Voice Conversion Fundamentals for Speaker De-identification Applications
Carmen García Mateo
Paula López Otero
Carmen Magariños
Laura Docío Fernández
Eduardo Rodríguez Banga
Outline

• Speaker De-identification
• Voice conversion overview
• Some examples of speaker de-identification systems
De-identification Definition and Applications

It’s Michael Laudrup!
and he says “blah blah blah”
and he sounds kind of sad
De-identification Definition and Applications

Someone said “blah blah blah” and sounds kind of sad, but I don’t know who he/she is…
De-identification Definition and Applications

Message and some other information remains, but identifying information disappears

- Privacy-preserving voice-driven transactions
  - Telephone banking
  - Call centers
- Privacy protection in sensitive information
  - Medical records
De-identification Definitions

• De-identification can be defined as a process by which a data custodian alters or removes identifying information, making it harder for users of the data to determine the identities of the data subjects.

• Speaker de-identification approaches must accomplish:
  1. Quality, since natural-sounding speech is highly convenient;
  2. Universality, allowing the de-identification of any speaker;
  3. Reversibility, making it possible for a trusted holder to recover the original speaker identity re-identification.
Voice Conversion (VC) Definitions and Framework

• Transformation of the voice characteristics of a (source) speaker into those of another (target) speaker while preserving:
  • Linguistic content or
  • Emotional state, or
  • Health condition ...

• Some popular applications:
  • Text-to-speech (TTS) synthesis → creation of new and personalized voices
  • Voice restoration
  • Security or privacy related usage: hiding the identity of the speaker

Where can we find ‘speaker identity’ in the speech signal?

- Linguistic versus non-linguistic information
- Non-linguistic information is more clearly linked to speaker individuality
  - Sociological factors (social class, region of birth/residence, age) -> affects speaking style, which acoustically realized predominantly in prosodic features (pitch contour, duration of words, rhythm, etc).
  - Physiological attributes of the speaker (e.g. anatomy of the vocal tract) affect the spectral content and determine the individual identity
- Perceptually, the most important acoustic features characterizing speaker individuality include the third and the fourth formant, the fundamental frequency and the closing phase of the glottal wave
- Voice conversion systems usually deal with the conversion of spectral features and a simple statistical mean and variance scaling of F0

Performance Evaluation of VC (De-ID Speech)

- “Listening tests” → Perceptual evaluation of quality / naturalness / similarity to a certain speaker by human listeners.

- Spectral Distance

- Speaker (dis)similarity can be measured by using a “Speaker verification system”

  objective measure of speaker de-identification performance

Classification of VC systems

• Based on speaker adaptation in HMM-based speech synthesis (HTS)

• Stand-alone voice conversion

In the training phase, HSMMs are generated using speech data from multiple speakers.

Model adaptation is applied to obtain HSMMs for a given target speaker.

The adapted HSMMs can be used in TTS synthesis for producing speech with the target voice.

The training phase generates conversion models based on training data.

In the conversion phase, the trained models are used for converting unseen utterances of source speech.

In most systems, parallel corpora are required.

Speech representation (Feature extraction)

- The most popular speech representations are based on the source-filter model.
  - **Filter (Vocal tract representation):** formants, line spectral frequencies (LSF), Mel-frequency cepstrum coefficients (MFCCs), Mel-cepstral coefficients (MCCs), combination of several (cepstral and linear prediction representations), etc.

- **Source:** Mixed mode excitation, sinusoidal modeling, etc.

![Diagram of the source-filter model](image)
<table>
<thead>
<tr>
<th>Feature</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSFs</td>
<td>Offer stability, good interpolation properties, and close relationship to formants. Model spectral peaks.</td>
</tr>
<tr>
<td>MFCCs</td>
<td>Model both spectral peaks and valleys. Reliable for measuring acoustic distances and thus useful especially for alignment.</td>
</tr>
<tr>
<td>MCCs</td>
<td>Perhaps the most widely used features for representing spectra both in stand-alone conversion and in HMM based synthesis. Benefits e.g. in alignment very similar to those of MFCCs.</td>
</tr>
<tr>
<td>Formants</td>
<td>Formant bandwidths, locations and intensities would be highly useful features in VC but reliable estimation is extremely challenging.</td>
</tr>
<tr>
<td>Spectral samples</td>
<td>Spectral domain samples can also be used as VC features. Typically used in warping based conversion.</td>
</tr>
<tr>
<td>$F_0$</td>
<td>$F_0$ or log $F_0$ are typically mean-shifted and scaled to the values of the target speaker.</td>
</tr>
<tr>
<td>Voicing</td>
<td>At least binary voicing or aperiodicity information is typically used. More refined voicing information may also be employed.</td>
</tr>
<tr>
<td>Excitation spectra</td>
<td>Sometimes details of the excitation spectra need to be modeled as well, for example when using sinusoidal modeling.</td>
</tr>
</tbody>
</table>
Speech Generation (re-synthesis)

• Vocoder are quite popular for generating the synthesized speech waveform
  • STRAIGHT vocoder (Kawahara et al., 1999) is a widely used analysis/synthesis framework. It decomposes speech into a spectral envelope without periodic interferences, F0, and relative voice aperiodicity
  • Harmonics-plus-noise (HNM) vocoder (Erro et al.)
VC approaches

- Codebook-based mapping
- Gaussian Mixture Models (GMM)
- Frequency Warping + Amplitude Scaling (FW+AS) transformations
- Hidden Markov Models
- Deep Neural Networks

Codebook-based mapping

• Training of a codebook of combined feature vectors. Then, during conversion, the source side of the vectors could be used for finding the closest codebook entry, and the target side of the selected entry could be used as the converted vector

• **Pros:** very simple and straightforward approach that can capture the speaker identity quite well

• **Cons:** the results suffers from frame-to-frame discontinuities and poor prediction capability on new data

Gaussian Mixture Models (GMM) based conversion

• The data is modeled using a GMM and converted by a function that is a weighted sum of local regression functions

• A GMM can be trained to model the density of source features only or the joint density of both source and target features

• **Pros:** very popular -> trade-off between quality and efficiency

• **Cons:** overfitting and oversmoothing

\[ F(x) = \sum_{i=1}^{n} p_i(x) \\left[ \alpha_i + \Sigma_i^{-1}(x - \mu_i) \right] \]

Frequency warping + Amplitude Scaling (FW+AS)

- A warping function is established between the source and target spectra

**Pros:**
- it can obtain very high speech quality
- Closer approximation to the target speaker
- Compared to GMM approach, avoids oversmoothing, formant broadening and effects of the conversion in the analysis/synthesis system

**Cons:**
- limitations regarding the success of identity conversion, due to problems in preserving the shape of modified spectral peaks and controlling the bandwidths of close formants
- **It requires a parallel corpus between the source and target speakers ➔ not universal**

"Voice conversion". Book Chapter, Nurminen et al. 2012
### GMM versus FW+AS: some examples

<table>
<thead>
<tr>
<th>Female-Male</th>
<th>Source</th>
<th>Target</th>
<th>Converted (de-identified)</th>
<th>Reconverted (re-identified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMMs</td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
</tr>
<tr>
<td>FW+AS</td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male-Female</th>
<th>Source</th>
<th>Target</th>
<th>Converted (de-identified)</th>
<th>Reconverted (re-identified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMMs</td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
</tr>
<tr>
<td>FW+AS</td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
<td><img src="image" alt="Voice" /></td>
</tr>
</tbody>
</table>

Transformations trained using 20 sentences and a simple DTW for alignment
### GMM versus FW+AS

**Objective evaluation**

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Traditional GMM</th>
<th>FW + AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Converted-Source</td>
<td>4.88 +/- 0.05</td>
<td>4.86 +/- 0.13</td>
</tr>
<tr>
<td>Reconverted-Source</td>
<td>3.75 +/- 0.06</td>
<td>0.62 +/- 0.02</td>
</tr>
<tr>
<td>Source-Target</td>
<td>6.32 +/- 0.05</td>
<td></td>
</tr>
</tbody>
</table>

**Equations:**

1. \[ MCD[dB] = \frac{1}{M} \sum_{n=1}^{M} MCD_m \]

2. \[ MCD_m = \frac{10}{\log_{10}} \sum_{d=1}^{24} (c_{m,d} - c_{m,d}^{conv})^2 \]

Where \( c_{m,d} \) and \( c_{m,d}^{conv} \) are the \( d^{th} \) dimension feature of the \( m^{th} \) frame of original target and converted MCEP vector, respectively; and \( M \) is the number of analysis frames.
GMM versus FW+AS

• GMM:
  - Greater similarity to the target speaker -> Not relevant in de-identification (we don’t want to mimic the target speaker)
  - Reconverted speech less similar to the source speaker -> Poor re-identification accuracy

• FW+AS:
  - Less quality degradation -> Higher quality in both de-identified and re-identified speech
  - Easily reversible -> Desirable for re-identification
  - Reconverted speech more similar to the source speaker -> Higher re-identification accuracy
  - Parameters trained for a specific pair of speakers can be applied to a third speaker (without risk of artefacts) -> Desirable in de-identification
Frequency warping + amplitude scaling (FW+AS): achieving universality

1. Acoustic analysis to extract speech parameters
2. Transformation of the spectral parameters using manually-defined or pre-trained conversion functions

\[ y = A \times x + b \]

where \( x \) is a Mel-cepstral vector, \( A \) denotes a FW matrix, \( b \) represents an AS vector, and \( y \) is the transformed version of \( x \).

3. F0 scaling for better de-identification

\[ \log \hat{f}_0^t = \frac{\sigma_t}{\sigma_s} (\log f_0^s - \mu_s) + \mu_t \]

4. Speech resynthesis from the transformed parameters

C. Magariños et. Al., "Piecewise linear definition of transformation functions for speaker de-identification". SPLINE 2016
Piecewise linear definition of FW+AS functions

Frequency warping

Three variables: \( f_a, f_b, k \)
\[ \beta = k\alpha \text{ (} 0 < k < 1 \text{)} \]

We can obtain different warping functions by modifying \( \alpha \) (\( |\alpha| < \pi/4 \))

\[
W(f) = \begin{cases} 
\frac{a_2}{a_1} \cdot f, & f < a_1 \\
\frac{b_2 - a_2}{b_1 - a_1} \cdot (f - a_1), & a_1 < f < b_1 \\
\frac{f_{nyq} - b_2}{f_{nyq} - b_1} \cdot (f - b_1), & f > b_1
\end{cases}
\]

\[ A = (a_1, a_2) = \left( f_a \cos \left( \frac{\pi}{4} + \alpha \right), f_a \sin \left( \frac{\pi}{4} + \alpha \right) \right) \]
\[ B = (b_1, b_2) = \left( f_b \cos \left( \frac{\pi}{4} + \beta \right), f_b \sin \left( \frac{\pi}{4} + \beta \right) \right) \]

\( f_a = 700 \text{ Hz} \)
\( f_b = 3000 \text{ Hz} \)
\( |\alpha| = \pi/12, \pi/24, \pi/30, \pi/36 \)

C. Magariños et. Al., "Piecewise linear definition of transformation functions for speaker de-identification“. SPLINE 2016
Piecewise linear definition of FW+AS functions

F0 and amplitude scaling

F0 scaling proportional to $\alpha$ by using a linear function

<table>
<thead>
<tr>
<th></th>
<th>Female to male</th>
<th>Male to female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle $\alpha$</td>
<td>F0 scaling</td>
<td>Angle $\alpha$</td>
</tr>
<tr>
<td>Trans1</td>
<td>$-\pi/12$</td>
<td>0.5000</td>
</tr>
<tr>
<td>Trans2</td>
<td>$-\pi/24$</td>
<td>0.6667</td>
</tr>
<tr>
<td>Trans3</td>
<td>$-\pi/30$</td>
<td>0.7143</td>
</tr>
<tr>
<td>Trans4</td>
<td>$-\pi/36$</td>
<td>0.7500</td>
</tr>
</tbody>
</table>

Amplitude scaling using weighted Hanning-like bands equally spaced in the Mel scale

- 15 bands
- Weights randomly selected from [-10, -5, 0, 5, 10]

C. Magariños et. Al., "Piecewise linear definition of transformation functions for speaker de-identification". SPLINE 2016
Objective evaluation by using a speaker identification system

• Characteristics of the speaker identification system:
  • Feature extraction: 19 MFCC+E+Δ+ Δ^2
  • i-vector representation (UBM with 1024 mixtures, i-vectors of dimension 100)
  • Dot-scoring

• Performance measure: de-identification accuracy

\[
de - id\ accuracy = \frac{\#\text{utt not assigned to source speaker}}{\#\text{utt}}
\]

• Database: Albayzín corpus in Spanish language
  • 40 test speakers, 10 utterances each (1 for enrollment, 9 for testing)

C. Magariños et. Al., "Piecewise linear definition of transformation functions for speaker de-identification". SPLINE 2016
## Results

<table>
<thead>
<tr>
<th>Transformation</th>
<th>FW</th>
<th>FW+F0</th>
<th>FW+F0+AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans1</td>
<td>82.5%</td>
<td>98.6%</td>
<td>96.9%</td>
</tr>
<tr>
<td>Trans2</td>
<td>53.9%</td>
<td>87.2%</td>
<td>88.1%</td>
</tr>
<tr>
<td>Trans3</td>
<td>30.6%</td>
<td>64.2%</td>
<td>68.3%</td>
</tr>
<tr>
<td>Trans4</td>
<td>4.4%</td>
<td>28.0%</td>
<td>36.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Original</th>
<th>FW</th>
<th>FW+F0+AS</th>
<th>FW+F0+AS</th>
<th>FW+F0+AS</th>
<th>FW+F0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trans1</td>
<td>Trans1</td>
<td>Trans2</td>
<td>Trans3</td>
<td>Trans4</td>
<td>Trans2</td>
</tr>
<tr>
<td>Female</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
</tr>
<tr>
<td>Male</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
<td>[Icon]</td>
</tr>
</tbody>
</table>

C. Magariños et. Al., "Piecewise linear definition of transformation functions for speaker de-identification". SPLINE 2016
Speaker De-identification Using Pre-Trained Transformation Functions

• The universality issue is achieved since no parallel corpus between the input and target speakers is needed; therefore, it does not require training the voice conversion parameters each time a new input speaker is proposed.

• First, a pool of voice conversion transformations is trained using a set of speakers with parallel corpora.

• Any time a new user has to be de-identified, the most similar speaker in the pool of speakers is found, and the pre-trained transformation between this speaker and its most dissimilar speaker in the corpus is applied to the new speaker.
Speaker De-identification Using Pre-Trained Transformation Functions

Speaker De-identification Using Pre-Trained Transformation Functions

• Source speaker: selected as the most similar to the input speaker
  • Given that the transformation function is not trained for the input speaker, we ensure that it was trained for a similar speaker

• Target speaker: selected as the most dissimilar to the source or input speaker
  • We want the de-identified speech to be as different to the input speaker as possible

\[ S_{\text{source}} = \{S_{\text{source}_1}, \ldots, S_{\text{source}_{ns}}\} \]
\[ S_{\text{target}} = \{S_{\text{target}_1}, \ldots, S_{\text{target}_{nt}}\} \]

"Reversible Speaker De-identification Using Pre-Trained Transformation Functions", Magariños et al. Submitted to CSL. 2016
Speaker De-identification Using Pre-Trained Transformation Functions

i-vector are extracted from the speech and compared using dot scoring or PLDA scoring

\[ S^*_{\text{source}} = \arg \max_{i=1,...,n_s} \text{score}(w_{\text{input}}, w_{\text{source}_i}) \]

\[ S^*_{\text{target}} = \arg \min_{i=1,...,n_t} \text{score}(w^*_{\text{source}}, w_{\text{target}_i}) \]

Method 1

\[ S^*_{\text{target}} = \arg \min_{i=1,...,n_t} \text{score}(w_{\text{input}}, w_{\text{target}_i}) \]

Method 2

Speaker De-identification Using Pre-Trained Transformation Functions: Objective Evaluation

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification accuracy on original speech</td>
<td>99.2%</td>
</tr>
<tr>
<td>De-identification accuracy, method 1</td>
<td>89.4%</td>
</tr>
<tr>
<td>De-identification accuracy, method 2</td>
<td>90.3%</td>
</tr>
<tr>
<td>Re-identification accuracy</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

Speaker De-identification Using Pre-Trained Transformation Functions: Objective Evaluation

Speaker De-identification Using Pre-Trained Transformation Functions: Confusion Matrices

Speaker De-identification Using Pre-Trained Transformation Functions: Subjective Evaluation

Speaker De-identification Using Pre-Trained Transformation Functions: Some examples

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>De-identified</th>
<th>Re-identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td><img src="image1" alt="Speaker" /></td>
<td><img src="image2" alt="Speaker" /></td>
<td><img src="image3" alt="Speaker" /></td>
</tr>
<tr>
<td>F2</td>
<td><img src="image4" alt="Speaker" /></td>
<td><img src="image5" alt="Speaker" /></td>
<td><img src="image6" alt="Speaker" /></td>
</tr>
<tr>
<td>M1</td>
<td><img src="image7" alt="Speaker" /></td>
<td><img src="image8" alt="Speaker" /></td>
<td><img src="image9" alt="Speaker" /></td>
</tr>
<tr>
<td>M2</td>
<td><img src="image10" alt="Speaker" /></td>
<td><img src="image11" alt="Speaker" /></td>
<td><img src="image12" alt="Speaker" /></td>
</tr>
</tbody>
</table>

A Step further:
Universality, Naturalness and Reversibility in a De-Id approach

• Piecewise-defined and pre-trained transformation functions are universal → they can be applied to any speaker

• The use of FW+AS leads to quite natural sounding speech with slight quality degradation

• FW+AS are easily invertible, making it possible to recover the original speech with little distortion

References


Thank you!