Deep Neural Network-Based Noise Estimation for Robust ASR in Dual-Microphone Smartphones

Iván López-Espejo, A. M. Peinado, A. M. Gomez, and J. M. Martín-Doñas
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain
{iloes,amp,amgg}@ugr.es, mdjuamart@correo.ugr.es

IberSPEECH '16 - IX Jornadas en Tecnologías del Habla
11-25-2016, Lisboa

Outline

Introduction

Motivation

New ASR upswing

The use of automatic speech recognition (ASR) applications has notably increased due to the latest mobile devices:

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

Noise-robust ASR in smartphones

- It is crucial to tackle with noisy environments.
- We can take benefit from the dual-mic feature to provide better noise estimates.

Introduction

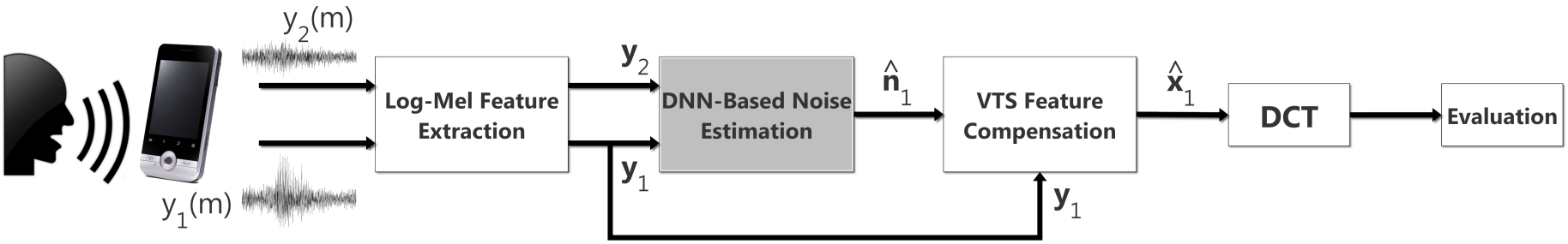
Objectives

In close-talk position:

1. To **estimate noise** for the primary channel by using the information contained in both channels.
   - We experiment with **deep neural networks** (DNNs).
2. To assess its quality when it is used by a feature compensation technique over a dual-channel noisy speech database (**AURORA2-2C**).

Proposed Method

Block diagram of the noise-robust ASR framework considered in this work

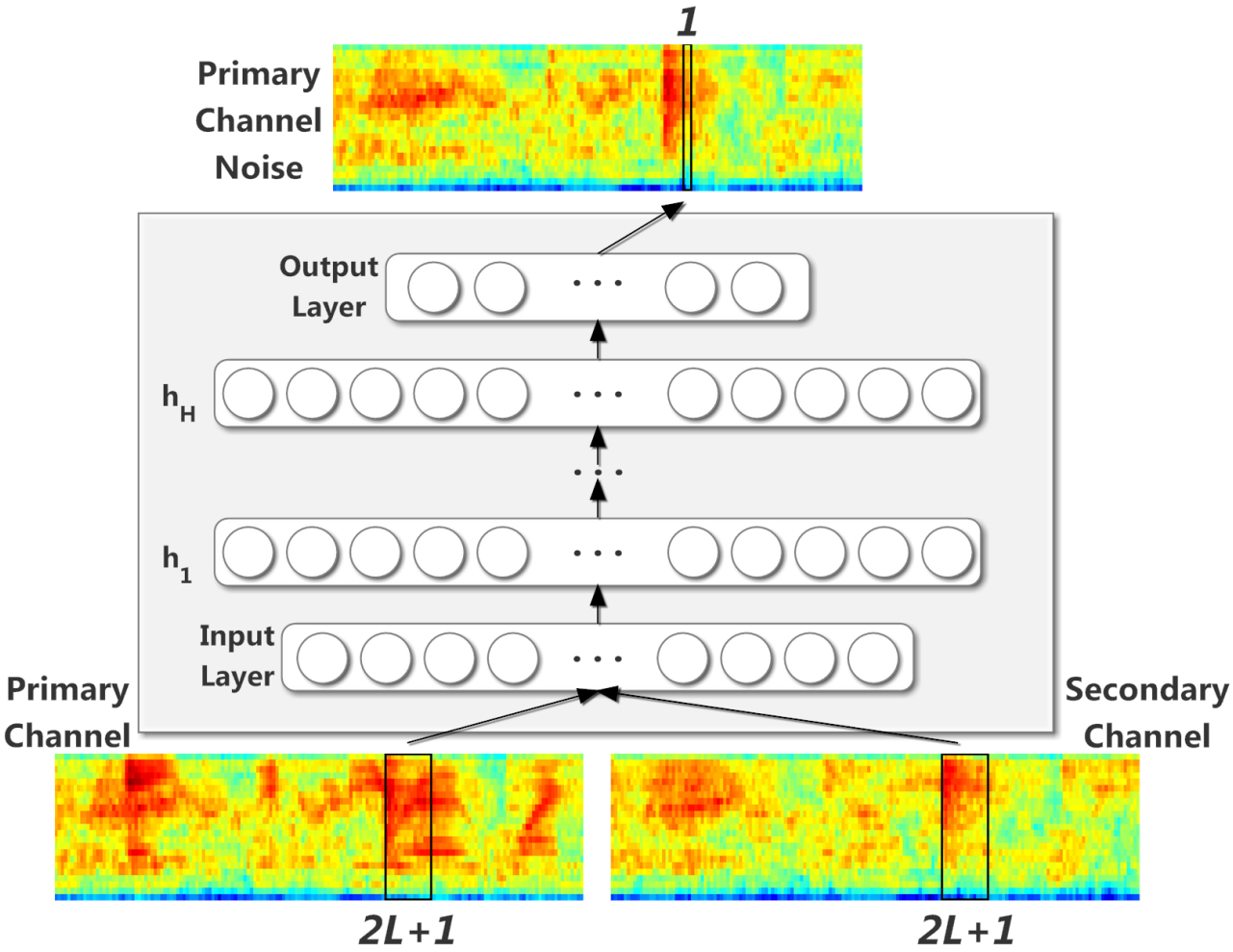


About this system...

- We try to exploit the **power level difference** between the two sensors of the device.
- This is a hybrid **DNN/signal processing** architecture.

Dual-Channel DNN-Based Noise Estimation

Basic System



Features:

$$\mathcal{Y}(t) = \begin{pmatrix} \mathbf{y}(t-L) \\ \vdots \\ \mathbf{y}(t+L) \end{pmatrix},$$

where

$$\mathbf{y}(t) = \begin{pmatrix} \mathbf{y}_1(t) \\ \mathbf{y}_2(t) \end{pmatrix}$$

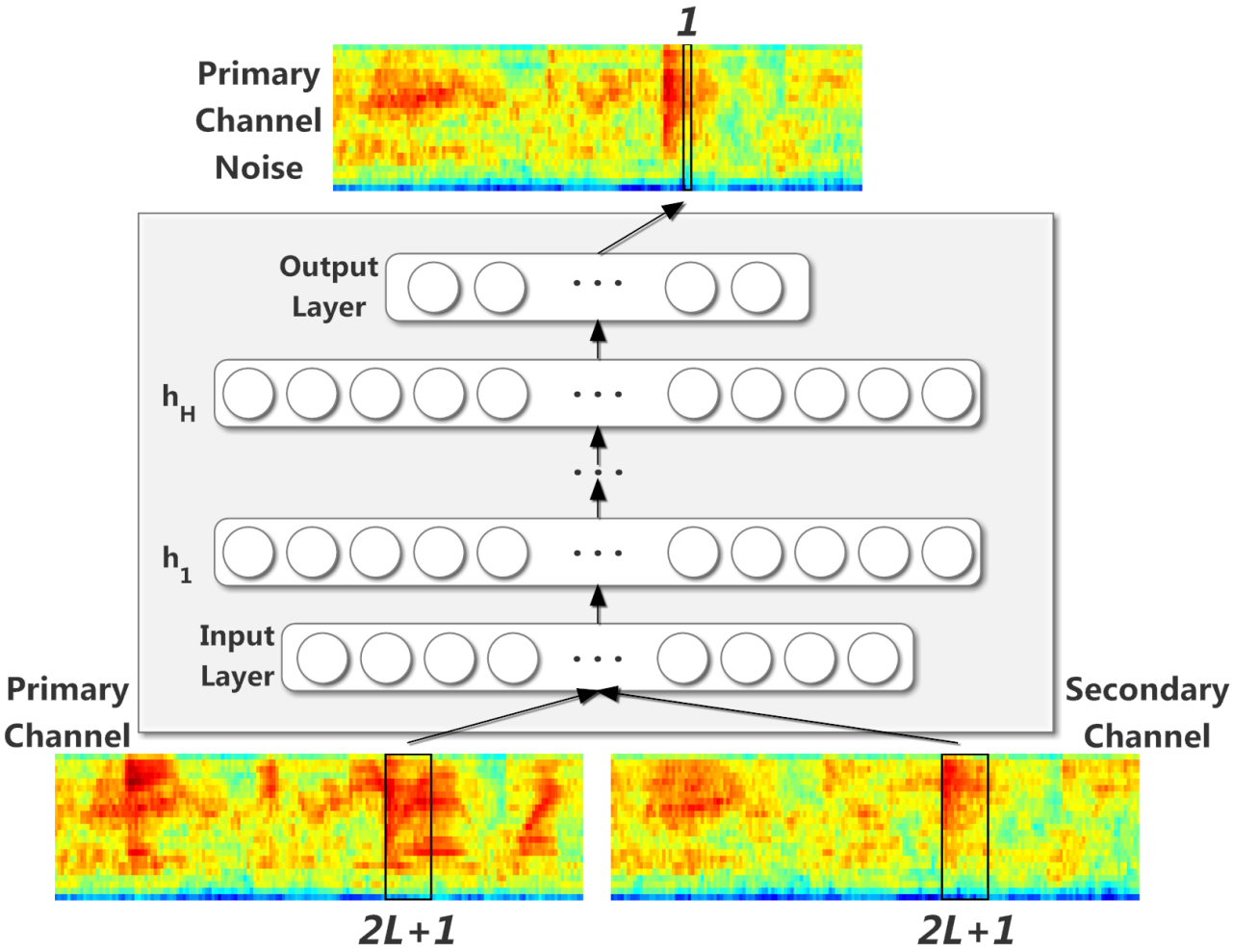
- Input dim.:
$\dim(\mathcal{Y}(t)) = 2\mathcal{M}(2L+1)$

Target:

- Actual noise vector $\mathbf{n}_1(t)$
- Output dim.: $\mathcal{M} \times 1$

Dual-Channel DNN-Based Noise Estimation

Basic System



Training issues:

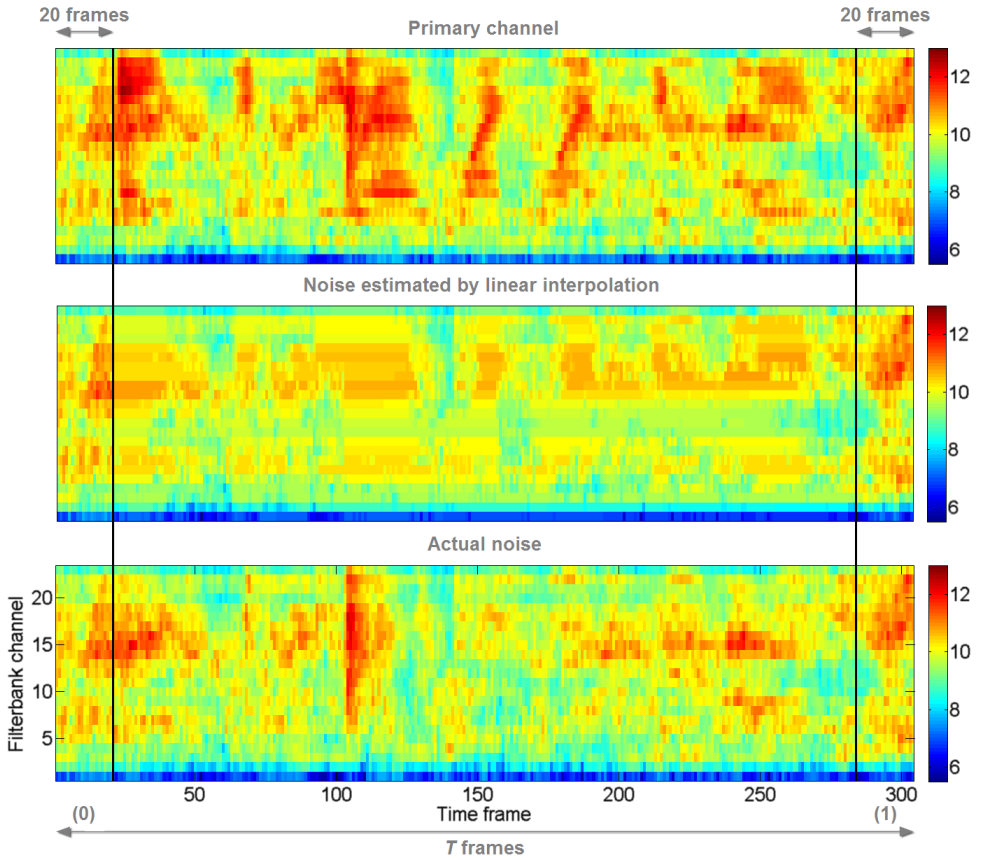
- The DNN is pre-trained by considering each pair of layers as RBMs

- The DNN is trained by using the backpropagation algorithm (**MSE criterion**)

Dual-Channel DNN-Based Noise Estimation

Noise-Aware Training

- Noise-aware training (NAT) first appeared to strengthen the DNN-based acoustic modeling for ASR.
- We want to imitate linear interpolation noise estimation.



Augmented features:

$$\mathcal{Y}_{NAT}(t) = \begin{pmatrix} \mathcal{Y}(t) \\ \bar{\mathbf{n}}_1^{(0)} \\ \bar{\mathbf{n}}_1^{(1)} \\ \boldsymbol{\sigma}_1^{(0)} \\ \boldsymbol{\sigma}_1^{(1)} \\ \tau(t) \end{pmatrix},$$

where

$$\tau(t) = t/(T-1);$$

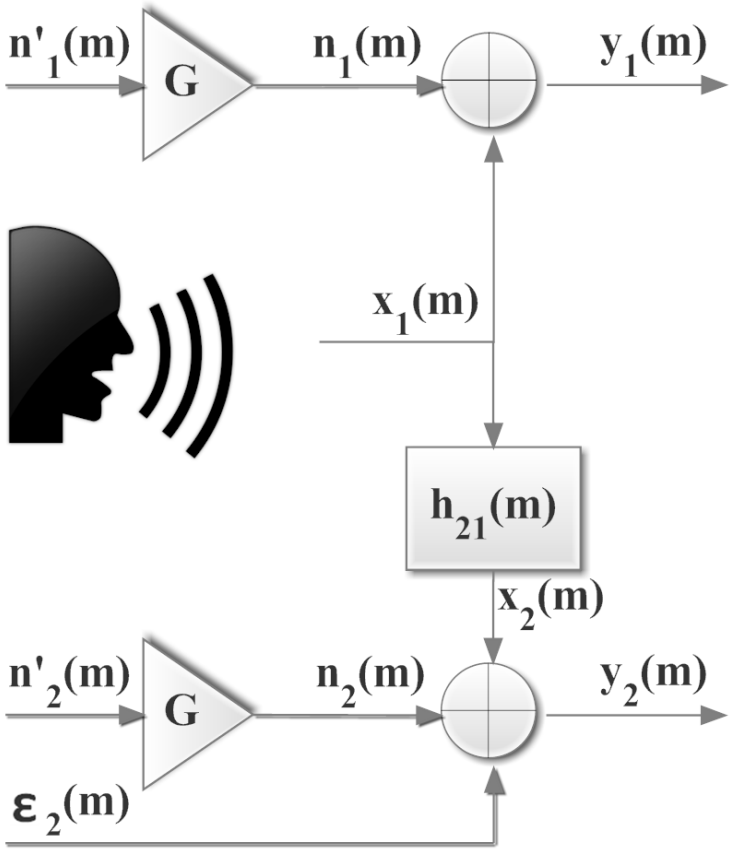
$$t = 0, 1, ..., T-1$$

- Input dim.:
$\dim(\mathcal{Y}_{NAT}(t)) =$
$\dim(\mathcal{Y}(t)) + 4\mathcal{M} + 1 =$
$4\mathcal{M}\left(L + \frac{3}{2}\right) + 1$

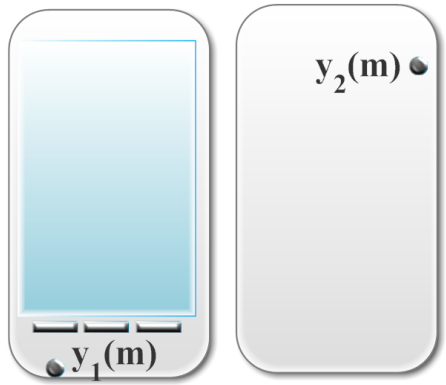
Target:
- It is the same! Actual $\mathcal{M} \times 1$ noise vector $\mathbf{n}_1(t)$

Experiments and Results

The AURORA2-2C 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

Experiments and Results

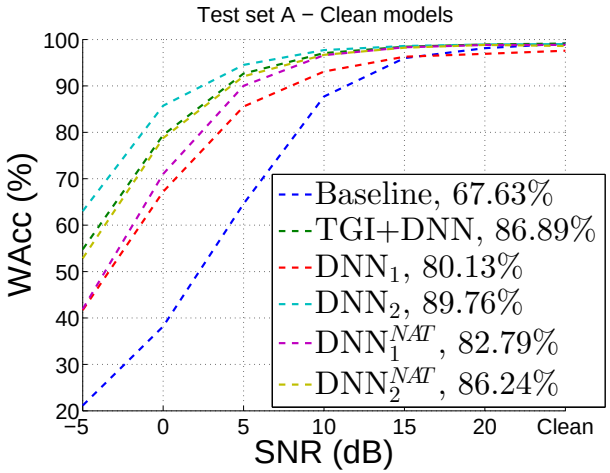
DNN Properties

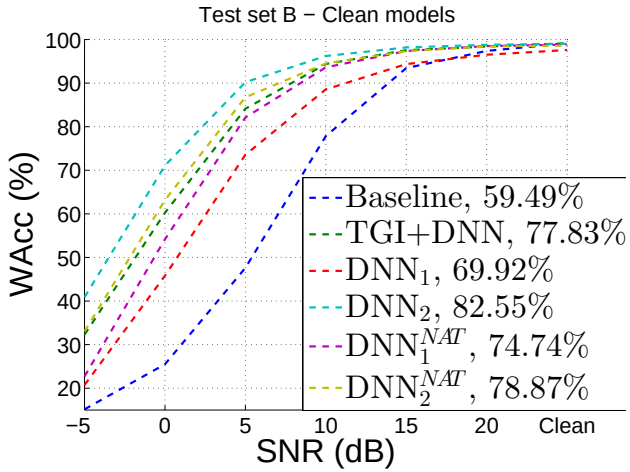
About the DNN configuration...

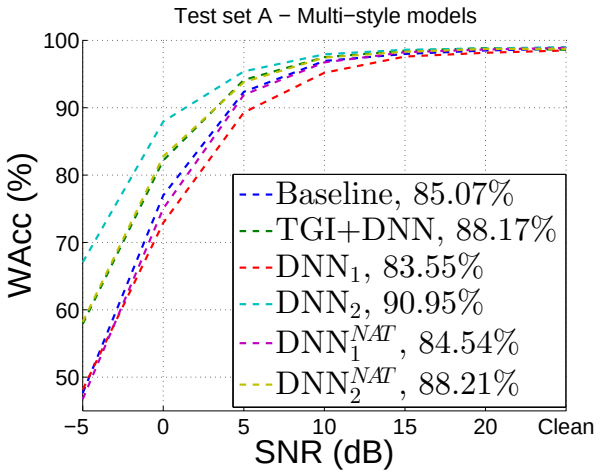
- Five hidden layers are used.
- It was trained using 25600 sample pairs of input-target vectors by just considering noises of test set A.
- $L = 2$ was chosen (and $\mathcal{M} = 23$):
  1. Input layer has 230 (323) nodes without (with) NAT.
  2. Hidden layers have 512 nodes.
  3. Output layer has $\mathcal{M} = 23$ nodes.

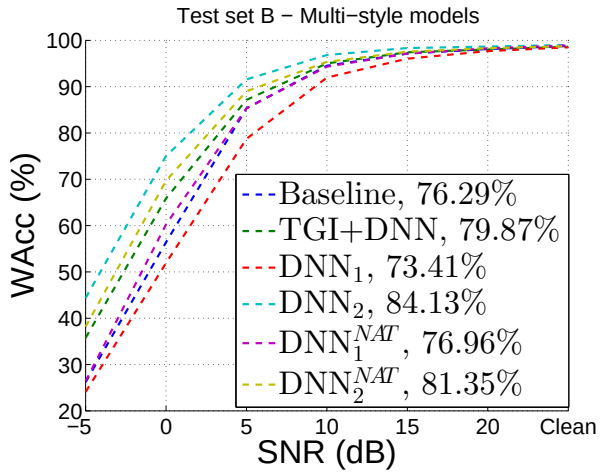
Experiments and Results

DNN-Based Noise Estimation Results







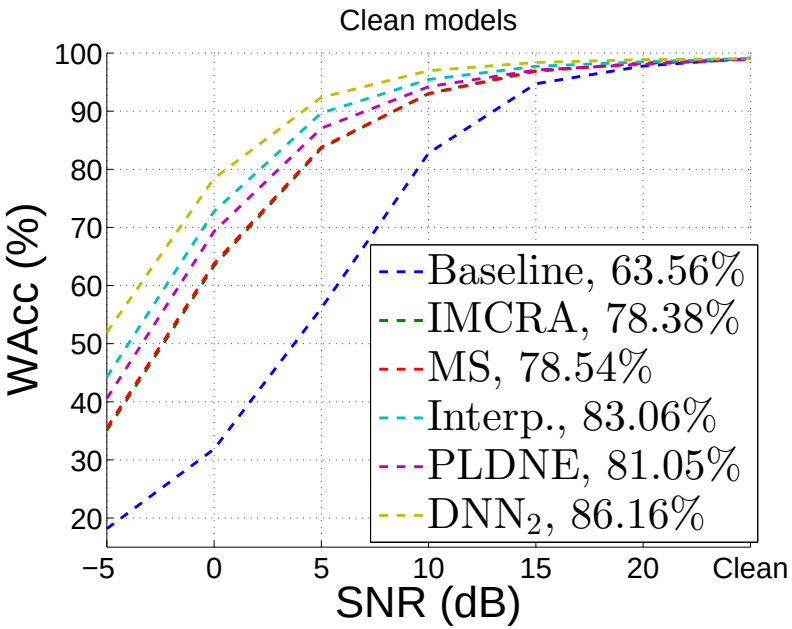


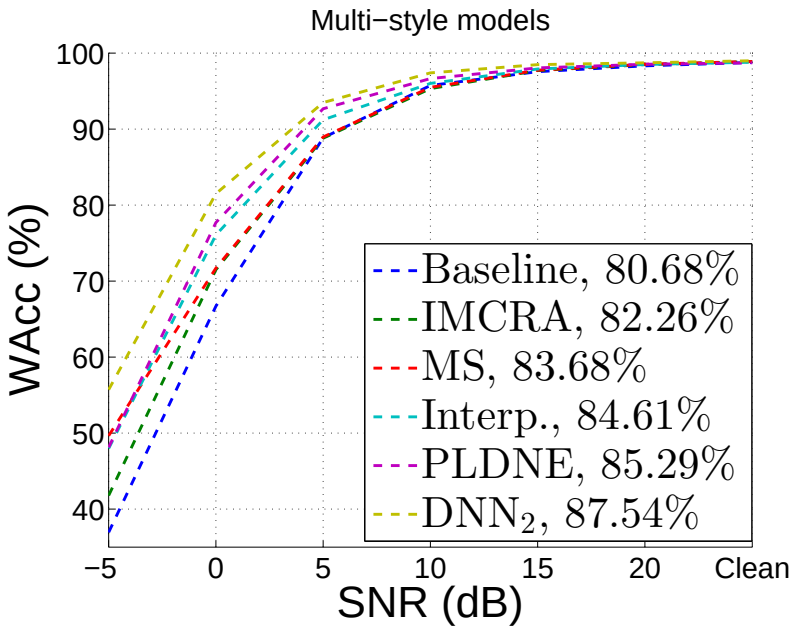
- DNN_1: Only the primary channel is used as input
- DNN_2: The dual-channel is used as input (as presented)
- DNN_1^{NAT}: DNN_1 with NAT
- DNN_2^{NAT}: DNN_2 with NAT

GMM-HMM-based ASR system (trained with both clean and multi-style data)

Experiments and Results

A Comparison with other Noise Estimation Methods





Noise estimation methods for comparison

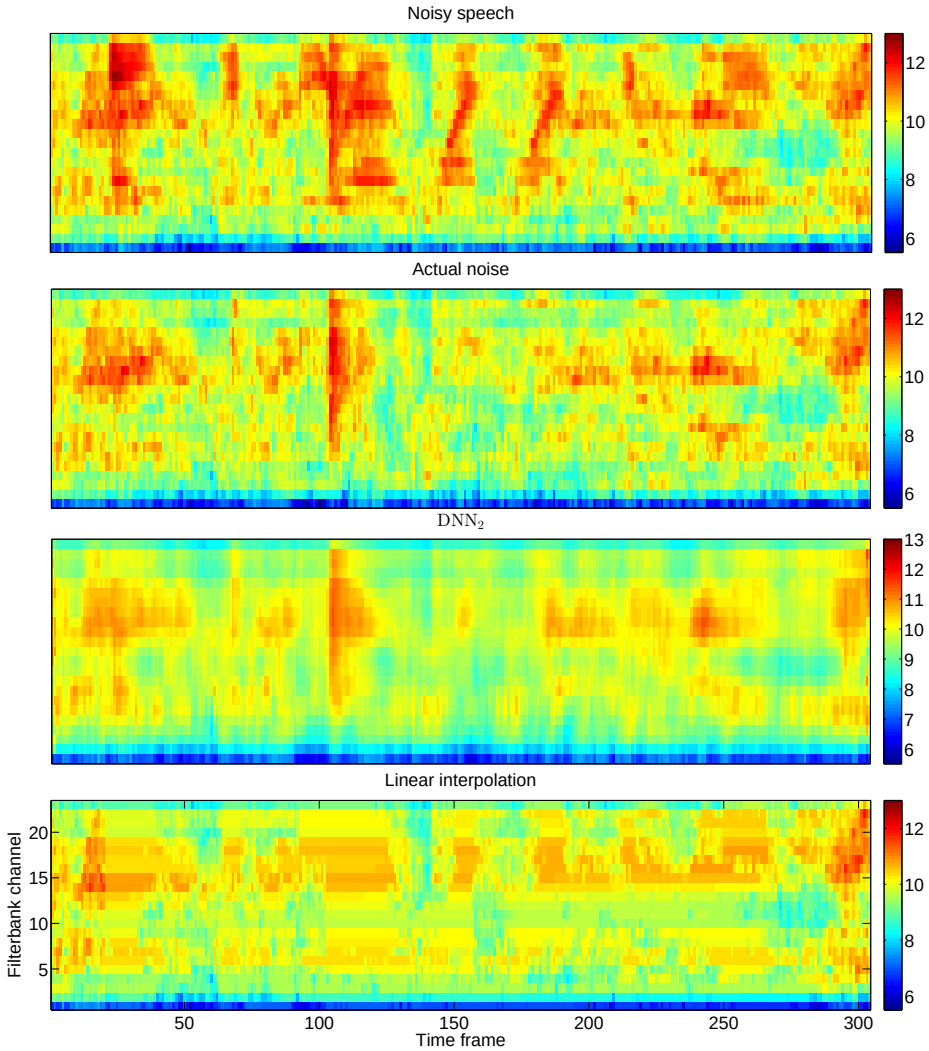
- **Single-channel methods**: Improved minima controlled recursive averaging (IMCRA), minimum statistics (MS) and linear interpolation (Interp.).
- **PLDNE** is for dual-mic smartphones (PLD and homogeneous noise field).

GMM-HMM-based ASR system (trained with both clean and multi-style data)

Experiments and Results

A Comparison with other Noise Estimation Methods

Example of noise estimation



Conclusions

Some conclusions and future work

- The DNN has been able to take advantage of the dual-channel information, providing significant improvements on performance.
- The use of the secondary channel can be seen as a sort of NAT.
- **Some benefits:**
  1. No assumptions are made.
  2. The DNN is able to learn complex non-linear dependencies between input and target.
  3. The dual-channel approach is efficient.
- **And as future work...**
  1. Exploring the performance of this approach with mobile devices with different small array configurations.
  2. We aim at studying how this method performs in far-talk conditions.

Thanks for your attention!

Contact:
Iván López-Espejo
Department of Signal Theory, Telematics and Communications
University of Granada
E-mail: iloes@ugr.es