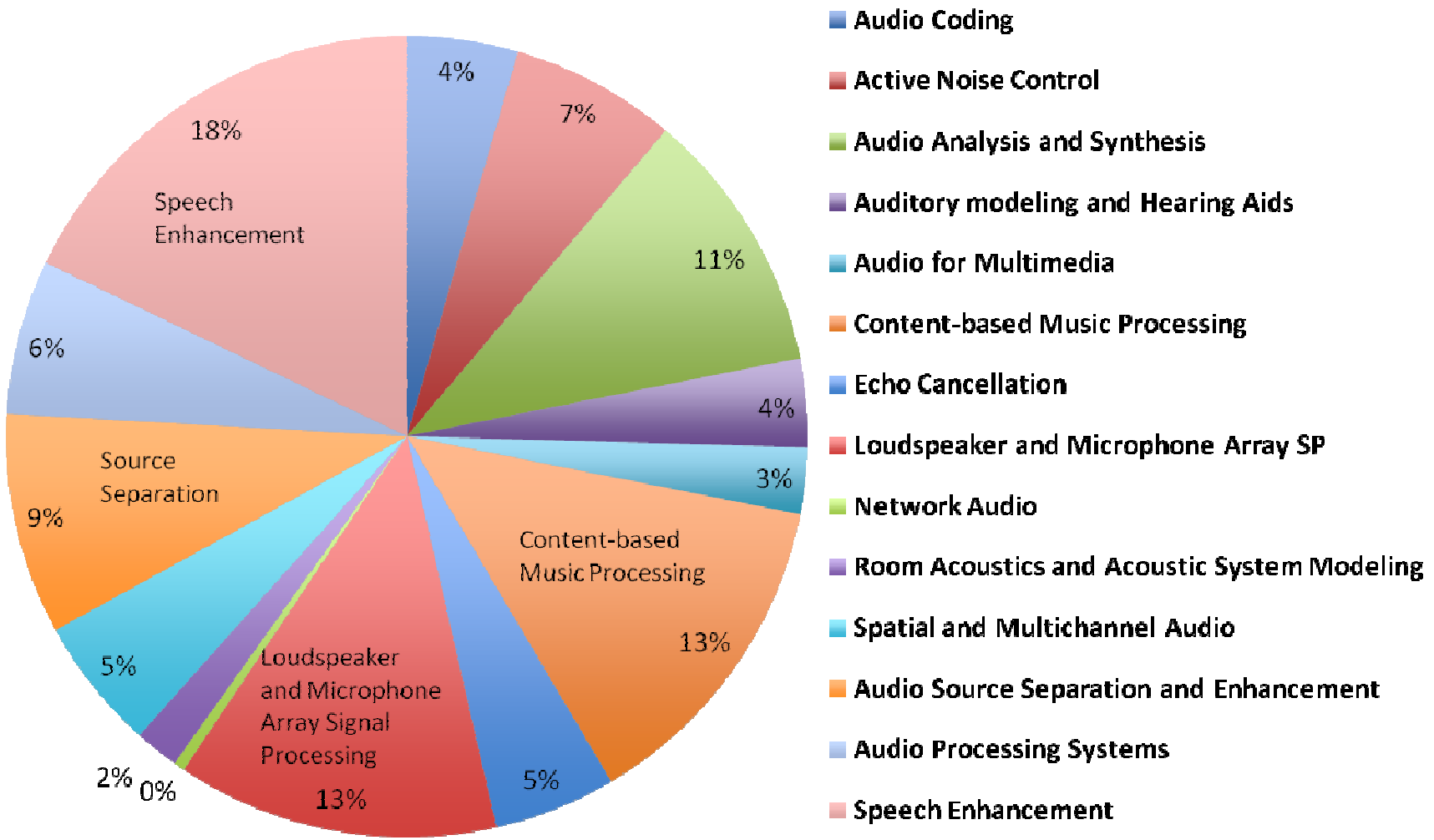
Trends
in
Audio and Acoustic Signal Processing

ICASSP 2011

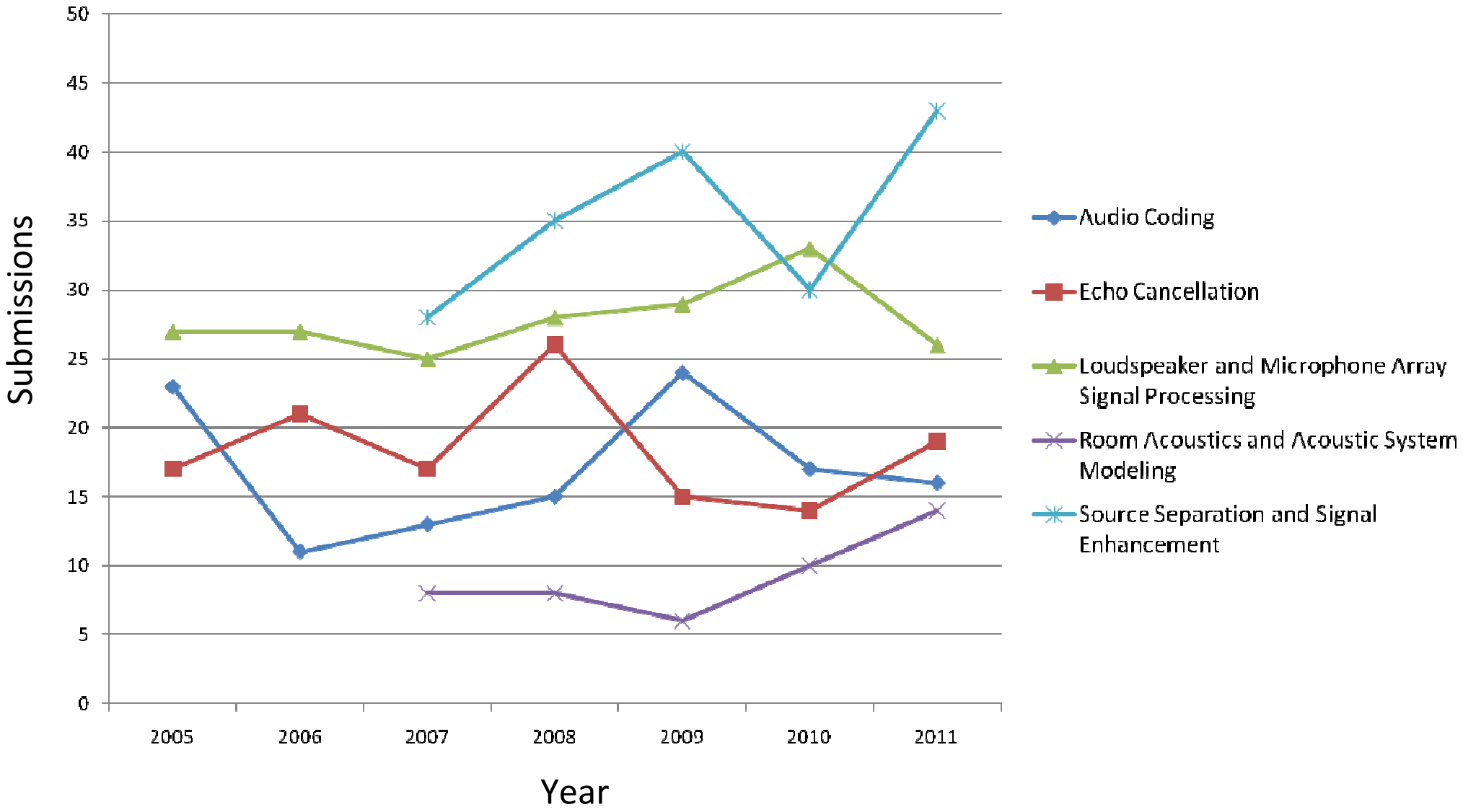
Malcolm Slaney, Yahoo! Research Silicon Valley, USA
Patrick A. Naylor, Imperial College London, UK

What do we mean by ‘Trends’?

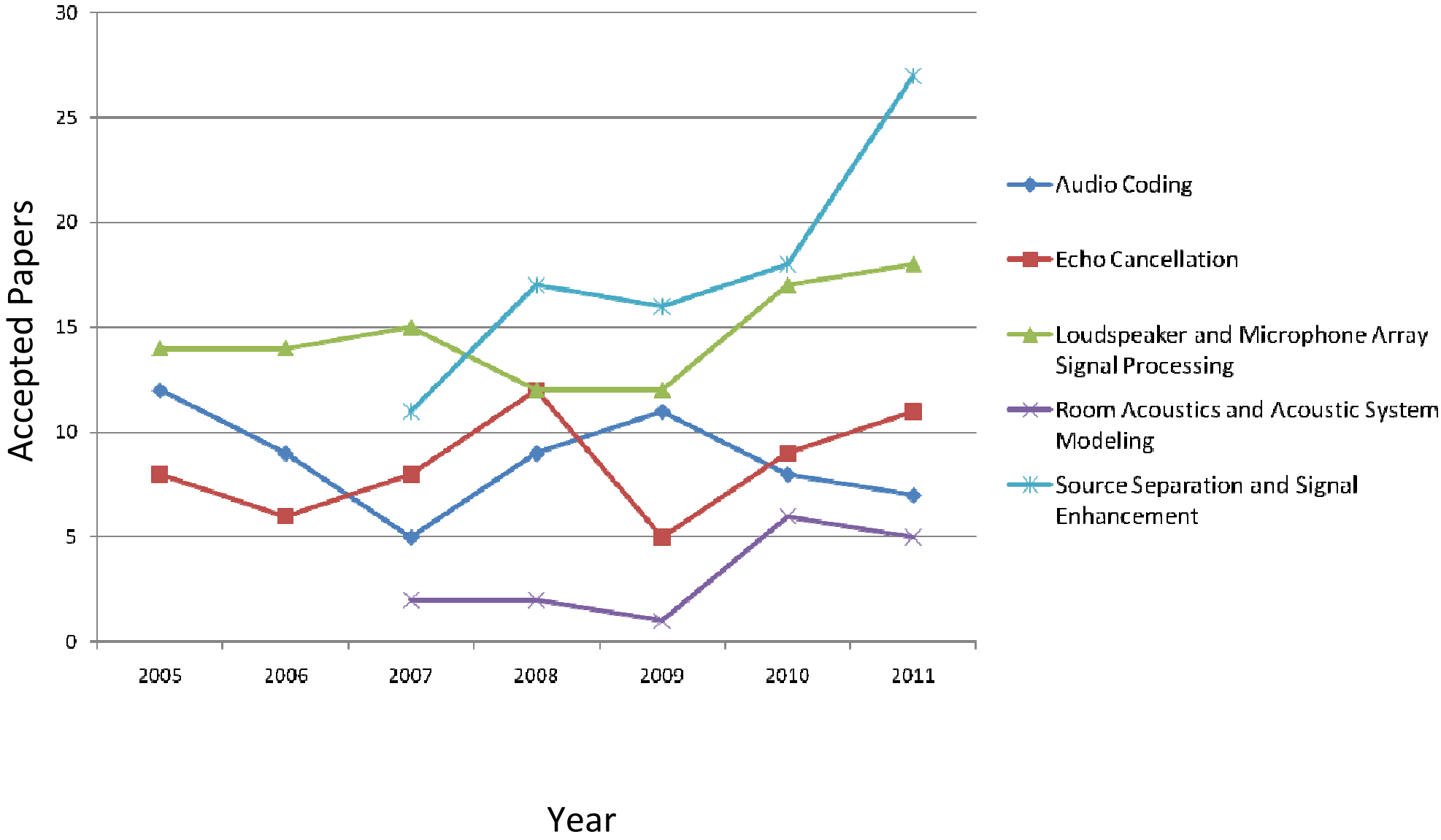
T-ASLP Submissions by EDICS, 2010



Historical Trends – ICASSP Submissions



Historical Trends – ICASSP Accepted Papers

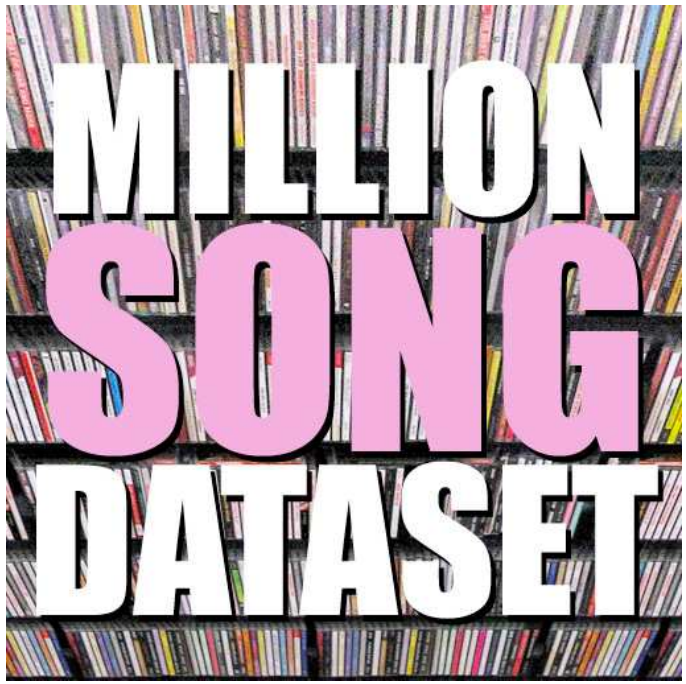


Music at ICASSP

- Three sessions
  - SS-L5: Music Signal Processing Exploiting Musical Knowledge
  - AE-L3: Music Signal Processing
  - AE-P7: Music Signal Processing
- Reasons
  - New EDICS
  - More content
  - Commercially relevant

Big Datasets

- Easy to rip CDs
  - Copyright issues
- Million-Song Dataset
  - Distribute features
  - Columbia and EchoNest



MIREX Competition

Audio Classification (Train/Test) Tasks

- –Audio Artist Identification
- –Audio US Pop Genre Classification
- –Audio Latin Genre Classification
- –Audio Music Mood Classification
- –Audio Classical Composer Identification

Audio Cover Song Identification

Audio Tag Classification

Audio Music Similarity and Retrieval

Symbolic Melodic Similarity

Audio Onset Detection

Audio Key Detection

Real-time Audio to Score Alignment (a.k.a Score Following)

Query by Singing/Humming

Audio Melody Extraction

Multiple Fundamental Frequency Estimation & Tracking

Audio Chord Estimation

Query by Tapping

Audio Beat Tracking

Structural Segmentation

Audio Tempo Estimation

Musical Separation

- Sound separation
  - Uses
    - Understanding (key, melody)
    - Transcriptions
    - Multipitch estimation
  - With better models
    - HMM
    - Scores
  - Techniques
    - NMF
    - Matching Pursuit
    - PLCA

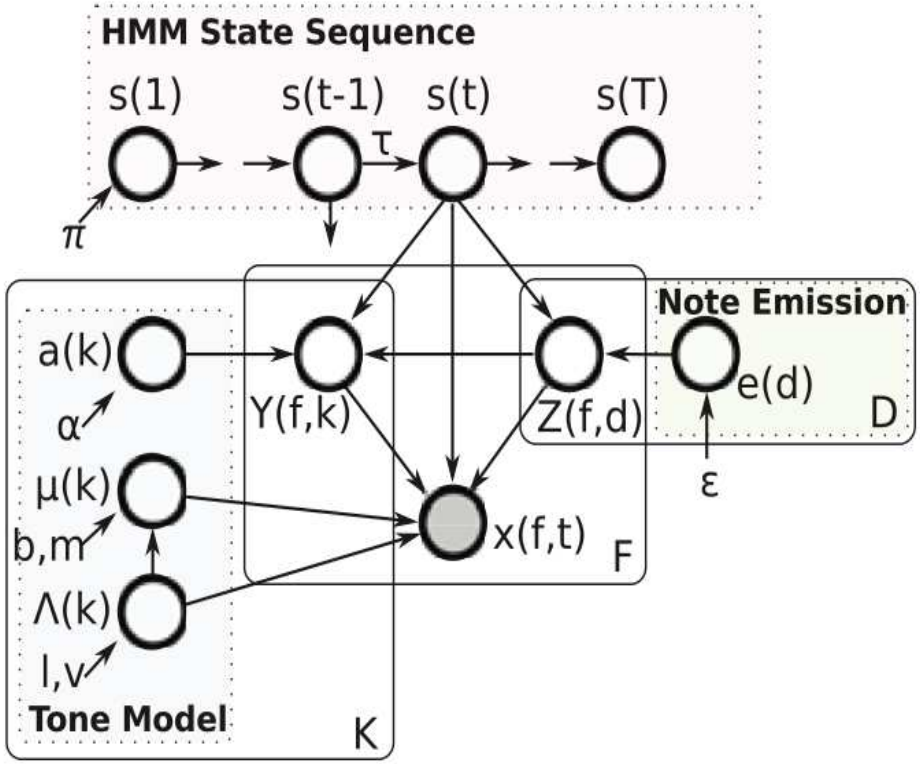
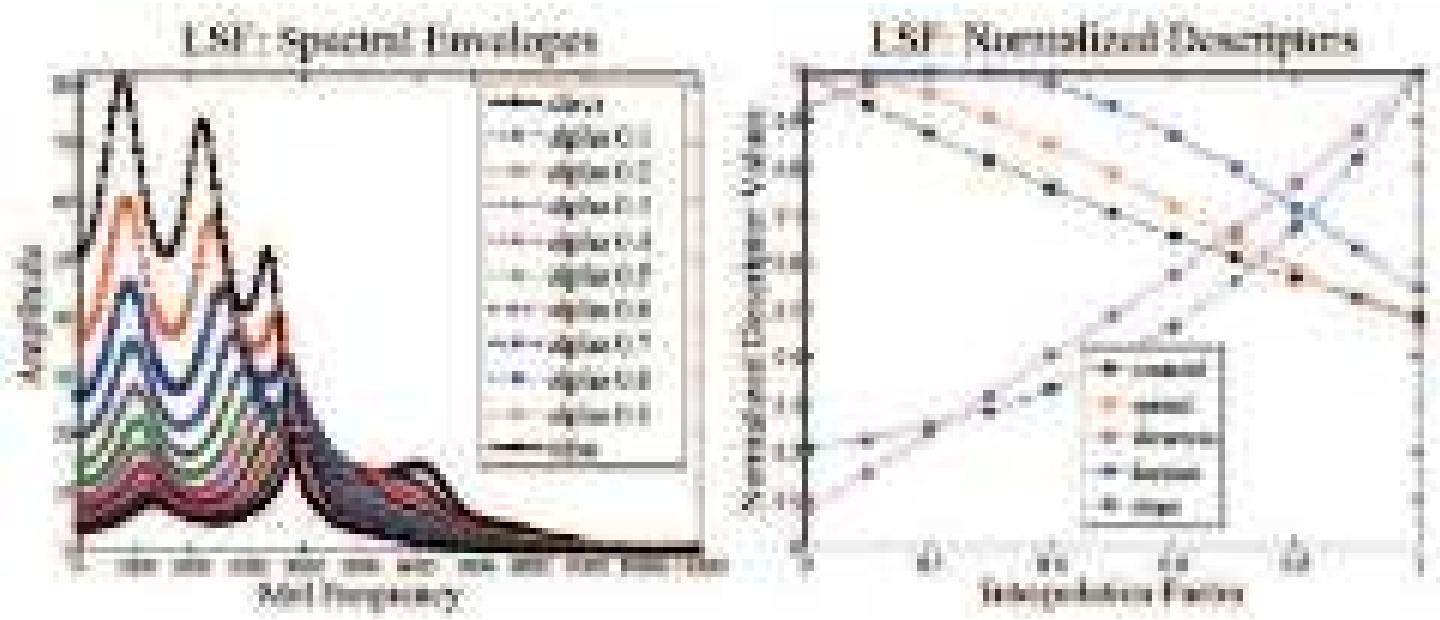


Fig. 1. Graphical model of our method.

From: POLYPHONIC AUDIO-TO-SCORE ALIGNMENT BASED ON BAYESIAN LATENT HARMONIC ALLOCATION HIDDEN MARKOV MODEL. Akira Maezawa, Hiroshi G. Okuno, Tetsuya Ogata, Kyoto University, Japan; Masataka Goto, National Institute of Advanced Industrial Science and Technology, Japan

Music Research

- Tagging
  - Genre
  - Emotion
- Miscellaneous
  - Morphing
  - Similarity



AASP-P2.1: SOUND MORPHING BY FEATURE INTERPOLATION, Marcelo Caetano, Xavier Rodet, Institut de Recherche et Coordination Acoustique/Musique, France

Applications vs. Algorithms?

If there is a trend towards things that look nice (applications), let’s not lose sight of the fundamental power behind them (algorithms).

Microphone Array Signal Processing

APPLICATIONS

- Hearing aids
- TV / Entertainment

GEOMETRY and DISTRIBUTION

- Linear, planar/cylindrical, spherical, distributed
- Spacing and orientation

TASKS

- Localization of sources
- Tracking
- Extraction/Separation
- Inference of room geometry

the Eigenmike®
– mh Acoustics

2 - 64 elements, 0.5 m, linear ‘wing’ array

skype

Features Get Skype Prices Accessories Business Support

[redacted] 3D LED [redacted] Series (40")

Call friends and family from the comfort of your couch

UN-40C7000

Enjoy widescreen Skype™ video calls from the comfort of your living room. With Skype on your [redacted]g TV you can make free Skype-to-Skype voice and video calls to friends and family, wherever they are in the world. Plus - you can also call landlines and cell phones from Skype at great rates.

Just plug in a Freetalk TV camera, sign in to Skype, and you're ready to talk face-to-face in HD.

This TV is compatible with: FREETALK TV Camera for [redacted] TVs.

Other Sizes

46" 55"

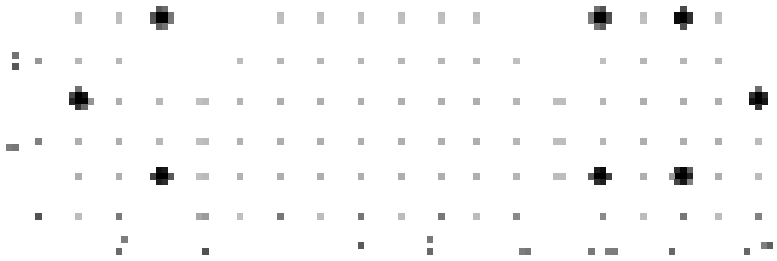
MICROPHONE POSITION OPTIMIZATION FOR PLANAR SUPERDIRECTIVE BEAMFORMING

Ina Kodrasi[1,2], *Thomas Rohdenburg*[2], *Simon Doclo*[1]

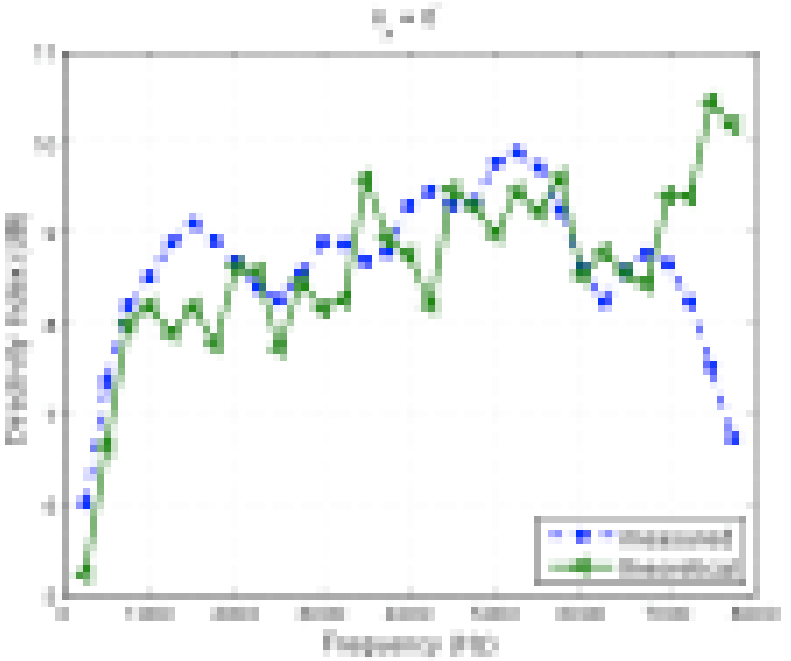
ina.kodrasi@uni-oldenburg.de

[1]University of Oldenburg, Institute of Physics, Signal Processing Group, Oldenburg, Germany

[2]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany



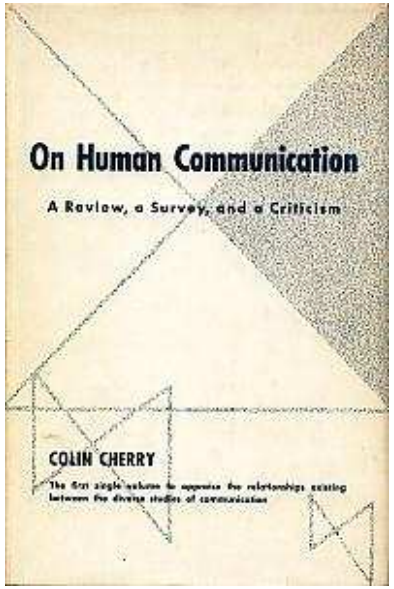
Planar Array Geometry



Directivity Index
measured vs theoretical

Source Separation

- ‘Cocktail party problem’
  - Colin Cherry in 1950s
  - Audio signals from multi-talker distant talking scenarios
  - Behavior of a listener presented with two speech signals simultaneously



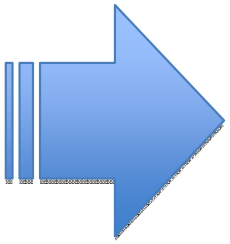
Colin Cherry
‘TRENDSETTER’

Source Separation

- Determined and underdetermined scenarios
  - Clustering based blind source separation
  - Permutation problem (EM)
  - Reverberation times of, say, 100 - 500 ms

Speech Enhancement

- Dereverberation technology

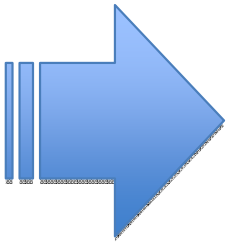


real-world applications

  - Single and multichannel
  - Acoustic channel inversion
  - Speech and Music

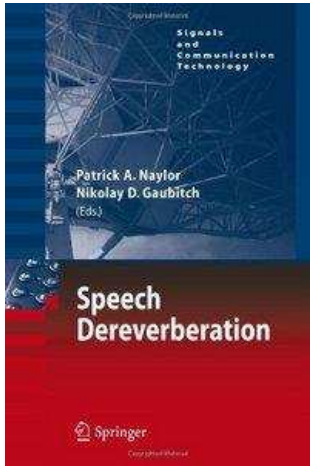
Speech Enhancement

- Dereverberation[1] technology



real-world applications

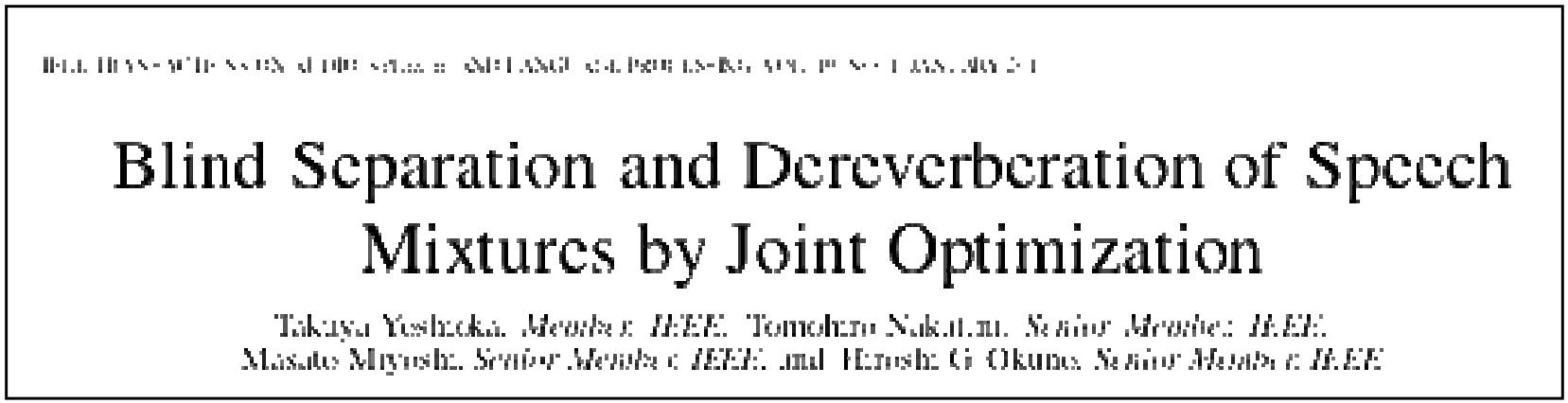
  - Single and multichannel
  - Acoustic channel inversion
  - Speech and Music



[1] P. A. Naylor and N. D. Gaubitch, Eds., Speech Dereverberation. Springer, 2010.

Synergies

- Joint dereverberation and blind source separation



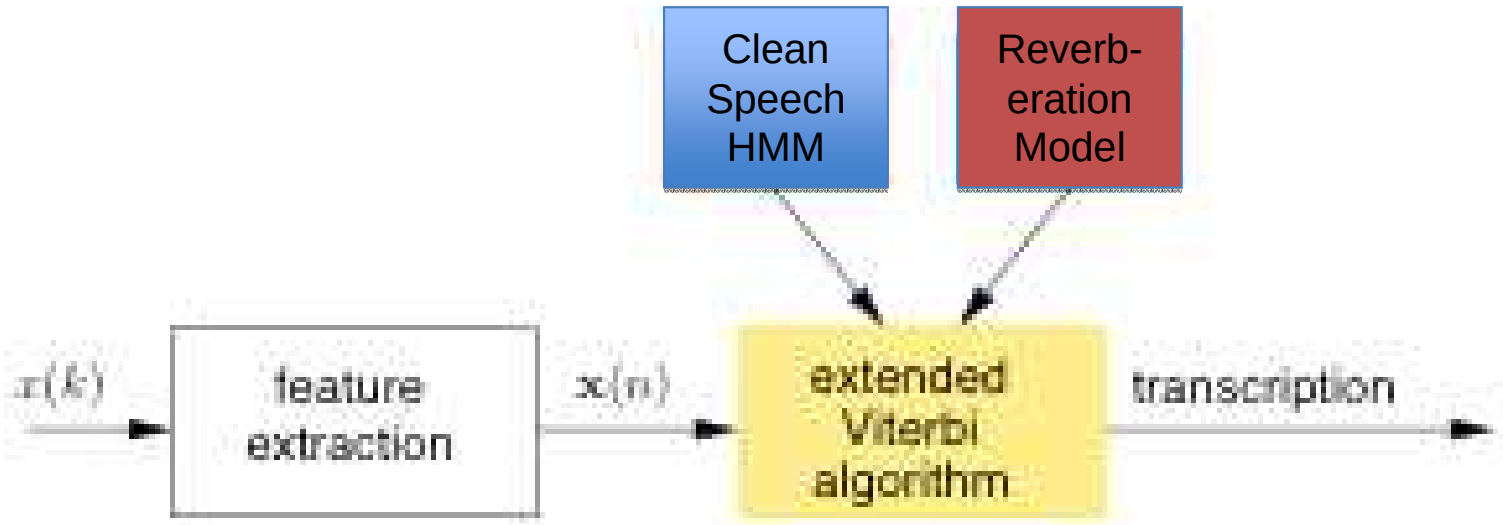
Blind Separation and Dereverberation of Speech Mixtures by Joint Optimization

Takuya Yoshioka, *Member, IEEE*, Tomohiro Nakatani, *Senior Member, IEEE*, Masato Miyoshi, *Senior Member, IEEE*, and Hiroshi G. Okuno, *Senior Member, IEEE*

- Speech recognition of reverberant speech

FRAME-WISE HMM ADAPTATION USING STATE-DEPENDENT REVERBERATION ESTIMATES

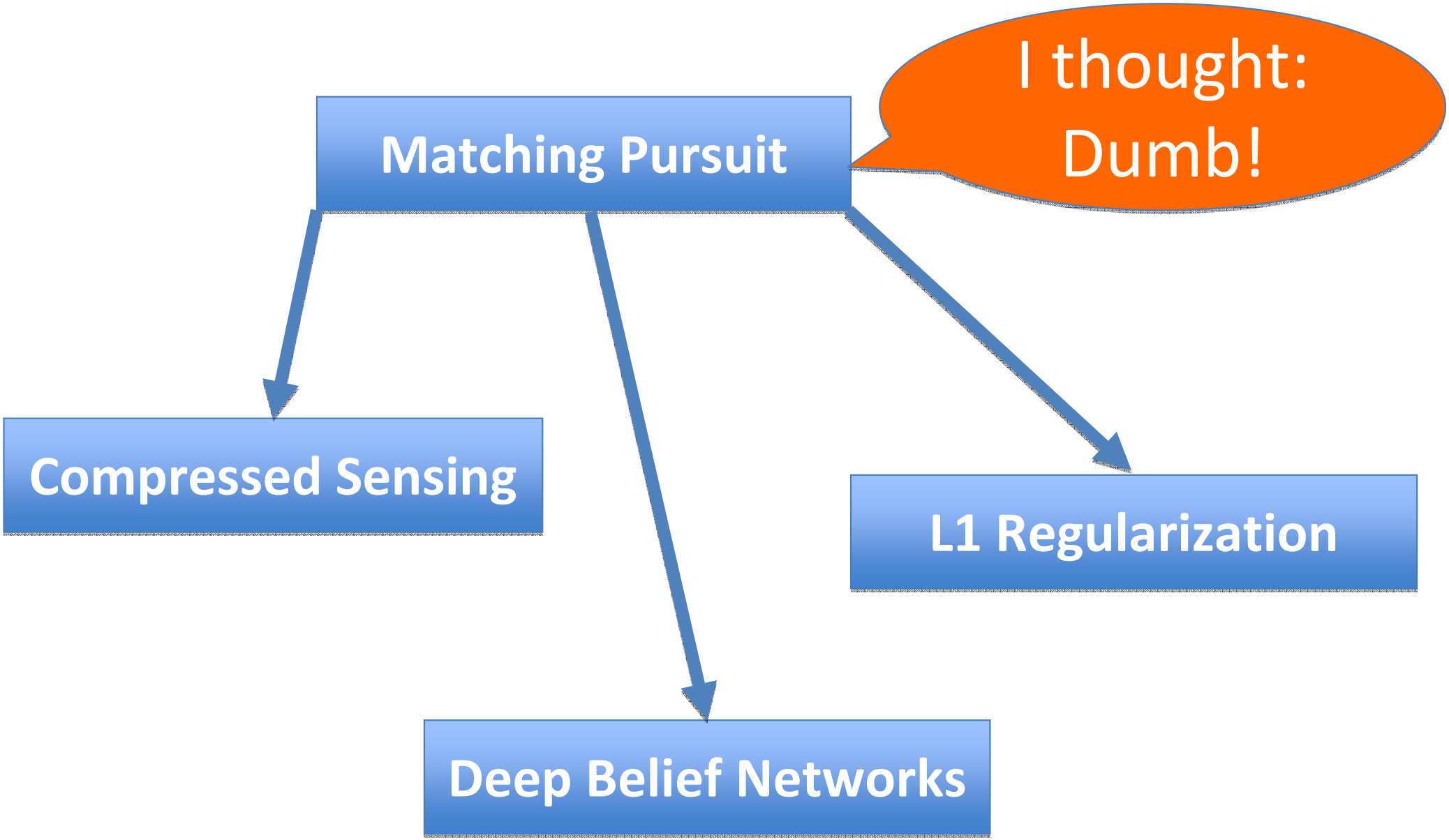
Armin Sehr, Roland Maas, and Walter Kellermann



Recurring Theme

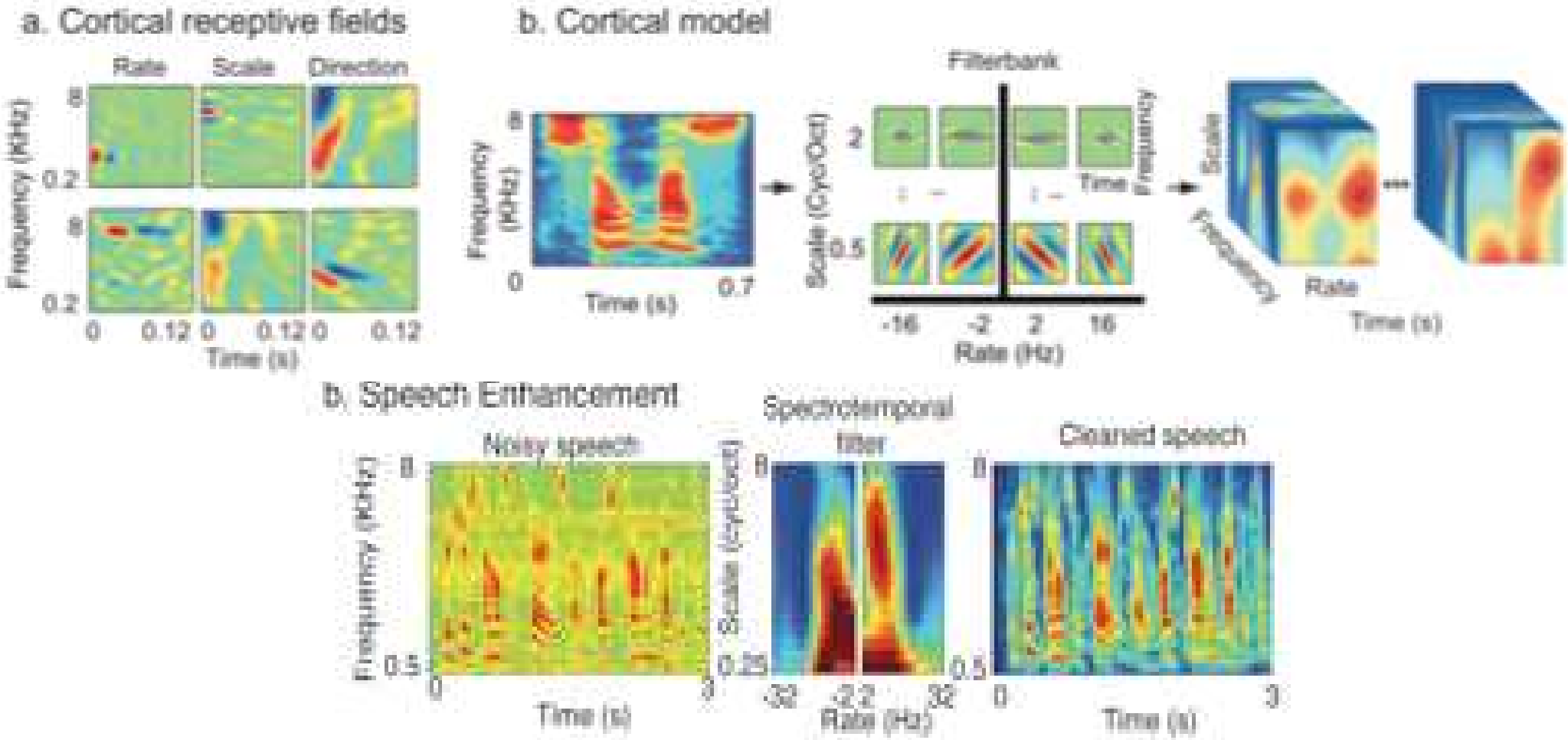
S P A R S I T Y !

History of Sparsity at ICASSP



Spatial-Temporal Receptive Fields

- Original sparse representation (spikes!)



Deep Belief Network

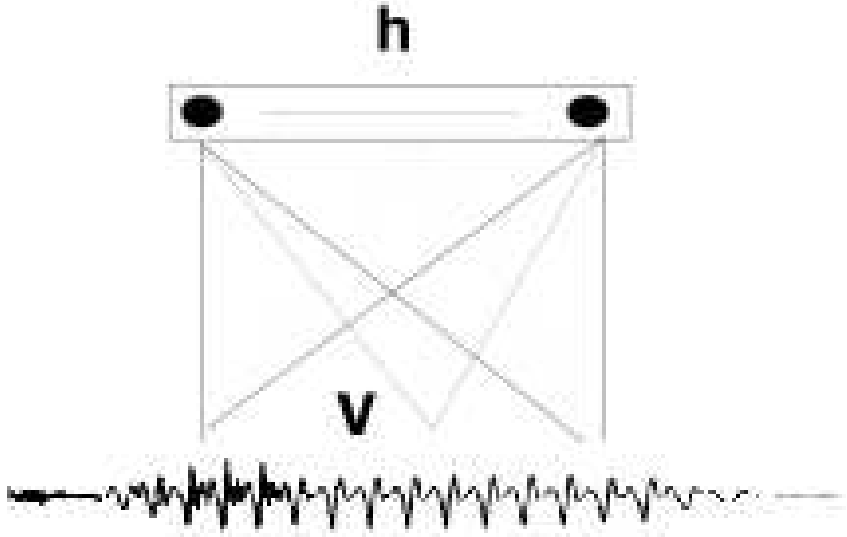
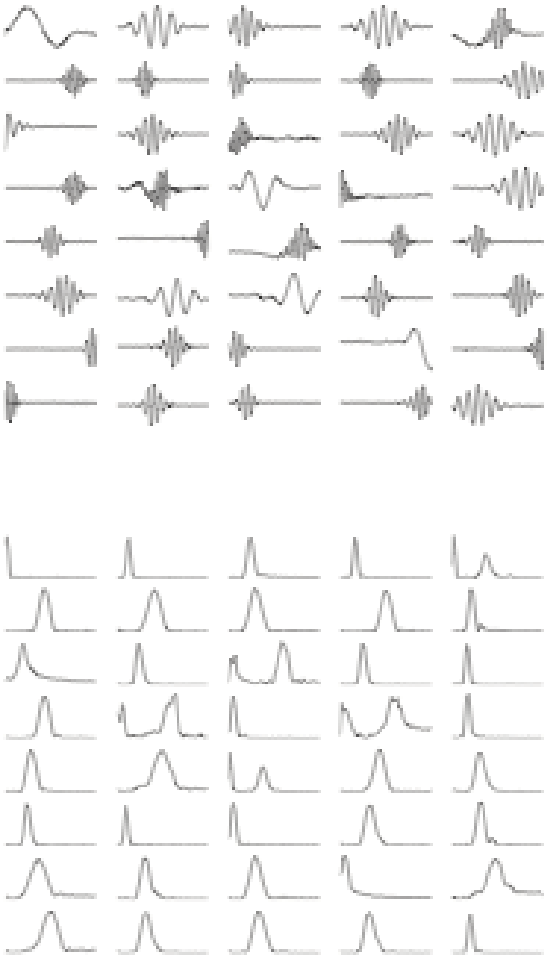


Fig. 1. RBM is used to model fragments of speech signals.

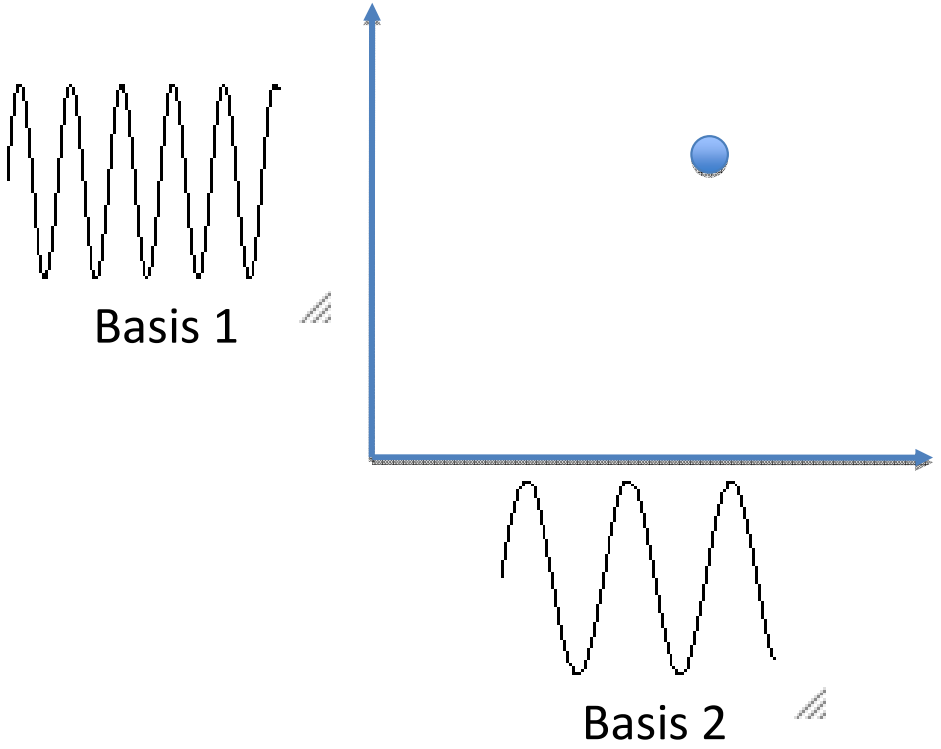


These are NOT your average wavelet/Gabor response!!

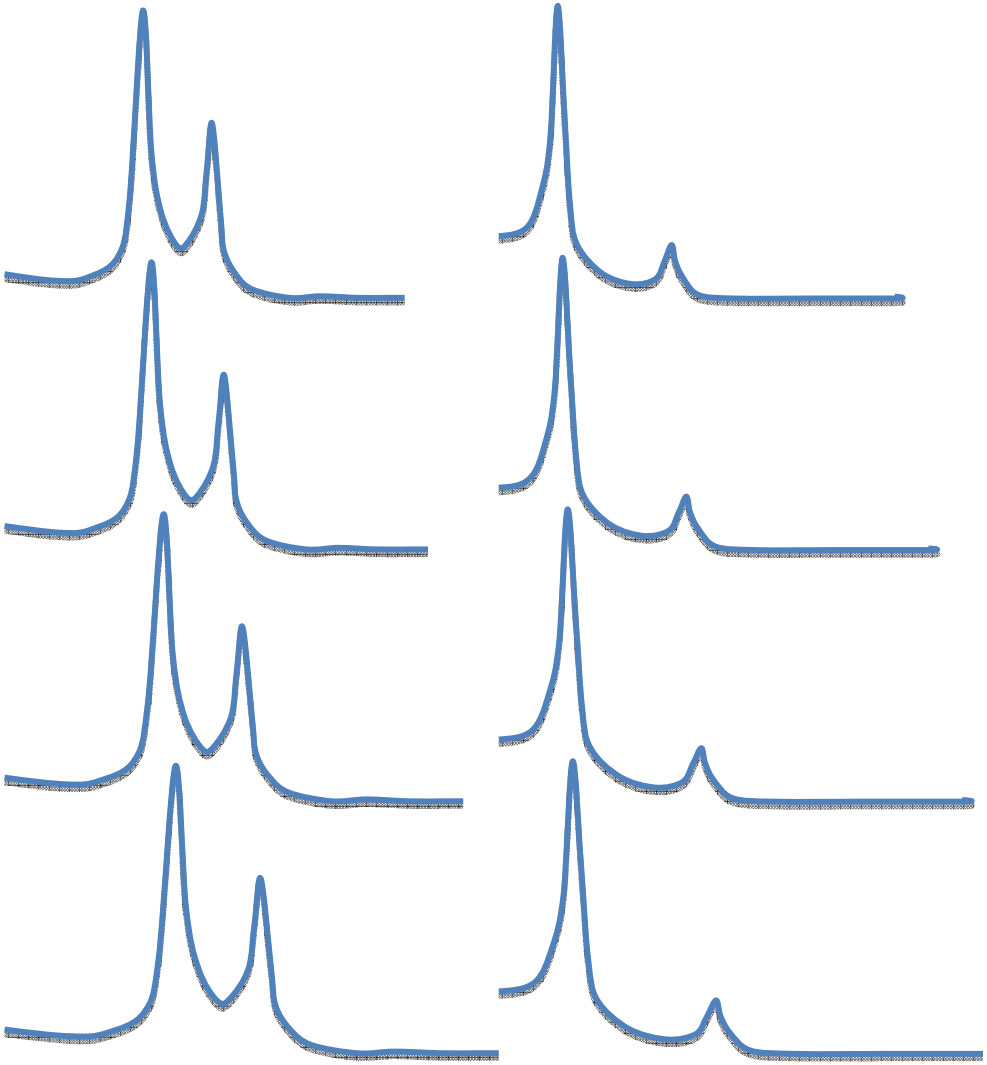
From: SS-L7.4: LEARNING A BETTER REPRESENTATION OF SPEECH SOUND WAVES USING RESTRICTED BOLTZMANN MACHINES, Navdeep Jaitly, Geoffrey Hinton, University of Toronto, Canada

Nonlinear Modeling via Sparsity

All-Possible Combinations



Over-complete Basis



Industrial Perspectives

“Remaining challenges [in source separation] could include BSS for unknown/dynamic number of sources.”

Tomohiro Nakatani NTT Communication Science Laboratories

Industrial Perspectives

Mixed-signal ICs for mobile phones

"Moore's Law is driving DSP speed and memory capacity … enabling implementation of sophisticated DSP functions that have resulted from years of research in acoustic signal processing. The end-user experience is one of natural wideband voice communication, devoid of acoustic background noise and unwanted artefacts."

Anthony Magrath Director of DSP Technology, Wolfson Microelectronics

Industrial Perspectives

“The applications of sound capture, speech enhancement, and audio processing technologies shift gradually from communications mostly, towards speech recognition and building natural human-machine interfaces for mobile devices, in cars, and in our living rooms.”

Ivan Tashev Microsoft Research

KINECT™
for XBOX 360.

Music Movies TV Shows App Store Podcasts Audiobooks iTunes U Ping

Welcome Malc

PING NEW

Discover new music
connecting with you
artists and friends.

Learn Mor

QUICK LINKS

Redeem
Buy iTunes Gifts
Power Search
Browse
Account
Support
iTunes LP
Inside iTunes

Recommendations
My Wish List
My Alerts
Complete My Album
iTunes Plus 4

Music

This Week | Recent Hits

See All >


Born This Way (...
Lady GaGa


Like My Mother ...
Lauren Alaina


American Idol Se...
Scotty McCreery

Glee: The Music,...
Glee Cast


Brighter Than th...
Colbie Caillat


Torches
Foster the People


Don't W
Jason D


This Is Country ...
Brad Paisley


I Love You This ...
Scotty McCreery


American Idol Se...
Lauren Alaina

Light Up the World
Glee Cast

Original Songs A...
Various Artists


MMG Presents: S...
Various Artists


Blacklig
Tedashi

TOP CHARTS

Singles

Songs | Music

After-thought

- Trend
  - Origin: Old English *trendan* ‘revolve, rotate’, of Germanic origin

- “What goes around, comes around” (?)

Texture

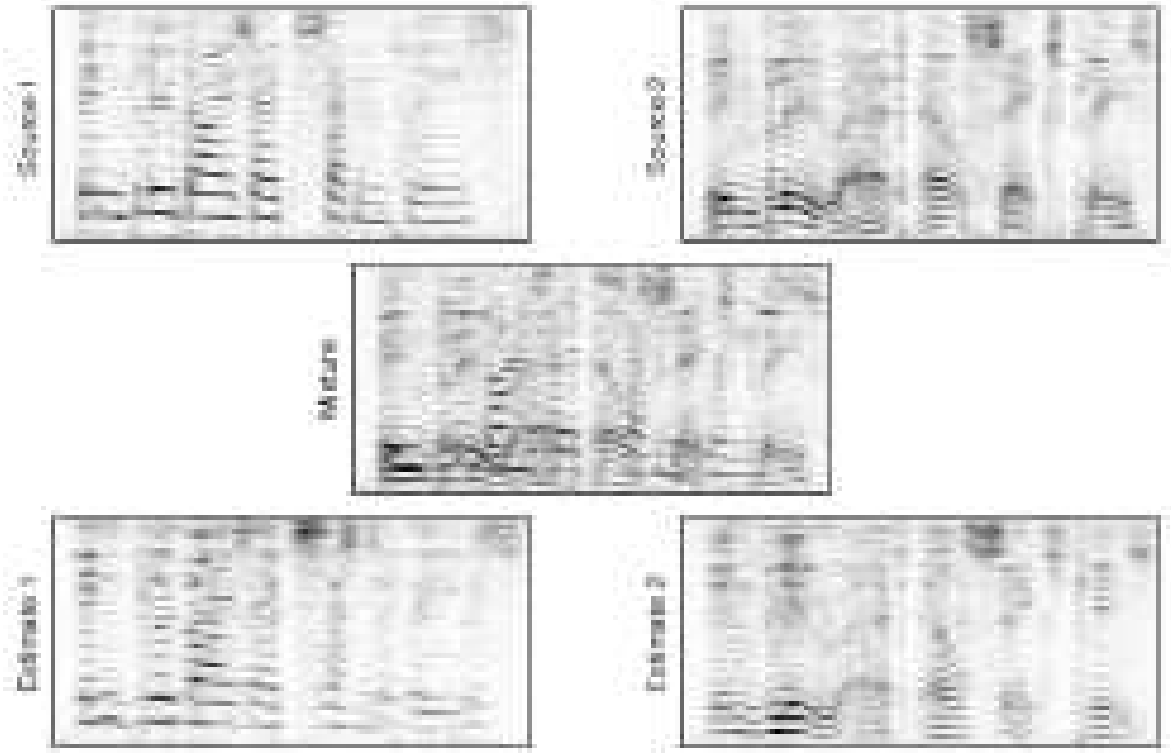


Fig. 6. *A result from a single source separation experiment. The top plots show two sentences spoken by two different speakers. The middle plot shows their observed mixture, and the bottom plots the estimates sources given training data from each speaker.*

Sparsity

- Better representations
  - Sparse
    - Matching pursuit
    - DBNs
    - features for recognition
  - New angles (cortical and textures)
  - Subspaces (latent and otherwise)