A brief introduction to multichannel noise reduction with deep neural networks
Romain Serizel

To cite this version:
Romain Serizel. A brief introduction to multichannel noise reduction with deep neural networks. SpiN 2020 - 12th Speech in Noise Workshop, Jan 2020, Toulouse, France. hal-02506387

HAL Id: hal-02506387
https://inria.hal.science/hal-02506387
Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A brief introduction to multichannel noise reduction with deep neural network.

Romain Serizel

LORIA, Univeristé de Lorraine, Inria, CNRS (France)

Thursday 9th, January 2020
Outline

1. Context and motivations
2. Multichannel speech enhancement
3. Introduction to artificial neural networks
4. DNN for multichannel speech enhancement
5. Conclusions and perspectives
Speech in real conditions

Problem “solved”

Close-up microphones, clean speech
Speech in real conditions

There is some work to do:

Moving microphones
Speech in real conditions

There is some work to do:

Distant microphones
Speech in real conditions

There is some work to do:
Distant microphones, mobile sources…
Multichannel speech enhancement?

Context
- Reduce the impact of the acoustic perturbations:
  - Cleaning the signal

In this talk
- Focus on noise removal/attenuation
  - Noise reduction
  - Noise suppression
  - De-noising
  - ...
1. Context and motivations

2. Multichannel speech enhancement

3. Introduction to artificial neural networks

4. DNN for multichannel speech enhancement

5. Conclusions and perspectives
Signal observed

\[ X = X_s - X_n \]
Filtering

Single-channel filter

Speech enhancement with DNN
Filtering

Single-channel filter

Multichannel filter
Single-channel enhancement

On single-mic data, perform spectral filtering/masking.

Source de parole

Mélange parole + bruit

Filtre de Wiener

Signal filtré
Multichannel enhancement (1)

On multi-mic data, exploit spatial localization.
Multichannel enhancement (2)

On multi-mic data, perform joint spatial and spectral masking.

Filtre de Wiener (anéchoïque)

Filtre de Wiener (réverb)
### MWF-based speech enhancement (1)

#### Speech distortion weighted MWF (MMSE criterion)

MMSE estimate of the (unknown) speech component:

$$\min_W E\{|X_{\text{Ref}} - W^H X_s|^2\} + \mu \{W^H X_n|^2\}$$

#### SDW-MWF

$$W = (R_s + \mu R_n)^{-1} R_s e_{\text{Ref}}$$

- $X_s, X_n$: speech, noise components of the microphone signals
- $W^H X$: output signal of the filter $W$
- $R_s$ and $R_n$: speech and noise correlation matrices
MWF-based speech enhancement (2)

Estimation of the speech correlation matrix

\[ R_s \] is estimated from speech+noise correlation matrix \( R_x \) and noise correlation matrix \( R_n \):

\[ R_s = R_x - R_n \]

→ Requires a voice activity detector!
Voice activity detector

Limitations

- Limited to full frequency band
- Energy-based VAD do not work well at low SNR
Voice activity detector

Limitations

- Limited to full frequency band
- Energy-based VAD do not work well at low SNR
Time-frequency masks

How can we obtain reliable masks?
Outline

1. Context and motivations
2. Multichannel speech enhancement
3. Introduction to artificial neural networks
4. DNN for multichannel speech enhancement
5. Conclusions and perspectives
A perceptron is a linear classifier:

\[ h = w \cdot x + b \]

\[ y = f(h) \text{ e.g. } \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{else} \end{cases} \]
Perceptron cannot model a XOR [Minsky et al., 1969] ⇒ Work on perceptron stopped
Perceptron cannot model a XOR [Minsky et al., 1969] → Work on perceptron stopped

At least single perceptron cannot!
Multilayer perceptrons (MLP):

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification
Multilayer perceptrons (MLP):

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification
Multilayer perceptrons

Multilayer perceptrons (MLP):

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

Hidden layer(s)
Multilayer perceptrons

Multilayer perceptrons (MLP):

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

Output layer
Multilayer perceptrons

Multilayer perceptrons (MLP):

- Feedforward network
- Perceptrons arranged in fully-connected layers
- Well suited for pattern classification

Deep neural networks have more than 1 hidden layers
Convolutional neural networks (CNN)

Keys ideas

- Detect patterns in images/spectrograms
- Apply sets of filters to obtain feature maps
- Subsample to control dimensionality

Image source: wikimedia
Recurrent neural networks (RNN)

Keys idea

- Model context

Image source: wikimedia

R. Serizel

Speech enhancement with DNN

09/01/2020
Recurrent neural networks (RNN)

Keys idea

- Model context

Problems

- Hard to train
- Exploding/vanishing gradient

Image source: wikimedia

R. Serizel
Speech enhancement with DNN
09/01/2020
Recurrent neural networks (RNN)

Keys idea
- Model context

Solutions
- LSTM, GRU...
1. Context and motivations
2. Multichannel speech enhancement
3. Introduction to artificial neural networks
4. DNN for multichannel speech enhancement
5. Conclusions and perspectives
DNN-base mask estimation for multichannel filtering

Use LSTM to estimate speech/noise masks [Heymann et al.].

speech mask [513]  noise mask [513]

feed-forward layers [512, 512]

BLSTM layer [256]

input magnitude spectrum vector [513]

DNN-base mask estimation for MWF

[Z. Wang]

Use the masks to estimate the correlation matrices and MWF filters.

Exemple performance

Evaluation in terms of word error rate (WER).

<table>
<thead>
<tr>
<th>Method</th>
<th>WER reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>WER baseline</td>
</tr>
<tr>
<td>Single-channel DNN</td>
<td>no WER reduc.</td>
</tr>
<tr>
<td>Delay-and-sum</td>
<td>21% rel. WER reduc.</td>
</tr>
<tr>
<td>DNN post-filter</td>
<td>20% rel. WER reduc.</td>
</tr>
<tr>
<td>Multichannel DNN</td>
<td>39% rel. WER reduc.</td>
</tr>
</tbody>
</table>

Greatly outperforms multichannel NMF (25% rel. WER reduc.) and DNN used to directly predict multichannel filters.


What input should we give to the network?
(Part of Phd work from L. Perotin)
[With E. Vincent and A. Guérin]
Multichannel speech enhancement for HOA

Represent the sound field in a basis of spherical harmonics
Multichannel speech enhancement for HOA

Use only the 4 channels from the first 2 orders

L. Perotin, R. Serizel, E. Vincent, A. Guérin. *Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings*. In proc. of ICASSP, 2018,
Multichannel speech enhancement for HOA [L. Perotin]

- HOA beamformer
- LSTM to estimate speech/noise masks
- GEVD-MWF filter

L. Perotin, R. Serizel, E. Vincent, A. Guérin. Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In proc. of ICASSP, 2018,
### Setup
- Tested on a subset of ESTER
- 20 sentences (4043 words)
- Acoustic conditions: $TR_{60}=350ms$, room-mic = 1.65m

### Results

<table>
<thead>
<tr>
<th>Interference</th>
<th>noise</th>
<th>speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25°</td>
</tr>
<tr>
<td>Reverberant source</td>
<td>9.80</td>
<td>10.80</td>
</tr>
<tr>
<td>Mixture</td>
<td>68.25</td>
<td>91.96</td>
</tr>
<tr>
<td>Simple beamformer</td>
<td>40.80</td>
<td>75.42</td>
</tr>
<tr>
<td>GEVD (w/o noise est.)</td>
<td>21.69</td>
<td>78.43</td>
</tr>
<tr>
<td>GEVD (w noise est.)</td>
<td>21.72</td>
<td>23.1</td>
</tr>
</tbody>
</table>
How can we deal with ad-hoc microphone arrays?
(Part of Phd work from N. Furnon)
[With I. Illina and S. Essid]
Heterogeneous unconstrained microphone arrays

Target:

- Use the microphones already available
Collaborative approach

**Targets:**

- Collaboration between the nodes (common target)
- Feedback from the common target to the node level
Distributed MWF with DNN [N. Furnon]

\[ \mathbf{z}_j \neq \mathbf{k} \quad \mathbf{y}_k \quad \mathbf{z}_k \quad \mathbf{s}_k \]

Filter

DNN

Node \( k \)

Nodes \( j \neq k \)

Distributed MWF with DNN [N. Furnon]

Filter

DNN

$z_j \neq k \quad y_k \quad z_k \quad M_s \quad s_k$

Node $k$

Nodes $j \neq k$

Distributed MWF with DNN (Results)


Speech enhancement with DNN
Outline

1. Context and motivations
2. Multichannel speech enhancement
3. Introduction to artificial neural networks
4. DNN for multichannel speech enhancement
5. Conclusions and perspectives
Conclusions and perspectives

Conclusions

• Deep learning is now the state-of-the-art in multichannel speech enhancement
• Important improvements compared to previous approaches

Perspectives

• Unconstrained microphone array processing,
• Distributed learning,
• Application to real-world conditions,
• Extension to other related domains...