Distributed speech processing in IoT devices

SOUNDS Seasonal School on Distributed Signal Processing and Optimization

Daniele Giacobello, 9/19/2022

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- Commercialization & Conclusions

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability

Part 1

- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
  - Distributed signal processing
    - Homogeneous networks
      - Node-Specific Noise Reduction
      - Node-Specific DOA Estimation
    - Heterogeneous networks
- Commercialization & Conclusions

Part 2

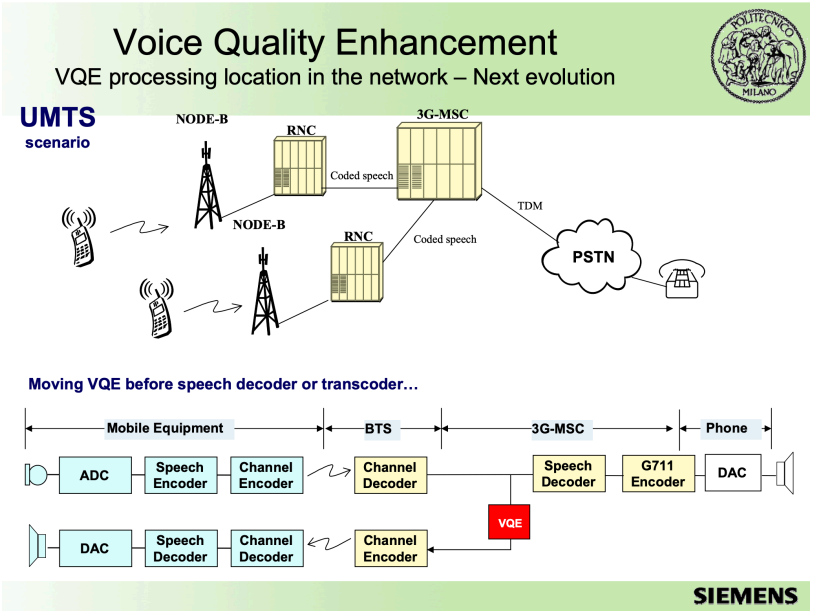
Outline

- **About me**
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- Commercialization & Conclusions

About me

Education

- B.Sc. and M.Sc. in Telecommunication Engineering from Politecnico di Milano
  - Specialization in Digital Communication
    - Master's thesis on "compressed" voice quality enhancement (with Siemens Communications)



PhD

Sparse Linear Prediction for Speech & Audio Processing

- Ph.D. in E&E Engineering from Aalborg University
  - ESR of Marie (Skłodowska-)Curie Action “SIGNAL”
  - Modeling, Analysis, Coding, Enhancement of speech signals
  - Thesis “Sparse Linear Prediction for Speech and Audio Processing”
  - Supervisor: Prof. S.H. Jensen and Prof. M.G. Christensen (AAU)
  - Co-supervision with Prof. Marc Moonen (KU Leuven)
  - Visiting stay at University of Miami and KU Leuven (worked with Toon)
- Bragging rights: IEEE Signal Processing Society 2014 Young Author Best Paper Award (150+ citations)

Sparse Linear Prediction for Speech Processing

- New formulation for LP coding allows for better tradeoffs between speech representations

D and R terms from operational R/D theory

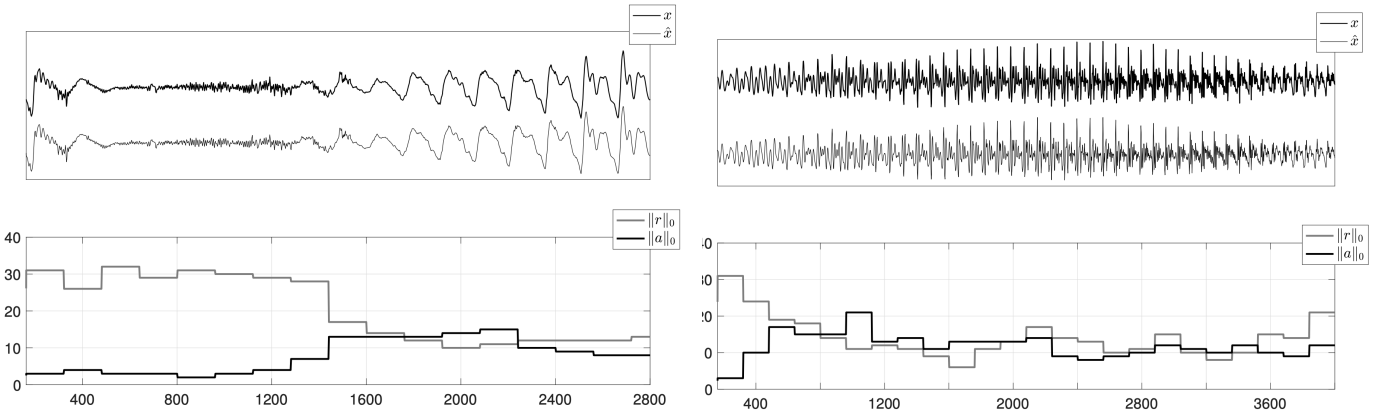
$$\begin{aligned} \text{minimize} \quad & D(\mathbf{x}, \hat{\mathbf{x}}), \\ \text{subject to} \quad & R(\hat{\mathbf{x}}) \leq R^*, \end{aligned}$$

$D(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{W}(\mathbf{x} - \mathbf{\Phi}(\hat{\mathbf{a}})\hat{\mathbf{r}})\|_2^2,$ Analysis-by-Synthesis equation

$R(\hat{\mathbf{x}}) = R(\hat{\mathbf{a}}) + R(\hat{\mathbf{r}}).$ $\longrightarrow$ $R(\hat{\mathbf{x}}) = \alpha\|\hat{\mathbf{a}}\|_0 + \beta\|\hat{\mathbf{r}}\|_0.$

$$\hat{\mathbf{a}}_k, \hat{\mathbf{r}}_k = \underset{\mathbf{a},\mathbf{r}}{\operatorname{argmin}} \|\mathbf{r} - \check{\mathbf{x}} - \check{\mathbf{X}}\mathbf{a}\|_2^2 + \alpha\|\mathbf{a}\|_0 + \beta\|\mathbf{r}\|_0,$$

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{r}_k - \check{\mathbf{x}}_k - \check{\mathbf{X}}_k\mathbf{a}_k\|_2^2, \\ \text{subject to} \quad & \|\mathbf{a}_k\|_1 \leq \delta, \\ & \|\mathbf{r}_k\|_1 \leq \gamma, \end{aligned}$$

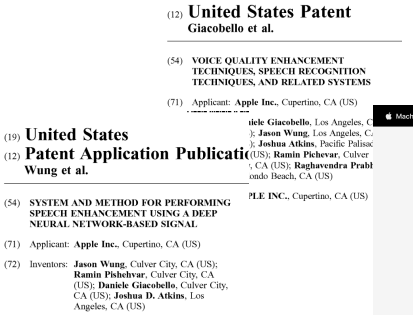


D. Giacobello, M. Murthi, T. L. Jensen, M. G. Christensen, “Revisiting the Linear Prediction Analysis-by-Synthesis Speech Coding Paradigm Using Real-Time Convex Optimization”, in *Rec. 52nd Asilomar Conference on Signals, Systems, and Computers*, October 2018.

About me

Relevant Work Experience (2010-2016)

- **Broadcom**
  - English ASR system for Broadcom bluetooth chips to accommodate multiple languages.
  - Speech and audio coding (OPUS)
  - Designed different algorithms for dual-mic voice processing and supported their integration into Broadcom baseband processor
- **Beats -> Apple**
  - Worked on Beats first proprietary voice communication and speech recognition solutions -> Homepod
  - Worked on spatial audio reproduction algorithms -> Apple Spatial Audio
- **FocusMotion**
  - Algorithm development for motion tracking In wearable devices



(12) **United States Patent**
Giacobello et al.

(54) **VOICE QUALITY ENHANCEMENT TECHNIQUES, SPEECH RECOGNITION TECHNIQUES, AND RELATED SYSTEMS**

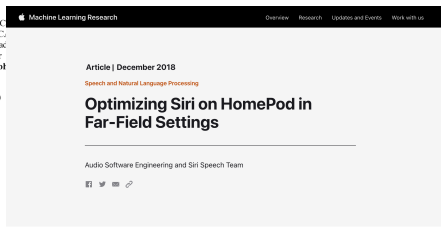
(71) Applicant: **Apple Inc.**, Cupertino, CA (US)

(19) **United States**
(12) **Patent Application Publicati**
Wung et al.

(54) **SYSTEM AND METHOD FOR PERFORMING SPEECH ENHANCEMENT USING A DEEP NEURAL NETWORK-BASED SIGNAL**

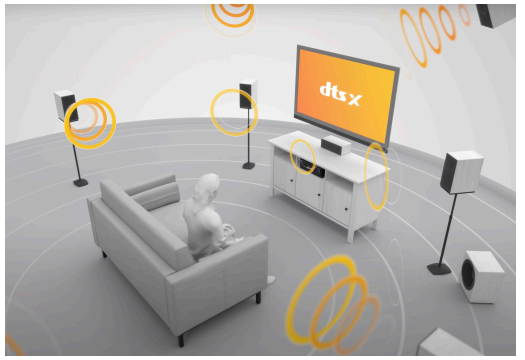
(71) Applicant: **Apple Inc.**, Cupertino, CA (US)

(72) Inventors: **Jason Wung**, Culver City, CA (US); **Ramin Pishehvar**, Culver City, CA (US); **Daniele Giacobello**, Culver City, CA (US); **Joshua D. Atkins**, Los Angeles, CA (US)

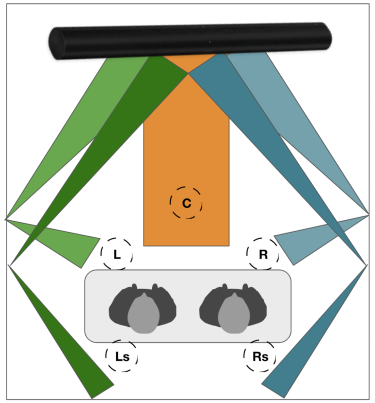


About me

Relevant Work Experience (2016-present)



- **DTS**
  - Design and implementation of new audio processing components for DTS new codecs.
  - Supported improvement of existing C/C++ codebase for DTS audio codecs (e.g., DTS-X).
- **Sonos**
  - Developed Sonos first proprietary far-field speech processing solution
  - Developed audio beamforming algorithms for soundbars
- **Apple**
  - ???
  - Immersive Media Group handles all things speech and audio for Apple products



About me

Extra-curricular activities

- **Technical Committees Member**
  - IEEE SPS Technical Committee on Audio and Acoustics (Challenges subcommittee)
  - IEEE SPS Technical Committee on Speech and Language Processing (Challenges subcommittee)
  - EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing
  - ISCA Special Interest Group on Security and Privacy in Speech Communication
- **Associate Editor**
  - IEEE Signal Processing Letters
  - EURASIP Journal on Audio, Speech, and Music Processing
- **ME-UYR mentor**
  - Mentoring Experiences for Underrepresented Young Researchers

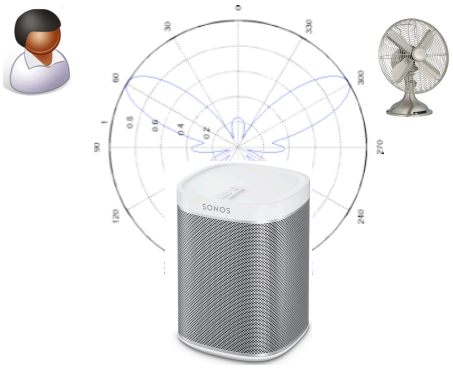
Outline

- About me
- **Introduction: Sonos Voice Solution**
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- Commercialization & Conclusions

Sonos Voice Solution

Sonos Voice Solution 2016-2018

- Initial voice solution (9/2016-10/2017 Sonos One launch) was based on a fixed beamforming strategy using the 6-mic circular array modeling Amazon Echo devices
- However,
  - Traditional beamforming techniques not ideal for ASR front-end
  - Poor noise reduction performance
- Initially my work focused on the (6) SC-AEC
  - Loudspeaker feedback can be very high (working at -35 dB ERL)



Modern Audio Platform

- WiFi couldn't reliably transmit audio.
  - Bridge to Sonos mesh.
- WiFi didn't provide adequate coverage
  - Peer-to-peer mesh network.
- Secure WiFi meant typing SSID and security keys.
  - Wireless secure setup with just a button press in 2004!
- Time synchronization for speakers

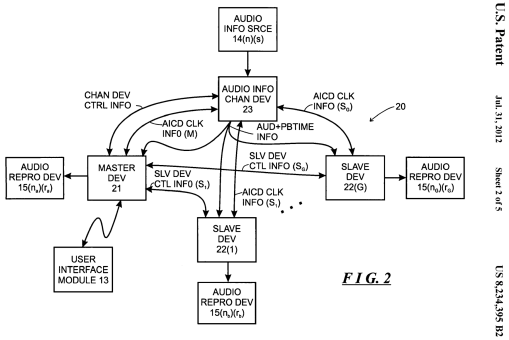


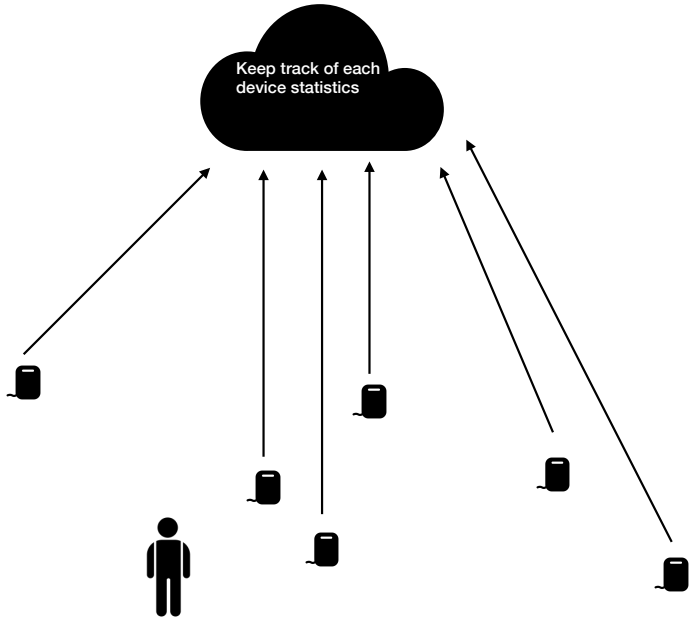
FIG. 2

SONOS™

Multi-device problem

Device Arbitration

Which one has the best SNR?

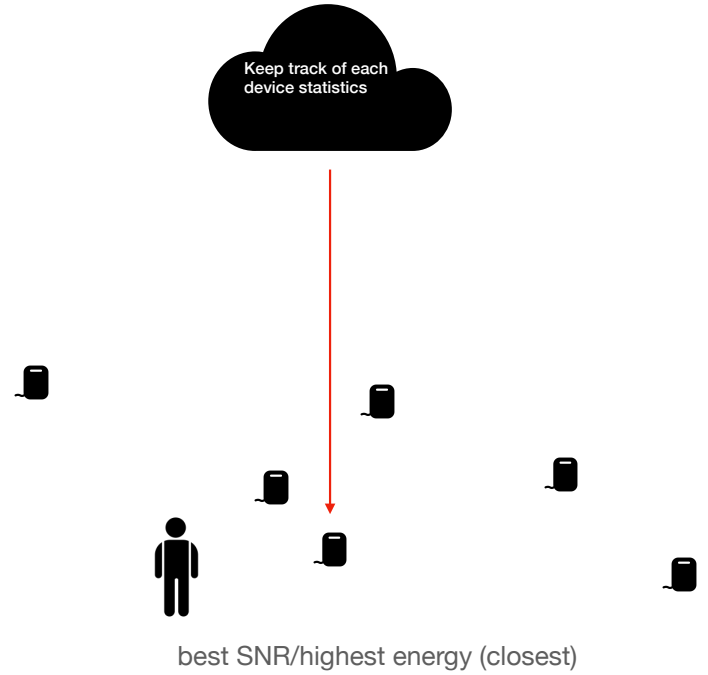


Devices don’t have access to other devices statistics

Energy-based methods exploit the free-field relationship
E~1/R^2 between energy and distance

Multi-device problem

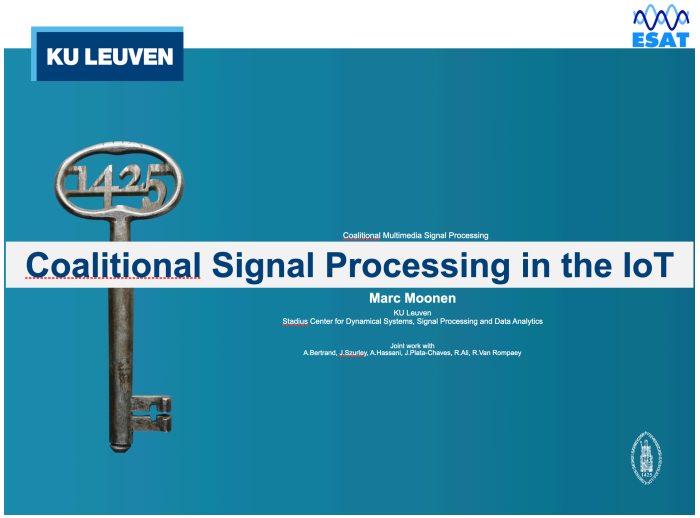
Device Arbitration

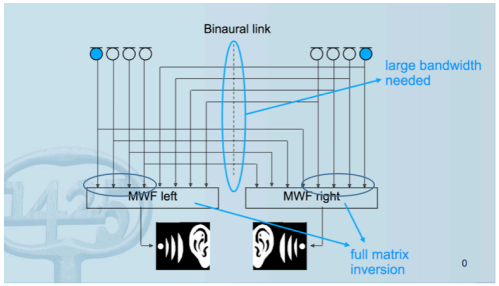


best SNR/highest energy (closest)

Difficulties when applied to home smart speakers in reverberant rooms with high levels of noise and interference (such as TVs and air conditioners)

At the same time…





It has been proven that full connection is not necessary as long as we are trying to estimate one source!
4 L + 1 aligned R
4 R + 1 aligned L

Theory has already been laid out for a series of algorithms, not just MCWF!

from Marc Moonen's presentation ''Distributed Adaptive Node-Specific Signal Estimation in Wireless Acoustic Sensor Networks''.
Joint work with A.Bertrand, J.Szurley, A. Hassani, R. Serizel

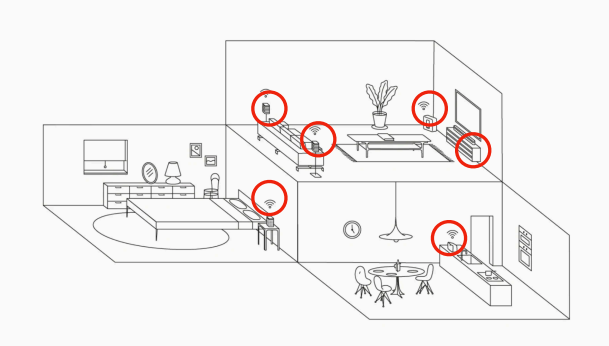
Real-time distributed speech enhancement [youtube]

Rank-1 approximation based multichannel wiener filtering algorithms for noise reduction [paper]

Signal Processing Algorithms for Wireless Acoustic Sensor Networks [phd]

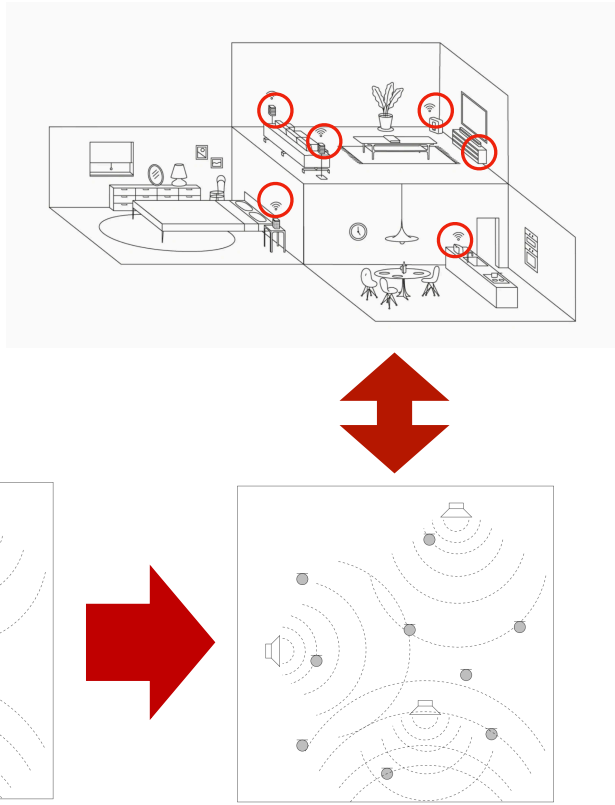
From single-device to multi-device processing

- Single-device processing with device arbitration hardly ideal…
- A Sonos home system is a voice-enabled multi-device system
- Ideally, we want to exploit the full capacity of our distributed system for voice processing and sensor data analysis.
  - Drop rigid requirements of mic location for fixed beampattern design
  - Remove array-size limitations
  - Improve scalability (processing group of mics with different algorithms)
  - Increase fault tolerance (what if one mic breaks? 10+ years of service!)



Wireless Acoustic Sensor Networks

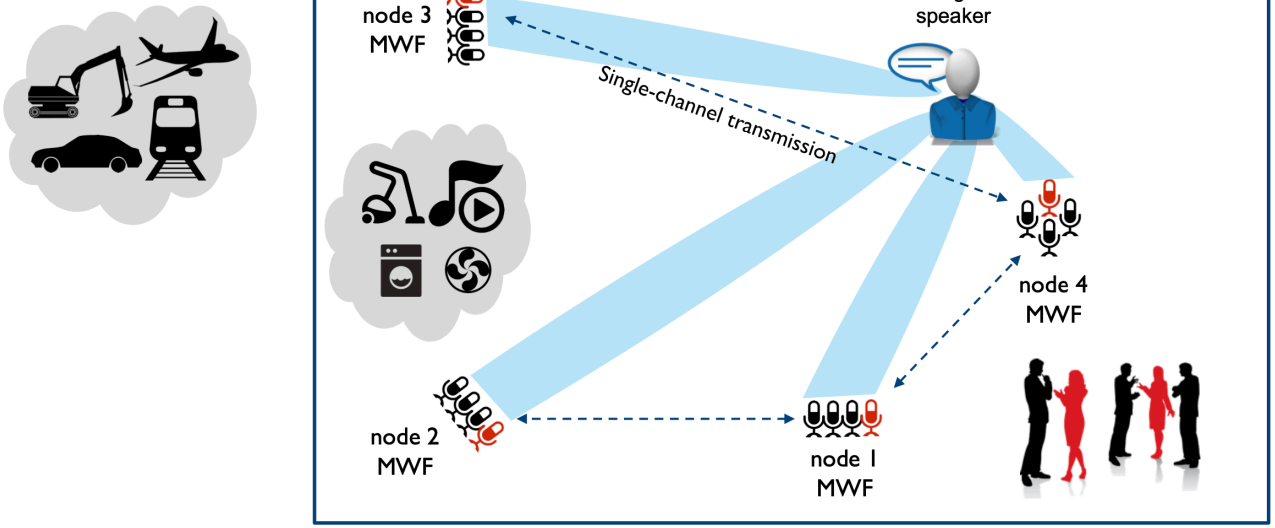
- Wireless Acoustic Sensor Networks offer the right solution!
  - How do device communicate with each other?
  - Focus on the signal processing level to relax the high-demanding constraints on the network layer design



A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective." B. *18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*. IEEE, 2011.

Wireless Acoustic Sensor Networks

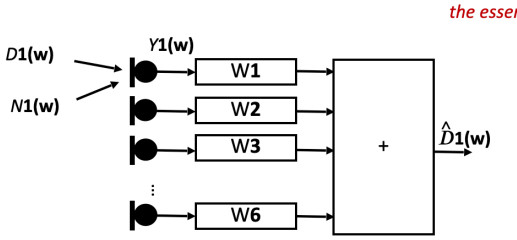
Distributed signal processing



- Each node uses an MWF to estimate the speech signal as locally observed at its own reference microphone (to preserve spatial cues)

Proposed improvements in Sonos voice solution

- First: let's start using the Multi-Channel Weiner Filter
  - Main selling point: no assumptions on geometry
  - Gateway to distributed DSP



the essence of it…

$Y(w)=D(w)+N(w)$

$\min E\{ |\hat{D}1(w)-W(w)Y(w)| \}$

$W(w) = inv(Ryy)ryd(w)$

where $ryd(w)=Rdd[1\ 0\ldots0]T$

Key is the estimation of noise and speech correlation matrices (spatial information)

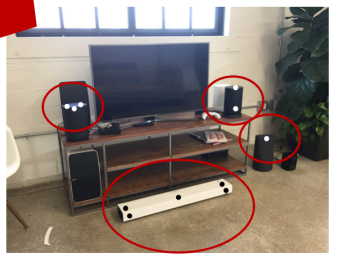
current solution vs. future solutions

Traditional Array

WASN on device

…potentially!

WASN on multiple SONOS devices



Coalitional Signal Processing

Distributed -> Coalitional
Homogeneous -> Heterogeneous

- Devices with different signal processing tasks (e.g. noise reduction, AEC, localization, ...)
  - cooperate/form coalitions
  - improve each device’s performance
- Every device uses other devices to form its own wireless sensor network and in return contributes to the sensor network of every other device

What can be enabled by WASN?

- Improvements in Voice Enhancement

- Acoustic monitoring

Improvements in Voice Enhancement

Multichannel Weiner Filter (MCWF) on WASN

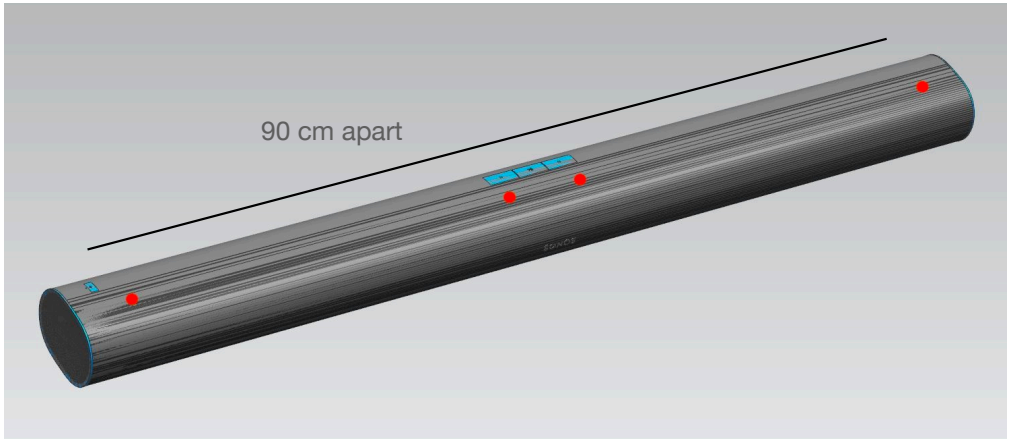
- MCWF produces a linear minimum mean squared error (MMSE) estimate of the desired speech component in the signal captured by one of the microphones.
- MCWF does not rely on a priori knowledge of the microphone or sound source locations
- Pragmatically interesting for us as it allows:
  - Different microphone-array geometries and configurations
  - Different industrial design, form factors, and HW modules
  - Different performance requirements and use cases

S. Bagheri, D. Giacobello, “Linear Filtering for Noise-Suppressed Speech Detection”, *U.S. Patent No. 10,847,178*, November 2020.

Improvements in Voice Enhancement

Example: MCWF implementation on Sonos ARC

- Not a WASN per se but WASN “inspired” offers improvements for all SNRs and conditions
- Including the satellite speakers for surround we can have a 12-mic WASN surrounding the user!



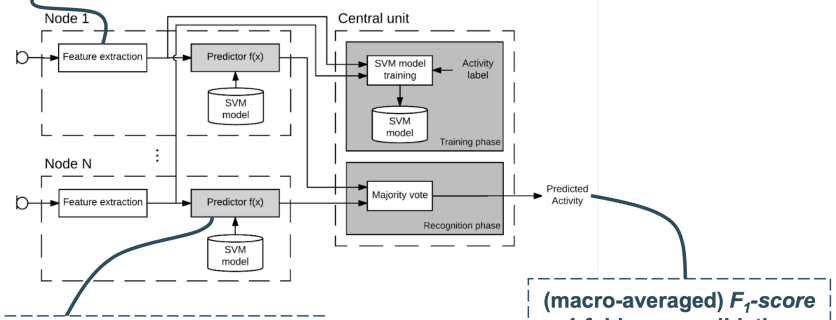
S. Bagheri, D. Giacobello, “Linear filtering for noise-suppressed speech detection via multiple network microphone devices”, *U.S. Patent No. 10,692,518*, June 2020.

Acoustic monitoring with WASN

- Most literature on WASN is in the context of acoustic monitoring of an environment:
  - event detection, sound classification, localization
- Microphones = cheap sensors with no line-of-sight constraints
- Work in collaboration with Gert Dekkers, KU Leuven

Acoustic monitoring with WASN

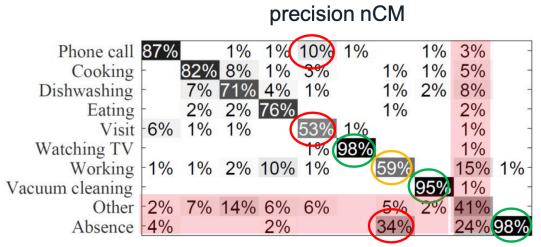
Mel-Frequency Cepstral Coefficients Δ- Δ Δ
Window: 30 ms, step: 10 ms, 26 melbands, 14 DCT coefficients
Mean and std of 15s window, 5s step

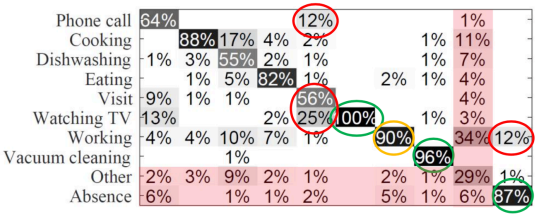


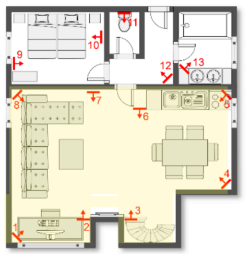
Support Vector Machine
RBF-kernel, 1-vs-1 coding, 3-fold cross-validation, class weighted
Majority Vote
For each room seperately

(macro-averaged) *F_1-score*
4-fold cross-validation

L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, and B. Vanrumste, "Energy efficient monitoring of activities of daily living using wireless acoustic sensor networks in clean and noisy conditions," in 2015 23rd European Signal Processing Conference (EUSIPCO), Aug 2015, pp. 449–453.





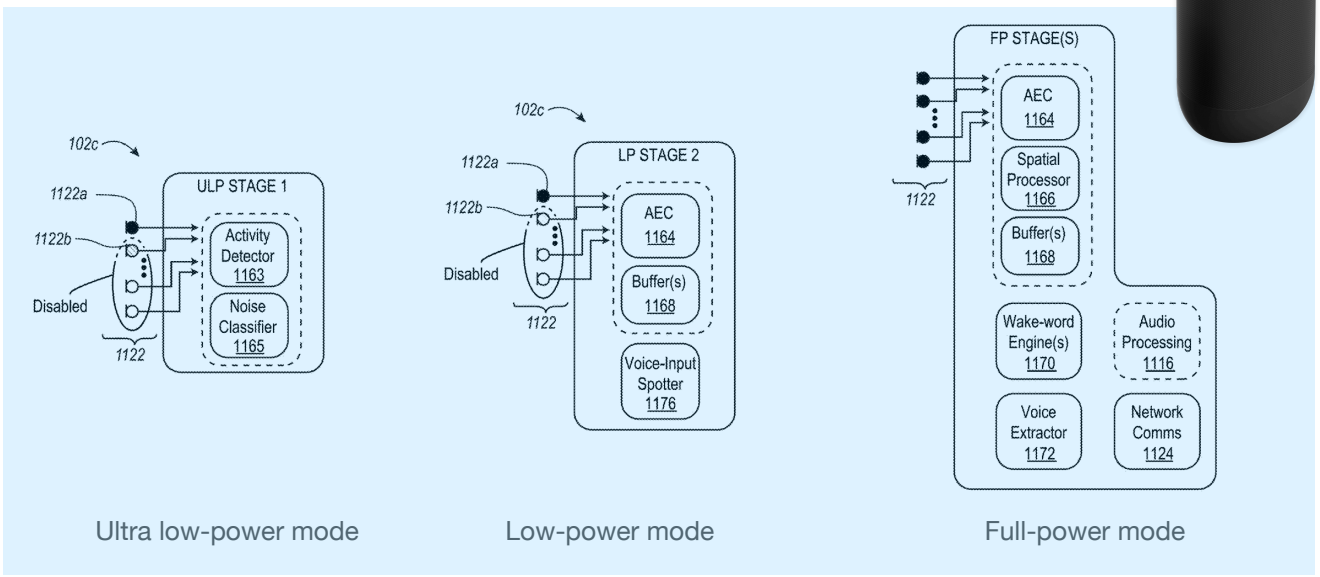


F_1-score: **82.3±2.2%**

Acoustic monitoring

Single Device

- Used to determine the logic for microphone subset selection
- Multistage logic for battery-enabled speaker



(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2020/0125162 A1**
D'Amato et al. (43) **Pub. Date: Apr. 23, 2020**

(54) **MULTIPLE STAGE NETWORK MICROPHONE DEVICE WITH REDUCED POWER CONSUMPTION AND PROCESSING LOAD**

(71) Applicant: **Sonos, Inc.**, Santa Barbara, CA (US)

(72) Inventors: **Nick D'Amato**, Santa Barbara, CA (US); **Daniele Giacobello**, Los Angeles, CA (US); **Joachim Fainberg**, Santa Barbara, CA (US); **Klaus Hartung**, Santa Barbara, CA (US)

(21) Appl. No.: **16/168,389**

(22) Filed: **Oct. 23, 2018**

Publication Classification

(51) **Int. Cl.**
G06F 1/32 (2006.01)
G06F 3/16 (2006.01)
H04R 3/00 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 1/3287*** (2013.01); ***H04R 3/00*** (2013.01); ***G06F 3/167*** (2013.01); ***G06F 1/3231*** (2013.01)

(57) **ABSTRACT**

Systems, methods, and devices with reduced power consumption in network microphone devices. In one embodiment, a network microphone device is configured to perform a method that includes (i) capturing audio content; (ii) using a first algorithm to perform a keyword detection process for determining whether the audio content includes a keyword; (iii) responsive to determining that the audio content includes the keyword, using a second, more computationally intensive algorithm to perform a wake-word detection process for determining whether the audio content includes a wake word; and (iv) responsive to performing the wake-word detection process, (a) causing a voice service corresponding to the wake word to process the audio content if the wake-word detection process confirms that the audio content includes the wake word or (b) ceasing performance of the wake-word detection process if the wake-word detection process disconfirms that the audio content includes the wake word.

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2022/0238120 A1**
Jones et al. (43) **Pub. Date: Jul. 28, 2022**

(54) **SYSTEMS AND METHODS FOR POWER-EFFICIENT KEYWORD DETECTION**

(71) Applicant: **Sonos, Inc.**, Santa Barbara, CA (US)

(72) Inventors: **Aaron Jones**, San Diego, CA (US); **Saeed Bagheri Sereshki**, Goleta, CA (US); **Daniele Giacobello**, Los Angeles, CA (US)

(21) Appl. No.: **17/248,427**

(22) Filed: **Jan. 25, 2021**

Publication Classification

(51) **Int. Cl.**
G10L 17/22 (2006.01)
G10L 17/02 (2006.01)
G10L 15/02 (2006.01)
G10L 15/05 (2006.01)
G10L 15/22 (2006.01)

(52) **U.S. Cl.**
CPC ***G10L 17/22*** (2013.01); ***G10L 17/02*** (2013.01); *G10L 2015/223* (2013.01); ***G10L 15/05*** (2013.01); ***G10L 15/22*** (2013.01); ***G10L 15/02*** (2013.01)

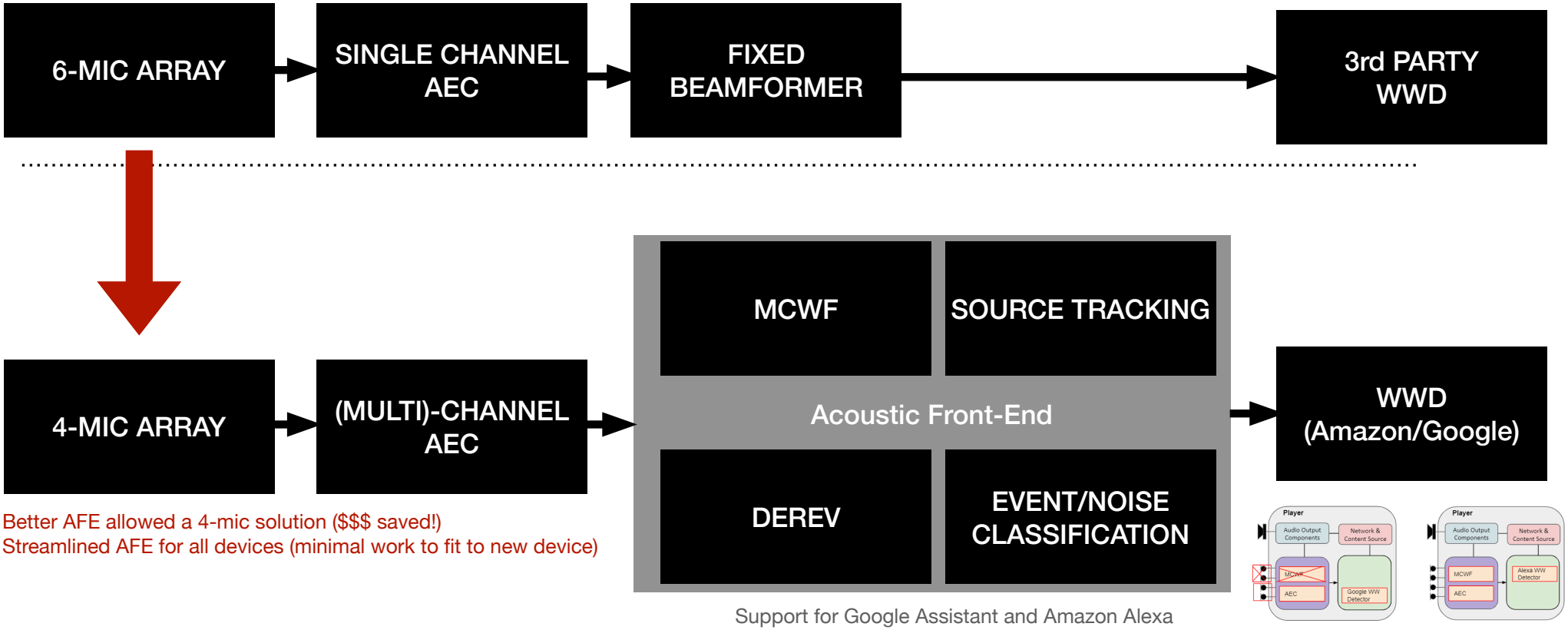
(57) **ABSTRACT**

Systems and methods for audio processing include capturing sound data via at least one microphone of a network microphone device (NMD) and determining whether the captured sound includes voice activity. While in a first stage, the NMD forgoes spatial processing of the captured sound data. If the NMD determines that the detected sound includes voice activity, the NMD transitions to a second stage. In this second stage, the NMD spatially processes the detected sound to produce filtered sound data and detects a wake word. After detecting the wake word, the NMD may determine an action to be performed based on the captured sound data.

Other speech processing components

- Dereverberation based on (Adaptive) Sparse Linear Prediction $\hat{\mathbf{g}}_m = \underset{\mathbf{g}_m}{\operatorname{argmin}} \|\mathbf{x}_m - \mathbf{X}_\tau \mathbf{g}_m\|_1 + \alpha \|\mathbf{g}_m\|_1, \quad m = 1, \ldots, M$
  - ADMM formulation can be interpreted as iterative “sparsification” of the ‘’classical” LP solution
- Source tracking
  - GMM modeling of the acoustic space, EM for tracking
- Keyword Spotting
  - Started as side project to eliminate 3rd party solution in 2017, now fully implemented for Sonos Spoken Language Understanding (SLU) solution
  - Simple FF architecture with ReLU activation (efficient embedded formulation)
  - Data augmentation with ISM, RT, FEM (pyroomacoustics)
  - Work in collaboration with Joachim Fainberg, U. Edinburgh
- Multi-channel AEC

Sonos Voice: 2017 vs 2019



Better AFE allowed a 4-mic solution ($$$ saved!)
Streamlined AFE for all devices (minimal work to fit to new device)

Support for Google Assistant and Amazon Alexa

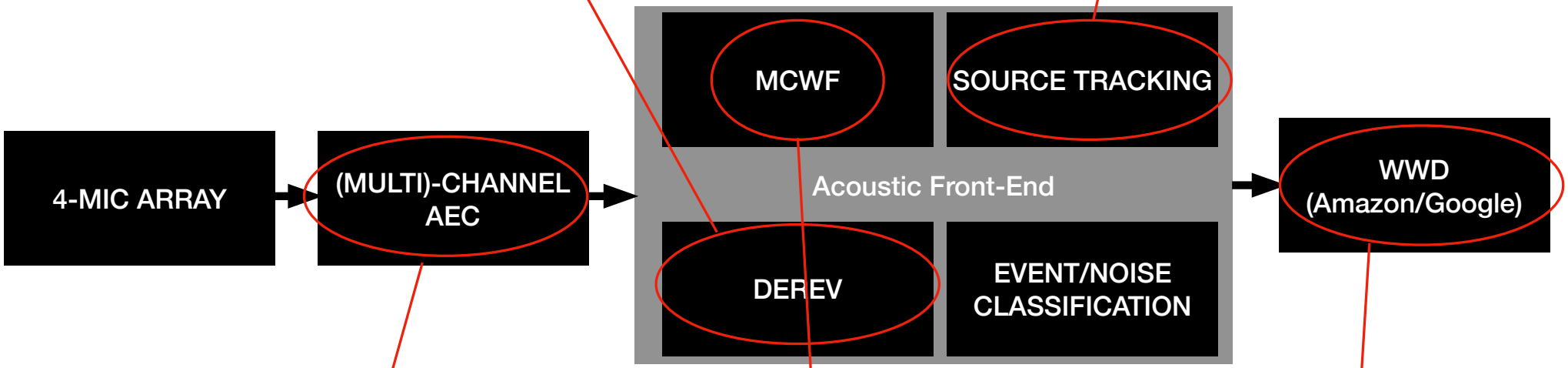
SPEECH DEREVERBERATION BASED ON CONVEX OPTIMIZATION ALGORITHMS FOR GROUP SPARSE LINEAR PREDICTION

Daniele Giacobello[1] *and Tobias Lindstrøm Jensen*[2]

[1]Sonos Inc., Santa Barbara, CA, USA
[2]Signal and Information Processing, Department of Electronic Systems, Aalborg, Denmark

An Online Expectation-Maximization Algorithm for Tracking Acoustic Sources in Multi-Microphone Devices during Music Playback

Daniele Giacobello
Sonos Inc.,
Santa Barbara, CA, USA
daniele.giacobello@sonos.com



ROBUST STFT DOMAIN MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION WITH ADAPTIVE DECORRELATION OF THE REFERENCE SIGNALS

Saeed Bagheri, Daniele Giacobello

Sonos Inc., Santa Barbara, CA, USA

Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter

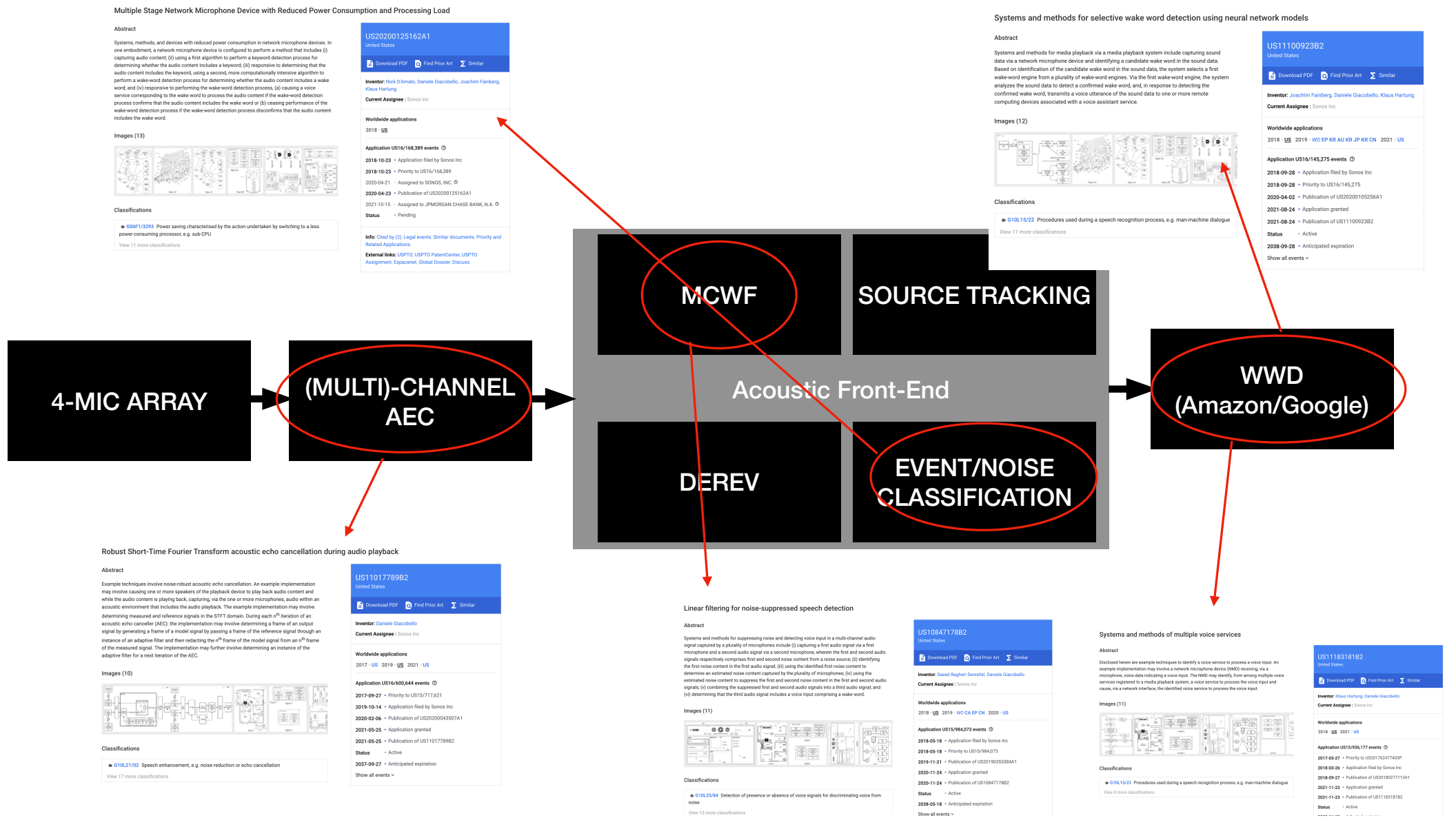
Saeed Bagheri, Daniele Giacobello

Sonos Inc., Santa Barbara, CA, USA
{saeed.sereshki,daniele.giacobello}@sonos.com

CONTINUITY OF CONTROL

Voice Assistant Coexistence on the Sonos Platform

Daniele Giacobello
Distinguished Research Engineer, Advanced Technology Team


Multiple Stage Network Microphone Device with Reduced Power Consumption and Processing Load
Abstract
Systems, methods, and devices with reduced power consumption in network microphone devices. In one embodiment, a network microphone device is configured to perform a method that includes (i) capturing audio content; (ii) using a first algorithm to perform a keyword detection process for determining whether the audio content includes a keyword; (iii) responsive to determining that the audio content includes the keyword, using a second, more computationally intensive algorithm to perform a wake-word detection process for determining whether the audio content includes a wake word; and (iv) responsive to performing the wake-word detection process, (a) causing a voice service corresponding to the wake word to process the audio content if the wake-word detection process confirms that the audio content includes the wake word or (b) ceasing performance of the wake-word detection process if the wake-word detection process disconfirms that the audio content includes the wake word.
Images (13)
Classifications
G06F1/3293 Power saving characterised by the action undertaken by switching to a less power-consuming processor, e.g. sub-CPU
View 11 more classifications
US20200125162A1
United States
Download PDF
Find Prior Art
Similar
Inventor: Nick D'Amato, Daniele Giacobello, Joachim Fainberg, Klaus Hartung
Current Assignee : Sonos Inc
Worldwide applications
2018 · US
Application US16/168,389 events
2018-10-23 • Application filed by Sonos Inc
2018-10-23 • Priority to US16/168,389
2020-04-21 • Assigned to SONOS, INC.
2020-04-23 • Publication of US20200125162A1
2021-10-15 • Assigned to JPMORGAN CHASE BANK, N.A.
Status • Pending
Info: Cited by (2), Legal events, Similar documents, Priority and Related Applications
External links: USPTO, USPTO PatentCenter, USPTO Assignment, Espacenet, Global Dossier, Discuss
Systems and methods for selective wake word detection using neural network models
Abstract
Systems and methods for media playback via a media playback system include capturing sound data via a network microphone device and identifying a candidate wake word in the sound data. Based on identification of the candidate wake word in the sound data, the system selects a first wake-word engine from a plurality of wake-word engines. Via the first wake-word engine, the system analyzes the sound data to detect a confirmed wake word, and, in response to detecting the confirmed wake word, transmits a voice utterance of the sound data to one or more remote computing devices associated with a voice assistant service.
Images (12)
Classifications
G10L15/22 Procedures used during a speech recognition process, e.g. man-machine dialogue
View 11 more classifications
US11100923B2
United States
Download PDF
Find Prior Art
Similar
Inventor: Joachim Fainberg, Daniele Giacobello, Klaus Hartung
Current Assignee : Sonos Inc
Worldwide applications
2018 · US 2019 · WO EP KR AU KR JP KR CN 2021 · US
Application US16/145,275 events
2018-09-28 • Application filed by Sonos Inc
2018-09-28 • Priority to US16/145,275
2020-04-02 • Publication of US20200105256A1
2021-08-24 • Application granted
2021-08-24 • Publication of US11100923B2
Status • Active
2038-09-28 • Anticipated expiration
Show all events
4-MIC ARRAY
(MULTI)-CHANNEL AEC
MCWF
SOURCE TRACKING
Acoustic Front-End
DEREV
EVENT/NOISE CLASSIFICATION
WWD (Amazon/Google)
Robust Short-Time Fourier Transform acoustic echo cancellation during audio playback
Abstract
Example techniques involve noise-robust acoustic echo cancellation. An example implementation may involve causing one or more speakers of the playback device to play back audio content and while the audio content is playing back, capturing, via the one or more microphones, audio within an acoustic environment that includes the audio playback. The example implementation may involve determining measured and reference signals in the STFT domain. During each nth iteration of an acoustic echo canceller (AEC): the implementation may involve determining a frame of an output signal by generating a frame of a model signal by passing a frame of the reference signal through an instance of an adaptive filter and then redacting the nth frame of the model signal from an nth frame of the measured signal. The implementation may further involve determining an instance of the adaptive filter for a next iteration of the AEC.
Images (10)
Classifications
G10L21/02 Speech enhancement, e.g. noise reduction or echo cancellation
View 17 more classifications
US11017789B2
United States
Download PDF
Find Prior Art
Similar
Inventor: Daniele Giacobello
Current Assignee : Sonos Inc
Worldwide applications
2017 · US 2019 · US 2021 · US
Application US16/600,644 events
2017-09-27 • Priority to US15/717,621
2019-10-14 • Application filed by Sonos Inc
2020-02-06 • Publication of US20200043507A1
2021-05-25 • Application granted
2021-05-25 • Publication of US11017789B2
Status • Active
2037-09-27 • Anticipated expiration
Show all events
Linear filtering for noise-suppressed speech detection
Abstract
Systems and methods for suppressing noise and detecting voice input in a multi-channel audio signal captured by a plurality of microphones include (i) capturing a first audio signal via a first microphone and a second audio signal via a second microphone, wherein the first and second audio signals respectively comprises first and second noise content from a noise source; (ii) identifying the first noise content in the first audio signal; (iii) using the identified first noise content to determine an estimated noise content captured by the plurality of microphones; (iv) using the estimated noise content to suppress the first and second noise content in the first and second audio signals; (v) combining the suppressed first and second audio signals into a third audio signal; and (vi) determining that the third audio signal includes a voice input comprising a wake word.
Images (11)
Classifications
G10L25/84 Detection of presence or absence of voice signals for discriminating voice from noise
View 15 more classifications
US10847178B2
United States
Download PDF
Find Prior Art
Similar
Inventor: Saeed Bagheri Sereshki, Daniele Giacobello
Current Assignee : Sonos Inc
Worldwide applications
2018 · US 2019 · WO CA EP CN 2020 · US
Application US15/984,073 events
2018-05-18 • Application filed by Sonos Inc
2018-05-18 • Priority to US15/984,073
2019-11-21 • Publication of US20190355384A1
2020-11-24 • Application granted
2020-11-24 • Publication of US10847178B2
Status • Active
2038-05-18 • Anticipated expiration
Show all events
Systems and methods of multiple voice services
Abstract
Disclosed herein are example techniques to identify a voice service to process a voice input. An example implementation may involve a network microphone device (NMD) receiving, via a microphone, voice data indicating a voice input. The NMD may identify, from among multiple voice services registered to a media playback system, a voice service to process the voice input and cause, via a network interface, the identified voice service to process the voice input.
Images (11)
Classifications
G10L15/22 Procedures used during a speech recognition process, e.g. man-machine dialogue
View 8 more classifications
US11183181B2
United States
Download PDF
Find Prior Art
Similar
Inventor: Klaus Hartung, Daniele Giacobello
Current Assignee : Sonos Inc
Worldwide applications
2018 · US 2021 · US
Application US15/936,177 events
2017-03-27 • Priority to US201762477403P
2018-03-26 • Application filed by Sonos Inc
2018-09-27 • Publication of US20190277113A1
2021-11-23 • Application granted
2021-11-23 • Publication of US11183181B2
Status • Active

2019-2021

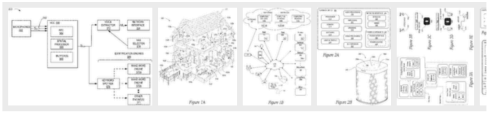
- Acquired Snips, embedded voice recognition startup in Paris
- Current challenges:
  - Joint design of Front-end and Back-end
  - Complexity-constrained DNN architectures
  - Realistic data augmentation

Systems and methods for selective wake word detection using neural network models

Abstract

Systems and methods for media playback via a media playback system include capturing sound data via a network microphone device and identifying a candidate wake word in the sound data. Based on identification of the candidate wake word in the sound data, the system selects a first wake-word engine from a plurality of wake-word engines. Via the first wake-word engine, the system analyzes the sound data to detect a confirmed wake word, and, in response to detecting the confirmed wake word, transmits a voice utterance of the sound data to one or more remote computing devices associated with a voice assistant service.

Images (12)



Classifications

G10L15/22 Procedures used during a speech recognition process, e.g. man-machine dialogue

View 11 more classifications

US11100923B2
United States

Download PDF | Find Prior Art | Similar

Inventor: Joachim Fainberg, Daniele Giacobello, Klaus Hartung
Current Assignee : Sonos Inc

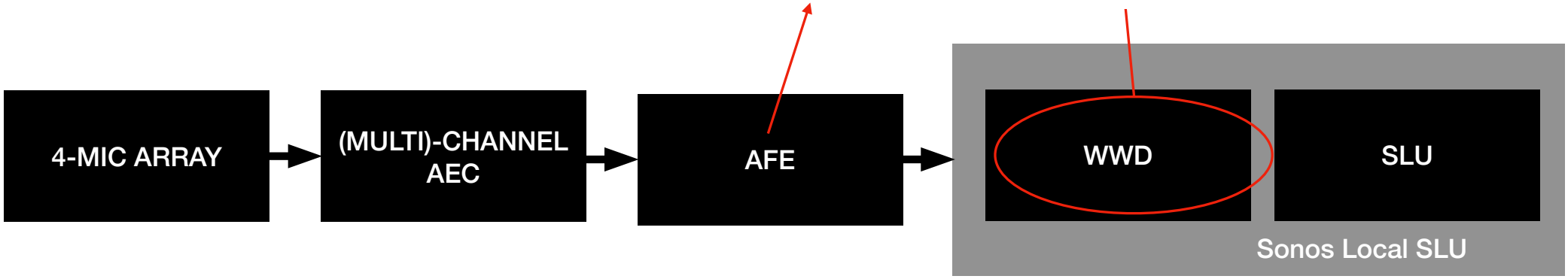
Worldwide applications
2018 · US 2019 · WO EP KR AU KR JP KR CN 2021 · US

Application US16/145,275 events
2018-09-28 • Application filed by Sonos Inc
2018-09-28 • Priority to US16/145,275
2020-04-02 • Publication of US20200105256A1
2021-08-24 • Application granted
2021-08-24 • Publication of US11100923B2
Status • Active
2038-09-28 • Anticipated expiration
Show all events

Noise Reduction for Distant Voice Recognition in Smart Speakers

Saeed Sereshki
Principal Audio Research Engineer, Advanced Technology Team

Daniele Giacobello
Distinguished Research Engineer, Advanced Technology Team



Outline

- About me
- Introduction: Sonos Voice Solution
- **Multichannel Weiner Filter and Speech Presence Probability**
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- Commercialization & Conclusions

Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter

Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter

Saeed Bagheri Daniele Giacobello

SONOS

INTERSPEECH 2019

Introduction

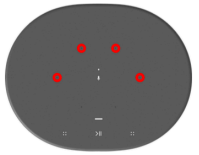
- Sonos voice enabled smart speakers

Sonos One

Sonos Beam

Sonos Move

- **Challenges**:
  - Different microphone-array geometries and configurations
  - Different industrial design, form factors, and HW modules
  - Different performance requirements and use cases



Introduction

- Sonos voice enabled smart speakers

Sonos One

Sonos Beam

Sonos Move

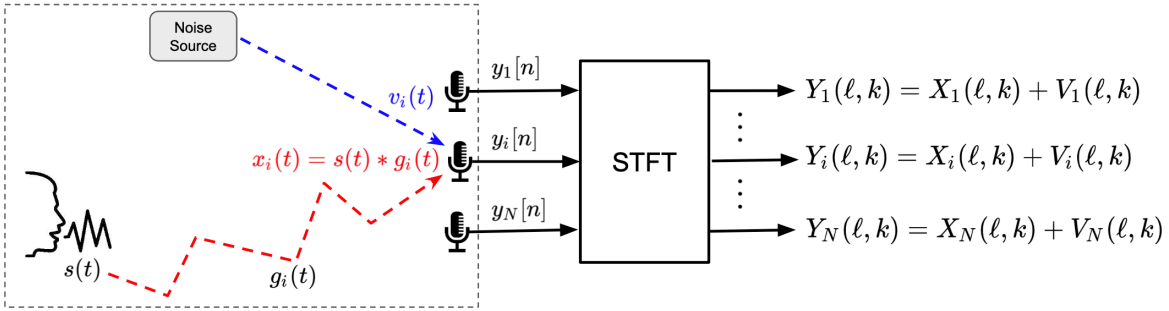
- **Objectives**:
  - A robust and scalable far-field multi-channel noise reduction method
  - Easy to deploy on different devices, and different microphone-array geometries
  - Fast prototyping, testing, and deployment
  - Trade-off noise reduction level to optimize device specific performance metrics
  - Long life-time $\longrightarrow$ fault tolerant!
  - Generalization to distributed applications (e.g., Sonos home sound system)

Relevant work

- Adaptive multi-channel noise reduction techniques
  - Frost beamformer [Frost, 1972]
  - Linearly constrained minimum variance (LCMV) beamformers [Er and Cantoni, 1983; Darlington, 1958]
  - Minimum variance distortion-less response (MVDR) beamformer [Capon, 1969]
  - Generalized side-lobe canceler (GSC) [Griffiths and Jim, 1982]
  - Multi-channel Wiener filter (MWF) [Benesty et al., 2008]
- A **common frequency-domain framework** is proposed in [Souden et al., 2010a]
  - MVDR, GSC, and the parametric multi-channel Wiener filter (**PMWF**) are formulated in the same framework
  - Trade-off between noise reduction and speech distortion
  - No assumptions on the geometry of the microphone array

Problem definition

- **Observation Model** (uncorrelated noise and speech signals)

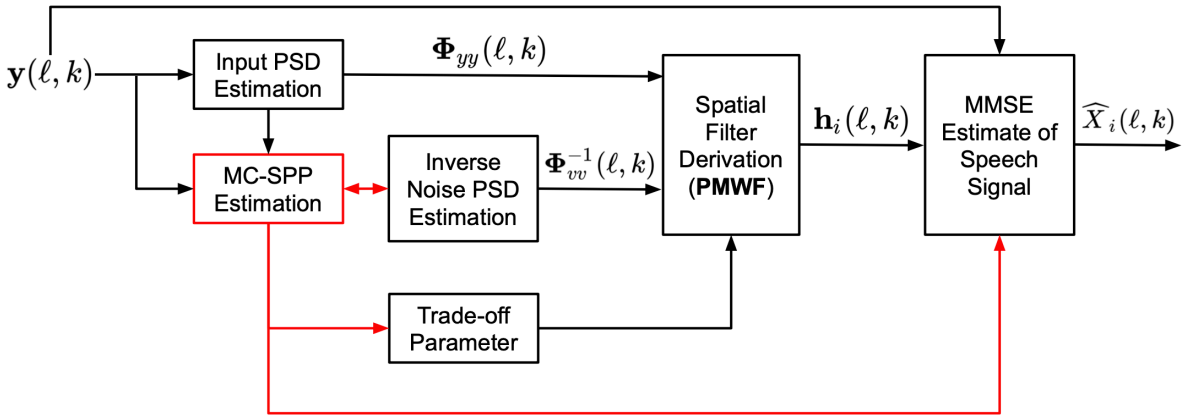


- **Objective**: Apply a **spatial linear filter** to estimate the speech signal

$$\widehat{X}_i(\ell,k) = \mathbf{h}_i^H(\ell,k) \quad \underbrace{\mathbf{y}(\ell,k)}_{\triangleq \begin{bmatrix} Y_1(\ell,k) \\ \vdots \\ Y_N(\ell,k) \end{bmatrix}}$$

Our contributions

- A practical implementation of PMWF which incorporates an estimate of the multi-channel speech presence probability (MC-SPP)



Review: Parametric Multi-Channel Weiner Filter

Constraint optimization formulation in frequency domain [Souden et al., 2010a]

$$\begin{aligned} \underset{\mathbf{h}_i(\ell,k)}{\arg\max} \quad & \overbrace{\zeta_{nr}(\mathbf{h}_i(\ell,k))}^{\text{local noise reduction factor}} \\ \text{subject to} \quad & \underbrace{\nu_{sd}(\mathbf{h}_i(\ell,k))}_{\text{local speech distortion}} \leq \sigma^2(\ell,k) \end{aligned}$$

- $\zeta_{nr}(\mathbf{h}_i(\ell,k))$ is a function of noise PSD matrix $\mathbf{\Phi}_{vv}(\ell,k)$
- $\nu_{sd}(\mathbf{h}_i(\ell,k))$ is a function of speech PSD matrix $\mathbf{\Phi}_{xx}(\ell,k)$

Review: Parametric Multi-Channel Weiner Filter

PMWF solution [Souden et al., 2010a; Doclo and Moonen, 2002; Spriet et al., 2004]

$$\mathbf{h}_i(\ell,k) = \frac{\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k) - \mathbf{I}_N}{\beta(\ell,k) + \xi(\ell,k)}\,\mathbf{e}_i$$

- Depends on the input and noise PSD matrices
- $\xi(\ell,k) \triangleq \text{tr}\Big\{\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\Big\} - N \longrightarrow$ Multi-Channel *a priori* SNR
- $\beta(\ell,k)$: trade-off parameter $\longrightarrow$ Inverse of the Lagrange multiplier
  - $\beta = 0 \Longrightarrow$ MVDR
  - $\beta = 1 \Longrightarrow$ MCWF

MC-SPP Estimation

Speech and noise are modeled as complex multivariate Gaussian random variables

- MC-SPP expression [Souden et al., 2010b]

$$\underbrace{p(\ell,k)}_{\text{MC-SPP}} = \left\{1 + \frac{q(\ell,k)}{1-q(\ell,k)}[1+\xi(\ell,k)]\exp\left[-\frac{\gamma(\ell,k)}{1+\xi(\ell,k)}\right]\right\}^{-1}$$

- $q(\ell,k)$: the *a priori* speech absence probability [Cohen, 2003; Souden et al., 2011]
- $\xi(\ell,k)$: Multi-Channel *a priori* SNR
- $\gamma(\ell,k) \triangleq \mathbf{y}^H(\ell,k)\left[\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\mathbf{\Phi}_{vv}^{-1}(\ell,k) - \mathbf{\Phi}_{vv}^{-1}(\ell,k)\right]\mathbf{y}(\ell,k)$

MC-SPP Estimation

Speech and noise are modeled as complex multivariate Gaussian random variables

- MC-SPP expression [Souden et al., 2010b]

$$\underbrace{p(\ell,k)}_{\text{MC-SPP}} = \left\{1 + \frac{q(\ell,k)}{1-q(\ell,k)}[1+\xi(\ell,k)]\exp\left[-\frac{\gamma(\ell,k)}{1+\xi(\ell,k)}\right]\right\}^{-1}$$

- $q(\ell,k)$: the *a priori* speech absence probability [Cohen, 2003; Souden et al., 2011]
- $\xi(\ell,k)$: Multi-Channel *a priori* SNR
- $\gamma(\ell,k) \triangleq \mathbf{y}^H(\ell,k)\left[\mathbf{\Phi}_{vv}^{-1}(\ell,k)\mathbf{\Phi}_{yy}(\ell,k)\mathbf{\Phi}_{vv}^{-1}(\ell,k) - \mathbf{\Phi}_{vv}^{-1}(\ell,k)\right]\mathbf{y}(\ell,k)$
- **Smoothing**:
  - $\overline{p}(\ell,k) = \alpha_p\,\overline{p}(\ell-1,k) + (1-\alpha_p)\,p(\ell,k)$
  - $\overline{p}(\ell,k) = \min\left\{\max\left\{\overline{p}(\ell,k),\, p_{\text{min}}\right\},\, p_{\text{max}}\right\}$

Utilization of MC-SPP - Part 1

- **Estimation of PSD Matrices**
  - Input PSD matrix: $\widehat{\mathbf{\Phi}}_{yy}(\ell,k) = \alpha_y\, \widehat{\mathbf{\Phi}}_{yy}(\ell-1,k) + (1-\alpha_y)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$
  - Noise PSD matrix: generalization of the IMCRA approach [Cohen, 2003]

    $\rightarrow\ \widehat{\mathbf{\Phi}}_{vv}(\ell,k) = \widetilde{\alpha}_v(\ell,k)\, \widehat{\mathbf{\Phi}}_{vv}(\ell-1,k) + \big(1-\widetilde{\alpha}_v(\ell,k)\big)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$

    $\rightarrow\ \widetilde{\alpha}_v(\ell,k) \triangleq \alpha_v + (1-\alpha_v)\underbrace{\overline{p}(\ell,k)}_{\text{MC-SPP}}$

Utilization of MC-SPP - Part 1

- **Estimation of PSD Matrices**
  - Input PSD matrix: $\widehat{\mathbf{\Phi}}_{yy}(\ell,k) = \alpha_y\, \widehat{\mathbf{\Phi}}_{yy}(\ell-1,k) + (1-\alpha_y)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$
  - Noise PSD matrix: generalization of the IMCRA approach [Cohen, 2003]

    $\rightarrow\ \widehat{\mathbf{\Phi}}_{vv}(\ell,k) = \widetilde{\alpha}_v(\ell,k)\, \widehat{\mathbf{\Phi}}_{vv}(\ell-1,k) + \big(1-\widetilde{\alpha}_v(\ell,k)\big)\, \mathbf{y}(\ell,k)\mathbf{y}^H(\ell,k)$

    $\rightarrow\ \widetilde{\alpha}_v(\ell,k) \triangleq \alpha_v + (1-\alpha_v)\, \underbrace{\overline{p}(\ell,k)}_{\text{MC-SPP}}$
- Both PMWF and MC-SPP need the inverse of the noise PSD matrix
  - Matrix inversion lemma

$$\widehat{\mathbf{\Phi}}_{vv}^{-1}(\ell,k) = \frac{1}{\widetilde{\alpha}_v(\ell,k)} \left[\widehat{\mathbf{\Phi}}_{vv}^{-1}(\ell-1,k) - \frac{\widetilde{\mathbf{y}}(\ell,k)\, \widetilde{\mathbf{y}}^H(\ell,k)}{g(\ell,k)} \right]$$

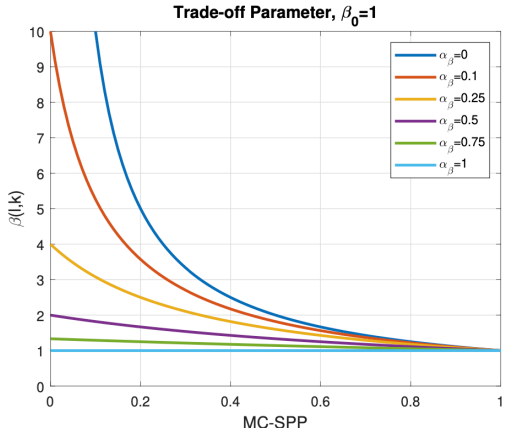
  - $\widetilde{\mathbf{y}}(\ell,k) \triangleq \widehat{\mathbf{\Phi}}_{vv}^{-1}(\ell-1,k)\, \mathbf{y}(\ell,k)$ and $g(\ell,k) \triangleq \frac{\widetilde{\alpha}_v(\ell,k)}{1-\widetilde{\alpha}_v(\ell,k)} + \mathbf{y}^H(\ell,k)\widetilde{\mathbf{y}}(\ell,k)$

Utilization of MC-SPP - Part 2

- **MC-SPP Controlled trade-off parameter**
  - The *a posteriori* SPP has been used to control the trade-off between noise reduction and speech distortion [Ngo et al., 2009]

$$\beta(\ell, k) = \frac{\beta_0}{\alpha_\beta + (1 - \alpha_\beta)\,\beta_0\;\overline{p}(\ell, k)}$$

  - Outperforms a fixed trade-off parameter
  - Flexible for device and application specific tuning



Utilization of MC-SPP - Part 3

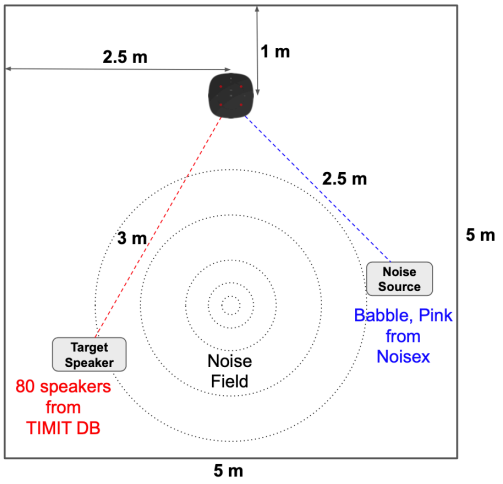
- **MMSE estimate of the desired speech signal**

$$\widehat{X}_i(\ell,k) = \overline{p}(\ell,k)\ \underbrace{\mathbf{h}_i^H(\ell,k)\,\mathbf{y}(\ell,k)}_{\text{PMWF output}} + (1-\overline{p}(\ell,k))\ G_{\text{min}}\ Y_i(\ell,k)$$

  - Reduces the speech distortion caused due to the estimation error in MC-SPP
  - G_{min} determines the maximum amount of noise suppression
  - G_{min} is tuned to optimize the performance metrics of interest (*e.g.*, word error rate in ASR systems, wake-word detection rate)

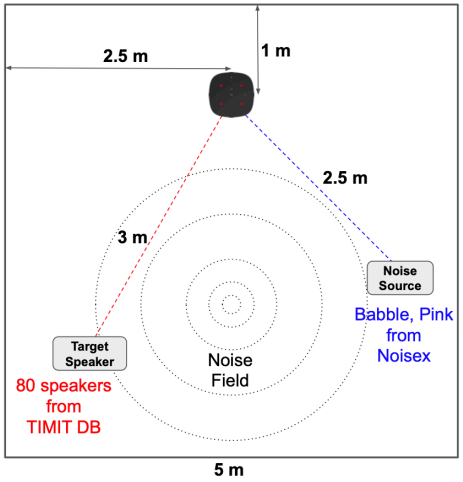
Simulation Setup

Sampling frequency	16 KHz
Microphone array	Sonos One
# of Microphones	4
Frame length	512
Frame overlap	50%
Window function	Hann
T_{60}	300 ms
RIR generation	Image source method



Simulation Setup

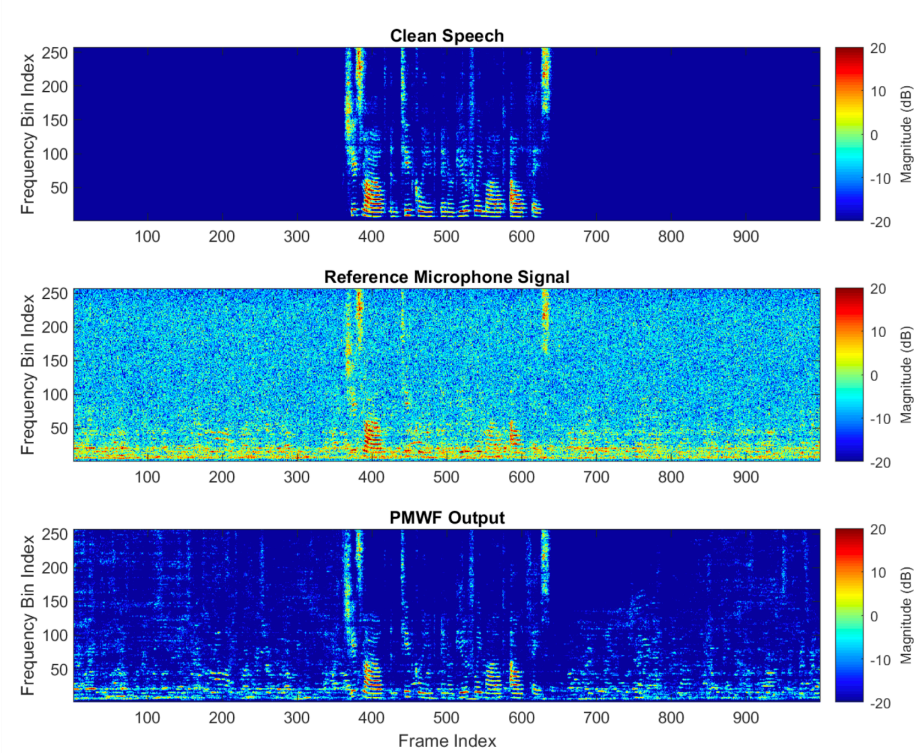
Sampling frequency	16 KHz
Microphone array	Sonos One
# of Microphones	4
Frame length	512
Frame overlap	50%
Window function	Hann
T_{60}	300 ms
RIR generation	Image source method



Parameters used to implement the proposed algorithm

$\alpha_v = 0.95$	$\alpha_y = 0.95$	$\alpha_p = 0.1$	$G_{\text{min}} = 0.1$
α_β varies	β_0 varies		
$p_{\text{max}} = 0.99$	$p_{\text{min}} = 0.01$	$q_0 = 0.5$	

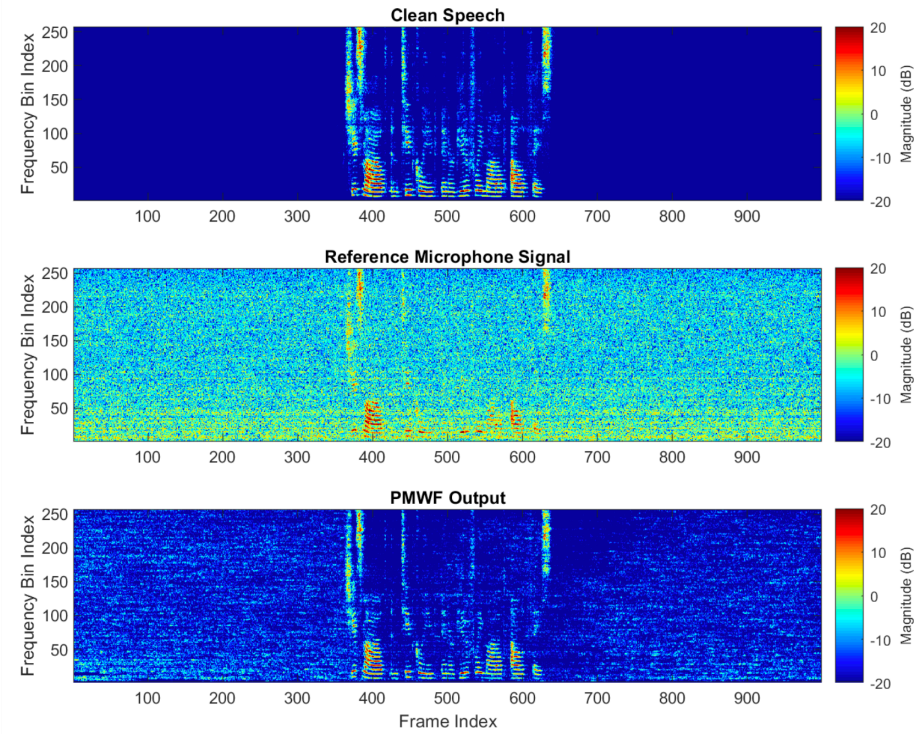
Example 1: Babble Noise, Input SINR = 0dB



Metrics

ΔSINR	11.85 dB
NR	16.12 dB
SD	−5.82 dB

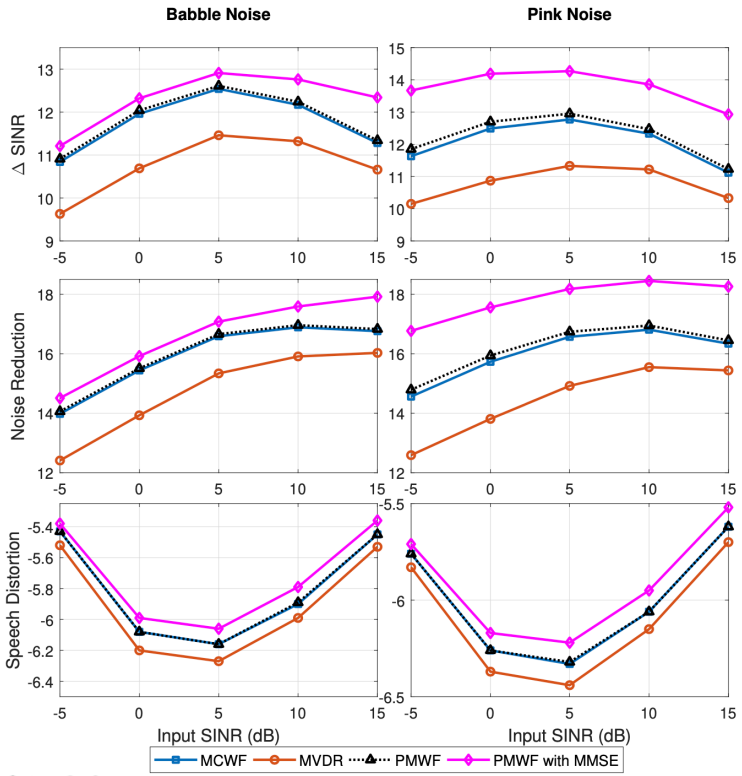
Example 2: Pink noise, Input SINR = 0dB



Metrics

ΔSINR	15.06 dB
NR	18.28 dB
SD	−6.49 dB

Simulation Results



Configurations:

Test Name	β_0	α_β	MMSE estimate
MVDR	0	1	No
MCWF	1	1	No
PMWF	1	0.75	No
PMWF with MMSE	1	0.75	Yes

PMWF Performance:

- Better NR, and ΔSINR
- Minor increase in SD
- In practice, trade-off NR by tuning β_0 and α_β to optimize the performance metrics of interest (*e.g.*, word error rate in ASR systems, wake-word detection rate)

Conclusions

- A **robust** and **scalable** far-field multi-channel noise reduction method
  - Improvement in SINR, and NR performance
  - Trade-off NR level to optimize device-specific performance metrics (PESQ, WWD, and WER)
  - Applicable to different microphone-array geometries
  - Easy to deploy on different devices after proper tuning of hyper-parameters

References

J. Benesty, J. Chen, and Y. Huang. *Microphone array signal processing*. Springer Science & Business Media, 2008.

J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8): 1408–1418, 1969. ISSN 0018-9219. doi: 10.1109/PROC.1969.7278.

I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475, 2003. ISSN 1063-6676. doi: 10.1109/TSA.2003.811544.

S. Darlington. Linear least-squares smoothing and prediction, with applications. *The Bell System Technical Journal*, 37(5):1221–1294, 1958. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1958.tb01550.x.

S. Doclo and M. Moonen. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9):2230–2244, Sep. 2002. ISSN 1053-587X. doi: 10.1109/TSP.2002.801937.

S. Doclo, S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters. Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):38–51, Jan 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2004291.

L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982. ISSN 0018-926X. doi: 10.1109/TAP.1982.1142739.

K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen. Incorporating the conditional speech presence probability in multi-channel wiener filter based noise reduction in hearing aids. *EURASIP Journal on Advances in Signal Processing*, 2009.

M. Souden, J. Benesty, and S. Affes. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):260–276, 2010a. ISSN 1558-7916. doi: 10.1109/TASL.2009.2025790.

M. Souden, J. Chen, J. Benesty, and S. Affes. Gaussian model-based multichannel speech presence probability. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):1072–1077, 2010b. ISSN 1558-7916. doi: 10.1109/TASL.2009.2035150.

M. Souden, J. Chen, J. Benesty, and S. Affes. An integrated solution for online multichannel noise tracking and reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2159–2169, 2011. ISSN 1558-7916. doi: 10.1109/TASL.2011.2118205.

A. Spriet, M. Moonen, and J. Wouters. Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction. *Signal Processing*, 84(12):2367 – 2387, 2004. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2004.07.028. URL `http://www.sciencedirect.com/science/article/pii/S0165168404002002`.

End of Part 1

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability

Part 1

- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
  - Distributed signal processing
    - Homogeneous networks
      - Node-Specific Noise Reduction
      - Node-Specific DOA Estimation
    - Heterogeneous networks
- Commercialization & Conclusions

Part 2

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- **Speech Enhancement via Multiple Network Microphone Devices**
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- Commercialization & Conclusions

Linear Filtering for Noise-Suppressed Speech Detection via Multiple Network Microphone Devices

- Information sharing across devices

(12) **United States Patent**
Sereshki et al.

(10) **Patent No.: US 10,692,518 B2**
(45) **Date of Patent: Jun. 23, 2020**

(54) **LINEAR FILTERING FOR NOISE-SUPPRESSED SPEECH DETECTION VIA MULTIPLE NETWORK MICROPHONE DEVICES**

(71) Applicant: **Sonos, Inc.**, Santa Barbara, CA (US)

(72) Inventors: **Saeed Bagheri Sereshki**, Goleta, CA (US); **Daniele Giacobello**, Los Angeles, CA (US)

(73) Assignee: **Sonos, Inc.**, Santa Barbara, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 9 days.

(21) Appl. No.: **16/147,710**

(22) Filed: **Sep. 29, 2018**

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,741,038 A 4/1988 Elko et al.
4,941,187 A 7/1990 Slater
(Continued)

FOREIGN PATENT DOCUMENTS

AU 2017100486 A4 6/2017
AU 2017100581 A4 6/2017
(Continued)

OTHER PUBLICATIONS

US 9,299,346 B1, 03/2016, Hart et al. (withdrawn)
(Continued)

Primary Examiner — Feng-Tzer Tzeng
(74) *Attorney, Agent, or Firm* — Fortem IP LLP; Mary L. Fox

Central node configuration

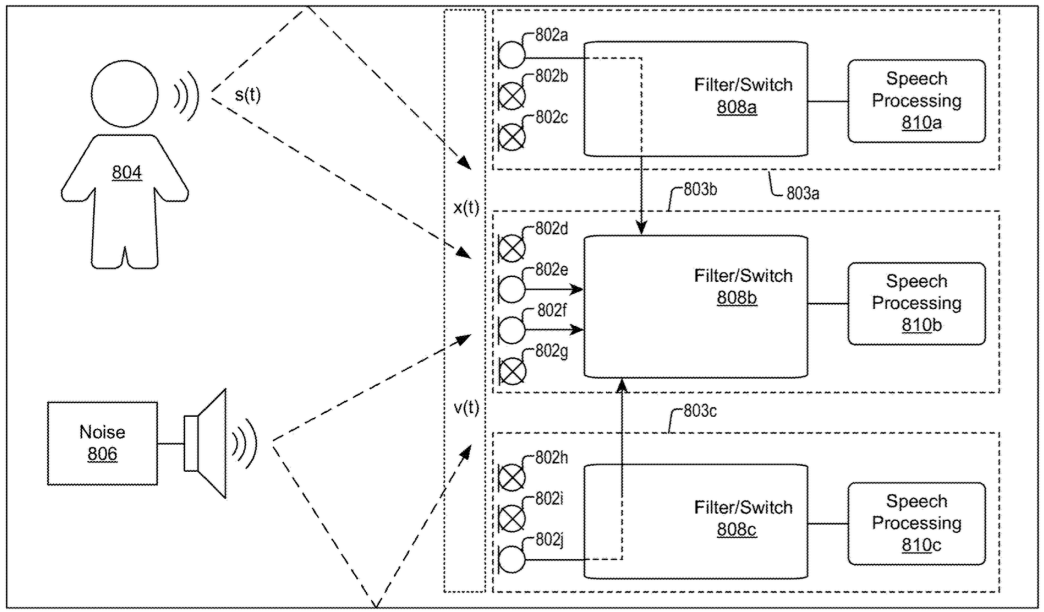


Figure 8D

- Introducing the concept of microphone or information sharing across devices

Two network microphone devices (NMD)

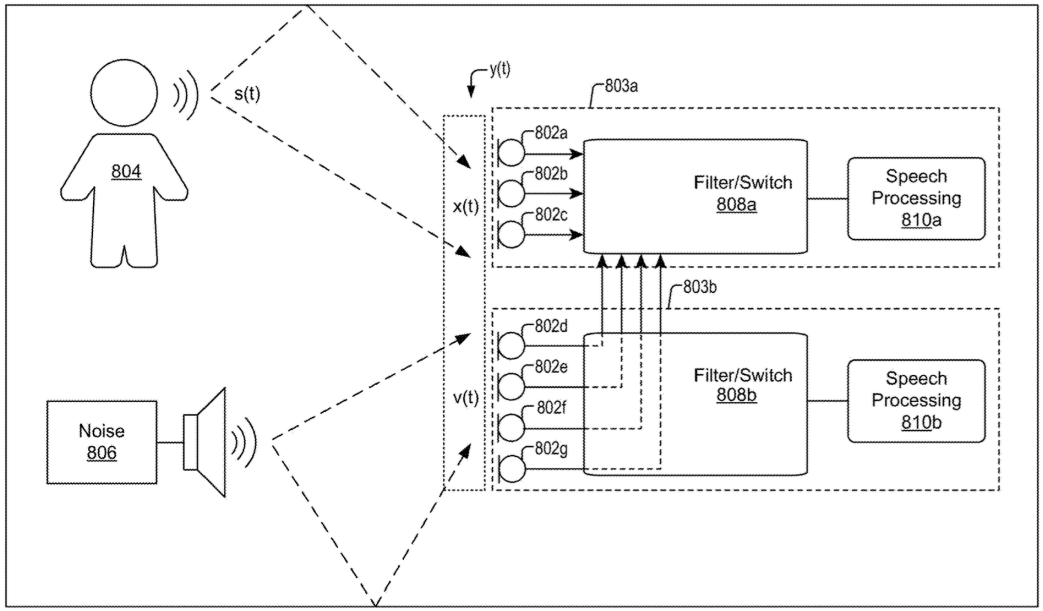


Figure 8A

Reduced-microphone configuration

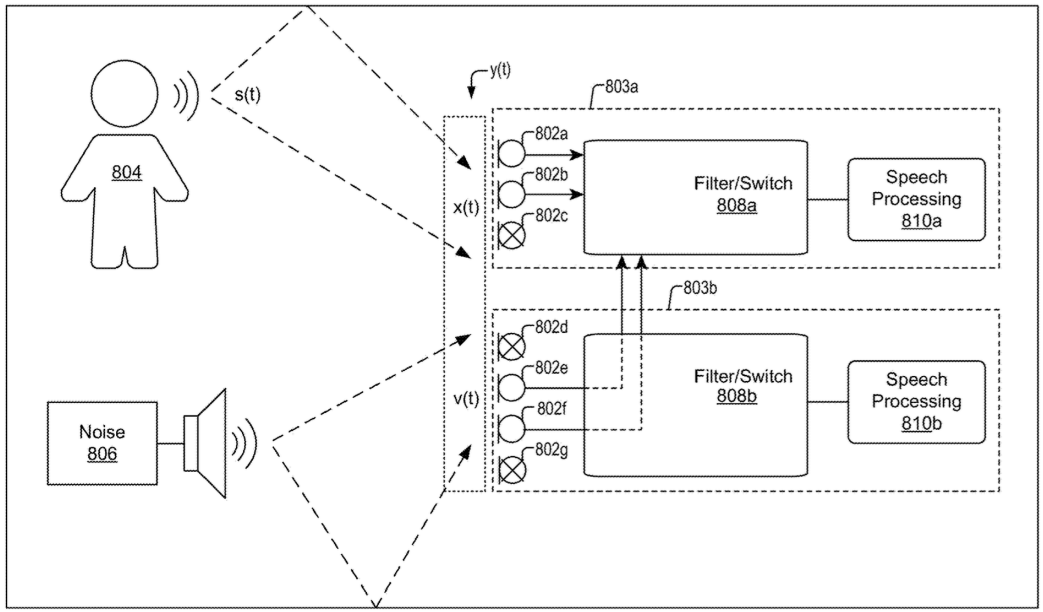


Figure 8B

Information Sharing

MCSPP based method

- The current system use a single channel speech mask to drive the PMWF.
- First, we replace the single channel speech masks with multichannel speech presence probability masks to drive the PMWF.
- Second, we combine the multichannel speech masks generated on two separate devices to drive the PMWF on the primary device.

MCSPP based method

Analytical expression and mask sharing

- Multichannel speech presence probability is computed using the following expression,

$$M(k,n) = \left\{ 1 + \frac{q(k,n)}{1-q(k,n)}[1+\epsilon(k,n)]\exp\left[-\frac{\beta(k,n)}{1+\epsilon(k,n)}\right] \right\}$$

$$\beta(k,n) = \mathbf{y}^H(k,n)\Phi_{vv}^{-1}(k,n)\Phi_{xx}(k,n)\Phi_{vv}^{-1}(k,n)\mathbf{y}(k,n)$$

$$\epsilon(k,n) = \text{tr}\left[\Phi_{vv}^{-1}(k,n)\Phi_{xx}(k,n)\right]$$

A-priori speech absence probability (SAP)

Audio data matrix

Noise PSD

Speech PSD

- MCSPP combined from two devices using the following relations,
  - Arithmetic Mean : $M(k,n) = \frac{M_1(k,n) + M_2(k,n)}{2}$
  - Geometric Mean: $M(k,n) = \sqrt{M_1(k,n)\,.\,M_2(k,n)}$
  - Product: $M(k,n) = M_1(k,n)\,.\,M_2(k,n)$

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- **Wireless acoustic sensor networks (WASNs)**
  - Distributed signal processing
    - Homogeneous networks
      - Node-Specific Noise Reduction
      - Node-Specific DOA Estimation
    - Heterogeneous networks
- Commercialization & Conclusions

Special Thanks to

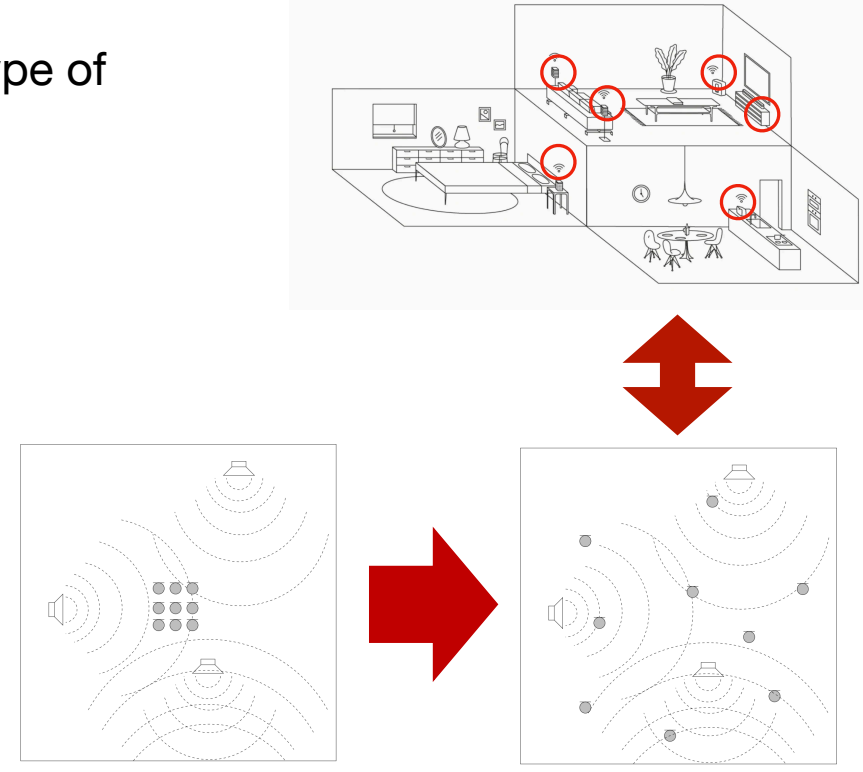
Dr. Amin Hassani (KU Leuven, Connexounds)

for providing some of the material used here!

Wireless Acoustic Sensor Networks (WASNs)

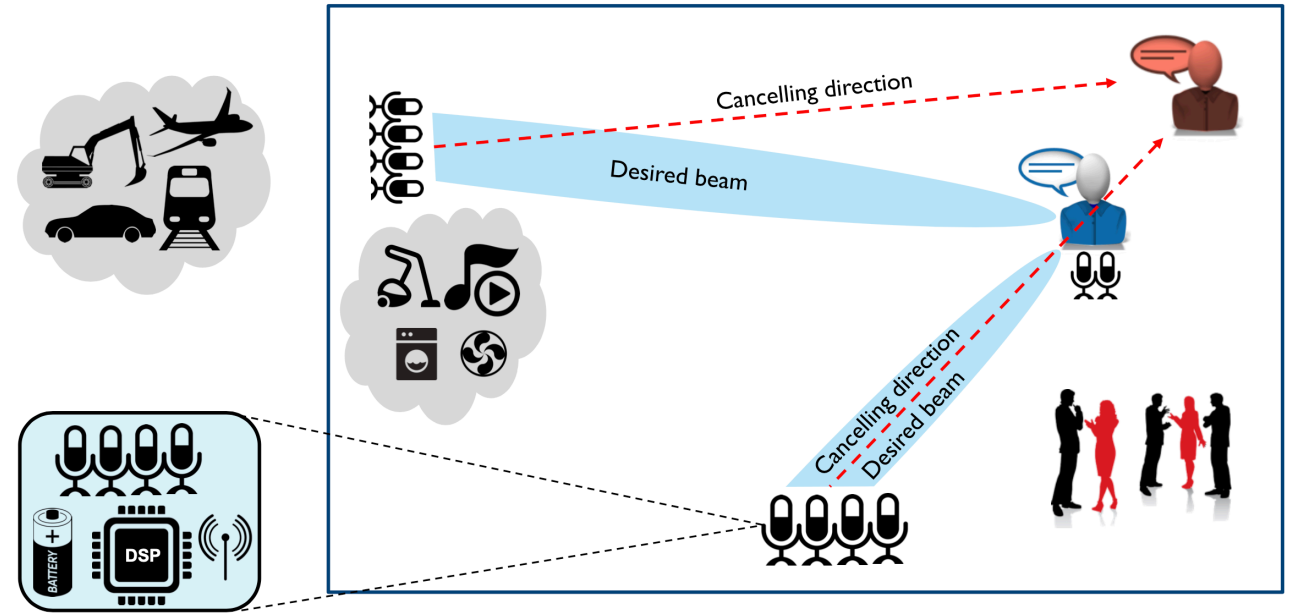
Wireless Acoustic Sensor Networks

- Wireless Acoustic Sensor Networks support the type of architecture common to the Sonos ecosystem



A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective." B. *18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*. IEEE, 2011.

WASN = Multiple Microphone Arrays



Possible candidates: smart phones, laptops, hearing aids

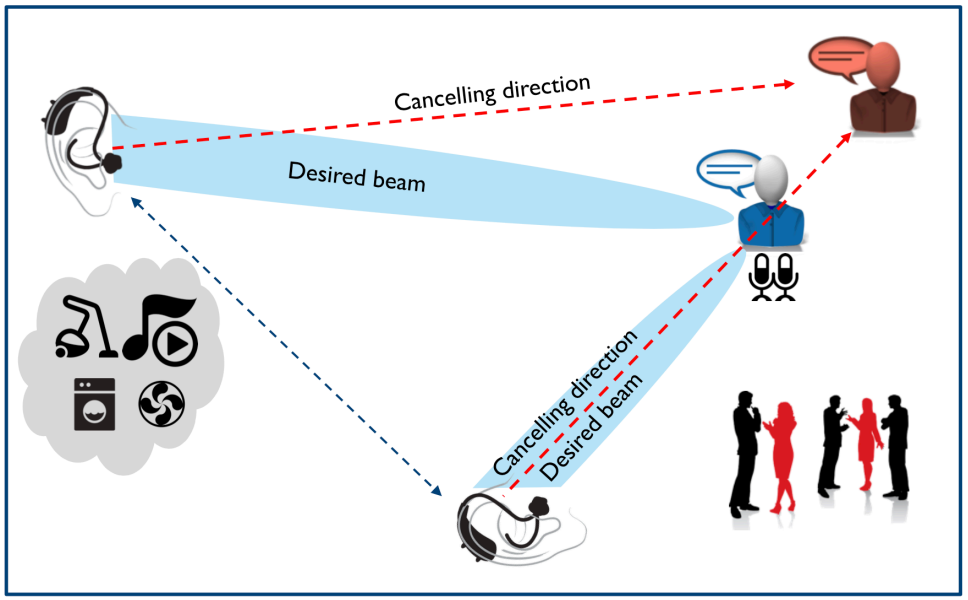
Performance of beamforming, MWF, DOA estimation improves with the **number** of microphones

Benefits and Applications of WASNs

- Benefits:
  - More microphones, better spacing ➔ more spatial information
  - High probability to find microphones close to the target source (with high SNR)
  - Wireless connectivity ➔ more flexible, suitable for wearable or mobile applications (e.g., hearing aids)

- Applications:
  - Cooperative hearing devices (e.g., binaural hearing aids, supported by external microphones or other audio devices)
  - Cooperative voice capture for multi-speaker (conferencing) systems
  - Cooperative meeting transcription systems
  - Cooperative speech activity presence detection

Node-Specific Signal and Parameter Estimation in WASN

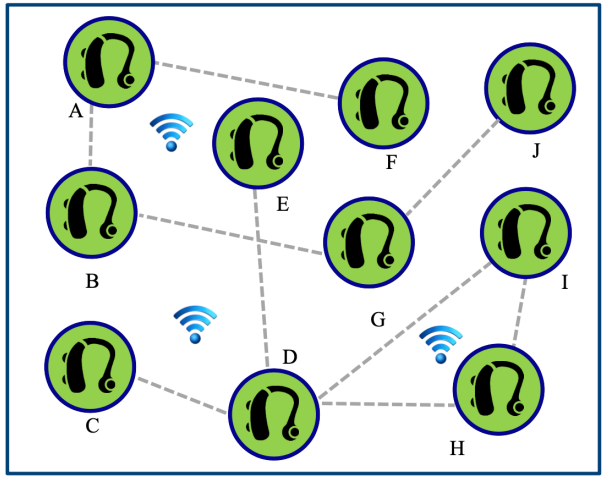


Why **node-specific**? E.g., to preserve binaural cues for directional hearing

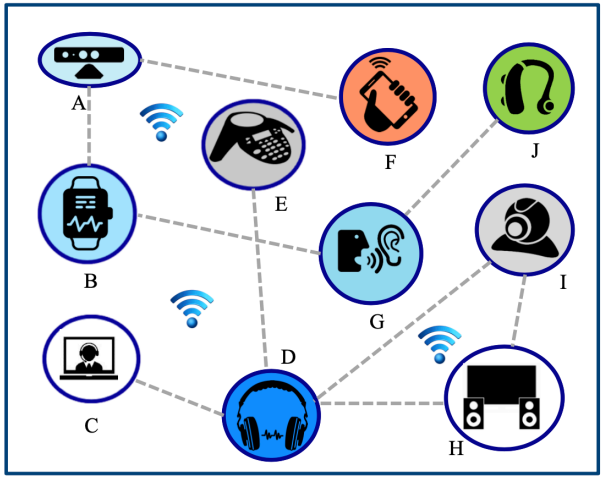
One node-specific task for each node ➔ **multi-task WASNs**

As opposed to single-task WASNs, e.g., for automatic meeting transcription

Devices in a Multi-Task WASN



- **Similar Devices:** e.g., connected hearing aids
- Special case: 2-node WASN: binaural hearing aids



- **Different Devices:** diverse hardware capabilities and resources (IoT-like)
- Different power budget, communication range, computational power.

Homogeneous vs. Heterogeneous Multi-Task WASNs

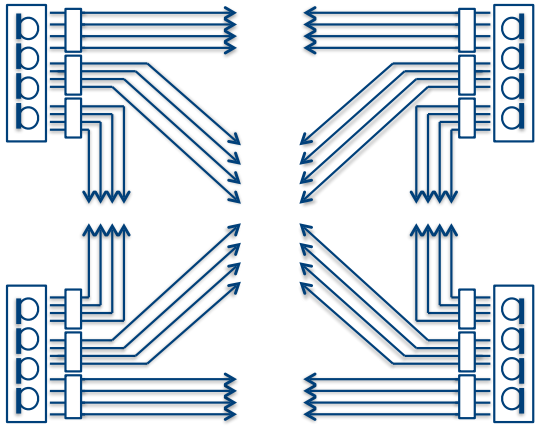
- **Homogeneous multi-task WASNs**
  - All nodes perform the same basic ASP technique (e.g., beamforming, MWF, or DOA estimation)
- **Heterogeneous multi-task WASNs**
  - Nodes perform different basic ASP techniques

Outline

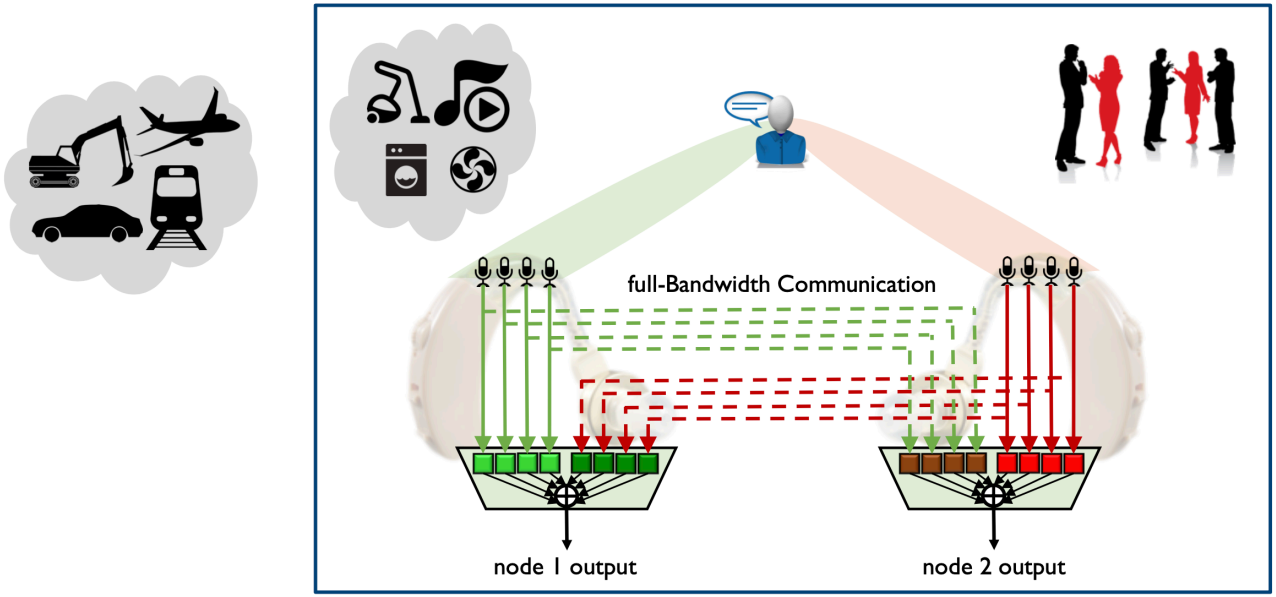
- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
  - **Distributed signal processing in WASNs**
    - Homogeneous networks
      - Node-Specific Noise Reduction
      - Node-Specific DOA Estimation
    - Heterogeneous networks
- Commercialization & Conclusions

Centralized Processing

- In a world without complexity constraints, there is no need for coalitional SP:
  - Every device sends all of its sensor signals to all other devices
  - Every device has access to all sensor signals in the network and with these solves its own SP problem
  - In a network with 10 devices and 10 sensors each, communication cost is O(100) (with broadcasting), O(1000)(without)



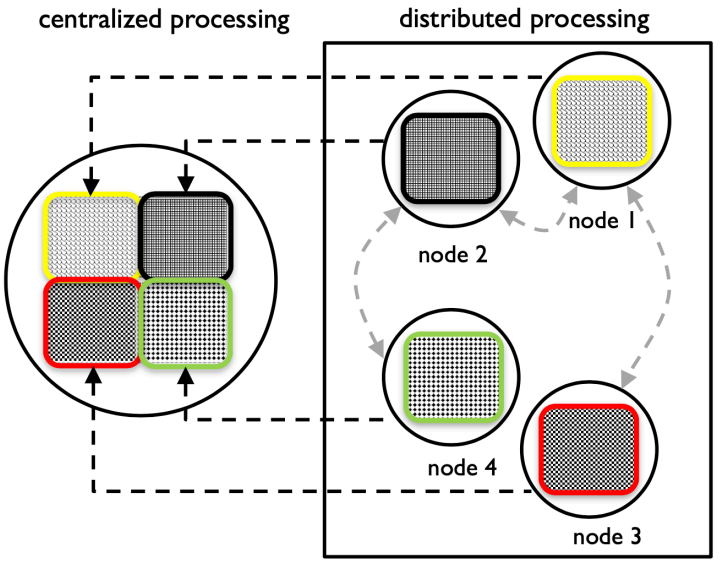
Centralized Processing in a Binaural HA



- Centralized estimation defines the 'optimal' performance: exploits full coherence between signals
- Drawbacks: large per-node communication bandwidth, heavy computational burden

Distributed Processing in a WASN

- **Distributes** the heavy processing of the between the nodes
- Reduces per-node communication and computational load
- Still optimal performance?? Challenging!



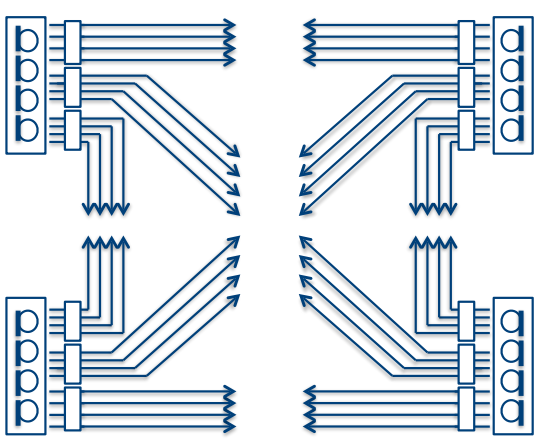
Distributed Processing in a WASN

- More realistically, under complexity constraints, need to reduce
  - Communication cost (number of signals, coding, …)
  - Computational cost (in each node)

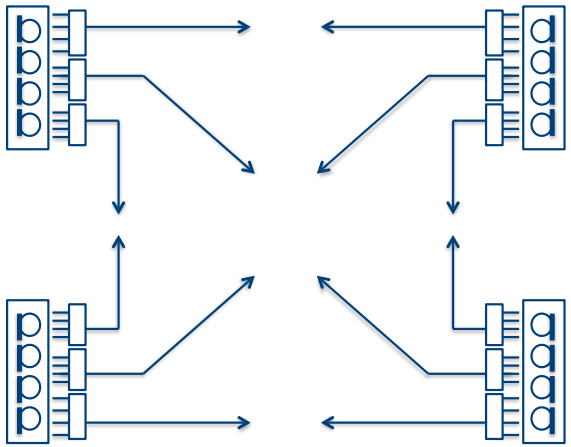
- Research focus on
  - reducing the number of signals communicated between devices (= communication cost).
  - Leads to computational cost reduction as well (cfr. O(M^2)…O(M^3) algorithms)

- Caveat
  - Signals estimated are linearly dependent on a common latent random process (i.e., dimensionality reduction possible!)

Distributed Processing in a WASN



#s of signals communicated between devices reduced by means of **FUSION RULES**

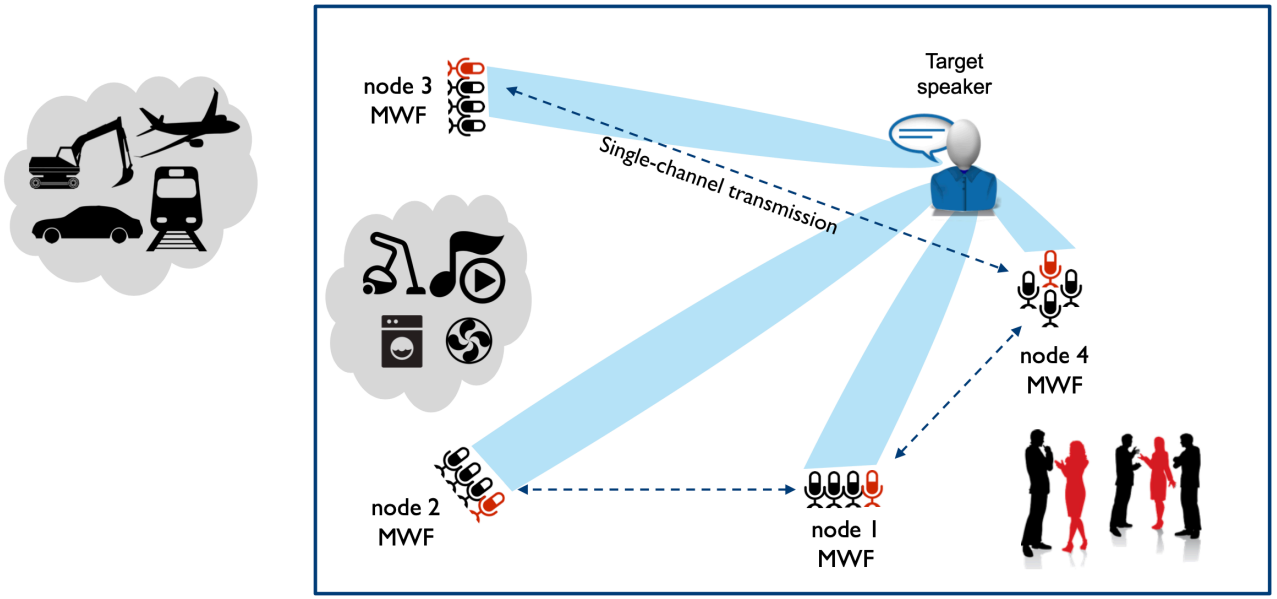


Distributed Processing in a WASN

- Fusion Rules
  - Each device transmits only (one or a few) ’fused’ (=linearly combined) versions of its sensor signals
  - The fusion rules (=linear weights) will be computed in an iterative procedure (spread over time)
- Upon convergence, each device achieves the same performance in its SP task as if it had access to all sensor signals in the network

Distributed Adaptive Node-specific Signal Estimation

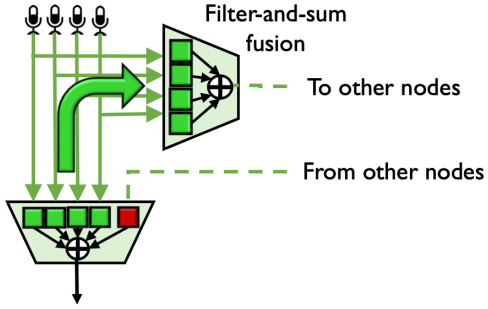
(DANSE) [A. Bertrand, 2010]



- Each node uses an MWF to estimate the speech signal as locally observed at its own reference microphone (to preserve spatial cues)
- One (centralized) node-specific MWF is defined for each node

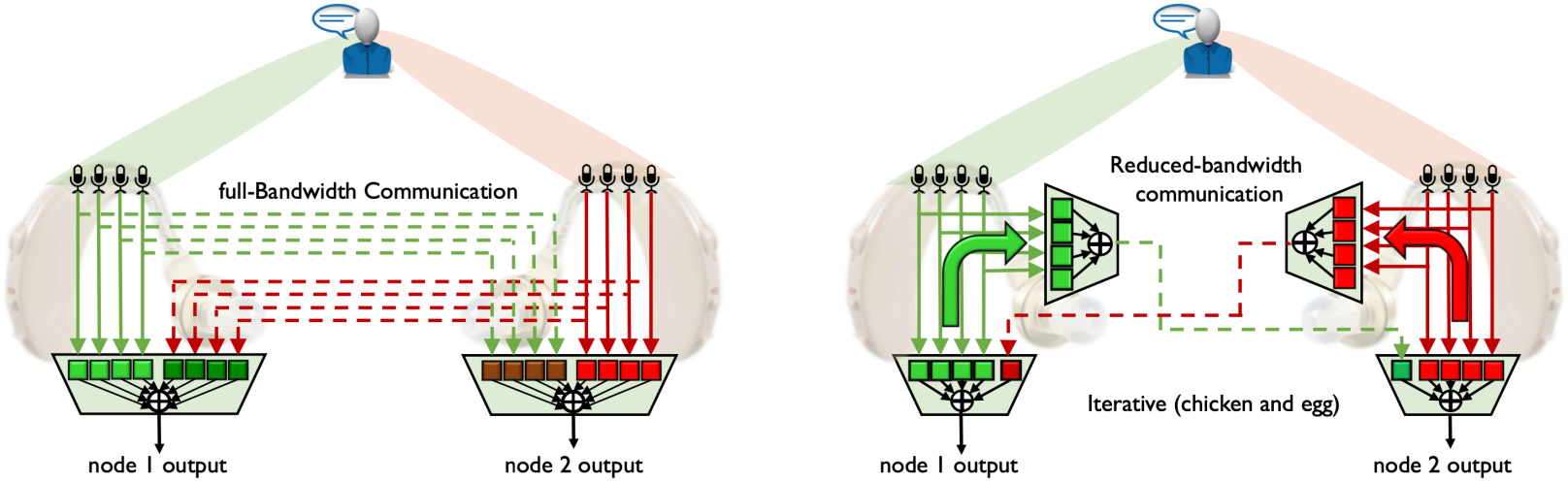
Summary of the DANSE Algorithm

- DANSE: distributed implementation of the centralized (=optimal) MWFs
- Nodes only exchange fused microphone signals
  - Single-speaker case: single-channel broadcast signal per node
  - S-speaker case: S-channel broadcast signals per node
- Fusion rules are part of the local MWF vector
- Iterative computation
- Converges to centralized (=optimal) node-specific MWFs
  (as if nodes have access to all microphone signals)



Distributed Processing with Filter-and-Sum compression

Hearing Aid Example



- Compression: linear **filter-and-sum** signal 'fusion', fusion rules: task-dependent
- For efficient communication scheme and fulfill the real-time processing requirements: each block of new microphone samples transmitted **only once** ➔ iterations **spread over time**

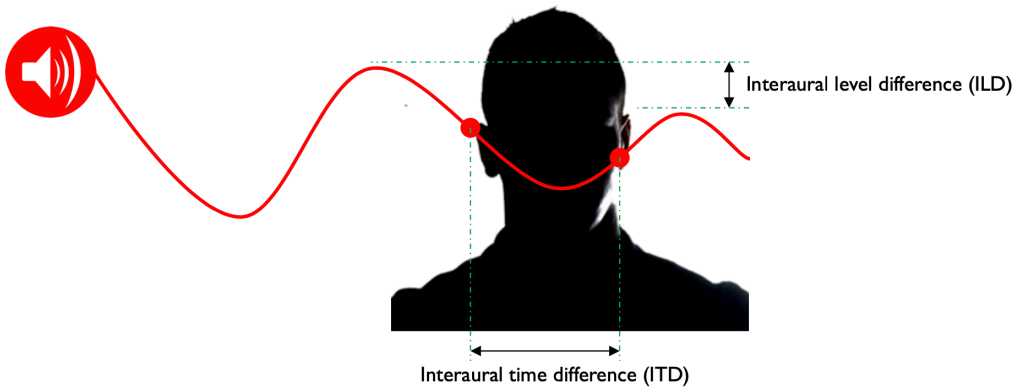
Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
  - Distributed signal processing
    - **Homogeneous networks**
      - **Node-Specific Noise Reduction**
      - Node-Specific DOA Estimation
    - Heterogeneous networks
- Commercialization & Conclusions

Homogeneous networks

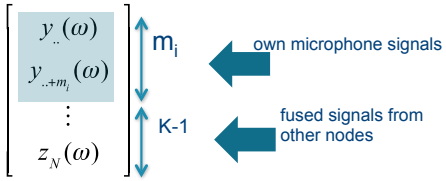
Node-specific noise reduction/speech enhancement

- Example: Binaural hearing aids
  - Two devices with wireless link & cooperation
  - Need ‘node specific’ signal estimation to preserve binaural cues



Node-specific noise reduction/speech enhancement

- Distributed algorithm
  - For each block of samples (=iteration),
    - Every node **fuses** own microphone signals into 1 signal (based on its current fusion vector) and transmits this to other nodes
    - One node is ‘updating node’ (round robin):
      - Perform **local reduced size MWF** with own microphone signals + fused signals received from other nodes, i.e.
        - Update local MWF
        - Update local fusion rule



      - Every node produces local output based on its current local MWF

Node-specific noise reduction/speech enhancement

- Converges to optimal (=centralized) MWF solution for each node, as if node has access to all microphone signals in the network
- Can be extended to (K-speaker) rank-K case
  - K signals transmitted per node, instead of 1
- Can be modified for tree topologies & ad-hoc topologies

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - **Homogeneous networks**
            - Node-Specific Noise Reduction
            - **Node-Specific DOA Estimation**
    - Heterogeneous networks
- Commercialization & Conclusions

Homogeneous networks

Node-specific DoA estimation

- Each node has multiple microphones
- Each node knows its own geometry but does not know the network-wide geometry
- Each node estimates DoA for a desired speech signal, under background noise
  - **Node-specific DOA estimation**
- Network topology: Fully connected

Node-specific DoA estimation

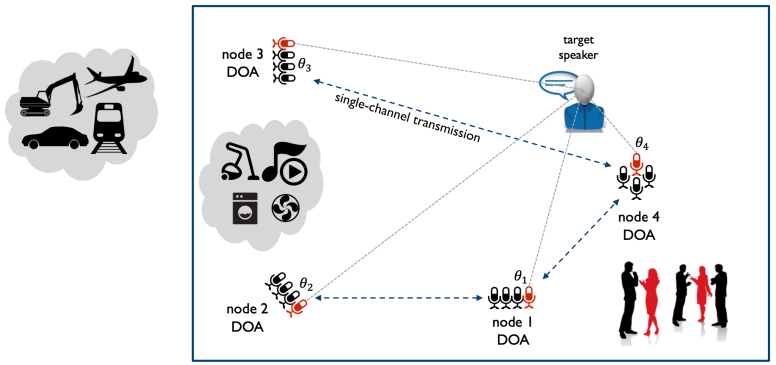
- Will assume same 1-speaker rank-1 data model

$$\begin{bmatrix} y_1(\omega) \\ y_{m_1}(\omega) \\ \vdots \\ y_M(\omega) \end{bmatrix} = \begin{bmatrix} d_1(\omega) \\ d_{m_1}(\omega) \\ \vdots \\ d_M(\omega) \end{bmatrix} s(\omega) + \begin{bmatrix} n_1(\omega) \\ n_{m_1}(\omega) \\ \vdots \\ n_M(\omega) \end{bmatrix}$$

node-1 part

node-K part

- Each node can estimate its node-specific DoA (for instance) by applying the MUSIC-algorithm, which is based on an estimate of its local steering vector (=local di's)
- Local steering vector is (best) obtained as subvector of network-wide steering vector (=all di's)



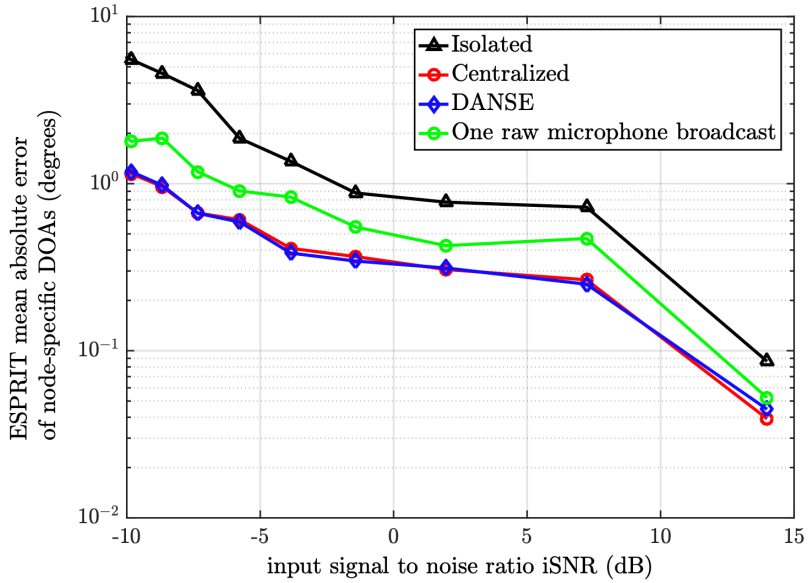
Node-specific DoA estimation

GEVD-based steering vector estimation

- Estimate Ryy, Rnn in speech+noise/noise-only periods
  - MCSPP
- Generalized Eigenvalue Decomposition (GEVD)
- Steering vector obtained through GEVD

Node-specific DoA estimation

- Distributed algorithm
  - For each block of samples (=iteration)
    - Every node **fuses** own microphone signals into 1 signal (based on its current fusion vector) and transmits this to other nodes
    - One node is ‘updating node’ (round robin):
      - Perform **local reduced size GEVD** with own microphone signals + fused signals received from other nodes
      - Update local steering vector estimate
      - Update local fusion rule
    - Every node computes DoA from local steering vector
- Converges to optimal (=centralized)
  - GEVD-based steering vector estimate for each node, as if node has access to all microphone signals in the network!
- Can be extended to (K-speaker) rank-K case
  - K signals transmitted per node, instead of 1
- Can be modified for tree topologies & ad-hoc topologies



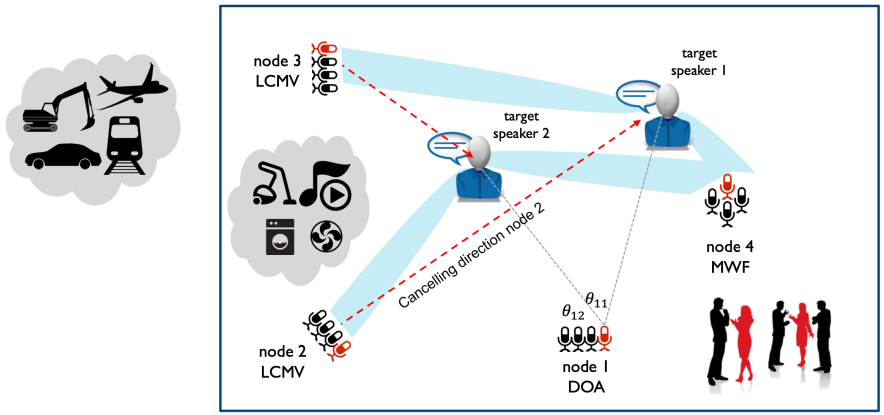
Hassani A., Bertrand A., Moonen M.,"Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network", Signal Processing, vol. 107, Feb. 2015, pp. 68-81

Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - **Heterogeneous networks**
- Commercialization & Conclusions

Heterogeneous multi-task networks

- Multiple target sources, multiple interfering sources: GEVD-based projection-based LCMV(*)+DOA+MWF



Multiple target sources, multiple interfering sources, multiple (localized/diffuse) noise sources

Node 1: DOAs of all the sources Node 2 and 3: LCMV with opposite interests

Node 4: MWF estimates mixture of both sources

Heterogeneous multi-task networks

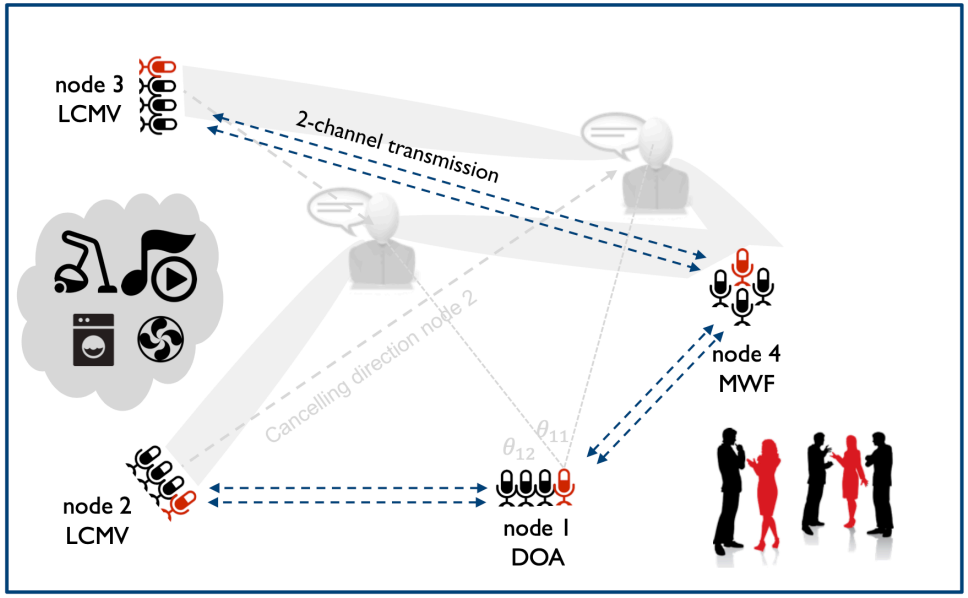
What if different devices have different SP tasks?

- Nodes start exchanging their fused signals to (possibly) enhance their tasks, without being aware of the tasks of other nodes.

- However, the fusion rules are task-dependent è computed/updated according to the basic task locally solved by the node.

- So under such conditions, can these nodes be compatible? Can they really improve each other’s tasks?

Heterogeneous multi-task networks



Solution: distributed algorithm which allows the nodes to obtain the centralized solution of their corresponding node-specific task, while cooperating with reduced-bandwidth per-node signal transmission.

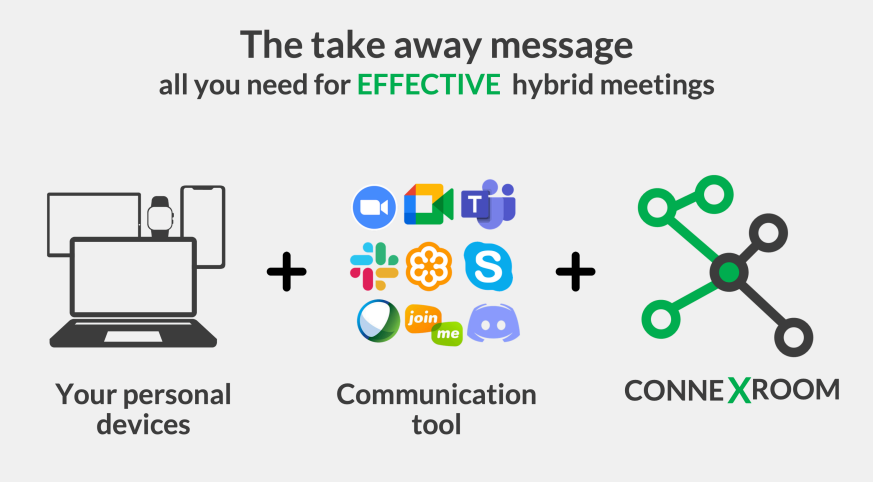
Hassani A., Plata-Chaves J., Bahari M.H., Moonen M., Bertrand A., "Multi-Task Wireless Sensor Network for Joint Distributed Node-Specific Signal Enhancement, LCMV Beamforming and DOA Estimation", IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 3, Apr. 2017, pp. 518 - 533.

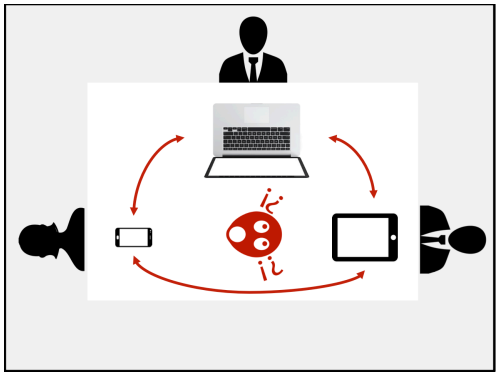
Outline

- About me
- Introduction: Sonos Voice Solution
- Multichannel Weiner Filter and Speech Presence Probability
- Speech Enhancement via Multiple Network Microphone Devices
- Wireless acoustic sensor networks (WASNs)
    - Distributed signal processing
        - Homogeneous networks
            - Node-Specific Noise Reduction
            - Node-Specific DOA Estimation
        - Heterogeneous networks
- **Commercialization & Conclusions**

Meeting Trends

- Many more (ad-hoc) meetings
- Many smaller meeting rooms
- Much more desire in using personal devices and flexible solutions



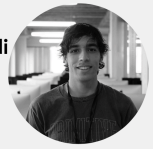


CONNEXOUNDS

CONNEXOUNDS Team

Amin Hassani
CEO & Co-Founder
PhD in Audio DSP

Giuliano Bernardi
CTO & Co-Founder
PhD in Audio DSP



Elias Lajimi
Audio Software developer

Matko Matic
Audio DSP Engineer

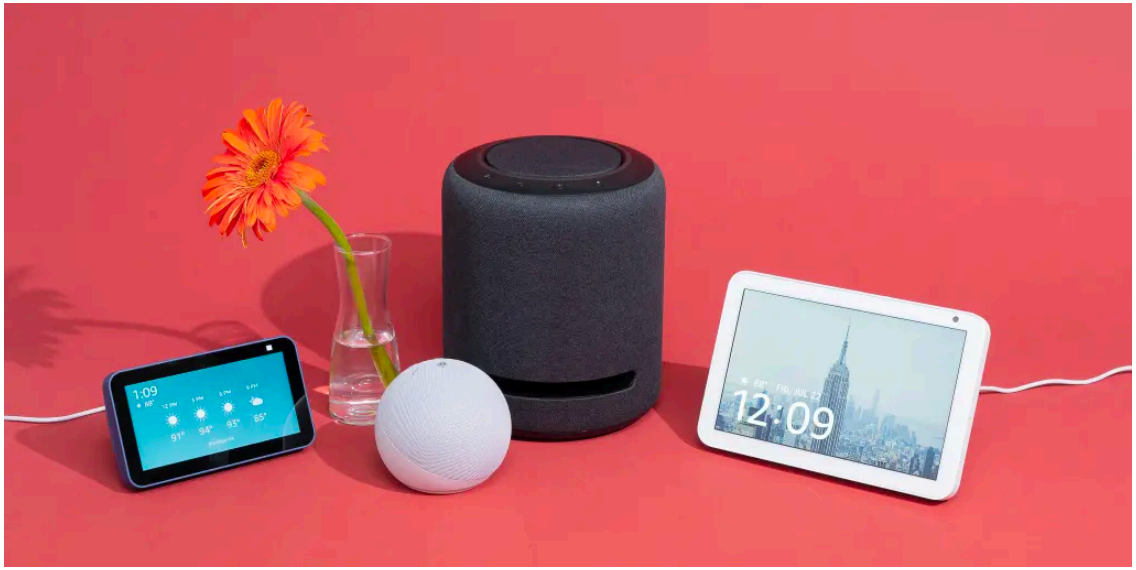
Marc Moonen
Board member & Co-Founder

Toon van Waterschoot
R&D Advisor & Co-Founder

KU LEUVEN

CONNEXOUNDS is a Spin-off of KU LEUVEN (Belgium)

Amazon



END-TO-END ALEXA DEVICE ARBITRATION

Jarred Barber[1]*, *Yifeng Fan*[2]*†, *Tao Zhang*[1]

[1]Amazon Alexa AI, USA
[2] University of Illinois, Urbana-Champaign, USA

FEDERATED LEARNING CHALLENGES AND OPPORTUNITIES: AN OUTLOOK

Jie Ding, Eric Tramel, Anit Kumar Sahu, Shuang Wu, Salman Avestimehr, Tao Zhang

Alexa AI, Amazon

Challenges and Opportunities in Multi-device Speech Processing

Gregory Ciccarelli, Jarred Barber, Arun Nair, Israel Cohen, Tao Zhang

Amazon Alexa AI
{gcciccar, barbjarr, aanair, isrcohen, taozhng}@amazon.com

Multiroom Speech Emotion Recognition

Erez Shalev Israel Cohen
Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering
Technion–Israel Institute of Technology, Haifa 3200003, Israel
{erezsh@ef, icohen@ee}.technion.ac.il

Conclusions

Multi-device processing

- Coordinate distributed devices into single virtual array
- Challenges:
  - Unknown relative microphone positions
  - Poor clock synchronization
  - Differences in gain
- Theory for “optimal” voice enhancement has been laid out
- Bridging the gap between theory and practice is the next big step

Distributed Signal Processing

- Distributed algorithms for multi-task WASNs
  - Significantly reduce the required communication bandwidth
  - Robust in adverse conditions (highly non-stationary noise, low SNR, erroneous VADs)
  - Let each node obtain the centralized (=optimal) performance
- Merits of GEVD-based DANSE have been experimentally assessed in real-time

Future Directions

- Smart homes and smart environment comprising of ad-hoc networks
- Processing not siloed to a single unit
- New layer of communication and processing protocols needs to be established between devices
- Concept of information broker (know-it-all units)
- New sensing capabilities (we only considered acoustics!) and modalities
- Maintaining user trust
  - Networks need to keep privacy!
  - Federated learning

Thank you!!

- Connect with me:
  - giacobello@gmail.com
  -