Voice conversion and speaker adaptation based on physically meaningful transforms

Daniel Erro

AHOLAB, University of the Basque Country
Basque Science Foundation
Bilbao, Spain
Introduction

• De-identification?
  Not yet
Introduction

• De-identification?
  Not yet

• What then?
  Personalization of natural and synthetic speech
Introduction

• De-identification?
  Not yet

• What then?
  Personalization of natural and synthetic speech

• Is this technology usable in de-identification?
  Yes, of course!
Introduction

• De-identification?
  Not yet

• What then?
  Personalization of natural and synthetic speech

• Is this technology usable in de-identification?
  Yes, of course!

• This talk
  Brief overview of our recent work
Outline

• Voice conversion

• Speaker adaptation

• Other research lines

• Forthcoming events
Outline

- Voice conversion
- Speaker adaptation
- Other research lines
- Forthcoming events
Voice conversion

source

Bla bla bla

VC system

Bla bla bla

target
Voice conversion
Voice conversion

source

\{x_t\}_{t=1}^{N_x}

target

\{y_t\}_{t=1}^{N_y}

bla
ble
bli
blo
blu

bla
ble
bli
blo
blu
Voice conversion

source

\{x_t\}_{t=1...N_x} \quad \{x_t, y_t\}_{t=1...N} \quad \{y_t\}_{t=1...N_y}

target

\{x_t\}_{t=1...N_x}

\{x_t, y_t\}_{t=1...N}

\{y_t\}_{t=1...N_y}
Voice conversion

source

\{x_t\}_{t=1}^{N_x}

\{x_t, y_t\}_{t=1}^{N}

F(x)

\{y_t\}_{t=1}^{N_y}

target

bla
ble
bli
blo
blu

bla
ble
bli
blo
blu
Voice conversion

- State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)
Voice conversion

- State of the art: “blind” transformations driven by data
  
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)
Voice conversion

• State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)
Voice conversion

- State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)
Voice conversion

- State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

In (Mel-) cepstral domain...

\[ F(x) = Ax + b \]

- FW
- AS
Voice conversion

- State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

In (Mel-) cepstral domain...

\[ F(x) = Ax + b \]
Voice conversion

- State of the art: “blind” transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

\[
F(x) = Ax + b \\
A = \sum_{i=1}^{M} p_i(x) \cdot A_i, \quad b = \sum_{i=1}^{M} p_i(x) \cdot b_i
\]
Voice conversion

• State of the art: "blind" transformations driven by data
  Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

\[
A_\alpha = \begin{bmatrix}
1 & \alpha & \alpha^2 & \cdots \\
0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots \\
0 & -\alpha + \alpha^3 & 1-4\alpha^2 + 3\alpha^4 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}
\]

\[F(x) = Ax + b\]

\[A = \sum_{i=1}^{M} p_i(x) \cdot A_i, \quad b = \sum_{i=1}^{M} p_i(x) \cdot b_i\]
Voice conversion

- State of the art: “blind” transformations driven by data

Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

\[ A_\alpha = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots \\ 0 & 1 - \alpha^2 & 2\alpha - 2\alpha^3 & \cdots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \]

In (Mel-) cepstral domain...

\[ F(x) = A_\alpha x + b \]

\[ \alpha = \sum_{i=1}^{M} p_i(x) \cdot \alpha_i, \quad b = \sum_{i=1}^{M} p_i(x) \cdot b_i \]
Voice conversion

- State of the art: “blind” transformations driven by data

Physically meaningful transformations based on Frequency Warping (FW) + Amplitude Scaling (AS)

\[
A_\alpha = \begin{bmatrix}
1 & \alpha & \alpha^2 & \cdots \\
0 & 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots \\
0 & -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots \\
\vdots & \vdots & \vdots & \ddots \\
\end{bmatrix}
\]

In (Mel-) cepstral domain...

\[
F(x) = A_\alpha x + b
\]

\[
\alpha = \sum_{i=1}^{M} p_i(x) \cdot \alpha_i, \quad b = \sum_{i=1}^{M} p_i(x) \cdot b_i
\]

BLFW+AS
Voice conversion

• Results

Better conversion

Better quality

Ideal performance

1,0 2,0 3,0 4,0 5,0
Voice conversion

- Examples

<table>
<thead>
<tr>
<th>SRC</th>
<th>TGT</th>
<th>GMM</th>
<th>ML+GV</th>
<th>BLFW+AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
<tr>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
<tr>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
<tr>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
<td>🎤</td>
</tr>
</tbody>
</table>
Voice conversion

• Related publications
Outline

• Voice conversion

• Speaker adaptation

• Other research lines

• Forthcoming events
Speaker adaptation

- Speaker-adaptive synthesis
Speaker adaptation

- Speaker-adaptive synthesis
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  
  BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  - BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- State of the art: CSMAPLR, 100 training utts + txt + ph
  BLFW+AS, 1 utt + txt (1 class)
Speaker adaptation

- Results

Higher means better!

Similarity to target BEFORE adaptation

Similarity to target AFTER adaptation

Quality AFTER adaptation
Speaker adaptation

- Examples

TGT 💿 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧

ADAP 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧 🎧
Speaker adaptation

- Related publications
Outline

• Voice conversion

• Speaker adaptation

• Other research lines

• Forthcoming events
Other research lines

• Detection of converted/synthetic speech

  Discover traces of the vocoder (phase)
Other research lines

• Detection of converted/synthetic speech
  Discover traces of the vocoder (phase)

• Data hiding in speech signals
  By manipulating signal phase
Other research lines

• Detection of converted/synthetic speech
  Discover traces of the vocoder (phase)

• Data hiding in speech signals
  By manipulating signal phase

• Speaker diarization in meetings
  Fusion of classifiers
Other research lines

• Detection of converted/synthetic speech
  Discover traces of the vocoder (phase)

• Data hiding in speech signals
  By manipulating signal phase

• Speaker diarization in meetings
  Fusion of clasifiers

Questions about any of these works?
Come and ask me (or Inma) later!
Outline

• Voice conversion

• Speaker adaptation

• Other research lines

• Forthcoming events
• Renowned experts propose research projects
• Students and early-stage researchers apply for participation
• Teams are built
• Work together for 4 weeks in Bilbao
- Multimodal signal analysis and synthesis
- Intuitive interfaces and personalized systems in real and virtual environments
- Assistive technologies for education and social inclusion
- Assistive and rehabilitation technologies
- Search in multimedia and multilingual documents
- Affective and social signal processing
- Multimodality for biometrics and security
- Innovative musical interfaces
- Augmented reality

- Embodied agents
- Human-robot and human-environment interactions in smart environments
- Multimodal conversational systems
- Self-learning and adapting systems
- Innovative modalities and modalities conversion
- Applications of Multimodal interfaces
- Performing arts applications
- Teleoperation and telerobotics
- ...
• November 30th, 2013  Notification of interest
• December 15th, 2013  Full project proposal
• January 10th, 2013  Notification of acceptance to project leaders
                      Start call for Participation
• February 28th, 2014  End call for participation
                      Team building
• March 28th, 2014  Notification of acceptance to participants
• June 9th - July 5th, 2014  eNTERFACE'14 Workshop
Looking forward to receiving your project proposals!!

http://aholab.ehu.es/eNERFACE14