Robust ASR on Mobile Devices with Small Array

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

The University of Sheffield
6-3-2015, Sheffield

Where I come from...

SigMAT (Signal Processing, Multimedia Transmission and Speech/Audio Technologies)

Main research lines of the group:

- **Robust speech recognition on mobile environments.**
    - **Robust ASR on mobile devices with small microphone array.**
- Robust transmission of speech and video.
- Ultrasonic non-destructive testing.
- Signal processing in proteomics.

Work done so far...

An overview of the work done so far

Work done by me et al. (in chronological order):

1. Feature Enhancement for Robust Speech Recognition on Smartphones with Dual-Microphone. *In Proc. of EUSIPCO, 2014*.
2. A Deep Neural Network Approach for Missing-Data Mask Estimation on Dual-Microphone Smartphones: Application to Noise-Robust Speech Recognition. *Lecture Notes in Computer Science, vol. 8854, 2014*.
3. Soft-Mask Spectral Weighting for Robust Speech Recognition in Smartphones with a Dual-Microphone. *Submitted to INTERSPEECH'15*.
4. Power Spectrum Enhancement for Noise-Robust Speech Recognition with Small Microphone Arrays. *About to submit it*.

Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Introduction and motivation

New ASR upswing

The use of ASR applications has notably increased due to the latest portable electronic devices:

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

Noise-robust ASR in smartphones

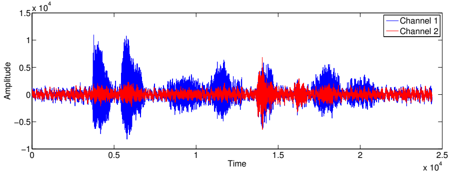
- It is crucial to tackle with noisy environments.
- We can take benefit from the novel dual-mic feature.

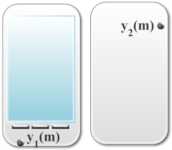
Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

The power level difference

In a close-talk 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 similar.
- **Our goal:** Estimating the clean speech power spectrum at the primary channel by using the information at both channels.

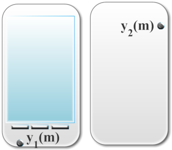




Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Dual-channel signal model

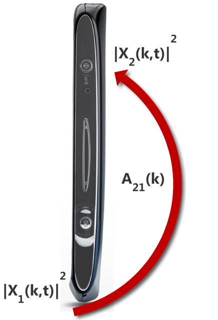
- We consider additive noise: $y_i(m) = x_i(m) + n_i(m)$, where $i = 1, 2$ indicates the mic (channel).
- Assuming that speech and noise are independent:
  $|Y_1(k,t)|^2 = |X_1(k,t)|^2 + |N_1(k,t)|^2$
  $|Y_2(k,t)|^2 = |X_2(k,t)|^2 + |N_2(k,t)|^2$



Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Minimum mean square noise (MMSN) feature enhancer

- Minimum mean square noise (**MMSN**) feature enhancer is defined as
$|\hat{X}_1(k,t)|^2 = \mathbf{w}_k^{\mathrm{T}} \begin{pmatrix} |Y_1(k,t)|^2 \\ |Y_2(k,t)|^2 \end{pmatrix}.$
- The speech power in the second channel is related with the speech power in the first one through a time-invariant factor $A_{21}(k)$:
$|Y_2(k,t)|^2 = A_{21}(k)|X_1(k,t)|^2 + |N_2(k,t)|^2.$
- Weights are computed by using the well-known MVDR (minimum variance distortionless response) approach: $\mathbf{w}_k = \frac{\mathbf{\Phi}_{N,k}^{-1}(1, A_{21}(k))^{\mathrm{T}}}{(1, A_{21}(k))\mathbf{\Phi}_{N,k}^{-1}(1, A_{21}(k))^{\mathrm{T}}}.$

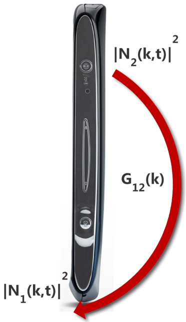


$|\mathbf{X_2}(\mathbf{k},\mathbf{t})|^2 = \mathbf{A_{21}}(\mathbf{k})|\mathbf{X_1}(\mathbf{k},\mathbf{t})|^2$

Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Dual-channel spectral subtraction (DCSS)

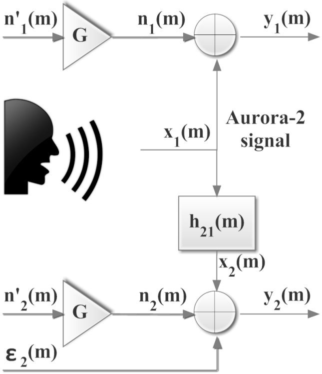
- We can also relate noise power spectra at both channels: $|Y_1(k,t)|^2 = |X_1(k,t)|^2 + G_{12}(k)|N_2(k,t)|^2$.
- Dual-channel spectral subtraction (**DCSS**) estimator: $|\hat{X}_1(k,t)|^2 = \frac{|Y_1(k,t)|^2 - G_{12}(k)|Y_2(k,t)|^2}{1 - G_{12}(k)A_{21}(k)}$.
- $G_{12}(k)$ is estimated by minimizing $\mathbb{E}\left[(|N_1(k,t)|^2 - G_{12}(k)|N_2(k,t)|^2)^2\right]$: $\hat{G}_{12}(k) = \frac{\hat{\phi}_{N,k}(1,2)}{\hat{\phi}_{N,k}(2,2)}$.

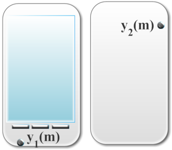


$|N_1(k,t)|^2 = G_{12}(k)|N_2(k,t)|^2$

Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

The AURORA2-2C-CT database

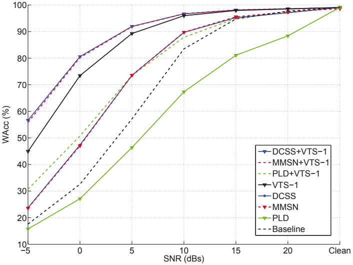




- **Test A:** Bus, babble, car and pedestrian street
- **Test B:** Cafe, street, bus and train stations
- **SNRs:** {-5,0,5,10,15,20} dB and clean

Feature Enhancement for Robust ASR on Smartphones with Dual-Mic

Results



GMM-HMM
(trained with clean speech)

- **PLD:** speech enhancer for smartphones with dual-microphone.
- MMSN and DCSS have a similar performance when $A_{21}(k) \to 0$:

$$\begin{cases} w_1(k)^{\{MMSN,DCSS\}} \to 1 \\ w_2(k)^{\{MMSN,DCSS\}} \to -\frac{\phi_{N,k}(1,2)}{\phi_{N,k}(2,2)} \end{cases}$$

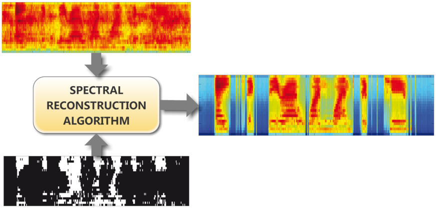
$$|\hat{X}_1(k,t)|^2 = w_1(k)|Y_1(k,t)|^2 + w_2(k)|Y_2(k,t)|^2$$

A DNN Approach for Mask Estimation on Dual-Mic Smartphones

Motivation for this work

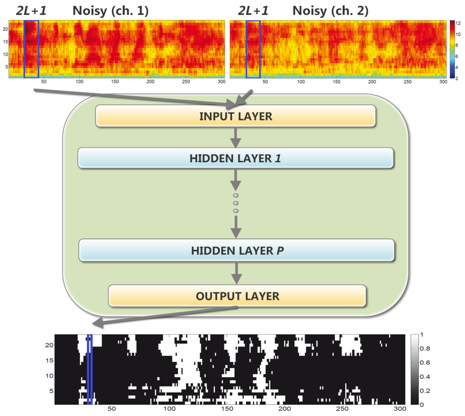
Another possible approach to noise-robust ASR:
spectral reconstruction

A BINARY MASK IS NEEDED



A DNN Approach for Mask Estimation on Dual-Mic Smartphones

DNN-based mask estimation system



Features:

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

where

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

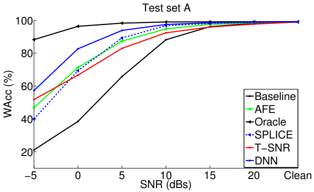
- Input dim.: $d = 2 \cdot \mathcal{M} \cdot (2L+1) \times 1$

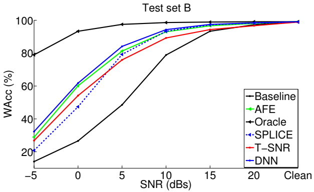
Target:

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

A DNN Approach for Mask Estimation on Dual-Mic Smartphones

Experiments and results





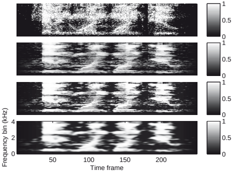
	WAcc (%)			Wrong mask bins (%)		
	Test A	Test B	Average	Test A	Test B	Average
Baseline	67.96	59.78	63.87	-	-	-
AFE	82.71	76.37	79.54	-	-	-
Oracle+TGI	96.67	94.41	95.54	0	0	0
SPLICE	82.03	72.72	77.38	-	-	-
T-SNR+TGI	81.21	72.87	77.04	17.97	19.89	18.93
DNN+TGI	**88.10**	**78.07**	**83.08**	**10.07**	**16.19**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

GMM-HMM (trained with clean speech)

Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Description of the approach

- We follow a Wiener filter approach: $|\hat{X}_1(k,t)|^2 = \hat{H}_1^2(k,t)|Y_1(k,t)|^2$.
- $\hat{H}_1^2(k,t) = \left(\hat{\xi}_1(k,t)/(\hat{\xi}_1(k,t)+1)\right)^2$ may be seen as a spectral weighting soft-mask (b).
- We exploit the PLD by assuming that $\varphi_{X_1}(k,t) \gg \varphi_{X_2}(k,t)$ and $\varphi_{N_2}(k,t) \approx \varphi_{N_1}(k,t)$: $\hat{\xi}_1(k,t) = \max\left(\frac{\varphi_{Y_1}(k,t)}{\varphi_{Y_2}(k,t)} - 1, 0\right)$.
- We apply a post-processing to improve the soft-mask:
  1. Slight contrast by using a sigmoid function (c).
  2. Median and Gaussian filtering to improve the spectro-temporal coherence (d).



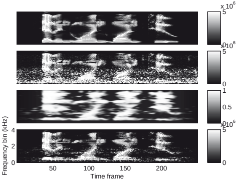
From top to bottom:

(a) Oracle
(b) $\hat{H}_1^2(k,t)$
(c) After sigmoid
(d) After filtering

Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Description of the approach

- We follow a Wiener filter approach: $|\hat{X}_1(k,t)|^2 = \hat{H}_1^2(k,t)|Y_1(k,t)|^2$.
- $\hat{H}_1^2(k,t) = \left(\hat{\xi}_1(k,t)/(\hat{\xi}_1(k,t)+1)\right)^2$ may be seen as a spectral weighting soft-mask.
- We exploit the PLD by assuming that $\varphi_{X_1}(k,t) \gg \varphi_{X_2}(k,t)$ and $\varphi_{N_2}(k,t) \approx \varphi_{N_1}(k,t)$: $\hat{\xi}_1(k,t) = \max\left(\frac{\varphi_{Y_1}(k,t)}{\varphi_{Y_2}(k,t)} - 1, 0\right)$.
- We apply a post-processing to improve the soft-mask:
  1. Slight contrast by using a sigmoid function.
  2. Median and Gaussian filtering to improve the spectro-temporal coherence.



From top to bottom:

1. Clean
2. Noisy
3. Soft-mask
4. Enhanced

Soft-Mask Weighting for Robust ASR on Smartphones with a Dual-Mic

Results

Tech./SNR (dB)	-5	0	5	10	15	20	Clean	Av. (-5 to 20)
Baseline	18.15	31.85	56.11	82.78	94.72	97.76	**99.13**	**63.56**
SMW	26.23	51.76	77.03	89.49	94.19	96.09	98.40	**72.47**
MMSN	24.16	46.31	74.78	90.66	96.14	97.87	98.90	**71.65**
DCSS	24.37	46.69	75.06	90.65	96.03	97.64	**99.13**	**71.74**
New	28.36	52.59	79.65	92.39	96.68	98.04	99.11	**74.62**
VTS-1	44.25	72.75	89.69	95.44	97.71	98.49	99.09	**83.06**
SMW+VTS-1	29.37	56.52	79.75	90.19	94.08	95.72	98.13	**74.27**
MMSN+VTS-1	56.15	81.05	**92.41**	**96.60**	98.15	**98.65**	98.99	**87.17**
DCSS+VTS-1	56.22	81.04	92.40	96.57	**98.21**	98.63	99.09	**87.18**
New+VTS-1	**61.14**	**81.43**	92.05	95.89	97.82	98.48	99.05	**87.80**

Table: *WAcc results in terms of percentage. Results are averaged across all types of noise in test sets A and B.*

Power Spectrum Enhancement for Noise-Robust ASR with Small Mic Arrays

Motivation

Facts...

- Several types of devices can be used.
- Devices can have more than two mics arranged in different ways.
- Devices can be used in different and variable positions.

Therefore...

- We generalize our previous work to $\mathcal{C}$ mics and a variable position.
  - MMSN $\rightarrow$ **P-MVDR** (power MVDR)
  - DCSS $\rightarrow$ **MSS** (multichannel spectral subtraction)

Power Spectrum Enhancement for Noise-Robust ASR with Small Mic Arrays

Speech gain vector and P-MVDR and MSS equations

P-MVDR:

$$|\hat{X}_1(k,t)|^2 = \left(\frac{\mathbf{\Phi}_{k,t}^{-1}\mathbf{A}_{k,t}}{\mathbf{A}_{k,t}^{\mathrm{T}}\mathbf{\Phi}_{k,t}^{-1}\mathbf{A}_{k,t}} \right)^{\mathrm{T}} \mathbf{Y}_{k,t}$$

MSS:

$$|\hat{X}_1(k,t)|^2 = \frac{\mathbf{Y}_{k,t}^{\mathrm{T}}\mathbf{\Gamma}_{k,t}\mathbf{A}_{k,t} - \mathbf{Y}_{k,t}^{\mathrm{T}}\mathbf{A}_{k,t} \cdot ||\mathbf{G}_{k,t}||^2}{(\mathbf{A}_{k,t}^{\mathrm{T}}\mathbf{G}_{k,t})^2 - ||\mathbf{A}_{k,t}||^2 \cdot ||\mathbf{G}_{k,t}||^2}$$

$$\mathbf{\Gamma}_{k,t} = \mathbf{G}_{k,t} \cdot \mathbf{G}_{k,t}^{\mathrm{T}}$$

$$\mathbf{G}_{k,t} = (1, G_{21}(k,t), ..., G_{\mathcal{C}1}(k,t))^{\mathrm{T}}$$

- Use position or acoustics may be variable.
- We developed an MMSE-based estimator to estimate $\mathbf{A}_{k,t} = (1, A_{21}(k,t), ..., A_{\mathcal{C}1}(k,t))^{\mathrm{T}}$ on a frame-by-frame basis.

Power Spectrum Enhancement for Noise-Robust ASR with Small Mic Arrays

About our experiments and results

- We created the AURORA2-2C-FT database emulating a smartphone with a dual-mic but in far-talk conditions.
- We created validation test datasets with real noisy data for both close-talk and far-talk conditions.
- Our recognition results showed the success of our developments in all cases.

Conclusions and future work

- Multichannel information can be exploited to improve ASR performance.
- There is little work on robust ASR with small mic arrays → We should be able to achieve further improvements.
- We are very interested in obtaining new and good results on the CHiME-3 database.

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