A Deep Neural Network Approach for Missing-Data Mask Estimation on Dual-Microphone Smartphones: Application to Noise-Robust Speech Recognition

Iván López-Espejo[*], J.A. González[†], A.M. Gomez[*], and A.M. Peinado[*]
[*]Dept. of Signal Theory, Telematics and Com., University of Granada, Spain
[†]Dept. of Computer Science, University of Sheffield, UK
{iloes,amgg,amp}@ugr.es, j.gonzalez@sheffield.ac.uk

IberSPEECH '14 - VIII Jornadas en Tecnologías del Habla
11-20-2014, Las Palmas de Gran Canaria

Outline

Introduction
Motivation

New ASR upswing

The use of automatic speech recognition (ASR) applications has notably increased due to 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 take benefit from the novel dual-mic feature.

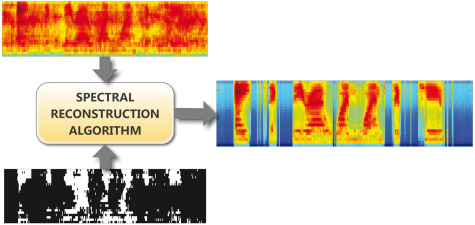
Introduction

Motivation

One possible approach to noise-robust ASR:

Spectral reconstruction

A BINARY MASK IS NEEDED



Introduction

Objectives

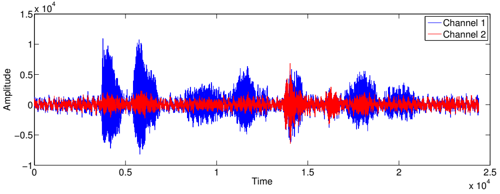
In a conversational position:

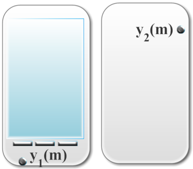
1. To **estimate missing-data masks** for the first channel by using the information contained in both channels.
    - We experiment with **deep neural networks** (DNNs).
2. To assess their quality when they are used by a spectral reconstruction technique over a dual-channel noisy speech database (**AURORA2-2C**).

The Power Level Difference

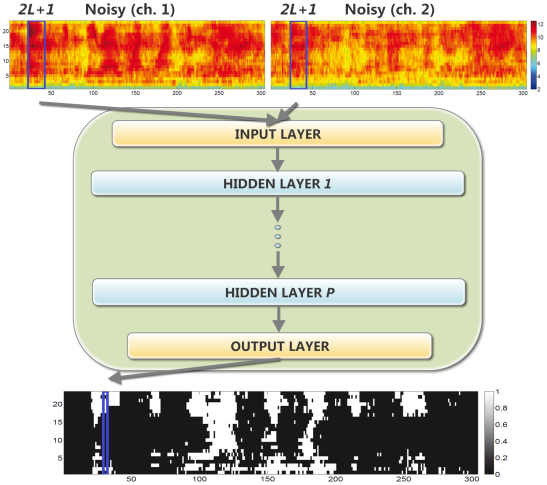
In a conversational position:

- Speech power at the primary mic tends to be greater than at the secondary one.
- **Far field noise:** Noise power received at both mics is almost the same.





DNN-Based Mask Estimation System



Features:

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

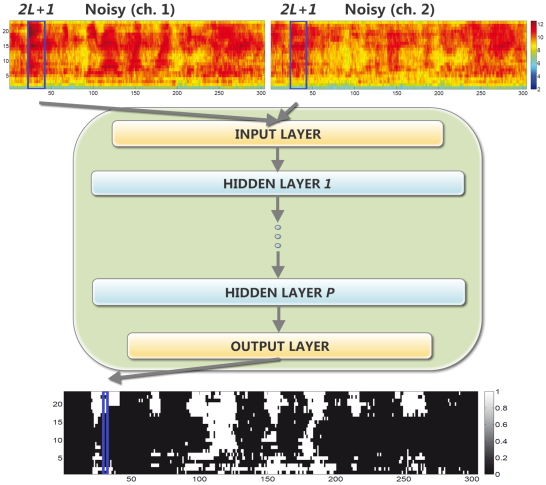
where

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

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

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

DNN-Based Mask Estimation System

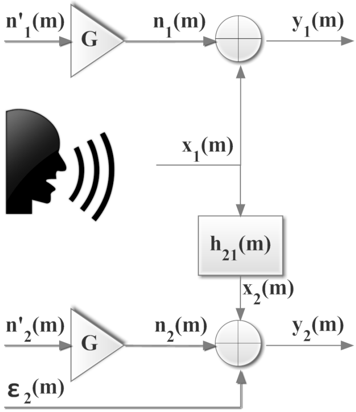


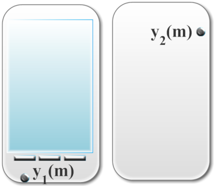
Training issues:

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

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

Experiments and Results
The AURORA2-2C Database





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

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

Experiments and Results

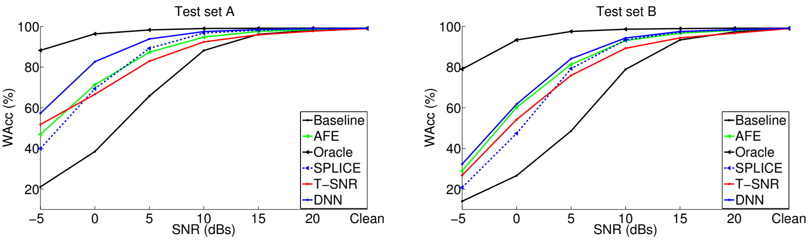
DNN Properties

About the DNN configuration...

- Two hidden layers are used.
- It was trained using 19200 sample pairs of input-output vectors by just considering noises of test set A.
- $L = 2$ was chosen (and $\mathcal{M} = 23$):
  1. Input layer has $d = 2 \cdot \mathcal{M} \cdot (2L + 1) = 230$ nodes.
  2. Hidden layers have 460 nodes.
  3. Output layer has $\mathcal{M} = 23$ nodes.

Experiments and Results

Results



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

GMM-HMM (trained with clean speech)

Conclusions

Some conclusions and future work

- The DNN has been able to take advantage of the dual-channel information, providing significant improvements on performance.
- **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. Exhaustive search regarding the architecture and training process of the DNN.
  2. We aim to extend this method in order to deal with a hands-free scenario.

Thanks for your attention and any questions?

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