DNN-Based Missing-Data Mask Estimation for Noise-Robust ASR in Dual-Microphone Smartphones

Iván López-Espejo*, José A. González†, Angel M. Gomez*, and Antonio M. Peinado*

*Dept. of Signal Theory, Telematics and Com., University of Granada, Spain

†Dept. of Computer Science, University of Sheffield, UK

Motivation

New ASR upswing

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

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

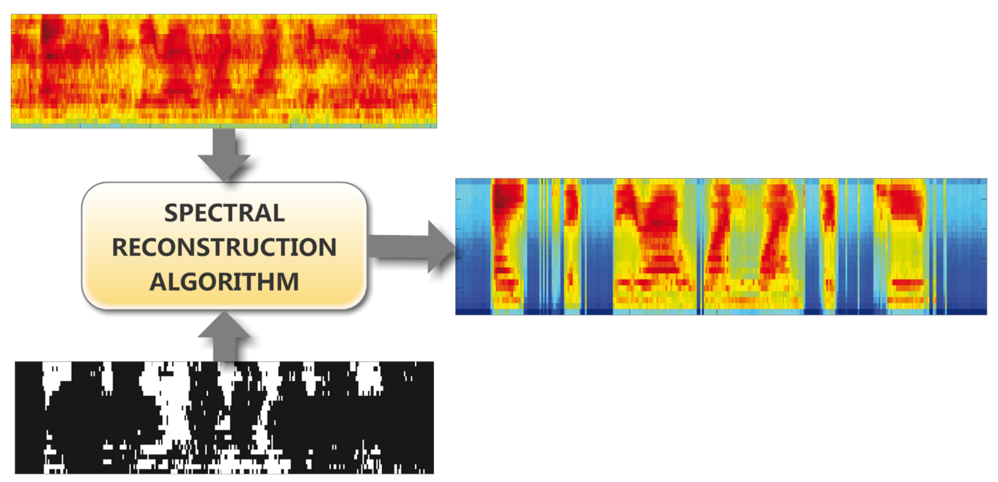
Noise-robust ASR in smartphones

- It is crucial to tackle with **noisy environments**.
- We can benefit from the novel dual-microphone feature.

Objectives

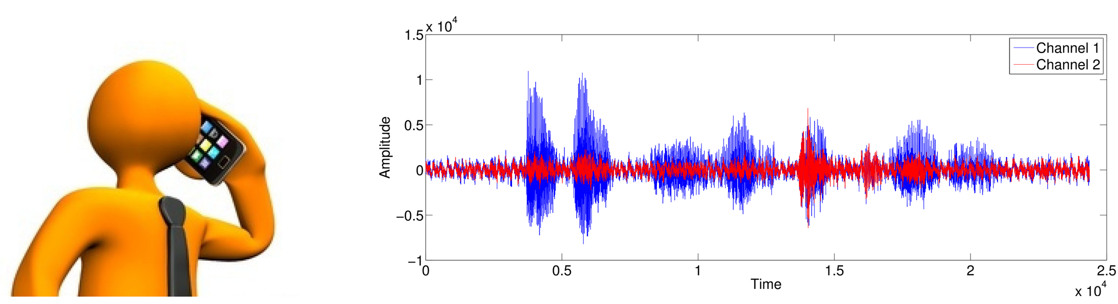
Our goals

- To **improve ASR performance** in noisy conditions by exploiting dual-mic configurations **on smartphones**.
- To use **spectral reconstruction** by means of TGI (truncated-Gaussian based imputation).
- To **estimate** missing-data **masks by using DNNs**.



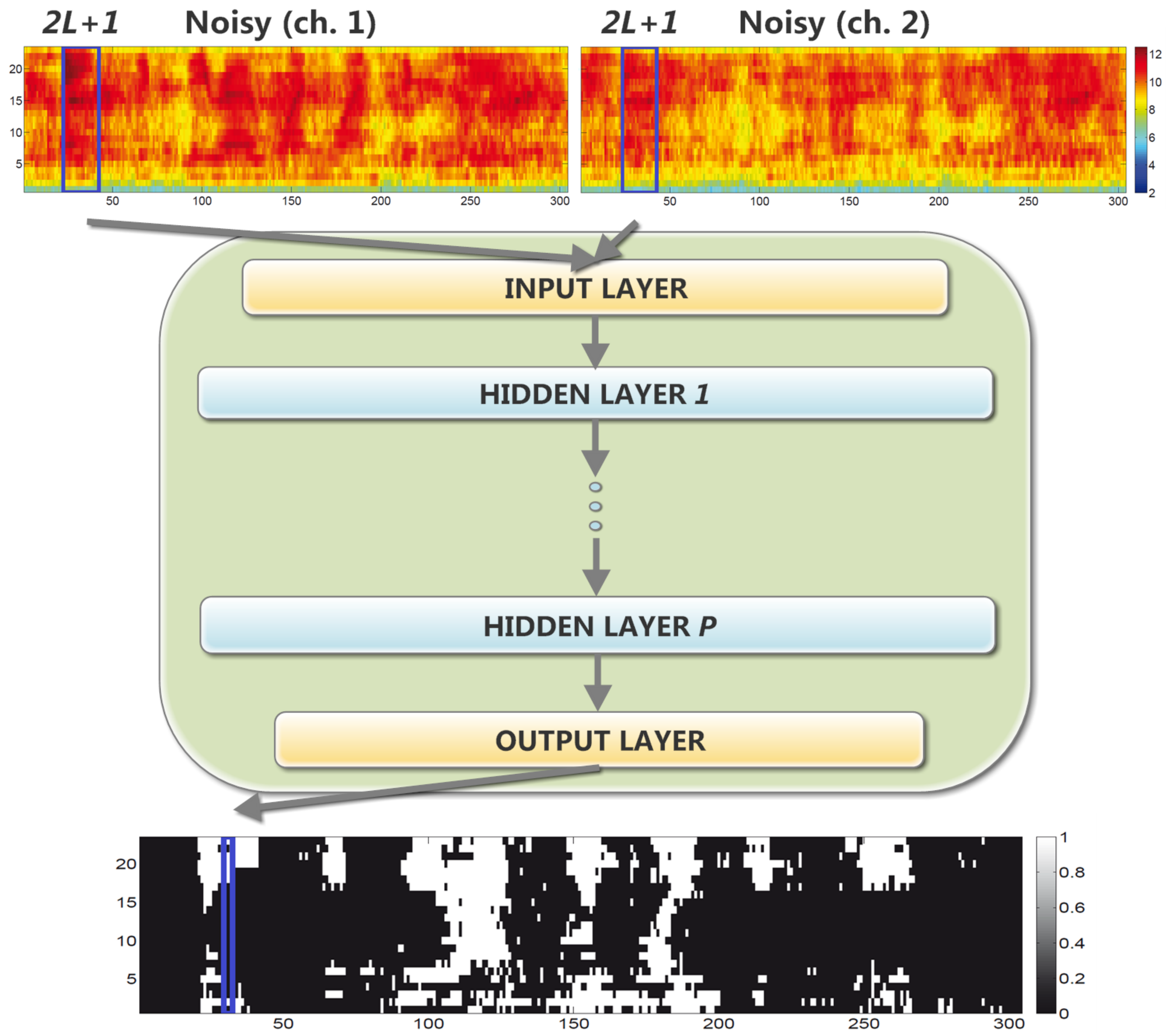
What we want the DNN exploits

- The dual-channel information provided by the smartphone.
- The power level difference (PLD) between the 2 mics in close-talk conditions.



Speech power is greater at the 1st ch. than at the 2nd one. Noise power is similar at both channels.

DNN-Based Proposed System



Features

$$\mathcal{Y} = \begin{pmatrix} \mathbf{y}(t-L) \\ \vdots \\ \mathbf{y}(t+L) \end{pmatrix}, \text{ where } \mathbf{y}(t) = \begin{pmatrix} \mathbf{y}_1(t) \\ \mathbf{y}_2(t) \end{pmatrix}$$

- Input dim.: $d_F = 2 \cdot \mathcal{M} \cdot (2L+1) \times 1$

Target

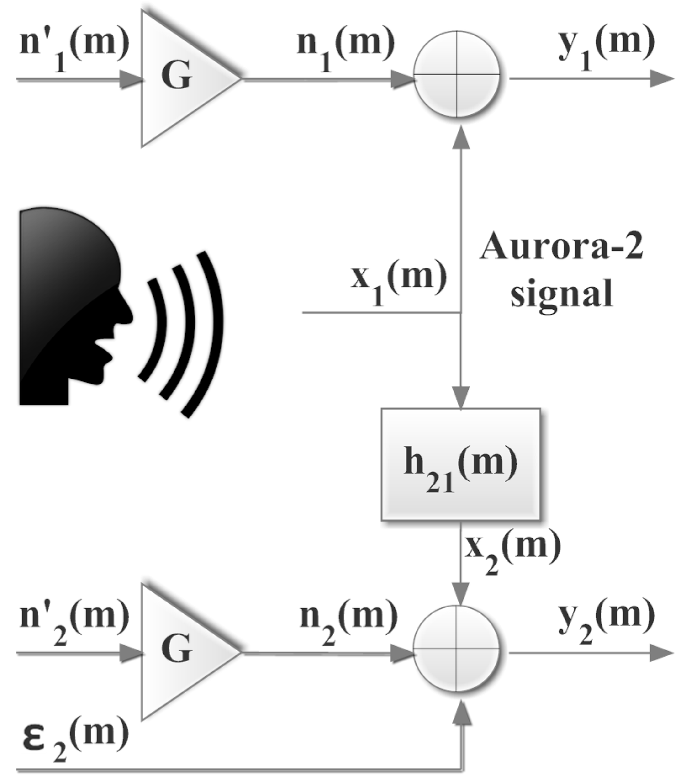
- Oracle binary mask vector for $\mathbf{y}_1(t)$
- 7 dB SNR threshold
- Output dim.: $d_T = \mathcal{M} \times 1$

Training issues

- The DNN is pre-trained by considering each pair of layers as RBMs.
- The DNN is trained by using the backpropagation algorithm (**cross-entropy criterion**).
- Since $\mathcal{M} = 23$ and $L = 2$, $d_F = 230$ and $d_T = 23$.
- We use 2 hidden layers with $2d_F = 460$ nodes in each one.

The AURORA2-2C-CT

The AURORA2 - 2 Channels - Close-Talk database



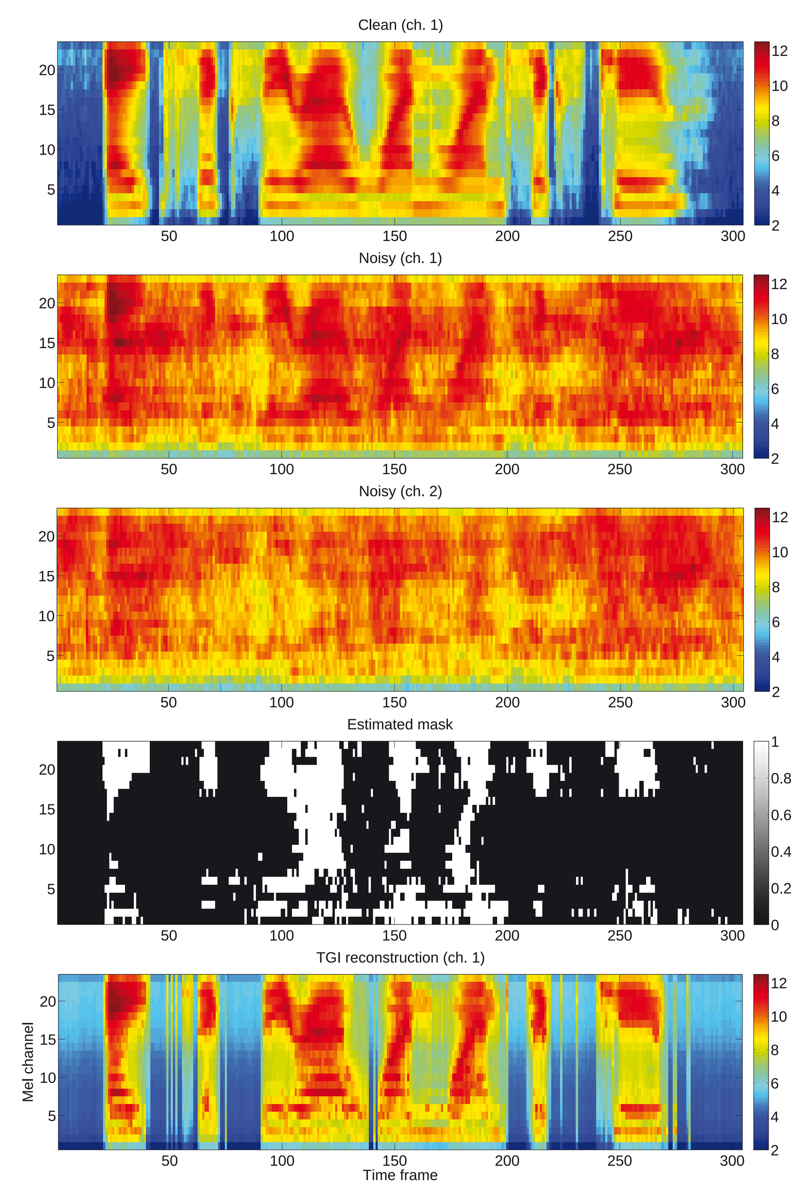
- The AURORA2-2C-CT emulates the acquisition of noisy speech data with a dual-microphone smartphone used in close-talk conditions.
- It is based on the well-known Aurora-2 database.
- It is used for speech recognition experiments in this work.

Design

- $h_{21}(m)$ is modeled as a time-invariant FIR filter trained from speech recorded with a smartphone.
- $\{n'_i; i = 1, 2\}$ were recorded with a smartphone and scaled by G to get a certain SNR for $y_1(m)$.
- The AURORA2-2C-CT is structured as Aurora-2.

Example of Application

- *Example of the TGI reconstruction of an utterance (all the spectrograms are in the log-Mel domain):*



- *From top to bottom:* clean utterance at the 1st ch., corrupted by bus noise at 0 dB at the 1st ch., corrupted by bus noise at 0 dB at the 2nd ch., mask estimated by the proposed DNN-based system and resulting TGI reconstruction (1st ch.).

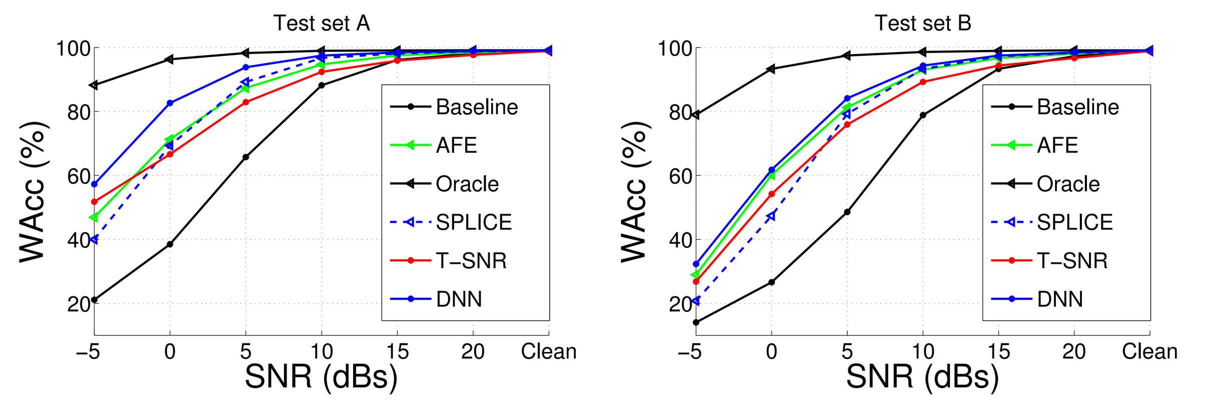
Experimental Results

Experimental framework

- We use the ETSI front-end to extract features.
- A GMM-HMM-based ASR back-end trained on clean speech is employed.
- The DNN is trained with the noises in test set *A*.
- TGI is combined with our proposal (DNN), with oracle masks and with masks estimated by thresholding an ML-based SNR estimate (T-SNR).
- Baseline, AFE and SPLICE are also included for comparison.

Results

	WAcc (%)			Wrong mask bins (%)		
	Test A	*Test B*	*Avg.*	*Test A*	*Test B*	*Avg.*
Baseline	67.96	59.78	63.87	-	-	-
AFE	82.71	76.37	79.54	-	-	-
Oracle	96.67	94.41	95.54	0	0	0
SPLICE	82.03	72.72	77.38	-	-	-
T-SNR	81.21	72.87	77.04	17.97	19.89	18.93
DNN	88.10	78.07	83.08	10.07	16.19	13.13



- The DNN exhibits some generalization ability to unseen noises during the training phase according to the results for test set *B*.

Conclusions

Conclusions

- The DNN has been able to take advantage of the dual-channel information.
- The DNN overcomes the analytical modeling capabilities and allows better performance.

Future work

- To extend this method to deal with a far-talk scenario, where the PLD assumptions are not completely valid.

Contact

Iván López-Espejo

Dept. of Signal Theory, Telematics and Com.
University of Granada, Spain
E-mail: iloes@ugr.es

ugr Universidad de Granada