SONOS

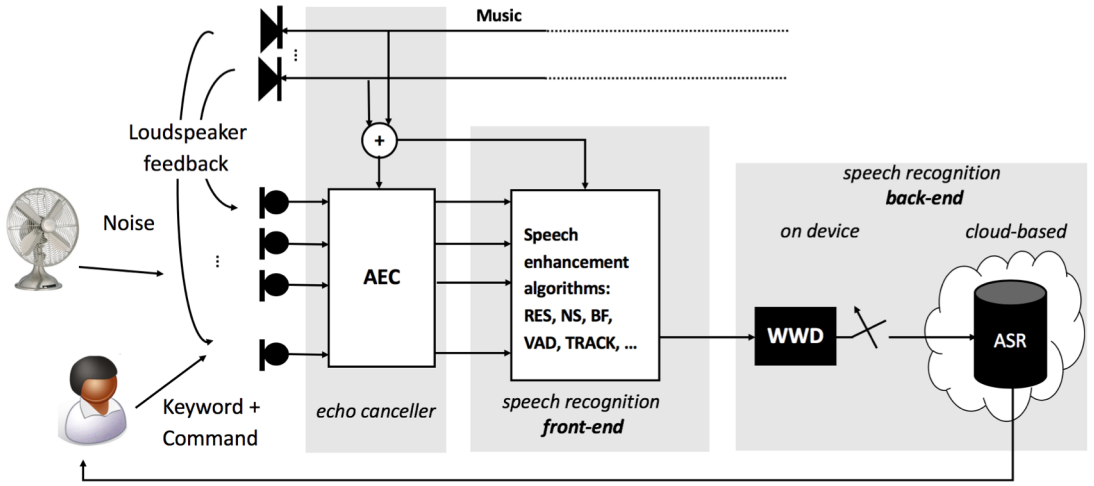
An Online Expectation-Maximization Algorithm for Tracking Acoustic Sources in Multi-Microphone Devices during Music Playback

Daniele Giacobello

Motivation (1/2)

A flexible statistical framework to track desired (DS) and interfering sources (IS) and to estimate the probability that each source is active.

Use case: distant control of a multi-microphone *smart* loudspeaker device often with music playing (signal-to-playback ratio as low as -40 dB).



Motivation (2/2)

Tracking multiple moving sources essential component of modern multi-microphone speech enhancement systems.

Classic spatial filtering approaches (e.g., beamforming) make often the filter unable to adapt fast enough to ever changing acoustic scenarios [Thiergart2014].

Parametric spatial filters based on instantaneous DOA allow the system to adapt quickly to the changing scenario but violate the underlying signal model [Thiergart2012, Thiergart2013].

***Informed* spatial filtering moves past the rigid assumptions of these two paradigms.**

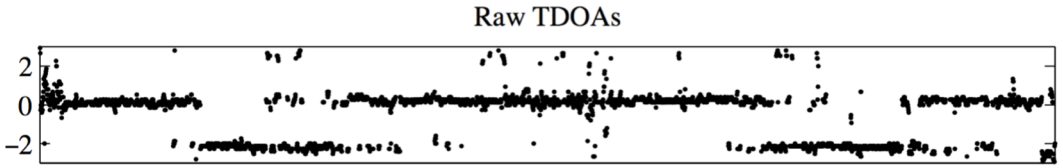
Spatial localization identifies prominent sources in the current acoustic scene based on the set of features chosen, e.g., TDOA.

Our contribution

1. A GMM is used to capture the nonstationary statistical behavior of an ensemble of sources where each source is seen as a multivariate Gaussian component.

> **The use of a mixture model provides a low-complexity method of tracking the spatial location of acoustic sources, while also maintaining estimates of their online statistical behavior.**

2. A GMM feature vector which should promote separation between each Gaussian component is designed comprising of a TDOA estimation.



Example of raw TDOA extracted from an acoustic scene

3. The source *membership* is then evaluated using the GMM model obtained.

4. After source classification, TDOAs of each DS can be extracted from the statistical model, along with the probability that each source is active.

Statistical Modeling of Acoustic Sources

Feature Vector Design (1/2)

Extracting meaningful features from the acoustic scenario to model and track the DSs using the statistical framework provided by the GMMs.

- ***Time Difference of Arrival (TDOA)*** **and** ***Correlation Measures***

Spatial information can be represented as the DOA or as the TDOA. Correlation-based approaches are generally used for finding the TDOA by maximizing a cost function designed to capture similarity between signals observed at different microphones.

The correlation-based cost functions can be leveraged for source tracking in an alternative way by measuring the maximum cost corresponding to the selected TDOA.

These can be expected to show small values for diffuse sources, but large values for point sources which are more consistent with DSs.

Feature Vector Design (2/2)

- ***Predictors of Speech Activity***

Voice activity detectors (VADs) provide measures which convey the likeness of a particular acoustic signal to speech.

Certain discriminative features used in VADs, i.e., pitch information, can be also leveraged for speaker identification.

- ***Predictors of Loudspeaker Activity***

The acoustic activity of the loudspeaker (i.e., if the loudspeaker playback is active), can be estimated from the residual of the acoustic echo cancellation, a necessary step to improve the signal at the microphone.

The coherence between the music playback and the output of the AEC (residual echo) can be used for this purpose.

Statistical Framework

A feature vector $\mathbf{x}_n$ is extracted from the observed acoustic scene and the statistical modeling is applied on a frame level.

A M-components GMM is fully parametrized by the set:

$$\Lambda = \{\underbrace{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M}_{\text{Mean vectors}}, \underbrace{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M}_{\text{Covariance matrices}}, \underbrace{w_1, \ldots, w_M}_{\text{Weights}}\}$$

The conditioned GMM likelihood is expressed as:

$$p\left(\mathbf{x}_n \,|\Lambda\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$$

Where $\mathcal{N}\left(\mathbf{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes a single Gaussian distribution.

Online/Recursive Parameter Estimation

We base the recursive estimation of the GMM parameters on the maximum a posteriori (MAP) criterion.

$\Lambda^{(N)}$: set of parameters estimated from $\mathcal{X}_{[1,N]} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$\mathcal{X}_{[N+1,N+K]} = \{\mathbf{x}_{N+1}, \ldots, \mathbf{x}_{N+K}\}$: new set of feature vectors measured in the acoustic space (when K=1, estimation for every new feature vector)

A MAP estimation of $\Lambda^{(N+K)}$ can be obtained from $\mathcal{X}_{[N+1,N+K]}$, known $p\left(\Lambda^{(N)} \middle| \mathcal{X}_{[1,N]}\right)$

Update rule for means, covariances, and weights:

$$
\begin{aligned}
\boldsymbol{\mu}_m^{(N+K)} &= a_m \mathbf{E}_{m,1}/E_{m,0} + (1-a_m)\,\boldsymbol{\mu}_m^{(N)}, \\
\boldsymbol{\Sigma}_m^{(N+K)} &= a_m \mathbf{E}_{m,2}/E_{m,0} \\
&\quad + (1-a_m)\left(\boldsymbol{\Sigma}_m^{(N)} + \boldsymbol{\mu}_m^{(N)}\boldsymbol{\mu}_m^{(N),T}\right) \\
&\quad - \boldsymbol{\mu}_m^{(N+K)}\boldsymbol{\mu}_m^{(N+K),T}, \\
w_m^{(N+K)} &= a_m E_{m,0}/K + (1-a_m)\,w_m^{(N)},
\end{aligned}
$$

Sufficient statistics are given by:

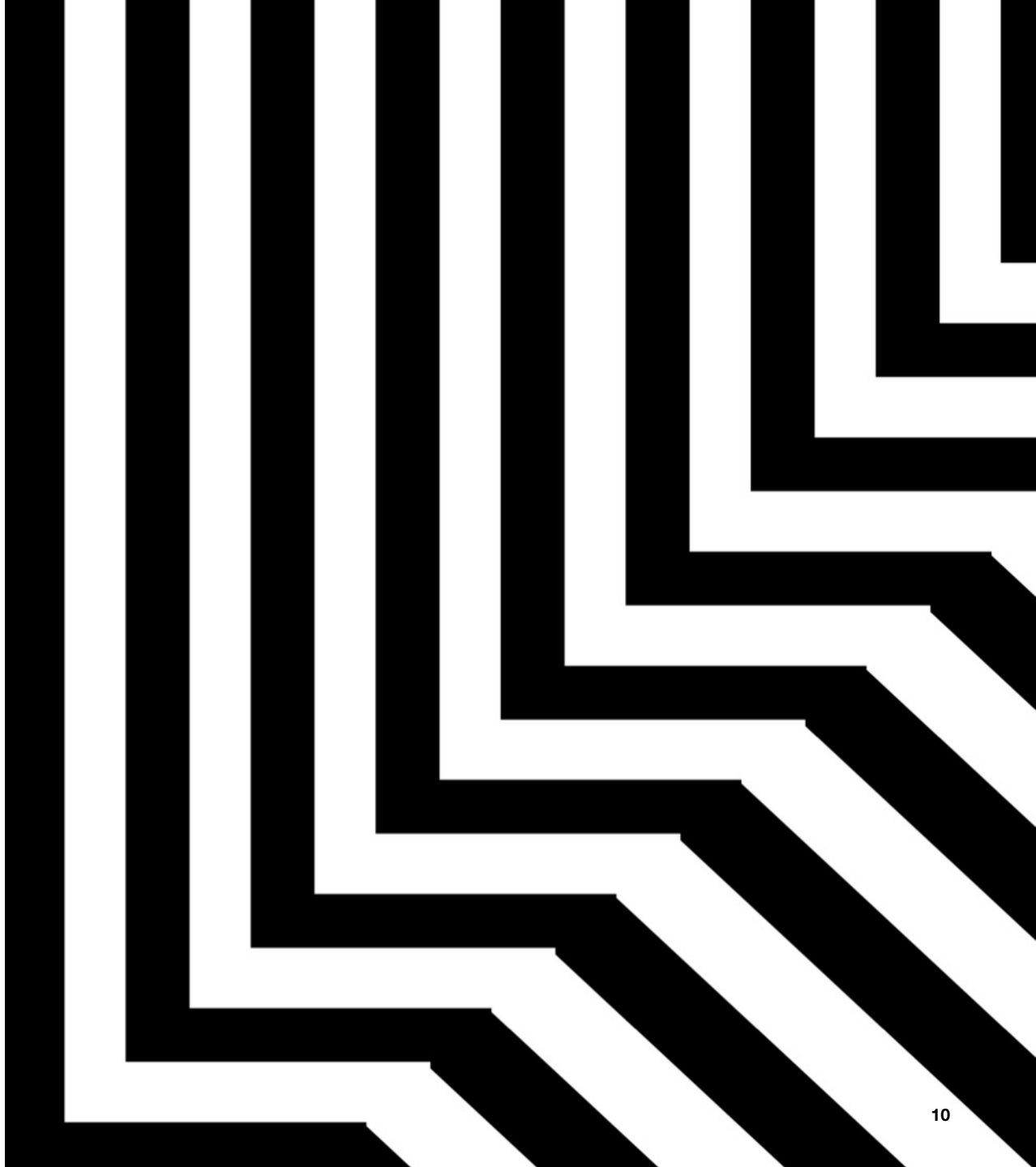
$$
\begin{aligned}
E_{m,0} &= \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right), \\
\mathbf{E}_{m,1} &= \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right)\mathbf{x}_{N+k}, \\
\mathbf{E}_{m,2} &= \sum_{k=1}^{K} P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right)\mathbf{x}_{N+k}\mathbf{x}_{N+k}^T,
\end{aligned}
$$

where $P\left(\Lambda_m^{(N+k)} \mid \mathbf{x}_{N+k}\right)$ is the posterior probability of mixture *m* and the adaptation step size is given by

$$a_m = E_{m,0}/\left(E_{m,0} + N w_m^{(N)}\right)$$

NB: Possible to reduce long-term dependencies by applying a max value

INFERENCE OF DESIRED SOURCE BEHAVIOR



Identifying the desired source(s)

The GMM obtained implicitly embeds the spatial location (LDS) of the sources and the posterior probability (PDS) that the sources are active.

We propose to use the minimum Mahalanobis distance between a point in the parameter space and the Gaussian mixtures representing the space at instant *n*:

$$\widehat{m}_{\mathrm{DS}} = \arg\min_{m} \sqrt{(\mathbf{z}_n - \boldsymbol{\mu}_m)\, \boldsymbol{\Sigma}_m\, (\mathbf{z}_n - \boldsymbol{\mu}_m)}$$

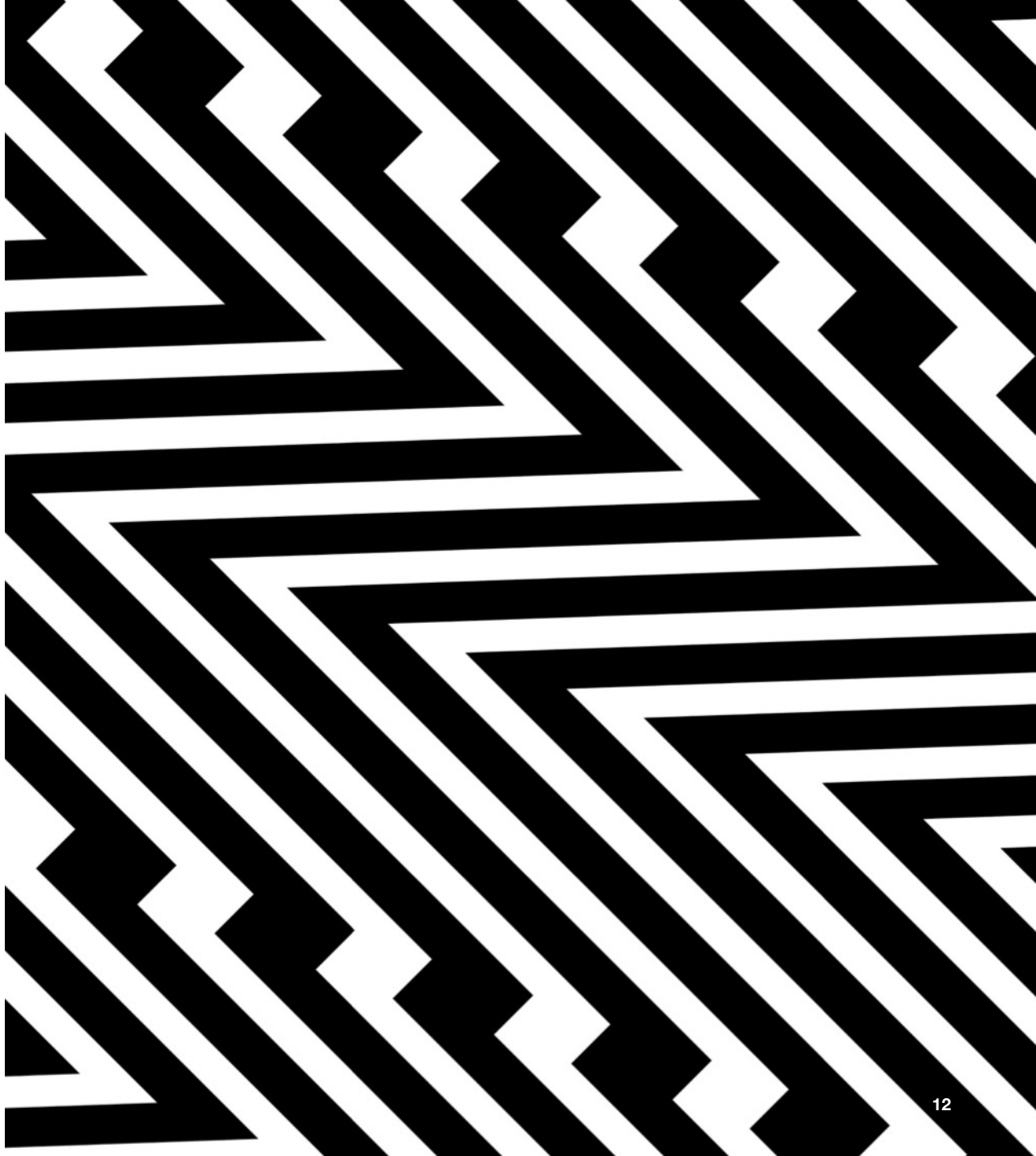
Once $\widehat{m}_{\mathrm{DS}}$, the distribution with smaller distance to $\mathbf{z}_n$, has been identified, the DS can be tracked and its level of activity can be inferred.

The spatial location of the DS, LDS, is determined as the TDOA element of the GMM mixture mean.

The probability of DS activity is estimated as the posterior probability of the DS mixture conditioned on $\mathbf{z}_n$

$$P_{\mathrm{DS}} = P\left(\widehat{m}_{\mathrm{DS}} \mid \mathbf{z}_n\right) = \frac{w_m \mathcal{N}\left(\mathbf{z}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)\big|_{m=\widehat{m}_{\mathrm{DS}}}}{\sum_{j=1}^{M} w_j \mathcal{N}\left(\mathbf{z}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

Experimental Evaluation



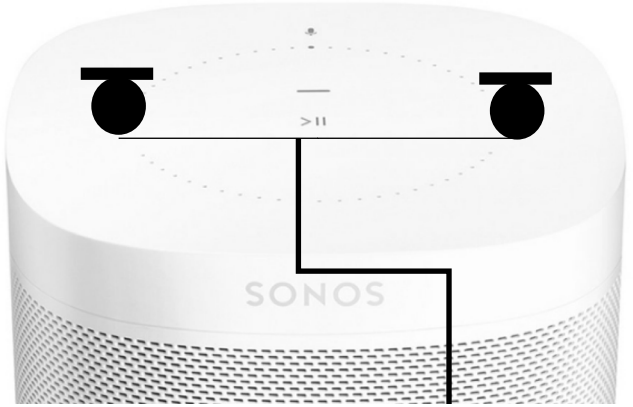
Experimental Setup (1/2)

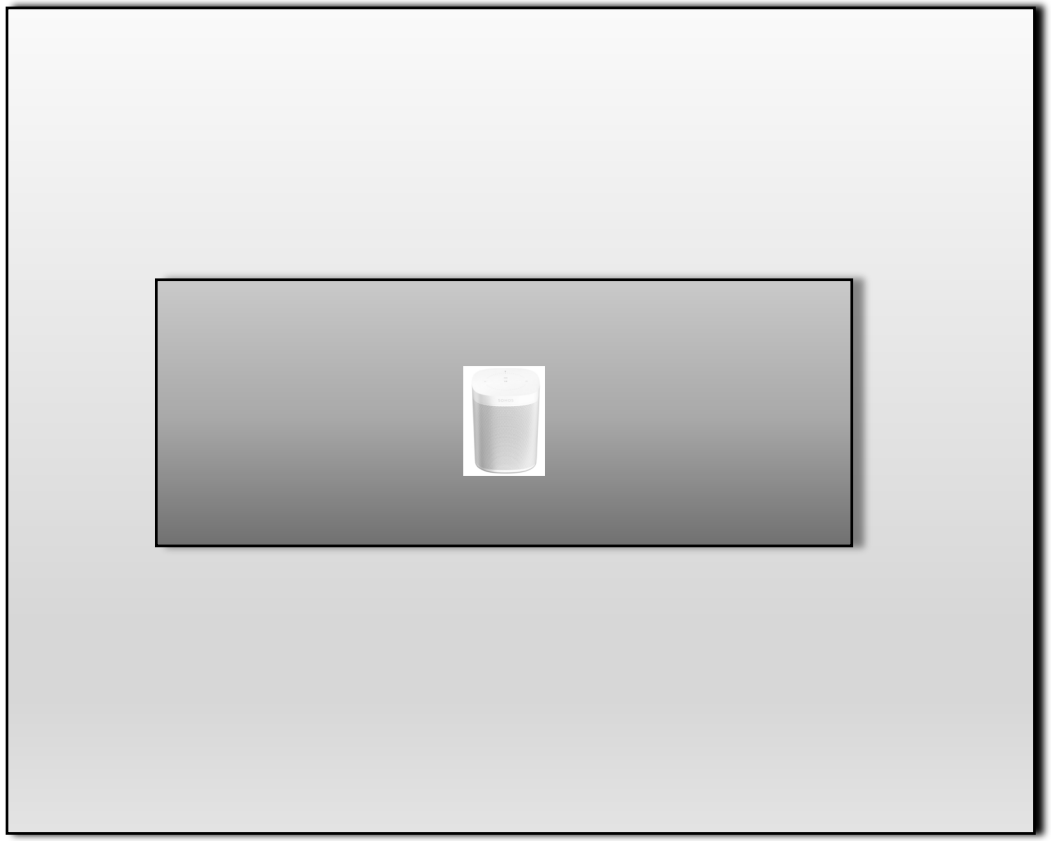
meeting room of size 7.5m×6.2m×2.6m with RT60 = 0.51s

centrally located 5m×2m rectangular table

Sonos™ One smart speaker (size approx. 16cm×12cm×12cm)

Two omnidirectional upward-facing mics at 72mm distance





fs=16 kHz

512 samples frames using a Hamming window with 50% overlap (512-point FFT).

A robust STFT-domain AEC is employed if music is playing (robust adaptation method to avoid using double-talk detection [Wada2009]).

Statistical model-based residual echo suppression to cope with the possibility of echo leakage.

Experimental Setup (2/2)

5-dimension feature vector:

- **TDOA and associated correlation measure (CDOA) obtained with the generalized correlation method of [Knapp76]**

- **VAD measure obtained through a combination of spectral entropy and energy [Giacobello2008]**

- **Likelihood ratio (LR) tests of the residual echo canceller (RES) [Lee2007]**

Autocorrelation-based pitch estimate, obtained using [Malah1982]

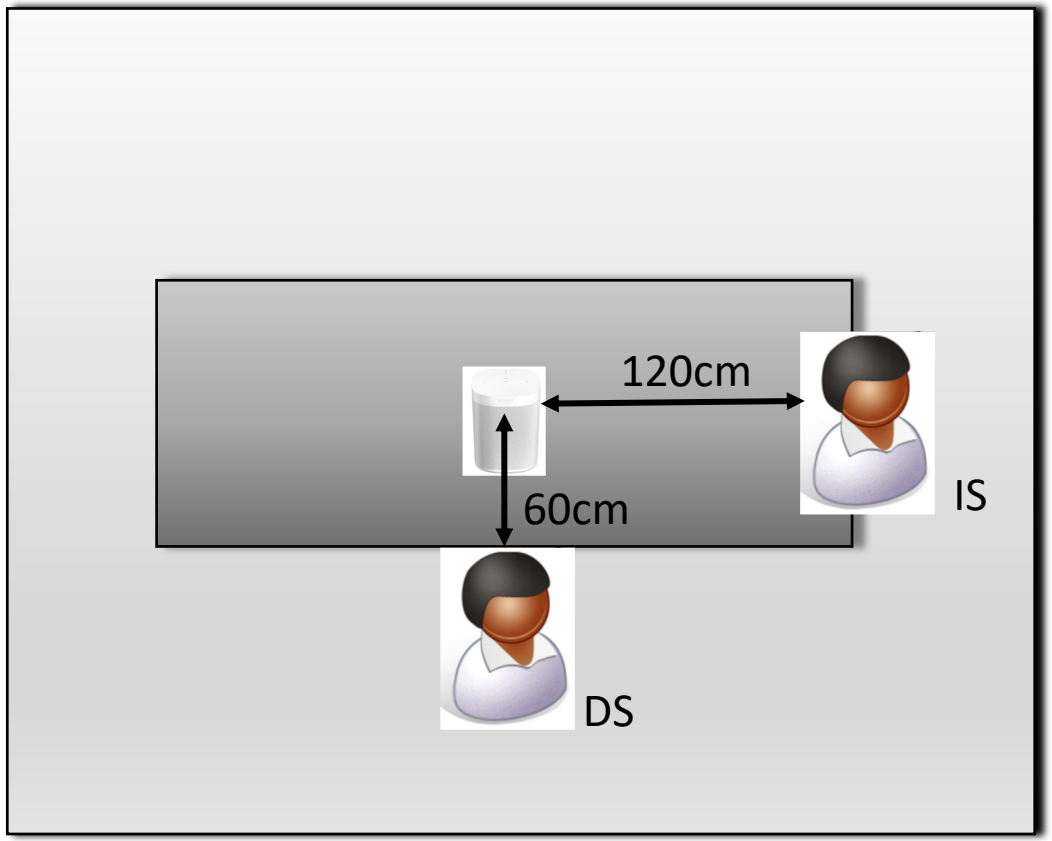
NB: calculation performed after the RES algorithm

The adaptation term in $a_m = E_{m,0} / \left(E_{m,0} + N w_m^{(N)} \right)$ **was set for a 250ms forgiving factor**

The parameter space was modeled initially using a GMM with three components empirically initiated. We then apply the frame-based recursive update.

To avoid local maxima in fitting the GMM model to our parameter space, we used the Gaussian splitting and merging criterion presented in [Ueda1998].

First Experiment: Static Talker (1/2)



To simulate the interaction with the device:

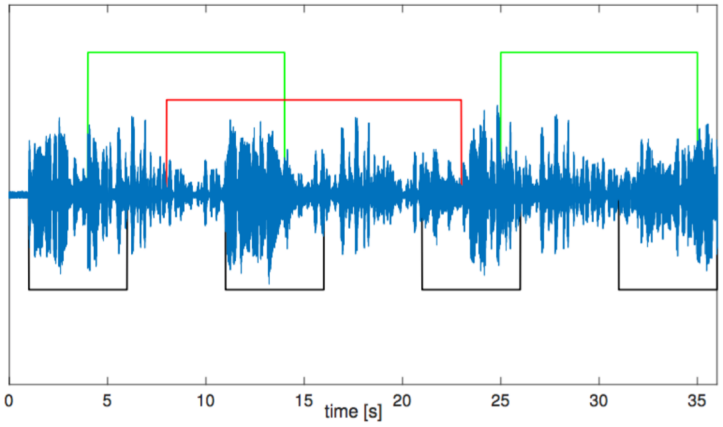
- DS active in 10s intervals

- music (PB) was played back from the loudspeaker in 5s intervals (randomly picked from TOP40 playlist)

- IS active in 15s intervals

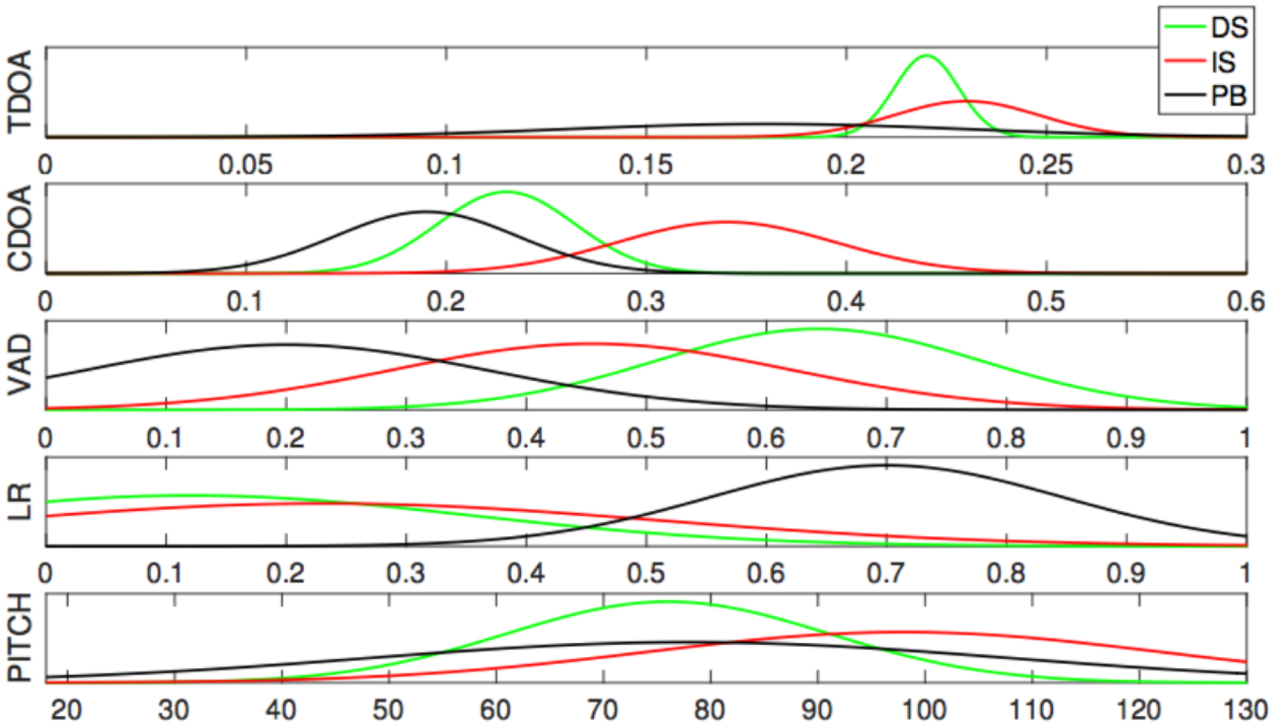
- SPL level of moderate listening, giving approximately a 20dB music-to-DS ratio (roughly 5dB after AEC)

DS was observed to be 5dB louder than the IS

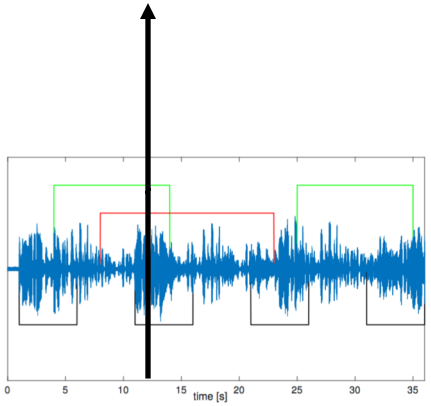


A 36s segment from the captured primary microphone signal. The highlighted part correspond to acoustic source activity, where green, red, and black correspond to DS, IS, and PB (after AEC and RES), respectively.

First Experiment: Static Talker (2/2)



A snapshot of model parameters from the GMM during the test signal. The panels show the projections of the multivariate mixture distributions onto individual feature subspaces.



Mixture 1 (green) is tracking the DS

Mixture 2 (red) is tracking the IS

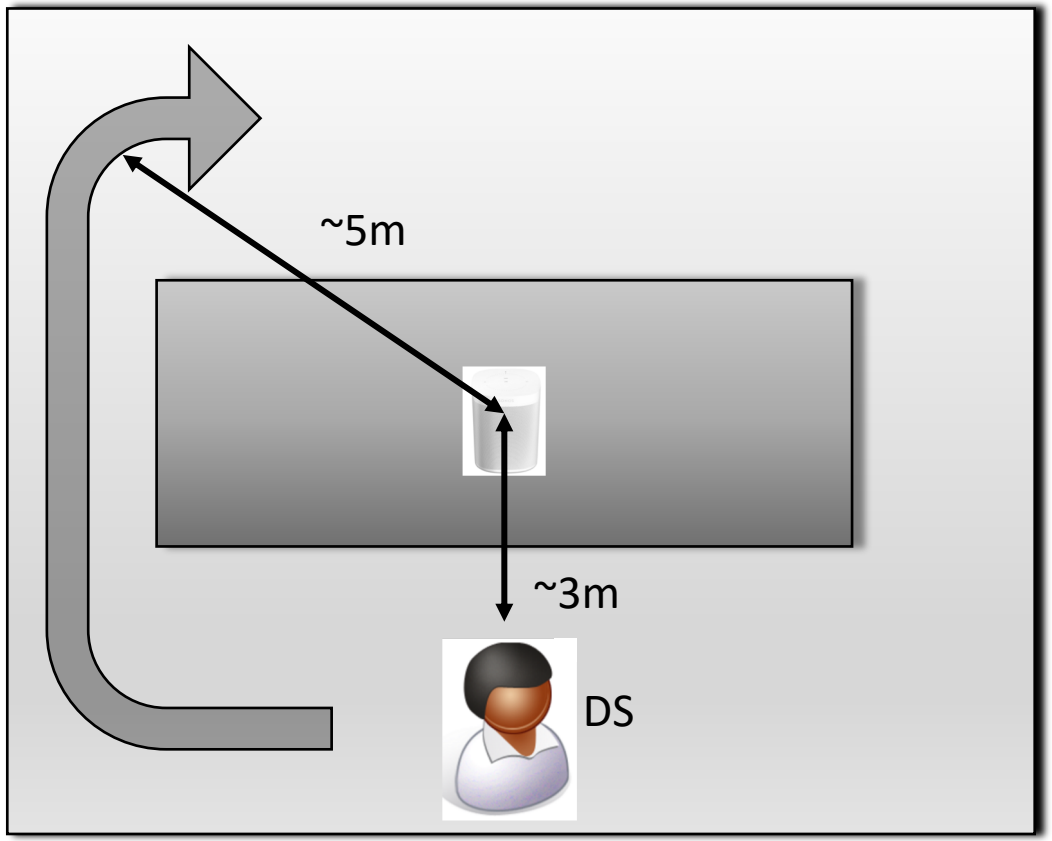
Mixture 3 (black) is capturing diffuse background noise and music playback (PB).

Individual features, especially the TDOA, were not able to provide clear separation between mixtures during the source classification phase.

The higher distance from the microphone of the IS makes the sound diffuse (top pane) but it is easier to discriminate using the CDOA (second pane).

Music and speech are easier to classify using the VAD values (consistent with music/speech differences in spectral entropy).

Second Experiment: Moving Talker



Talker reciting the alphabet while moving slowly through the room (one full rotation of the room in 45s and then 25s).

Ground truth angle calculated with a laser digital angle finder. Process was filmed with a camera to manually label the ground truth angle (similar to the AV16.3 dataset).

The music was again picked randomly from a TOP40 playlist with approximately a -20dB speaker-to-music ratio at the microphones at 60cm (up to -45dB at the furthest point).

The experiment was repeated with and without music playing from the device.

We repeated these recordings 5 times with 10 different speakers, for a total of 100 trials.

Front-back uncertainty of the measurement was neglected as not particularly problematic in our controlled scenario.

The measurement bias of the ground truth angle was also considered negligible.

Second Experiment: Results

Compared three GMM-based methods for tracking with a different number of features:

- GMM2D: TDOA and CDOA (the most intuitively features to perform tracking)
- GMM4D, TDOA, CDOA, VAD and RES
- GMM5D: TDOA, CDOA, VAD, RES, and PITCH

Comparison with two other popular nonlinear spatial tracking algorithms [16], [17].

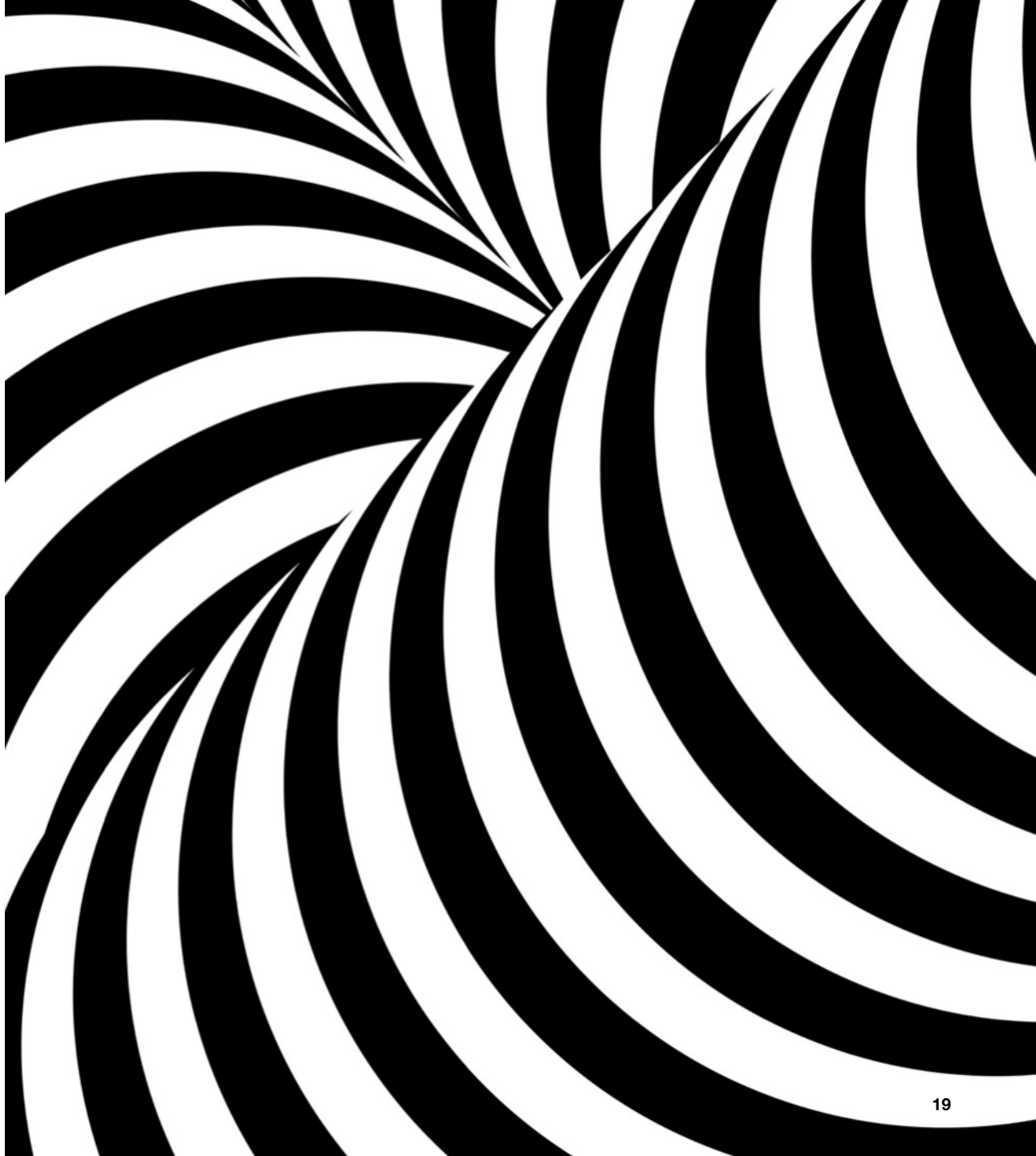
Including terms different from the TDOA helps the EM procedure by providing speech activity information directly into the multivariate GMM.

	RMSE DOA [°]			
	Slow Moving		Fast Moving	
	Music Off	Music On	Music Off	Music On
GMM_{2D}	9.3	13.5	8.8	18.2
GMM_{4D}	5.4	6.9	5.8	6.7
GMM_{5D}	4.3	**4.1**	**4.2**	**4.9**
[16]	**4.2**	7.4	5.1	10.2
[17]	5.4	8.7	7.3	11.4

LOCALIZATION ACCURACY IN TERMS OF RMSE OF THE DOA.

[16] J. Vermaak and A. Blake, “Nonlinear filtering for speaker tracking in noisy and reverberant environments,” ICASSP 2001. (SMC method)
[17] B. Ward and R. C. Williamson, “Particle filter beamforming for acoustic source localization in a reverberant environment,” ICASSP 2002.

Conclusions



We proposed a source tracking algorithm based on the GMM modeling of features extracted from the acoustic space.

The use of different features introduced complementary information in the multivariate model and increases mixture separation.

- **Carefully selecting a feature vector allows for better discriminating between RES, DS, and IS and grants a flexible amount of separation of the acoustic sources and system complexity.**

- **The use of voice activity metrics (residual echo likelihood and pitch) enriched the GMM model in the discrimination allowing for a relatively low RMSE in the DOA estimate during music playback from -20 to, almost, -50 dB of signal-to-music ratio with a talker moving fairly fast through the room.**

- **The pitch measure helped discriminating among the two talkers (while the music mixture resembles a flat distribution).**

- **The LR, calculated at the RES, works really well in discriminating PB, while gives roughly same likelihood values for the DS and IS sources.**

Thank you!

SONOS