Unscented Transform-Based Dual-Channel Noise Estimation: Application to Speech Enhancement on Smartphones

Iván López-Espejo[1], Juan M. Martín-Doñas[2], Angel M. Gomez[2], and Antonio M. Peinado[2]
[1]VeriDas | das-Nano, Navarre, Spain
[2]Dept. of Signal Theory, Telematics and Communications, University of Granada, Granada, Spain
ilopez@das-nano.com, {mdjuamart,amgg,amp}@ugr.es

41st International Conference on Telecommunications and Signal Processing
7-2018, Athens (Greece)

Outline

1. Introduction
   - Motivation
   - Objectives
2. UT-Based Dual-Channel Noise Estimation
   - Dual-Channel MMSE Estimation
   - Dual-Channel Distortion Model
   - Unscented Transform Application
   - Implementation Issues
3. Experimental Evaluation
   - The AURORA2-2C-CT Database
   - Estimation Error Results
   - Speech Quality Results
   - Speech Intelligibility Results
4. Conclusions

Introduction
Motivation

New speech processing boom

The use of speech processing applications has notably increased due to the latest smartphones:

- Great amount of apps (search-by-voice, dictation, voice biometrics, etc.).

Noise-robust speech processing in smartphones

- It is crucial to deal with noisy environments.
- We can take benefit of the dual-mic feature.

Introduction
Objectives

1. **To estimate the noise power spectrum** at the primary channel by exploiting the information at both channels.
   - Dual-channel MMSE estimation.
   - Unscented transform (UT) vs. vector Taylor series (VTS).
2. To explore a particular application of interest: **speech enhancement on a dual-mic smartphone** in close-talk conditions.

UT-Based Dual-Channel Noise Estimation

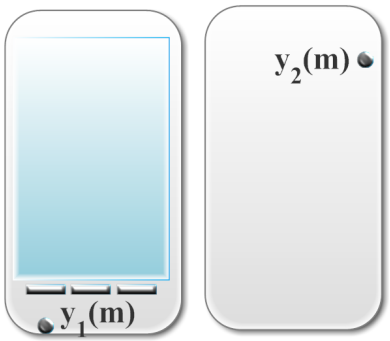
Dual-Channel MMSE Estimation

- $k = 1$ is the primary mic while $k = 2$ is the extra mic:
  $\mathbf{y}_k(t) = \left(|Y_k(0,t)|^2, ..., |Y_k(\mathcal{M}-1,t)|^2\right)^\top$
  $\mathbf{n}_k(t) = \left(|N_k(0,t)|^2, ..., |N_k(\mathcal{M}-1,t)|^2\right)^\top$

- The MMSE estimate of $\mathbf{n}_1(t)$ is

$$\hat{\mathbf{n}}_1(t) = \boldsymbol{\mu}_{n_1}(t) + \boldsymbol{\Sigma}_{n_1 y_1}(t)\boldsymbol{\Sigma}_{y_1}^{-1}(t)\left(\mathbf{y}_1(t) - \boldsymbol{\mu}_{y_1}(t)\right)$$

Through the unscented transform!



- $\mathbf{y}_2(t)$ will play here a supporting role.

UT-Based Dual-Channel Noise Estimation

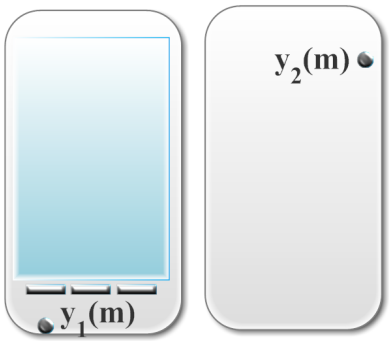
Dual-Channel Distortion Model

- Dual-channel speech distortion model:
  $\mathbf{y}_1(t) = \mathbf{x}_1(t) + \mathbf{n}_1(t)$
  $\mathbf{y}_2(t) = \mathbf{x}_2(t) + \mathbf{n}_2(t)$

- We relate the speech power at both channels through speech gains:
  $\mathbf{x}_2(t) = \mathbf{a}_{21}(t) \odot \mathbf{x}_1(t)$

- We sample the next $\mathbf{y}_1(t)$ model through **UT** to get $\boldsymbol{\mu}_{y_1}(t)$, $\boldsymbol{\Sigma}_{y_1}(t)$ and $\boldsymbol{\Sigma}_{n_1 y_1}(t)$:

$$\begin{aligned} \mathbf{y}_1(t) &= \mathbf{f}\left(\mathbf{a}_{21}(t), \mathbf{n}_1(t), \mathbf{n}_2(t); \mathbf{y}_2(t)\right) \\ &= \mathbf{a}_{21}^{-1}(t) \odot \left(\mathbf{y}_2(t) - \mathbf{n}_2(t)\right) + \mathbf{n}_1(t) \end{aligned}$$



UT-Based Dual-Channel Noise Estimation

Unscented Transform Application

Augmented signal

$$\mathbf{z}(t) = \left(\mathbf{a}_{21}^{\top}(t), \mathbf{n}_1^{\top}(t), \mathbf{n}_2^{\top}(t)\right)^{\top}$$

Sigma point generation

$$\begin{aligned}\mathbf{z}_i(t) &= \mathbf{z}_0(t) + \left(\sqrt{(3\mathcal{M}+\eta)\,\boldsymbol{\Sigma}_z(t)}\right)_i \\ \mathbf{z}_{i+3\mathcal{M}}(t) &= \mathbf{z}_0(t) - \left(\sqrt{(3\mathcal{M}+\eta)\,\boldsymbol{\Sigma}_z(t)}\right)_i \\ & i = 1, 2, ..., 3\mathcal{M}\end{aligned}$$

$$\mathbf{z}_i(t) = \begin{pmatrix} \mathbf{z}_{a_{21}}^{(i)}(t) \\ \mathbf{z}_{n_1}^{(i)}(t) \\ \mathbf{z}_{n_2}^{(i)}(t) \end{pmatrix}$$

Nonlinear function evaluation

$$\mathbf{z}_{y_1}^{(i)}(t) = \mathbf{f}\left(\mathbf{z}_{a_{21}}^{(i)}(t), \mathbf{z}_{n_1}^{(i)}(t), \mathbf{z}_{n_2}^{(i)}(t); \mathbf{y}_2(t)\right)$$
$$i = 0, 1, ..., 6\mathcal{M}$$

Statistics computation

$$\boldsymbol{\mu}_{y_1}(t) = \sum_{i=0}^{6\mathcal{M}} w_i \mathbf{z}_{y_1}^{(i)}(t)$$

$$\begin{aligned} w_0 &= \eta/(3\mathcal{M}+\eta) \\ w_i &= w_{i+3\mathcal{M}} = \frac{1}{2\,(3\mathcal{M}+\eta)} \\ & i = 1, 2, ..., 3\mathcal{M}\end{aligned}$$

$$\boldsymbol{\Sigma}_{y_1}(t) = \sum_{i=0}^{6\mathcal{M}} w_i \left(\mathbf{z}_{y_1}^{(i)}(t) - \boldsymbol{\mu}_{y_1}(t)\right)\left(\mathbf{z}_{y_1}^{(i)}(t) - \boldsymbol{\mu}_{y_1}(t)\right)^{\top}$$

$$\boldsymbol{\Sigma}_{n_1 y_1}(t) = \sum_{i=0}^{6\mathcal{M}} w_i \left(\mathbf{z}_{n_1}^{(i)}(t) - \boldsymbol{\mu}_{n_1}(t)\right)\left(\mathbf{z}_{y_1}^{(i)}(t) - \boldsymbol{\mu}_{y_1}(t)\right)^{\top}$$

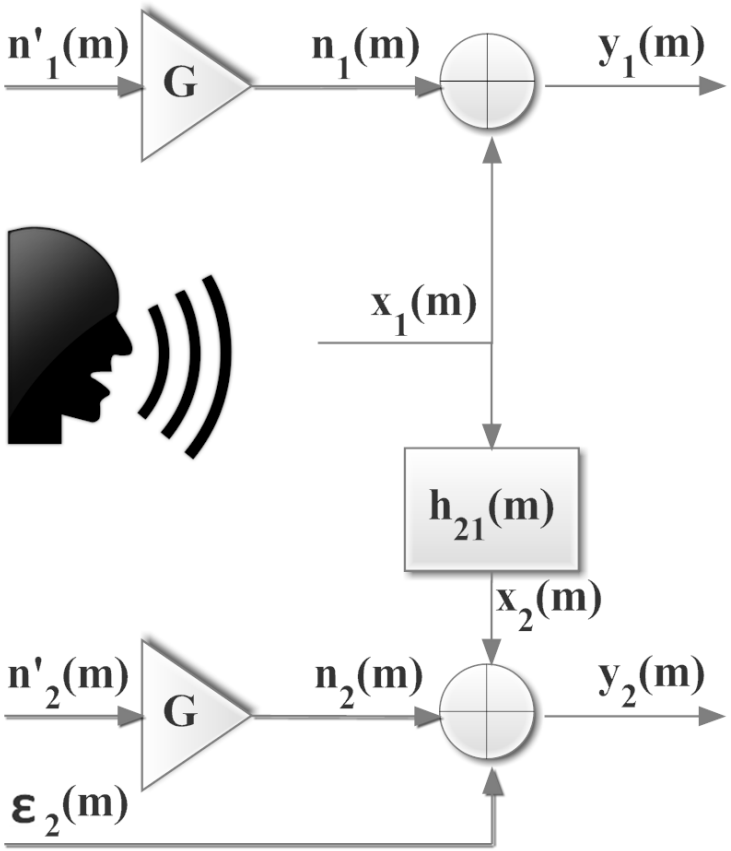
UT-Based Dual-Channel Noise Estimation

Implementation Issues

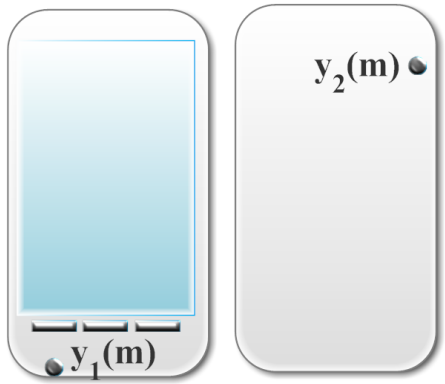
Relevant practical issues

- We assume that both $\mathbf{a}_{21}(t)$ and $\mathbf{n}_k(t)$ $(k = 1, 2)$ are wide-sense stationary random processes:
  - $\boldsymbol{\mu}_{a_{21}}$ and $\boldsymbol{\Sigma}_{a_{21}}$ are obtained from a development dataset.
  - Noise statistics are obtained from the first and last frames of each utterance.
- η is set to 0.
- Negative estimated bins are replaced by those from a noise estimation based on linear interpolation (around 2% of cases).

Experimental Evaluation
The AURORA2-2C-CT Database

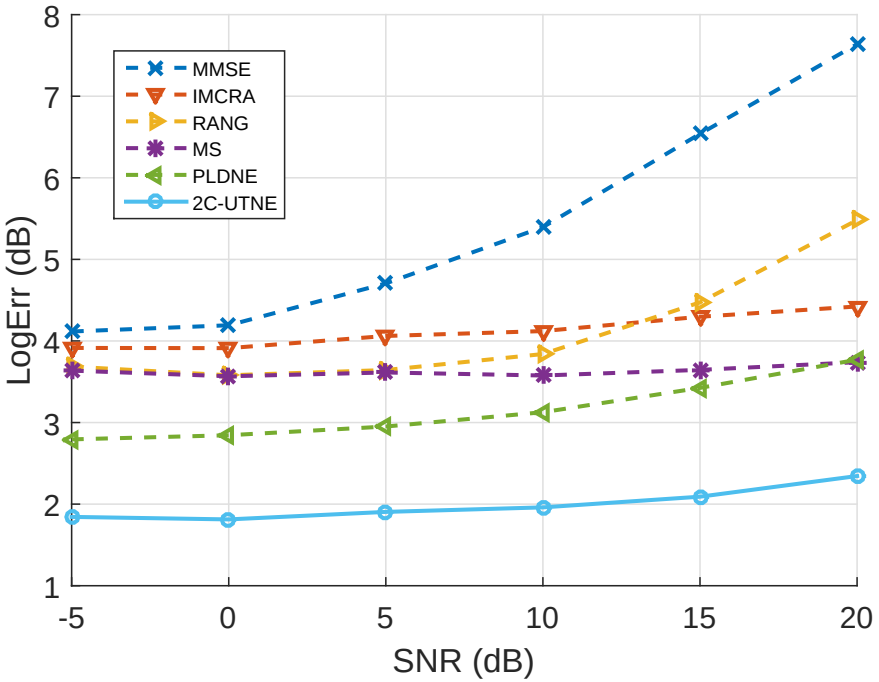


López-Espejo I., et al.: "Feature Enhancement for Robust Speech Recognition on Smartphones with Dual-Microphone". *In: EUSIPCO*, Lisbon (2014)



- **Test A:** Bus, babble, car and pedestrian street
- **Test B:** Café, street, bus and train stations
- **SNRs:** {-5,0,5,10,15,20} dB and clean

Experimental Evaluation
Estimation Error Results

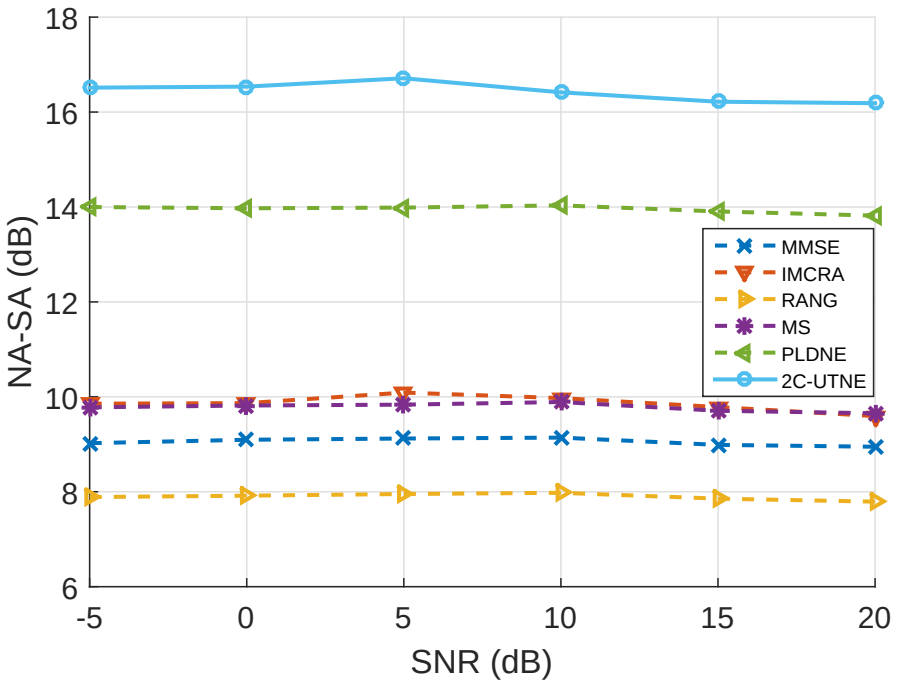


$$\text{LogErr} = \frac{1}{\mathcal{M}T}\sum_{f=0}^{\mathcal{M}-1}\sum_{t=0}^{T-1}\left|10\log_{10}\left[\frac{\Phi_{n_1}(f,t)}{\hat{\Phi}_{n_1}(f,t)}\right]\right|$$

$$\Phi_{n_1}(f,t) = 0.9\Phi_{n_1}(f,t-1) + 0.1|N_1(f,t)|^2$$

- **2C-UTNE - Our proposal**
- **PLDNE** - Power Level Difference Noise Estimator
- **MS** - Minimum Statistics
- **IMCRA** - Improved Minima Controlled Recursive Averaging
- **RANG** - Rangachari's algorithm
- **MMSE** - MMSE noise estimator

Experimental Evaluation
Speech Quality Results



- Noise estimation methods are combined with **dual-channel Wiener filtering**.
- **NA-SA**: Noise Attenuation minus Speech Attenuation.

- **2C-UTNE - Our proposal**
- **PLDNE** - Power Level Difference Noise Estimator
- **MS** - Minimum Statistics
- **IMCRA** - Improved Minima Controlled Recursive Averaging
- **RANG** - Rangachari's algorithm
- **MMSE** - MMSE noise estimator

Experimental Evaluation
Speech Intelligibility Results

	SNR (dB)						
	-5	0	5	10	15	20	Avg.
Noisy	0.280	0.415	0.556	0.693	0.819	0.909	0.612
MMSE	0.321	0.478	0.629	0.761	0.867	0.936	0.665
IMCRA	0.313	0.470	0.621	0.753	0.858	0.928	0.657
RANG	0.323	0.477	0.627	0.760	0.866	0.935	0.665
MS	0.319	0.476	0.627	0.760	0.867	0.936	0.664
PLDNE	0.346	0.496	0.640	0.767	0.868	0.933	0.675
2C-UTNE	**0.350**	**0.506**	**0.653**	**0.780**	**0.879**	**0.941**	**0.685**

- Noise estimation methods are combined with **dual-channel Wiener filtering**.
- **CSII**: Coherence Speech Intelligibility Index.

- **2C-UTNE - Our proposal**
- **PLDNE** - Power Level Difference Noise Estimator
- **MS** - Minimum Statistics
- **IMCRA** - Improved Minima Controlled Recursive Averaging
- **RANG** - Rangachari's algorithm
- **MMSE** - MMSE noise estimator

Conclusions

Some conclusions and future work

- We have achieved accurate noise estimates by taking advantage of...
  1. the unscented transform.
  2. the dual-channel observation (avoiding the use of a clean speech model while keeping a simple formulation).
- Results have shown the higher performance of our proposal with respect to state-of-the-art noise estimation methods.
- We will investigate on temporal dynamics modeling to further improve the performance of our method.

Thanks for your attention!

Iván López-Espejo
VeriDas | das-Nano, Navarre, Spain
ilopez@das-nano.com