-Margin Classification of Spectrographic Features

Ren Gang, Mark F. Bocko, Dave Headlam, and Justin Lundberg, *University of Rochester, USA*

In this paper we present a transcription method for polyphonic music. The short time Fourier transform is used first to decompose an acoustic signal into sonic partials in a time-frequency representation. In general the segmented partials exhibit distinguishable features if they originate from different “voices” in the polyphonic mix. We define feature vectors and utilize a max-margin classification algorithm to produce classification labels to serve as grouping cues, i.e., to decide which partials should be assigned to each voice. These classification labels are then used in statistical optimal grouping decisions and confidence levels are assigned to each decision. This classification algorithm shows promising results for the musical source separation.

PM–11 Towards a Musical Beat Emphasis Function

Matthew E.P. Davies and Mark D. Plumbley, *Queen Mary University of London, UK*, Douglas Eck, *University of Montreal, Canada*

We present a new method for generating input features for musical audio beat tracking systems. To emphasise periodic structure we derive a weighted linear combination of sub-band onset detection functions driven a measure of sub-band beat strength. Results demonstrate improved performance over existing state of the art models, in particular for musical excerpts with a steady tempo.

PM–12 Towards Co-Channel Speaker Separation by 2-D Demodulation of Sectrograms

Tianyu T. Wang and Thomas F. Quatieri, *MIT Lincoln Laboratory, USA*

This paper explores a two-dimensional (2-D) processing approach for co-channel speaker separation of voiced speech. We analyze localized time-frequency regions of a narrowband spectrogram using 2-D Fourier transforms and propose a 2-D amplitude modulation model based on pitch information for single and multi-speaker content in each region. Our model maps harmonically-related speech content to concentrated entities in a transformed 2-D space, thereby motivating 2-D demodulation of the spectrogram for analysis/synthesis and speaker separation. Using a priori pitch estimates of individual speakers, we show through a quantitative evaluation: 1) utility of the model for representing speech content of a single speaker and 2) its feasibility for speaker separation. For the separation task, we also illustrate benefits of the model’s representation of pitch dynamics relative to a sinusoidal-based separation system.

PM–13 Separation by “Humming”: User-Guided Sound Extraction from Monophonic Mixtures

Paris Smaragdis, *Adobe Systems Inc., USA*, Gautham J. Mysore, *Stanford University, USA*

In this paper we present a novel approach for isolating and removing sounds from dense monophonic mixtures. The approach is user-based, and requires the presentation of a guide sound that mimics the desired target the user wishes to extract. The guide sound can be simply produced from a user by vocalizing or otherwise replicating the target sound marked for separation. Using that guide as a prior in a statistical sound mixtures model, we propose a methodology that allows us to efficiently extract complex structured sounds from dense mixtures.

PM–14 Semi-Blind Disjoint Non-Negative Matrix Factorization for Extracting Target Source from Single Channel Noisy Mixture

So-Young Jeong, Kyuhong Kim, Jae-Hoon Jeong and Kwang-Cheol Oh, *Samsung Electronics Co., Korea*

We present a semi-blind non-negative matrix factorization(NMF) approach to suppress interference noises from a single channel mixture signal. By enforcing a disjointness constraint into the NMF error criterion under the semi-blind denoising framework, it is possible to decompose the mixture spectrogram into target and noise components by minimizing overlaps in the time-frequency domain. Experimental results show that the proposed semi-blind disjoint NMF algorithm can significantly suppress non-stationary noise components in the noisy mixture.

PM–15 Improving Separation of Harmonic Sources with Iteration Estimation of Spatial Cues

Jinyu Han and Bryan Pardo, *Northwestern University, USA*

Recent work in source separation of two-channel mixtures has used spatial cues (cross-channel amplitude and phase difference coefficients) to estimate time-frequency masks for separating sources. As sources increasingly overlap in the time-frequency domain or the angle between sources decreases, these spatial cues become unreliable. We introduce a method to re-estimate the spatial cues for mixtures of harmonic sources. The newly estimated spatial cues are fed to the system to update each source estimate and the pitch estimate of each source. This iterative procedure is repeated until the difference between the current estimate of the spatial cues and the previous one is under a pre-set threshold. Results on a set of three-source mixtures of musical instruments show this approach significantly improves separation performance of two existing time-frequency masking systems.

PM–16 A Nonlocally Weighted Soft-Constrained Natural Gradient Algorithm for Blind Separation of Reverberant Speech

Jack Xin, Meng Yu, Yingyong Qi, Hsin-I Yang and Fan-Gang Zeng, *University of California at Irvine, USA*

A nonlocally weighted soft-constrained natural gradient iterative method is introduced for robust blind separation in reverberant environment. The nonlocal weighting of the iterations promotes stability and convergence of the algorithm for long demixing filters. The scaling degree of freedom is controlled by soft-constraints built into the auxillary difference equations. The small divisor problem of iterations in silence durations of speech is resolved. Computations on synthetic speech mixtures based on measured binaural room impulse responses show that the algorithm achieves higher signal-to-inteference ratio improvement than existing methods in an office size room with reverberation time over 0.5 second.

PM–17 The Ideal Interaural Parameter Mask: A Bound on Binaural Separation Systems

Michael I. Mandel and Daniel P.W. Ellis, *Columbia University, USA*

We introduce the Ideal Interaural Parameter Mask as an upper bound on the performance of source separation algorithms that are based purely on the differences between two channels. With two additions to our Model-based EM Source Separation and Localization system, its performance approaches that of the IIPM upper bound to within 0.9 dB. These additions battle the effects of reverberation by absorbing reverberant energy and by forcing the ILD estimate to be larger than it might otherwise be. An oracle reliability measure was also added, in the hope that estimating parameters from more reliable regions of the spectrogram would improve separation, but it was not consistently useful.

PM–18 Source Enumeration of Speech Mixtures Using Pitch Harmonics

Keith D. Gilbert and Karen L. Payton, *University of Massachusetts Dartmouth, USA*

This paper proposes a method to simultaneously estimate the number, pitches, and relative locations of individual speech sources within instantaneous and non-instantaneous linear mixtures containing additive white Gaussian noise. The algorithm makes no assumptions about the number of sources or the number of sensors, and is therefore applicable to over-, under-, and precisely-determined scenarios. The method is hypothesis-based and employs a power-spectrum-based FIR filter derived from probability distributions of speech pitch harmonics. This harmonic windowing function (HWF) dramatically improves time-difference of arrival (TDOA) estimates over standard cross-correlation for low SNR. The pitch estimation component of the algorithm implicitly performs voiced-region detection and does not require prior knowledge about voicing. Cumulative pitch and TDOA estimates from the HWF form the basis for robust source enumeration across a wide range of SNR.

PM–19 Unsupervised Single-Channel Source Separation Using Bayesian NMF

Onur Dikmen and A. Taylan Cemgil, *Boğaziçi University, Turkey*

We propose a prior structure for single-channel audio source separation using Non-Negative Matrix Factorisation. For the tonal and percussive signals, the model assigns different prior distributions to the corresponding parts of the template and excitation matrices. This partitioning enables not only more realistic modeling, but also a deterministic way to group the components into sources. This also prevents the possibility of not detecting/assigning a component and remove the need for a dataset and training. Our method only needs the number of components of each source to be set, but this does not play a crucial role in the performance. Very promising results can be obtained using the model with too few design decisions and moderate time complexity.

PM–20 An Investigation of Discrete-State Discriminant Approaches to Single-Sensor Source Separation

Valentin Emiya, Emmanuel Vincent, and Rémi Gribonval, *INRIA-IRISA, France*

This paper investigated a new scheme for single-sensor audio source separation. This framework is introduced comparatively to the existing Gaussian mixture model generative approach and is focusing on the mixture states rather than on the source states, resulting in a discrete, joint state discriminant approach. The study establishes the theoretical performance bounds of the proposed scheme and an actual source separation system is designed. The performances are computed on a set of musical recordings and a discussion is proposed, including the question of the source correlation and the possible drawbacks of the method.

PM–21 On the Non-Uniqueness Problem and the Semi-Blind Source Separation

Francesco Nesta, *Fondazione Bruno Kessler-Irst, Italy*, Ted S. Wada and Biing-Hwang (Fred) Juang, *Georgia Institute of Technology, USA*, Shigeki Miyabe, *University of Tokyo, Japan*

Semi-blind source separation (SBSS) is a special case of the well-known source separation problem when some partial knowledge of the source signals is available to the system. In particular, a batch-wise adaptation in the frequency domain based on the independent component analysis (ICA) can be effectively used to jointly perform source separation and multi-channel acoustic echo cancellation (MCAEC) without double-talk detection. However, the non-uniqueness problem due to the correlated far-end reference signals still affects the SBSS approach. In this paper, we analyze the structure of the SBSS de-mixing matrix and the behavior of a batch on-line adaptation algorithm under two most common far-end mixing conditions. We show that with a proper constraint on the de-mixing matrix, high echo reduction can be achieved just as the misalignment remains relatively low even for the worst-case scenario of single far-end talker and also without any pre-processing procedure to decorrelate the far-end signals.

PM–22 Coherent Spectral Estimation for a Robust Solution of the Permutation Problem

Francesco Nesta, *Fondazione Bruno Kessler-Irst,, Italy*, Ted S. Wada and Biing-Hwang (Fred) Juang, *Georgia Institute of Technology, USA*

In this paper, we propose a new method based on a coherent source spectral estimation for solving the permutation problem of frequency-domain blind source separation (BSS). It combines the robustness of the State Coherence Transform (SCT) to recursively estimate a smooth phase spectrum associated with each source and the precision of the inter-frequency correlation to solve for a correct permutation. Namely, the TDOAs estimated by the SCT are used to constrain the permutation correction process in order to force the resulting filters to be coherent across frequency. This intrinsic interconnection between the TDOA information and the spectral correlation makes the new approach robust even when the signal is short in duration and spatial aliasing is substantial. Experimental results show that the proposed method is able to drastically reduce the number of permutation errors for three sources recorded in a short time block using microphones with large spacing.

12:20 pm – 4:00 pm
Lunch & Afternoon Break

Lecture Session — LM2: Monday 4:00 am – 6:00 pm

Microphone Array and Source Separation

Session Chair:
Gary W. Elko
mh acoustics, USA

LM2–1 *4:00 pm – 4:20 pm*
On Optimal Beamforming for Noise Reduction and Interference Rejection

Mehrez Souden, Jacob Benesty, and Sofiène Affes, *Université du Québec, Canada*

In this paper, we study the performance of the minimum variance distortionless response (MVDR) and linearly constrained minimum variance (LCMV) noise reduction filters when a source of interference and ambient noise coexist with the target signal. We demonstrate that both filters are related as we decompose the MVDR filter into the LCMV and a matched filter (MVDR solution in the absence of interference). Both components are properly weighted to achieve maximum interference-plus-noise reduction at each frequency bin. Furthermore, we elaborate new closed-form expressions for the signal-to-interference ratio (SIR) and signal-to-noise ratio (SNR) at the output of the LCMV, MVDR, and matched filters. These expressions theoretically prove that a trade-off between noise reduction and interference rejection has to be made. In fact, the total removal of the interference may severely amplify the output ambient noise. Conversely, totally focusing on noise reduction leads to increased level of residual interference. The proposed study is finally supported by numerical examples.

LM2–2 *4:20 pm – 4:40 pm*

Robust Spherical Microphone Array Beamforming with Multi-Beam-Multi-Null Steering, and Sidelobe Control

Haohai Sun, Shefeng Yan, and U. Peter Svensson, *Norwegian University of Science and Technology, Norway*

A spherical harmonics domain microphone array beamforming approach is proposed, especially for voice pick-up application in video conferencing and tele-presence systems. It unifies 3D multi-beam forming with tractable mainlobe levels, automatic multi-null steering, sidelobe control, and robustness control into one optimization framework, using a single spherical microphone array. The optimum array weights are designed by maintaining distortionless responses in multiple mainlobe directions and guaranteeing all sidelobes below given threshold values, while minimizing the beamformer output power. A weight vector norm constraint is also employed to improve the robustness of the beamformer. A convex optimization formulation is derived, and implemented by second order cone programming (SOCP) method. Design examples demonstrate a satisfactory performance.

LM2–3 *4:40 pm – 5:00 pm*

Panoramic Recording and Reproduction of Multichannel Audio Using a Circular Microphone Array

Hüseyin Hacıhabiboğlu and Zoran Cvetković, *King's College London, UK*

Multichannel audio reproduction generally suffers from one or both of the following problems: i) the recorded audio has to be artificially manipulated to provide the necessary spatial cues, which reduces the consistency of the reproduced sound field with the actual one, and ii) reproduction is not panoramic, which degrades realism when the listener is not seated in a desired ideal position facing the center channel. A recording method using a circularly symmetric array of differential microphones, and a reproduction method using a corresponding array of loudspeakers is presented in this paper. Design of microphone directivity patterns to achieve a panoramic auditory scene is discussed. Objective results in the form of active intensity diagrams are presented.

LM2–4 *5:00 pm – 5:20 pm*

Factorial Scaled Hidden Markov Model for Polyphonic Audio Representation and Source Separation

Alexey Ozerov, Cédric Févotte, and Maurice Charbit, *Telecom ParisTech CNRS LTCI, France*

We present a new probabilistic model for polyphonic audio termed Factorial Scaled Hidden Markov Model (FS-HMM), which generalizes several existing models, notably the Gaussian scaled mixture model and the Itakura-Saito Nonnegative Matrix Factorization (NMF) model. We describe two expectation-maximization (EM) algorithms for maximum likelihood estimation, which differ by the choice of complete data set. The second EM algorithm, based on a reduced complete data set and multiplicative updates inspired from NMF methodology, exhibits much faster convergence. We consider the FS-HMM in different configurations for the difficult problem of speech / music separation from a single channel and report satisfying results.

LM2–5 *5:20 pm – 5:40 pm*

Finding Similar Acoustic Events Using Matching Pursuit and Locality-Sensitive Hashing

Courtenay Cotton and Daniel P.W. Ellis, *Columbia University, USA*

There are many applications for the ability to find repetitions of perceptually similar sound events in generic audio recordings. We explore the use of matching pursuit (MP) derived features to identify repeated patterns that characterize distinct acoustic events. We use locality-sensitive hashing (LSH) to efficiently search for similar items. We describe a method for detecting repetitions of events, and demonstrate performance on real data.

LM2–6 *5:40 pm – 6:00 pm*

Spatial Covariance Models for Under-Determined Reverberant Audio Source Separation

Ngoc Q.K. Duong, Emmanuel Vincent, and Rémi Gribonval, *IRISA-INRIA, France*

The separation of under-determined convolutive audio mixtures is generally addressed in the time-frequency domain where the sources exhibit little overlap. Most previous approaches rely on the approximation of the mixing process by complex-valued multiplication in each frequency bin, which is equivalent to assuming that the spatial covariance matrix of each source has rank 1. In this paper, we propose to represent each source via a full-rank spatial covariance matrix instead, which better approximates reverberation. We also investigate a possible parameterization of this matrix stemming from the theory of statistical room acoustics. We illustrate the potential of the proposed approach over a stereo reverberant speech mixture.

6:00 pm – 8:00 pm

Dinner (West Dining Room)

KEYNOTE ADDRESS — TUESDAY 8:00 AM – 8:45 AM

Nonlinear Cochlear Signal Processing and Phoneme Perception

Jont B. Allen
University of Illinois, USA

The most important communication signal is human speech. It is helpful to think of speech communication in terms of Claude Shannon's information theory channel model. When thus viewed, it immediately becomes clear that the most complex part of speech communication channel is in auditory system (the receiver). In my opinion, even after years of work, relatively little is known about how the human auditory system decodes speech. Given cochlear damaged, speech scores are greatly reduced,even with tiny amounts of noise. The exact reasons for this SNR-loss presently remain unclear, but I speculate that the source of this must be cochlear outer hair cell temporal processing, not central processing. Specifically, "temporal edge enhancement" of the speech signal and forward masking could easily be modified in such ears, leading to SNR-Loss. What ever the reason, SNR-Loss is the key problem that needs to be fully researched.

Biosketch

Jont B. Allen received his BS in Electrical Engineering from the University of Illinois, Urbana-Champaign, in 1966, and his MS and PhD in Electrical Engineering from the University of Pennsylvania in 1968 and 1970 respectively. After graduation he joined Bell Laboratories, and was in the Acoustics Research Department in Murray Hill, NJ from 1974 to 1996, as a Distinguished Member of Technical Staff. Since 1996 Dr. Allen was a Technology Leader at AT&T Labs-Research. Since Aug. 2003 Allen is a tenured Associate Professor in ECE, at the University of Illinois, and on the research staff of the Beckman Inst., Urbana, IL.

During his 32 year AT&T career Prof. Allen has specialized in cochlear and middle ear modeling and auditory signal processing. In the last 10 years, he has concentrated on the problem of human speech recognition. His expertise spans the areas of signal processing, physical acoustics, acoustic power flow and impedance measurements, cochlear modeling, auditory neurophysiology, auditory psychophysics, and human speech recognition.

Prof. Allen is a Fellow (May 1981) of the Acoustical Society of America (ASA) and Fellow (January 1985) of the Institute of Electrical and Electronic Engineers (IEEE). In 1986 he was awarded the IEEE Acoustics Speech and Signal Processing (ASSP) Society Meritorious Service Award, and in 2000 received an IEEE Third Millennium Medal. He is a past member of the Executive Council of the ASA, the Administration Committee (ADCOM) of the IEEE ASSP, has served as Editor of the ASSP Transactions, as Chairman of the Publication Board of the ASSP Society, as General Chairman of the International Conference on ASSP (ICASSP-1988), and on numerous committees of both the ASA and the ASSP. He is presently a member of ASA Publications Policy Board. He has organized several workshops and conferences on hearing research and signal processing. In 1984 he received funding from NIH to host the 2nd International Mechanics of Hearing Workshop. He has a strong interest in electronic publications and has produced several CDROM publications, including suggesting, and then overseeing technical details of, the publication of the J. Acoust. Soc. Am. in DjVu format, and developed the first subject classification system for the IEEE Transactions of the ASSP, as well as the ASSP Magazine.

In 1986-88 Prof. Allen participated in the development of the AT&T multi-band compression hearing aid, later sold under the ReSound and Danavox name, and served as a member of the ReSound and SoundID Scientific advisory boards. Since 1987 he has been an Adjunct Associate Research Scientist in the Department of Otolaryngology at Columbia University, and on the CUNY speech and Hearing Faculty (Adjunct). In 1990 he was an Osher Fellow at the Exploratorium museum in San Francisco. In 1991-92 he served as an International Distinguished Lecturer for the IEEE Signal Processing Society. In 1993 he served on the Dean's Advisory Council at the University of Pennsylvania. In 1994 Allen spent 5 weeks as Visiting Scientist and Distinguished Lecturer at the University of Calgary. From 1994-present Allen is the CTO of Mimosa Acoustics, which manufactures diagnostic equipment of cochlear and middle ear disorders, based on distortion product and power flow and acoustic impedance measurements. In 2000 he received the IEEE third Millennium Award, and in 2004, an IBM faculty award. Prof. Allen has more than 90 publications (36 peer reviewed) and 16 patents in the areas of speech noise reduction, speech and audio coding, and hearing aids. In 2006 Allen published a 150-page research monograph "Articulation and Intelligibility" that reviews the literature on human speech recognition from 1900 to the present, in the context of models of human speech perception.

Lecture Session — LT1: Tuesday 8:45 am – 10:25 am

Speech Enhancement, Dereverberation, and Echo Control

Session Chair:
Rudolf Rabenstein
University of Erlangen-Nuremberg, Germany

LT1–1 *8:45 am – 9:05 am*

Two-Stage Binaural Speech Enhancement with Wiener Filter Based on Equalization-Cancellation Model

Junfeng Li and Masato Akagi, *Japan Advanced Institute of Science and Technology, Japan*, Shuichi Sakamoto and Yôiti Suzuki, *Tohoku University, Japan*, Satoshi Hongo, *Miyagi National College of Technology, Japan*

The equalization-cancellation (EC) model has been extensively studied for expressing spatial release from masking (SRM) in psychoacoustics. Few research focuses on applying this psychoacoustic model to speech processing applications, such as speech enhancement. In this paper, we propose a two-stage binaural speech enhancement with Wiener filter (TS-BASE/WF) based on the EC model. In this proposed TS-BASE/WF, interfering signals are first estimated by equalizing and cancelling the target signal based on the EC model, and a time-variant Wiener filter is then applied to enhance the target signal given noisy mixture signals. The main advantages of the proposed TS-BASE/WF are: (1) effective in dealing with non-stationary multiple-source interfering signals; (2) successful in localizing the target sound source after processing. These advantages were confirmed by the comprehensive experiments in different spatial scenarios in terms of speech enhancement and sound localization.

LT1–2 *9:05 am – 9:25 am*

A Spatio-Temporal Power Method for Time-Domain Multi-Channel Speech Enhancement

Malay Gupta, Sylvain Angrignon, Chris Forrester, and Sean Simmons, *Research in Motion Corporation, Canada*, Scott C. Douglas, *Southern Methodist University, USA*

We present a new multi-stage iterative technique for enhancing noisy speech under low signal-to-interference-ratio (SNR) environments. In the present paper, the speech is enhanced in two stages, in the first stage the noise component of the observed signal is whitened, and in the second stage a spatio-temporal power method is used to extract the desired speech component. In both the stages, the coefficient adaptation is performed using the multichannel spatio-temporal correlation sequences of the observed data. The technique is mathematically equivalent and is computationally simpler than the existing generalized eigenvalue decomposition (GEVD) or the generalized singular value decomposition (GSVD) based techniques. Simulation results under low SNR diffuse noise scenarios indicate significant gains in SNR without introducing musical noise artifacts.

LT1–3 *9:25 am – 9:45 am*

On the Application of the LCMV Beamformer to Speech Enhancement

Emanuël A.P. Habets and Patrick A. Naylor, *Imperial College, UK*, Jacob Benesty, *University of Quebec, Canada*, Sharon Gannot, *Bar-Ilan University, Israel*, Israel Cohen, *Technion-IIT, Israel*

In theory the linearly constrained minimum variance (LCMV) beamformer can achieve perfect dereverberation and noise cancellation when the acoustic transfer functions (ATFs) between all sources (including interferences) and the microphones are known. However, blind estimation of the ATFs remains a difficult task. In this paper the noise reduction of the LCMV beamformer is analyzed and compared with the noise reduction of the minimum variance distortionless response (MVDR) beamformer. In addition, it is shown that the constraint of the LCMV can be modified such that we only require relative transfer functions rather than ATFs to achieve perfect coherent noise reduction. Finally, we evaluate the noise reduction performance achieved by the LCMV and MVDR beamformers for two coherent sources: one desired and one undesired.

LT1–4 *9:45 am – 10:05 am*

Statistical Models for Speech Dereverberation

Takuya Yoshioka, Hirokazu Kameoka, and Tomohiro Nakatani, *NTT Corporation, Japan*, Hiroshi Okuno, *Kyoto University, Japan*

This paper discusses a statistical-model-based approach to speech dereverberation. With this approach, we first define parametric statistical models of probability density functions (pdfs) for a clean speech signal and a room transmission channel, then estimate the model parameters, and finally recover the clean speech signal by using the pdfs with the estimated parameter values. The key to the success of this approach lies in the definition of the models of the clean speech signal and room transmission channel pdfs. This paper presents several statistical models (including newly proposed ones) and compares them in a large-scale experiment. As regards the room transmission channel pdf, an autoregressive (AR) model, an autoregressive power spectral density (ARPSD) model, and a moving-average power spectral density (MAPSD model are considered. A clean speech signal pdf model is selected according to the room transmission channel pdf model. The AR model exhibited the highest dereverberation accuracy when a reverberant speech signal of 2 sec or longer was available while the other two models outperformed the AR model when only a 1-sec reverberant speech signal was available.

LT1–5 *10:05 am – 10:25 am*

Adaptive FIR Filters with Automatic Length Optimization by Monitoring a Normalized Combination Scheme

Marcus Zeller and Walter Kellermann, *University of Erlangen-Nuremberg, Germany*, Luis A. Azpicueta-Ruiz, *Universidad Carlos III de Madrid, Spain*

This paper presents a novel strategy of adaptive filtering which provides an automatic self-configuration of the filter structure in terms of memory length. By monitoring the adaptive mixing of a normalized combination of two com-

peting filters with a different number of coefficients, an online estimate of the optimum filter length is obtained and used to dynamically scale the size of the employed filters. Furthermore, a more efficient, simplified version of this approach is proposed and shown to be equally effective while significantly reducing the required complexity. Experimental results for high-order real-world systems as well as stationary noise and speech signals demonstrate the good performance and the robust tracking behavior of the outlined algorithms in the context of realistic system identification scenarios.

10:25 am – 10:40 am
Coffee Break (East Dining Room)

Poster Session — PT: Tuesday 10:40 am – 12:20 pm

Speech Enhancement, Multichannel Audio Processing, and Hearing Aids

Session Chair:
Jingdong Chen
WeVoice Inc., USA

PT–1 Maximum Directivity Beamformer for Spherical-Aperture Microphones
Morag Agmon, Boaz Rafaely and Joseph Tabrikian, *Ben-Gurion University of the Negev, Israel*

Spherical microphone arrays have been recently studied for sound analysis and beamforming. These arrays have the advantage of spherical symmetry facilitating three-dimensional analysis. Performance of microphone arrays at the high-frequency range is typically limited by aliasing, which is a result of the spatial sampling process. A potential approach to avoid spatial aliasing is by using continuous sensors, in which spatial sampling is not required. This paper presents an optimal beamforming technique for the spherical-aperture microphone, which is based on a continuous sensor. The proposed beamforming technique is used to compute the optimal real aperture weighting function. Real aperture weighting functions are required to ensure the realizability of the sensor.

PT–2 On Robustness of Multi-Channel Minimum Mean-Squared Error Estimators under Super-Gaussian Priors
Richard C. Hendriks and Richard Heusdens, *Delft University of Technology, The Netherlands*, Jesper Jensen, *Oticon A/S, Denmark*

The use of microphone arrays in speech enhancement applications offer additional features, like directivity, over the classical single-channel speech enhancement algorithms. An often used strategy for multi-microphone noise reduction is to apply the multi-channel Wiener filter, which is often claimed to be mean-squared error optimal. However, this is only true if the estimator is constrained to be linear, or, if the speech and noise process are assumed to be Gaussian. Based on histograms of speech DFT coefficients it can be argued that optimal multi-channel minimum mean-squared error (MMSE) estimators should be derived under super-Gaussian speech priors instead. In this paper we investigate the robustness of these estimators when the steering vector is affected by estimation errors. Further, we discuss the sensitivity of the estimators when the true underlying distribution of speech DFT coefficients deviates from the assumed distribution.

PT–3 Blind Alignment of Asynchronously Recorded Signals for Distributed Microphone Array
Nobutaka Ono, Hitoshi Kohno, Nobutaka Ito, and Shigeki Sagayama, *The University of Tokyo, Japan*

In this paper, aiming to utilize independent recording devices as a distributed microphone array, we present a novel method for alignment of recorded signals with localizing microphones and sources. Unlike conventional microphone array, signals recorded by independent devices have different origins of time, and microphone positions are generally unknown. In order to estimate both of them from only recorded signals, time differences between channels for each source are detected, which still include the differences of time origins, and an objective function defined by their square errors is minimized. For that, simple iterative update rules are derived through auxiliary function approach. The validity of our approach is evaluated by simulative experiment.

PT–4 Artifacts in the Sound Field of a Moving Sound Source Reconstructed from a Microphone Array Recording
Jens Ahrens and Sascha Spors, *Deutsche Telekom Laboratories, Technische Universität Berlin, Germany*

We present an analysis of the sound field of a moving sound source reconstructed from recordings of a virtual dual-radius open-sphere microphone array. As a consequence of the discrete property of such microphone distributions artifacts arise, most notably spatial aliasing and spatial bandwidth limitation artifacts. We show that these artifacts are much more pronounced for moving sound sources than for static ones. We analyze the artifacts with a focus on a possible perceptual impairment when such recordings are used for audition purposes.

PT–5 Dolph-Chebyshev Radial Filter for the Near-Field Spherical Microphone Array
Etan Fisher and Boaz Rafaely, *Ben-Gurion University of the Negev, Israel*

A radial filter is developed for the near-field spherical microphone array, based on the Dolph-Chebyshev beampattern. The filter coefficients are reached analytically, given the Dolph-Chebyshev design parameters. Radial filtering in this method enables control over the attenuation of far-field sources, with significant improvement in far-field attenuation relative to the natural point source decay.

PT–6 Enhancement of Speech Intelligibility Using Transients Extracted by Wavelet Packets

Daniel M. Rasetshwane, J. R. Boston, Ching-Chung Li, John D. Durrant, and Greg Genna, *University of Pittsburgh, USA*

Speech transients have been shown to be important cues for identifying and discriminating speech sounds. We previously described a wavelet packet-based method for extracting transient speech. The algorithm uses a 'transitivity function' to characterize the rate of change of wavelet coefficients and it can be implemented in real-time to process continuous speech. Psycho-acoustic experiments to select parameters for and to evaluate this method are presented. Results show that modified speech created by amplifying transient speech and adding it to original speech has higher percent word recognition scores than original speech in the presence of background noise.

PT–7 Single-Microphone Wind Noise Reduction by Adaptive Postfiltering

Elias Nemer and Wilf Leblanc, *Broadcom Corp, USA*

This paper presents a novel time-domain algorithm for detecting and attenuating the acoustic effect of wind noise in speech signals originating from mobile terminals. The detection part makes use of metrics that exploits the properties of the spectral envelop of wind noise as well as its non-periodic and non-harmonic nature. LPC analyses of various orders are carried out and the results used to distinguish between wind and speech frames and to estimate the magnitude and location of the wind noise 'resonance'. The suppression part entails constructing a parameterized post-filter of an appropriate order having a 'null' where the wind noise 'resonance' is. Wind-only frames are used to estimate the wind noise energy, from which the emphasis parameters of the post-filter are adjusted to provide an appropriate attenuation. The proposed scheme may be combined with background-noise suppression algorithms, or with speech-formant-enhancing post-filters in the context of a speech codec.

PT–8 An Auditory-Based Transform for Audio Signal Processing

Qi (Peter) Li, *Li Creative Technologies, Inc., USA*

An auditory-based transform is presented in this paper. Through an analysis process, the transform coverts time-domain signals into a set of filter bank output. The frequency responses and distributions of the filter bank are similar to those in the basilar membrane of the cochlea. Signal processing can be conducted in the decomposed signal domain. Through a synthesis process, the decomposed signals can be synthesized back to the original signal through a simple computation. Fast algorithms for discrete-time signals are presented for both the forward and inverse transforms. The fast algorithm can also be used to compute continuous wavelet transform. Experiments are provided on noise reduction and synthesis for audio signal processing. The proposed transform is an alternative selection to the Fourier and wavelet transforms.

PT–9 Gain Adaptation Based on Signal-to-Noise Ratio for Noise Suppression

Devangi N. Parikh and David V. Anderson, *Georgia Institute of Technology, USA*, Sourabh Ravindran, *Texas Instruments Inc., USA*

We propose a technique of noise suppression using an automatic gain control system. This method involves expands the signal so that the noise floor of the signal is pushed down in low SNR regions and hence the effect of noise is reduced. This method does not require a VAD system and is of low computational complexity. The processing is based on the model of human perception and so the resulting noise suppressed speech is natural sounding. The algorithmic delay of this method is depends only on the group delay of the filters since the processing is done entirely in the time-domain. The average group delay at approximately the middle of the auditory band (1-2 kHz) is about 375-170 microseconds. We will show that our algorithm performs at par and in certain cases it outperforms other standard noise suppression techniques. Such a low-delay algorithm would be highly useful in hearing aids, Bluetooth devices, PA systems, teleconferencing equipment, etc.

PT–10 Improved A Priori SNR Estimation with Application in Log-MMSE Speech Estimation

Nils Höglund and Sven Nordholm, *Curtin University of Technology, Australia*

A speech enhancement method utilizing the harmonic structure of speech is presented. The method is an extension of the well known minimum mean square errorlog-spectral amplitude estimator(Log MMSE) method for speech enhancement. The improvement lies specifically on *a priori* SNR estimation by utilizing harmonic structure of speech. The method is based on a conditional averaging operation over adjacent frequency bands for each processed data block. The actual frequency bands used in the conditional averaging is determined by a pitch detector. Thus voiced segments are averaged over frequency according to the pitch and the corresponding harmonic structure of voiced speech. Non-voiced segments are averaged over frequency according to a random number depending on the pitch value. The result is overall better SNR and SNRSeg values in white noise over the standard Log MMSE reference method. In babble noise, the estimator rendered similar SNR and SNRSeg values as the Log-MMSE reference method. Subjectively the residue background noise sounded more natural when using the suggested method.

PT–11 Robust Audio Precompensation With Probabilistic Modeling of Transfer Function Variability

Lars-Johan Brännmark, *Uppsala University, Sweden*

A new approach to the single-channel loudspeaker equalization problem is presented. A scalar discrete-time mixed-phase precompensation filter is designed to be spatially robust, meaning that equalization performance should be insensitive to listener movements within a predefined spatial region. The problem is posed in a single-input multiple-

output (SIMO) feedforward control framework and a polynomial solution is derived, based on a set of room transfer functions (RTFs) measured at a number of control points in the region, and a multipoint mean-square error (MSE) criterion. Spatial robustness is obtained by the introduction of two novel strategies. Firstly, a probabilistic model is used to describe the RTF variability around each control point, and the MSE criterion is averaged with respect to this variability. Secondly, the pre-response errors, normally associated with mixed-phase equalizer design, are alleviated by restricting the compensator to have a certain structure. The proposed method is shown to produce filters with excellent time- and frequency-domain performance.

PT–12 Variable Control of the Pre-Response Error in Mixed Phase Audio Precompensation

Lars-Johan Brännmark and Anders Ahlén, *Uppsala University, Sweden*

We introduce a method for controlling the spatial robustness of a mixed phase loudspeaker equalizer design. The emphasis is on time domain behaviour and the pre-response error that is always present in mixed phase design. Based on measurements from a small spatial region, a mixed phase compensator is regularized to be valid also over a large spatial region. The regularization can be applied gradually to match any size of listener region, while fulfilling a set of constraints for the pre-response error. The compensation thus avoids the unacceptably high pre-response error levels that generally occur outside the measurement region. The proposed compensator design, which is validated on measurements in both small and large spatial regions, is shown to produce excellent results.

PT–13 Dynamic Impulse Response Model for Nonlinear Acoustic System and Its Application to Acoustic Echo Canceller

Shoichiro Saito, Akira Nakagawa, and Yoichi Haneda, *NTT Corporation, Japan*

We propose a dynamic impulse response model for a nonlinear acoustic system and a nonlinear acoustic echo canceller based on this model. In an acoustic system with small loudspeakers and low-quality enclosures, acoustic signal paths between loudspeakers and microphones often include nonlinearity. These responses cannot be expressed with a conventional linear system. We investigate the impulse responses with various input signal levels, and find that the responses differ with regularity according to the level of the input signal. Thus, we propose our dynamic impulse response model based on this fact. In this model, the system response varies according to the input signal levels, and the nonlinearity can be approximated with the set of linear filters. We applied this model to an acoustic echo canceller, and experimental evaluations using a small acoustic system show that the echo cancellation of the proposed method exceeds that of the conventional method by 4 dB on average and 10 dB at maximum.

PT–14 Acoustic Echo Cancellation Based on Independent Component Analysis and Integrated Residual Echo Enhancement

Ted S. Wada and Biing-Hwang (Fred) Juang, *Georgia Institute of Technology, USA*

This paper examines the technique of using a memoryless noise-suppressing nonlinearity in the adaptive filter error feedback-loop of an acoustic echo canceler (AEC) based on normalized least-mean square (NLMS) when there is additive noise at the near-end. It will be shown that introducing the nonlinearity to "enhance" the filter estimation error is well-founded in the information-theoretic sense and has a deep connection to the independent component analysis (ICA). The paradigm of AEC as a problem that can be approached by ICA leads to new algorithmic possibilities beyond the conventional LMS family of techniques. In particular, a right combination of the error enhancement procedure and a properly implemented regularization procedure enables the AEC to be performed recursively and continuously in the frequency domain when there are both ambient noise and double-talk even without the use of a double-talk detector (DTD) or a voice activity detector (VAD).

PT–15 A Phase-Based Dual Microphone Method to Count and Locate Audio Sources in Reverberant Rooms

Zaher El Chami and Alexandre Guerin, *Orange Labs, France*, Antoine Pham and Christine Servière, *CNRS-INPG, France*

This paper presents an off-line method to estimate the mixing conditions, characterized by the number of audio sources and their Time Difference Of Arrival (TDOA). The proposed method is based on the assumption of having, at least, one time frame where each one of the sources is dominant. From such frames, the TDOA of the correspondent dominant source is estimated by maximizing a novel coherence function similar to the linear regression proposed in [3]. Finally, the number of sources will be derived by counting the number of active directions of arrival (DOA). Experimental results on real-life recordings are presented, showing the good performance of the algorithm in reverberant environments.

PT–16 Stochastic Particle Filtering: A Fast SRP-PHAT Single Source Localization Algorithm

Hoang Do and Harvey F. Silverman, *Brown University, USA*

Computational cost has been an issue for the proven robust source localization algorithm, steered response power (SRP) using the phase transform (SRP-PHAT). Some proposed computation reduction algorithms degrade under high noise and reverberant conditions. Some require at least 10% the cost of a full SRP-PHAT gridsearch. In ICASSP 2007, we introduced a robust, low-cost global optimization technique, stochastic region contraction (SRC). In this paper, we present another algorithm, stochastic particle filtering (SPF), which uses SRC's initialization and is a kind of Sampling Importance Resampling (SIR) particle filtering. In this paper, the SRP is computed using a modification to the conventional PHAT, namely β-PHAT. Extensive experiments using real data and simulated data are shown. The results indicate that, while maintaining the desirable accuracy of the full search, this method reduces the cost to about half the cost of SRC (0.03% the cost of full search), thus making SRP-PHAT more practical for real-time applications.

PT–17 Multiple Sound Source Tracking Method Based on Subspace Tracking

Noboru Ohwada and Kenji Suyama, *Tokyo Denki University, Japan*

In this paper, we propose a novel method for tracking two talkers.The tracking is performed using a microphone array and is based on algorithms implemented successively, namely, the PAST (Projection Approximation Subspace Tracking) and IPLS (Interior Point Least Square) algorithms. When a large number of sound sources exist, the order of eigenvectors estimated by the PAST may change at each point in the time–frequency domain. This prevents us from separating and localizing each source. In addition, the appropriate initial values for the IPLS algorithm are required even for the silent-speech sections or in the low-energy domain because of the sparseness of speech signals. To overcome such difficulties, we propose a new method for assigning eigenvectors in the appropriate order and a method for setting the suitable initial value in the IPLS. Several results of experiments performed in an actual room environment show the effectiveness of the proposed method.

PT–18 Coherent Signals Direction-of-Arrival Estimation Using a Spherical Microphone Array: Frequency Smoothing Approach

Dima Khaykin and Boaz Rafaely, *Ben-Gurion of the Negev, Israel*

Direction-of-arrival (DOA) estimation of coherent signals is considered of great importance in signal processing. To estimate both azimuth and elevation angle with the same accuracy, 3-dimensional (3-D) array must be used. Spherical arrays have the advantage of spherical symmetry, facilitating 3-D DOA estimation. To apply high resolution subspace DOA estimation algorithms, such as MUSIC, in a coherent environment, a smoothing technique is required. This paper presents the development of a smoothing technique in the frequency domain for spherical microphone arrays. We show that frequency smoothing can be efficiently performed using spherical arrays due to the decoupling of frequency and angular components. Experimental comparison of DOA estimation between beamforming and MUSIC with frequency smoothing is performed with data measured in a real auditorium.

PT–19 Multichannel Voice Activity Detection with Spherically Invariant Sparse Distribution

Bowon Lee and Ton Kalker, *Hewlett-Packard Laboratories, USA*

We propose a statistical multichannel voice activity detection algorithm by modeling the frequency components of speech signals as sparse multivariate complex distributions. In particular, we formulate a likelihood ratio test by modeling a multichannel speech observation as a spherically invariant random process with a parameter governing its sparseness. In addition, we consider reverberation as a component of the statistical model. Experimental results show that our proposed method significantly reduces false-alarm rate for reverberation tails and that sparse distributions provide higher detection accuracy compared to the traditional Gaussian distribution.

PT–20 A Zone of Quiet Based Approach to Integrated Active Noise Control and Noise Reduction in Hearing Aids

Romain Serizel, Marc Moonen, and Jan Wouters, *Katholieke Universiteit Leuven, Belgium*, Søren Holdt Jensen, *Aalborg University, Denmark*

This paper presents an integrated approach to active noise control and noise reduction in hearing aids which is based on an optimization over a zone of quiet generated by the active noise control. A basic integrated scheme has been introduced previously to tackle secondary path effects and effects of noise leakage through an open fitting. This scheme, however, only takes the sound pressure at the ear canal microphone into account. In practice, it is desired to achieve noise control in a zone not limited to a single point. A scheme based on an average mean squared error criterion over the desired zone of quiet is presented here and compared experimentally with the original scheme.

PT–21 A Wiener-Based Implementation of Equalization-Cancellation Pre-Processing for Binaural Speech Intelligibility Prediction

Nicolas N. Ellaham, Christian Giguère, and Wail Gueaieb, *University of Ottawa, Canada*

This paper presents a precursor to an objective measure to predict speech intelligibility in binaural listening conditions. Such measures typically consist of a binaural preprocessing stage followed by intelligibility prediction using a monaural measure such as the Speech Intelligibility Index. In this work, an implementation of the equalization-cancellation process using Wiener filters is presented as a binaural pre-processing stage. The model is tested in simulated sound-field listening. Preliminary assessment is performed by comparison with recent work in this field. Speech intelligibility measurements from the literature are also used to qualitatively assess the improvements in signal-to-noise ratio obtained. This work will form the basis for a complete binaural intelligibility prediction system involving nonlinear input signals common in hearing aids.

Lecture Session — LT2: Tuesday 4:00 pm – 6:00 pm

Spatial Sound Perception, Analysis, and Reproduction

Session Chair:
Michael Goodwin
Audience, Inc., USA

LT2–1 *4:00 pm – 4:20 pm*

Generalized State Coherence Transform for Multidimensional Localization of Multiple Sources

Francesco Nesta, *Università di Trento, Italy*, Maurizio Omologo, *Fondazione Bruno Kessler-irst, Italy*

In our recent work an effective method for multiple source localization has been proposed under the name of cumulative State Coherence Transform (cSCT). Exploiting the physical meaning of the frequency-domain blind source separation and the sparse time-frequency dominance of the acoustic sources, multiple reliable TDOAs can be estimated with only two microphones, regardless of the permutation problem and of the microphone spacing. In this paper we present a multidimensional extension of the cSCT which allows one to localize several sources in the P-dimensional space. An important approximation is made in order to perform a disjoint TDOA estimation over each dimension which reduces the localization problem to linear complexity. Furthermore the approach is invariant to the array geometry and to the assumed acoustic propagation model. Experimental results on simulated data show a precise 2-D localization of 7 sources by only using an array of three elements.

LT2–2 *4:20 pm – 4:40 pm*

A Probabilistic Speaker Clustering for DOA-based Diarization

Katsuhiko Ishiguro, Takeshi Yamada, Shoko Araki, and Tomohiro Nakatani, *NTT Communication Science Laboratories, Japan*

We present a probabilistic speaker clustering and diarization model. Speaker diarization determines "who spoke when" from the recorded conversation of unknown number of people. We formulate this problem as the clustering of sequential auditory features generated by an unknown number of latent mixture components (speakers). We employ a probabilistic model which automatically estimates the number of speakers and time-varying speaker proportions. Experiments with synthetic and real sound recordings confirm that the proposed model can successfully infer the number and features of speakers and obtained better speaker diarization results than conventional models.

LT2–3 *4:40 pm – 5:00 pm*

Acoustic Reflection Path Tracing Using a Highly Directional Loudspeaker

Sakari Tervo, Jukka Pätynen, and Tapio Lokki, *Helsinki University of Technology, Finland*

A room impulse response measurement setup is proposed for the analysis of reflections. An omnidirectional loudspeaker that is traditionally used in acoustic measurements is replaced with a highly directional loudspeaker and which is rotated over a set of angles. The approach is shown to have benefits, as reflections can be found later in time than with traditional techniques. Secondly, it is shown that the proposed approach improves the separability of individual reflections in comparison to omnidirectional loudspeaker. This is achieved, since the impulse response of the proposed approach is spatially more separable.

LT2–4 *5:00 pm – 5:20 pm*

Feature Selection for Room Volume Identification from Room Impulse Response

Noam R. Shabtai, Yaniv Zigel, and Boaz Rafaely, *Ben-Gurion University of the Negev, Israel*

The room impulse response (RIR) can be used to calculate many room acoustical parameters, such as the reverberation time (RT). However, estimating the room volume, another important room parameter, from the RIR is typically a more difficult task requiring the use of other features from the RIR. Most of the existing fully-blind methods for estimating the room volume from the RIR do not combine features from different feature sets. This can be one reason to the fact that these methods are sensitive to differences in source-to-receiver distance and wall reflection coefficients. We propose a new approach in which hypothetical-volume room models are trained with room volume features from different feature sets. Estimation is performed by identifying the hypothesis with maximum likelihood (ML) using background model normalization. The different feature sets are compared using equal error rate (EER) of hypothesis verification. A combination of features from the different feature sets is selected so that minimum EER is achieved. Using the selected features, we achieve average detection rate of 98.8% with a standard deviation (STD) of 1.5% for eight rooms with different volumes, source-to-receiver distances, and wall reflection coefficients.

LT2–5 *5:20 pm – 5:40 pm*

Parsimonious Sound Field Synthesis Using Compressive Sampling

Georgios N. Lilis, *Cornell University, USA*, Daniele Angelosante and Georgios B. Giannakis, *University of Minnesota, USA*

Reproducing a sampled sound field over a two-dimensional area using an array of loudspeakers is a problem with well appreciated applications to acoustics and ultrasound treatment. Loudspeaker signal design has traditionally relied on a (possibly regularized) least-squares criterion. The fresh look advocated here, permeates benefits from advances in variable selection and compressive sampling by casting the sound field synthesis as a sparse linear regression problem that is solved by the least absolute shrinkage and selection operator (Lasso). Analysis and simulations demonstrate that the novel approach exhibits superb performance even for under-sampled sound fields, where least-squares methods yield inconsistent field reproduction. In addition, Lasso-based synthesis enables judicious placement of parsimonious loudspeaker arrays.

LT2–6 *5:40 pm – 6:00 pm*

Regularized HRTF Fitting Using Spherical Harmonics

Dmitry N. Zotkin, Ramani Duraiswami, and Nail A. Gumerov, *University of Maryland, USA*

By the Helmholtz reciprocity principle, the head-related transfer function (HRTF) is equivalent to an acoustic field created by a transmitter placed at the ear location. Therefore, it can be represented as a spherical harmonics spectrum – a weighted sum of spherical harmonics. Such representations are useful in theoretical and computational analysis. Many different (often severely undersampled) grids are used for HRTF measurement, making the spectral reconstruction difficult. In this paper, two methods of obtaining the spectrum are presented and analyzed both on synthetic (ground-truth data available) and real HRTF measurements.

6:00 pm – 8:00 pm

Dinner (West Dining Room)

KEYNOTE ADDRESS — WEDNESDAY 8:00 AM – 8:45 AM

Scalable Audio Coding for Heterogeneous Networks

Bastiaan Kleijn
KTH School of Engineering, Sweden

The increasingly heterogeneity of communication networks requires audio coders that can adapt instantaneously to both changing rate constraints and changing channels. This talk discusses a methodology aimed at always having the right coder configuration for the scenario at hand.

To facilitate adaptivity, it is advantageous to have statistical models of all components in a coding process: the source, the encoder, the channel, the decoder, and the receiver. The model parameters for the source and channel are estimated and the receiver is described by an auditory model. Given the source, channel, and auditory models for a signal segment, the parameters for the encoder and decoder are computed to maximize the performance perceived by the receiver, while satisfying network and user constraints. The approach implies that analytic relations must replace empirical findings. We briefly introduce high-rate quantization theory, on which derivations of such relations can be based. We then discuss a number of specific techniques. We derive the optimal rate distribution between the signal model and the signal given the model. We design quantizers that allow trading rate variation against distortion variation. We describe a scalable multiple-description coding (MDC) method with an arbitrary number of descriptions that allows the instantaneous computation of the optimal MDC configuration for any packet loss rate. We conclude by describing the architecture and performance of a complete system based on these principles.

Biosketch

Bastiaan Kleijn is a Professor at the School of Electrical Engineering at KTH (the Royal Institute of Technology) in Stockholm, Sweden and Head of the Sound and Image Processing Laboratory. He is also a co-founder of Global IP Solutions, where he remains Chief Scientist. He holds a M.S. and Ph.D. in Electrical Engineering from Stanford and Delft University of Technology, respectively, and a Ph.D. in Soil Science and an M.S. in Physics from the University of California. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996 and has since worked at KTH. Visiting positions he has held include Delft University of Technology and Vienna University of Technology. He has been and is on the Editorial Board of five journals in the general area of signal processing. He is a Fellow of the IEEE.

Lecture Session — LW: Wednesday 8:45 am – 10:20 am

Speech and Audio Coding

Session Chair:
Peter Kroon
Infineon Technologies, USA

LW–1 *8:45 am – 9:05 am*

Temporal Quantization of Spatial Information Using Directional Clustering for Multichannel Audio Coding

Shigeki Miyabe, *The University of Tokyo, Japan*, Keisuke Masatoki, Hiroshi Saruwatari, and Kiyohiro Shikano, *Nara Institute of Science and Technology, Japan*, Toshiyuki Nomura, *NEC Corporation, Japan*

Binaural cue coding, which is a representing low bit-rate coding of multichannel audio, generates large distortion when the audio data have complex spatial information, such as symphony. Such distortion caused by the low frequency resolution of spatial information because BCC quantizes the parameters of localization. In this paper we propose a new coding framework by quantizing the spatial information temporally. The single-channel sum signal is panned to the multiple channels by selecting the prototypes of the spatial filter. Optimization of the prototypes with minimal coding error is given by a k-means-like clustering of the angles whose centroids are given by the first principal components of the covariances in the classes. The efficiency of the proposed coding with high quality is verified both in the objective and subjective evaluations.

LW–2 *9:05 am – 9:25 am*

ITU-T G.719: A New Low-Complexity Full-Band (20 kHZ) Audio Coding Standard for High-Quality Conversational Applications

Minjie Xie and Peter Chu, *Polycom Inc., USA*, Anisse Taleb and Manuel Briand, *Ericsson Research – Multimedia Technologies, Sweden*

This paper describes a new low-complexity full-band (20 kHz) audio coding algorithm which has been recently standardized by ITU-T as Recommendation G.719. The algorithm is designed to provide 20 Hz - 20 kHz audio bandwidth using a 48 kHz sample rate, operating at 32 - 128 kbps. This codec features very high audio quality and low computational complexity and is suitable for use in applications such as videoconferencing, teleconferencing, and streaming audio over the Internet. After an overview of the ITU-T G.719 codec, two important modules of the codec - adaptive time-frequency transform and lattice vector quantization of transform coefficients are described in detail. Subjective test results from the Optimization/Characterization phase of G.719 are also summarized in the paper. It has been proven that the G.719 codec achieves transparent audio quality at 128 kbps.

LW–3 *9:25 am – 9:45 am*

IIR QMF-Bank Design for Speech and Audio Subband Coding

Heinrich Löllmann, Matthias Hildenbrand, Bernd Geiser, and Peter Vary, *RWTH Aachen University, Germany*

A new speech and audio codec has been submitted recently to ITU-T by a consortium of Huawei and ETRI as candidate proposal for the super-wideband and stereo extensions of ITU-T Rec. G.729.1 and G.718. This hierarchical codec with bit rates from 8-64 kbit/s relies on a subband splitting by means of a quadrature-mirror filter-bank (QMF-bank). For this, an allpass-based QMF-bank is used whose design and implementation is presented in this contribution. This IIR filter-bank allows to achieve a significantly lower signal delay in comparison to the traditional FIR QMF-bank solution without a compromise for the speech and audio quality.

LW–4 *9:45 am – 10:05 am*

Nested Microphone Array Processing for Parameter Estimation in Directional Audio Coding

Giovanni Del Galdo, Oliver Thiergart, and Fabian Küch, *Fraunhofer Institute for Integrated Circuits IIS, Germany*

Directional Audio Coding (DirAC) is an efficient technique to capture and reproduce spatial sound on the basis of a downmix signal and side information, i.e., direction of arrival and diffuseness of the sound field expressed in time-frequency domain. The main drawback of using arrays of omnidirectional microphones to obtain these parameters is that reliable estimates are available only in a certain frequency range, which depends on the aperture. To overcome this problem and cover large audio bandwidths nested microphone arrays can be used. In this paper, we derive optimal estimators which combine the microphone signals to achieve the best possible estimation of the DirAC parameters with respect to the mean squared error.

LW–5 *10:05 am – 10:25 am*

Coding of Spatio-Temporal Audio Spectra Using Tree-Structured Directional Filterbanks

Francisco Pinto and Martin Vetterli, *Ecole Polytechnique Fédérale de Lausanne, Switzerland*

We address the problem of integrating directional analysis of sound into the filterbank of a spatial audio coder, with the purpose of processing and coding with some degree of independence the plane waves traveling in different directions. A plane wave represents an elementary waveform in the spatio-temporal analysis of the sound field, the same way a complex exponential is an elementary waveform in the time domain analysis of signals. Since a two-dimensional separable filterbank is not flexible enough for this purpose, we propose a non-separable approach based on the quincunx filterbank with diamond-shaped filters, cascaded with a base transform filterbank. This solution provides an invertible and critically sampled decomposition of the spatio-temporal spectra into subbands representing the different directions of wave propagation.

10:25 am – 10:40 am

Coffee Break (East Dining Room)

Poster Session — PW: Wednesday 10:40 pm – 12:20 pm

Spatial Sound Perception and Reproduction Speech and Audio Coding

Session Chair:
Shoji Makino
University of Tsukuba, Japan

PW–1 Perfect Sequence LMS for Rapid Acquisition of Continuous-Azimuth Head Related Impulse Responses

Christiane Antweiler, *RWTH Aachen University, Germany*, Gerald Enzner, *Ruhr-University Bochum, Germany*

In recent publications, continuous-azimuth inference of head related impulse responses (HRIRs) was treated as a time-varying system identification problem on the basis of dynamical measurements. The system identification thus can be handled by LMS-type adaptive filters for which we have the freedom to choose the excitation signal in this application. In order to provide the perspective of reducing the measurement time to a minimum, we now suggest the optimal excitation signal in terms of the rate of convergence. This excitation signal is given by perfect sequences (PSEQs) out of the larger family of periodic pseudo-noise signals. After the discussion of specific implications of perfect sequences, we compare the performances of our perfect-sequence LMS algorithm (PSEQ-LMS) to the results of white noise processing. We demonstrate a uniform improvement by PSEQ-LMS in terms of instrumental mean-square error analysis as well as subjective listening to dynamic HRIRs. Both measures turn out to be consistent.

PW–2 Diffuseness Estimation Using Temporal Variation of Intensity Vectors

Jukka Ahonen and Ville Pulkki, *Helsinki University of Technology, Finland*

In the energetic analysis of sound field, the diffuseness of it is traditionally estimated as a ratio between magnitude of the sound intensity vector and energy density, which represent the net flow and the total amount of the energy in one point of sound field, respectively. In this article, a novel method for estimating diffuseness in energetic analysis is presented, where the temporal variation of sound intensity vectors is utilized. In proposed temporal-variation-based method, the length of time-averaged time intensity vectors is divided with the time-averaged length of the same vectors. Temporal-variation-based method was evaluated, and the results are presented in this article. Evaluation is performed in the context of a spatial sound processing technique, Directional Audio Coding (DirAC), in which the directional information of the sound field, direction-of-arrival (DOA) and diffuseness, is analyzed from microphone signals at frequency bands and in short time windows using the energetic analysis. In the synthesis phase, the audio is actively steered according to the analyzed DOA and diffuseness. In teleconferencing application of DirAC, only one- or two-dimensional representation of the sound field is typically available instead of three-dimensional representation. However, the traditional method to compute diffuseness, which is utilized so far in DirAC analysis, is intended to have three-dimensional information about sound field, and diffuseness is thus estimated more or less wrongly with one- or two-dimensional representations. The issues with the traditional method are also discussed in this article. In the presented evaluation, it is shown that the proposed temporal-variation -based diffuseness computation produces more reliable diffuseness estimation than the traditional method in two-dimensional, and especially in one-dimensional sound field analysis. This is an important result, as typical domestic stereophonic microphones can only deliver one-dimensional representation of sound field, and the new diffuseness formula enables the use of such microphones in the telecommunication applications of DirAC.

PW–3 Estimating Pressure at Eardrum with Pressure-Velocity Measurement from Ear Canal Entrance

Marko Hiipakka, Matti Karjalainen, and Ville Pulkki, *Helsinki University of Technology, Finland*

There are many applications where it would be important to know the sound pressure at the eardrum. Unfortunately, direct measurement is possible only in special studies where extreme care can be taken not to harm the fragile membrane or other parts of the ear canal. For example audiological applications and particularly in binaural sound reproduction such direct measurements are in most cases out of the question. Therefore, sound signals have to be measured in other points of the ear canal, or often at the open or blocked canal entrance. Then the problem is to use such these measurements to estimate the sound pressure at the eardrum, because it is the basic reference point for example in binaural headphone reproduction using head-related transfer functions or binaural recordings. This study shows that when both the sound pressure and the velocity are measured at the canal entrance using a pressure-velocity probe, the pressure signal at the eardrum can be estimated with much higher accuracy than from the pressure-only measurement. The method is demonstrated and validated by using physical simulators and computational modeling. A micromachined particle velocity transducer that makes volume velocity measurements possible has recently become available. We are suggesting the use of this device for ear canal measurements as a new approach to estimate the eardrum pressure. Measurements were realized with the Microflown PU Match Probe, a custom-made ear canal simulator and custom made a dummy head. The pressure frequency responses at the eardrum of the ear canal simulator and the dummy head were measured in free field conditions in an anechoic chamber. The pressure and velocity frequency responses at the entrance were measured directly after the eardrum pressure had been measured. The obtained velocity and pressure frequency responses were used to calculate estimates of the eardrum pressure of the simulator with different canal lengths. The estimated and measured pressure frequency responses at the eardrum agree well with different canal lengths and different eardrum impedances. Two different estimation methods were used. An acoustic transmission line-based model and a power-based estimation showed similar result. An important advantage with the power-based approach is that it gives accurate results even without the information of the canal length.

PW–4 HRTF Interpolation in the Wavelet Transfom Domain
Julio C. B. Torres and Mariane R. Petraglia, *Federal University of Rio de Janeiro, Brazil*

This paper presents a new HRTF (Head Related Transfer Function) interpolation technique for three-dimensional sound generation. The proposed approach is based on the determination of the optimal weights to be applied to the HRTF coefficients of neighboring positions in the wavelet domain in order to obtain the HRTF at a given point. The proposed method is compared to conventional interpolation methods through the error analysis of the HRTFs in the frequency and time domains. It is verified that the proposed method presents smaller interpolation errors for all HRIRs (Head Related Impulse Responses) of an available database.

PW–5 Sound Texture Synthesis Via Filter Statistics
Josh H. McDermott and Eero P. Simoncelli, *New York University, USA*, Andrew J. Oxenham, *University of Minnesota, USA*

Many natural sounds, such as those produced by rainstorms, fires, or insects at night, result from large numbers of rapidly occurring acoustic events. We hypothesize that humans encode these 'sound textures' with statistical measurements that represent their constituent features and the relationship between them. We explore this hypothesis using a synthesis algorithm that measures statistics in a real sound and imposes them on a sample of noise. Simply matching the marginal statistics (variance, kurtosis) of individual frequency subbands was generally necessary, but insufficient, to yield good results. Imposing various joint envelope statistics (correlations between bands, and autocorrelations within each band) greatly improved the results, frequently producing synthetic textures that sounded natural and that subjects could reliably recognize. The results suggest that such statistical representations could underlie sound texture perception, and that the auditory system may use fairly simple statistics to recognize many natural sound textures.

PW–6 An Overall Optimization Method for Arbitrary Sample Rate Converters Based on Integer Rate SRC and Lagrange Interpolation
Andreas Franck and Karlheinz Brandenburg, *Fraunhofer IDMT, Germany*

Combinations of integer rate sample rate conversion and polynomial interpolators are widely used structures for arbitrary or asynchronous sample rate conversion (ASRC). However, since these two components are typically designed independently, the performance of the structure is not optimal with respect to a given error norm such as weighted least squares or Chebyshev norms. To overcome this drawback, we propose a design method for this class of ASRC filters that enables an optimization of the whole structure. For this reason, an analytical description of the continuous frequency response of the system, derived from a closed formula for the continuous frequency response of Lagrange interpolators, is introduced. Based on this model, we propose a design method that minimizes the error in the Chebyshev sense. ASRC filters designed according to this design procedure exhibit considerable reduced error norms compared to existing designs, whereas the improvement depends on several factors such as signal bandwidth, the prototype filter order, the integer interpolation ratio and the order of polynomial interpolation.

PW–7 Spectral HRTF Enhancement for Improved Vertical-Polar Auditory Localization
Douglas S. Brungart and Griffin D. Romigh, *Air Force Research Laboratory, USA*

Under ideal laboratory conditions, individualized head-related transfer functions HRTFs can produce virtual sound localization performance approaching the level achieved with real sound sources in the free field. However, in real-world applications of virtual audio, practical issues such as fitrefit variability in the headphone response and non-individualized HRTFs generally lead to much worse localization performance, particularly in the up-down and front-back dimensions. Here we present a new technique that 'enhances' the localizability of a virtual sound source by increasing the spectral contrast of the acoustic features that are relevant for spatial perception within a set of locations with nearly identical binaural cues i.e., a 'cone-of-confusion'. Validation experiments show that this enhancement technique can improve localization accuracy across a broad range of conditions, with as much as a 33% reduction in vertical-polar localization error for nonindividualized HRTFs measured on a KEMAR manikin; a 25% reduction in vertical-polar error for nonindividualized HRTFs measured on other human listeners; and a 33% reduction in vertical-polar error for individualized HRTFs presented under nonideal laboratory conditions i.e., with headphone fit-refit variability. These results suggest that the proposed technique could provide benefits across a wide range of real-world virtual audio display applications.

PW–8 Multizone 2D Soundfield Reproduction via Spatial Band Stop Filters
Yan Jennifer Wu and Thushara D. Abhayapala, *The Australian National University, Australia*

Any attempt to create multiple independent soundfields in separate zones over an extended region of open space results in unintended interference in a given zones from other zones. In this paper, we design spatial band stop filters to suppress interzone interference in the regions of interests and pass the desired soundfields with no distortion. This is achieved by using the higher order spatial harmonics of one zone to cancel the undesirable effects of the lower order harmonics of the same zone on the other zones. We illustrate the work by designing and simulating a 2D two-zone soundfield.

PW–9 Soundfield Rendering with Loudspeaker Arrays Through Multiple Beam Shaping
Fabio Antonacci, Alberto Calatroni, Antonio Canclini, Andrea Galbiati, Augusto Sarti, and Stefano Tubaro, *Politecnico di Milano, Italy*

This paper proposes a method for the acoustic rendering of a virtual environment based on a geometric decomposition of the wavefield into multiple elementary acoustic beams, all reconstructed with a loudspeaker array. The point of origin, the orientation and the aperture of each beam is computed according to the geometry of the virtual environment that we want to render and to the location of the sources. Space-time filters are computed with a Least Squares approach to render the desired beam. Experimental results show the feasibility as well as the critical issues of the proposed algorithm.

PW–10 Wave Field Analysis Using Multiple Radii Measurements

Achim Kuntz and Rudolf Rabenstein, *University of Erlangen-Nuremberg, Germany*

Wave field analysis of stationary sound fields can be performed with high spatial resolution when sequential measurements on a circle are employed. Applying a circular harmonics decomposition to wideband audio data requires a careful setup of the measurement process to avoid noise amplification at certain frequencies. To this end, measurements with several microphones at multiple radii are considered here. Strategies from communication theory are used to combine their measurement signals for low noise amplification.

PW–11 Controlling a Spatialized Environmental Sound Synthesizer

Charles Verron and Grégory Pallone, *Orange Labs, France*, Mitsuko Aramaki and Richard Kronland-Martinet, *CNRS, France*

This paper presents the design and the control of a spatialized additive synthesizer aiming at simulating environmental sounds. First the synthesis engine, based on a combination of an additive signal model and spatialization processes, is presented. Then, the control of the synthesizer, based on a hierarchical organization of sounds, is discussed. Complex environmental sounds (such as a water flow or a fire) may then be designed thanks to an adequate combination of a limited number of basic sounds consisting in elementary signals (impacts, chirps, noises). The mapping between parameters describing these basic sounds and high level descriptors describing an environmental auditory scene is finally presented in the case of a rainy sound ambiance.

PW–12 3D-Continuous-Azimuth Acquisition of Head Related Impulse Responses Using Multi-Channel Adaptive Filtering

Gerald Enzner, *Ruhr-University Bochum, Germany*

A new method for fast acquisition of head related impulse responses (HRIRs) for 3D-continuous-azimuth representation of the auditory sphere is presented. While the continuous HRIR representation in the azimuth direction is important for the rendering of moving sources in binaural sound systems, an HRIR discretization in the elevation is more tolerable. Basically, the paper suggests a multi-channel adaptive system identification technique which is eventually suitable for online inference of the HRIRs. In a first step, several discrete elevations of the sphere are equipped with sound sources and a subject of interest is rotated in the center of the sphere during simultaneous reproduction of multi-channel probe noise. On the basis of the simultaneous binaural recording of all sources, the rotating HRIRs are then extracted gradually by the adaptive system identification. This concept overcomes discrete-azimuth HRIR tables and their traditional need for interpolation. While the quality of HRIRs was formerly expressed in terms of the azimuth-spacing of the HRIR, we now feature quasi-infinite resolution.

PW–13 A Perceptually Enhanced Scalable-to-Lossless Audio Coding Scheme and a Trellis-Based Approach for Its Optimization

Emmanuel Ravelli, Vinay Melkote, and Kenneth Rose, *University of California, USA*

Scalable-to-Lossless (SLS) audio compression, as standardized by MPEG, provides a lossy base layer compatible with the Advanced Audio Coding (AAC) format, ensuring state-of-the-art quality in the base layer, and additional fine grained enhancements that eventually provide a lossless compressed version of the signal. While SLS offers highly efficient lossless compression, the perceptual quality of its intermediate lossy layers has been observed to be suboptimal. This paper proposes a modified SLS audio coding scheme that provides enhanced perceptual quality at an intermediate bit-rate, at the expense of an additional parameter per frequency band as side-information. This scheme when coupled with a trellis-based optimization algorithm is demonstrated to outperform, in terms of quality at the intermediate bit-rate, both standard SLS and a recent perceptually enhanced variant, with minimal degradation in lossless coding performance.

PW–14 Parametric AM/FM Decomposition for Speech and Audio Coding

Tom Bäckström, *Fraunhofer Institute for Integrated Circuits (IIS), Germany*, Sascha Disch, *Leibnitz Universität Hannover, Germany*

Audio transform coding methods generally employ an assumption of locally stationary signals. However, as demands on coding efficiency increase, the assumption of stationarity becomes increasingly challenging. In this article we study a parametric method for modeling variability in an audio signal, especially with voiced speech in mind. Namely, we model modulation in fundamental frequency and amplitude, with the objective of allowing normalization of the signal to satisfy the stationarity requirements of transform based coders.

PW–15 Binaural Reproduction for Directional Audio Coding

Mikko-Ville Laitinen and Ville Pulkki, *Helsinki University of Technology, Finland*

A recently proposed method for spatial sound reproduction is Directional Audio Coding (DirAC). In DirAC spatial sound is recorded typically using a B-format microphone, and reproduced using an arbitrary number of loudspeakers. The direction and diffuseness is analyzed in frequency bands depending on time, forming the transmitted metadata. The underlying assumption is, that at one time instant and at one critical band the spatial resolution of auditory system is limited to decoding one cue for direction and another for inter-aural coherence. The reproduction principle is that the sound in one frequency band is simply presented with two cross-fading streams: a non-directional diffuse stream, and a directional non-diffuse stream. A method for binaural reproduction for DirAC is developed in this article. The binaural realization of DirAC is simply based on existing loudspeaker-based method, however, each loudspeaker signal is reproduced over a virtual loudspeaker implemented with HRTFs measured from a human subject. A method to utilize head tracking information in binaural DirAC was also developed. In the method, the metadata is transformed to match the orientation of listener's head. Subjective listening tests were conducted in an anechoic room having a multichannel loudspeaker setup visible. Different realizations of binaural DirAC were tested, with and without head tracking. As anchors, monophonic and stereophonic techniques with and without head tracking were also used. The resulting MOS scores for overall quality and spatial quality for binaural DirAC were very high, receiving the highest grades. With head tracking, the listeners consistently externalized the sound scenes, and the spatial impression was plausible.

PW–16 Applications of Signal Analysis Using Autoregressive Models for Amplitude Modulation

Sriram Ganapathy, Samuel Thomas, and Hynek Hermansky, *Johns Hopkins University, USA*, Petr Motlicek, *Idiap Research Institute, Switzerland*

Frequency Domain Linear Prediction (FDLP) represents an efficient technique for representing the long-term amplitude modulations (AM) of speech/audio signals using autoregressive models. For the proposed analysis technique, relatively long temporal segments (1000 ms) of the input signal are decomposed into a set of sub-bands. FDLP is applied on each sub-band to model the temporal envelopes. The residual of the linear prediction represents the frequency modulations (FM) in the sub-band signal. In this paper, we present several applications of the proposed AM-FM decomposition technique for a variety of tasks like wide-band audio coding, speech recognition in reverberant environments and robust feature extraction for phoneme recognition.

PW–17 Theoretical and Practical Comparisons of the Reassignment Method and the Derivative Method for the Estimation of the Frequency Slope

Brian Hamilton and Philippe Depalle, *McGill University, Canada*, Sylvain Marchand, *SCRIME/LaBRI-CNRS, France*

In the context of non-stationary analysis for additive synthesis, the theoretical comparison of the reassignment method (RM) and the derivative method (DM) for the estimation of the frequency slope is investigated. It is shown that the DM differs from the RM in that it does not consider the group delay. Theoretical equivalence is shown to be possible with a refinement of the DM. The refinement is evaluated on synthetic signals and shown to improve the estimation of the frequency slope. The differences between the two methods in terms of window and signal constraints are discussed to show when each method is more appropriate to use.

PW–18 Sinewave Parameter Estimation Using the Fast Fan-Chirp Transform

Robert B. Dunn, Thomas F. Quatieri, and Nicolas Malyska, *MIT Lincoln Laboratory, USA*

Sinewave analysis/synthesis has long been an important tool for audio analysis, modification and synthesis. The recently introduced Fan-Chirp Transform (FChT) has been shown to improve the fidelity of sinewave parameter estimates for a harmonic audio signal with rapid frequency modulation. A fast version of the FChT reduces computation but this algorithm presents two factors that affect sinewave parameter estimation. The phase of the fast FChT does not match the phase of the original continuous-time transform and this interferes with the estimation of sinewave phases. Also, the fast FChT requires an interpolation of the input signal and the choice of interpolator affects the speed of the transform and accuracy of the estimated sinewave parameters. In this paper we demonstrate how to modify the phase of the fast FChT such that it can be used to estimate sinewave phases, and we explore the use of various interpolators demonstrating the tradeoff between transform speed and sinewave parameter accuracy.

PW–19 Realization of Arbitrary Filters in the STFT Domain

Michael Goodwin, *Audience, Inc., USA*

It is well known that FIR filters can be efficiently realized in the short-time Fourier transform domain, namely using sliding windows and FFTs. In some scenarios, however, the standard overlap-add or overlap-save block convolution algorithms may be impractical due to the need for large transforms to implement long filters. In this paper, we show how to realize arbitrary FIR and IIR filters in an STFT framework whose block size, hop size, and transform size are not constrained by the target filter response. First, we review a subband filtering approach which enables implementation of arbitrarily long FIR filters in such a fixed-size STFT system via an overlap-add decomposition of the target filter impulse response. We then extend the subband filtering framework to the case of IIR filtering: first, we show how to map a single-pole filter into single-pole subband filters; then, we show how to map any arbitrary IIR filter into a subband filtering scheme. These approaches are useful for applications such as filter realization in modular frequency-domain systems, filter implementation on restricted hardware, compressed-domain processing in frequency-domain coding systems, and subband adaptive filtering.

WASPAA'09 REVIEWERS

We wish to thank the following people for taking time out of their busy schedules to review papers for this workshop. The selection process would have been impossible without their help. Furthermore, their comments and suggestions for improvement were instrumental in the paper revision process, which helped the authors achieve uniformly high quality throughout this proceedings.

Sofiène Affes
Robert Aichner
Shoko Araki
Jacob Benesty
Jörg Bitzer
Michael Brandstein
Herbert Buchner
Michael Casey
Jingdong Chen
Eric Diethorn
Jacek Dmochowski
Scott Douglas
Ramani Duraiswami
Gary Elko
Daniel Ellis
Cumhur Erkut
Gianpaolo Evangelista
Christof Faller
Tomas Gaensler
Sharon Gannot
Nikolay Gaubitch
Ralf Geiger
Simon Godsill
Alberto Gonzalez
Michael Goodwin
Steve Grant
Emanuel Habets
Wolfgang Herbordt
Jürgen Herre
Robert Hoeldrich
James Hopgood
Yiteng (Arden) Huang
Jyri Huopaniemi
Fred Juang
Matti Karjalainen
Walter Kellermann
Youngmoo Kim
Anssi Klapuri
Bastiaan Kleijn
Dorothea Kolossa
Peter Kroon
Fabian Küch
Eric Lehmann
Yuangqing Lin
Zicheng Liu
Shoji Makino
Rainer Martin
Jens Meyer
Marc Moonen
Meinard Müller
Tomohiro Nakatani
Patrick Naylor
Philip Nelson
Ryouichi Nishimura
Sven Nordholm
Maurizio Omologo
Karen Payton
Stefan Petrausch
Mark Plumbley
Thomas Quatieri
Rudolf Rabenstein
Boaz Rafaely
Sean Ramprashad
Mark Sandler
Augusto Sarti
Hiroshi Saruwatari
Gerhard Schmidt
Gerald Schuller
Michael Seltzer
Malcolm Slaney
Paris Smaragdis
Mehrez Souden
Sascha Spors
Sriram Srinivasan
Akihiko Sugiyama
Ivan Tashev
Heinz Teutsch
Vesa Välimäki
Peter Vary
Wieslaw Woszczyk
Udo Zoelzer
Dmitry Zotkin

List of Authors