Statistical Voice Conversion
Technologies for Alaryngeal Speech Enhancement

Tomoki Toda

Nara Institute of Science and Technology
Oct. 12, 2010
Vocal Disorder Due to Laryngectomy

- Laryngectomee
  - Removal of Larynx
  - Around 20,000 people in Japan
  - Separated trachea from vocal tract

Unable to produce speech in spite of holding ability of articulation

Unable to generate excitation without vibration of vocal folds

Deterioration of QOL (Quality of Life) caused by lack of speech production!
Purpose

- Development of techniques for enhancing speech production in laryngectomees

**Traditional techniques**

Alternative speaking methods for producing alaryngeal speech

- Compared with normal speech...
  - lower naturalness
  - lower intelligibility
  - lack of speaker individuality

**Techniques to be developed**

Enhancement to make alaryngeal speech sound more natural

- Compared with original alaryngeal speech...
  - improved naturalness
  - improved intelligibility
  - desired speaker individuality
Approach

- Applying statistical voice conversion to alaryngeal speech enhancement for achieving much better quality

**Traditional**

- Alaryngeal speech
  - Feature extraction
  - Signal processing
  - Waveform synthesis
  - Enhanced speech

**Proposed**

- Alaryngeal speech
  - Feature extraction
  - Statistical processing
  - Waveform synthesis
  - Enhanced speech

- Extracting features effective for conversion
- Training of conversion functions
- Exhibiting similar acoustic properties to normal speech

- Training data
  - Alaryngeal speech and normal speech
Approach

- Applying statistical voice conversion to alaryngeal speech enhancement for achieving much better quality

**Traditional**

1. Alaryngeal speech
2. Feature extraction
3. Signal processing
4. Waveform synthesis
5. Enhanced speech

**Proposed**

1. Alaryngeal speech → Feature extraction
2. Statistical processing
3. Waveform synthesis
   - Training of conversion functions
   - Exhbiting similar acoustic properties to normal speech
4. Enhanced speech
5. Extraction of features effective for conversion
6. Training data
   - Alaryngeal speech and normal speech
Types of Alaryngeal Speech

<table>
<thead>
<tr>
<th>Devices</th>
<th>Unnecessary</th>
<th>Necessary</th>
<th>Necessary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning difficulty</td>
<td>Very hard</td>
<td>Easy</td>
<td>Relatively easy</td>
</tr>
<tr>
<td>Conveyed sounds</td>
<td>Alaryngeal speech</td>
<td>Alaryngeal speech and excitation signals leaked from electrolarynx</td>
<td>Alaryngeal speech from loud speaker</td>
</tr>
</tbody>
</table>

- **Esophageal speech (ES)**
  - Pseudo glottis
  - [Hashiba et al., 2001]

- **Electrolaryngeal speech (EL)**
  - Electrolarynx
  - [Nakamura et al., 2007]

- **Body-conducted silent EL (Silent EL)**
  - Loud speaker
  - Silent sound source unit
  - Non-audible murmur (NAM) microphone
  - [Nakajima et al., 2004]
Acoustic Properties of ES

- Similar to real voice
  - Slightly unnatural but relatively intelligible
  - Poor speaker individuality
  - Including pitch information

Waveform
- Power unstably varying according to phonemes

\( F_0 \)
- Hard to extract \( F_0 \) corresponding to pitch information

Aperiodic factors
- Constantly high

Spectral envelop Useful
- Unstably varying according to phonemes
- Including specific noise sounds
- Including pitch information
Acoustic Properties of EL

- Mechanical sounds
  - Unnatural but quite intelligible
  - Poor speaker individuality
  - Artificial pitch information depending on $F_0$ of external excitation

### Waveform
- Power varying but basically depending on ON/OFF of electrolarynx

### $F_0$
- Almost constant during speaking

### Aperiodic factors
- Varying but basically depending on ON/OFF of electrolarynx

### Spectral envelop
- Varying according to phonemes
- Including leaked excitation signals
Acoustic Properties of Silent EL

- Mechanical and noisy sounds
  - Very unnatural and unintelligible
  - Poor speaker individuality
  - Artificial pitch information depending on $F_0$ of external excitation

**Waveform**
- Power slightly varying but basically depending on ON/OFF of external unit
- Low signal-to-noise ratio

**$F_0$**
- Almost constant during speaking

**Aperiodic factors**
- Slightly varying but basically depending on ON/OFF of external unit

**Spectral envelop** Useful
- Varying according to phonemes
- Attenuation of high frequency components due to body conduction
Approach

- Applying statistical voice conversion to alaryngeal speech enhancement for achieving much better quality

**Traditional**

1. Alaryngeal speech
2. Feature extraction
3. Signal processing
4. Waveform synthesis
5. Enhanced speech

**Proposed**

1. Alaryngeal speech
2. Feature extraction
3. Statistical processing
4. Waveform synthesis
5. Enhanced speech

- Extracting features effective for conversion
- Training of conversion functions
- Alaryngeal speech and normal speech
- Exhibiting similar acoustic properties to normal speech

- Training data
Statistical Voice Conversion (VC) [Abe et al., 1989]

- Technique for converting source speaker’s voices into target speaker’s voices while keeping linguistic contents unchanged.

1. Training with parallel data (around 50 utterance pairs)

   - Source speaker
   - Target speaker

   Conversion model from the source into the target

2. Conversion of any utterance
VC with Gaussian Mixture Model

[Stylianou et al., 1998]

- Training of joint probability density function (p.d.f.) of the source feature \( x_t \) and the target feature \( y_t \) given by GMM

\[
P(x_t, y_t | \lambda) = \sum_{m=1}^{M} \alpha_m N\left( \begin{bmatrix} x_t \\ y_t \end{bmatrix}; \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \right)
\]

- Conversion by maximizing conditional p.d.f. \( P(y_t | x_t, \lambda) \)
  - Significantly improvements yielded by using maximum likelihood estimation (MLE) of parameter trajectory [Toda et al., 2007]
One-to-Many Eigenvoice Conversion

[Toda et al., 2006]

- Technique for converting source speaker’s voices into arbitrary speaker’s voices
  - Using many pre-stored target speakers’ voices as prior information
  - Developing a conversion model flexibly adapted to arbitrary speakers
Eigenvoice-GMM (EV-GMM)

- Training of joint p.d.f. of the source feature $x_t$ and the $S$ pre-stored targets' features $y_t^{(1)}, \cdots, y_t^{(S)}$ given by EV-GMM

$$P(x_t, y_t^{(s)} | \lambda_{EV}, w^{(s)}) = \sum_{m=1}^{M} \alpha_m N \left( \begin{bmatrix} x_t \\ y_t^{(s)} \end{bmatrix}; \begin{bmatrix} \mu_m^{(x)} \\ \sum_m \end{bmatrix} \right) \begin{bmatrix} x_t \\ y_t^{(s)} \end{bmatrix} + B_m w^{(s)} + b_m^{(y)}, \left( \begin{bmatrix} \sum^{(xx)} \\ \sum^{(xy)} \end{bmatrix} \right) \begin{bmatrix} \sum^{(xx)} \\ \sum^{(yy)} \end{bmatrix}$$
Unsupervised Adaptation of EV-GMM

- Adaptation using only speech data with target voice quality
  - Arbitrary sentences available for adaptation
  - Robust adaptation using only one or two sentences

\[
\hat{w}' = \arg \max_{w'} \prod_{t=1}^{T} \int P(x_t, y'_t | \lambda_{EV}, w') \, dx_t
\]

\[
= \arg \max_{w'} \prod_{t=1}^{T} P(y'_t | \lambda_{EV}, w')
\]

Adaptation data: \( y'_1, y'_2, \ldots, y'_T \)

EV-GMM parameter set

Adaptive parameter

Likelihood of adapted EV-GMM at frame \( t \)

Marginal likelihood of adapted GMM

Source features as hidden variables
Approach

• Applying statistical voice conversion to alaryngeal speech enhancement for achieving much better quality.
Speech Features

• **Source feature** (from alaryngeal speech)
  - Mel-cepstral segment
    - Removing unstable fluctuations
    - Compensating collapsed phonemes

• **Target feature** (from normal speech)
  - Mel-cepstrum with power coefficient
  - Aperiodic factors
  - Log-scaled $F_0$ (including voiced/unvoiced information)

• **Training of three GMMs with parallel data of alaryngeal speech and normal speech**

Normal speech features:
- Mel-cepstrum
- Aperiodic factors
- Log-scaled $F_0$

Alaryngeal speech feature:
- Mel-cepstral segment
- Aperiodic factors
- Log-scaled $F_0$
Alaryngeal Speech-to-Speech with VC

Alaryngeal speech

- Spectral analysis
- Mel-cepstrum sequence
- Segment feature extraction
- Mel-cepstral segment sequence

Enhanced speech

- Filtering
- Excitation
- Excitation generation

GMM for mel-cepstrum estimation

- MLE of trajectory
- Mel-cepstrum sequence

GMM for aperiodic factor estimation

- MLE of trajectory
- Aperiodic factor sequence

GMM for log-scaled $F_0$ estimation

- MLE of trajectory
- Log-scaled $F_0$ / UV symbol sequence
Voice Quality Control with EVC

- Alaryngeal speech-to-speech based on traditional VC
  - Enhanced voice quality determined by the target speaker in training
  - Needing around 50 utterance-pairs to develop a conversion model

Further implementing one-to-many EVC
- Adapting model to desired voice quality
- One arbitrary utterance available for adaptation
- Manual control

Parallel data → Training → GMM

Multiple parallel data sets → Training → EV-GMM → Adaptation → Adapted EV-GMM

Target speaker’s voice
ES-to-Speech

[Doi et al., 2009–2010]

- $F_0$ estimation process
  - Estimating $F_0$ corresponding to pitch information of ES
  - As target data in training, using normal speech uttered by a non-disabled speaker mimicking pitch of ES as accurately as possible
  - Statistical $F_0$ estimation from spectrum of ES

Target voice quality:

Waveform

$F_0$

Aperiodic factors

Spectral envelope

ES-to-Speech with EVC
EL-to-Speech

- $F_0$ estimation process
  - Estimating $F_0$ of normal speech
  - Avoiding difficulties of $F_0$ estimation by converting to whisper
  - Still effective for converting to normal speech
  - Slight improvements yielded by using $F_0$ control of external excitation signals with air pressure [Uemi et al., 1995]

[Doi et al., 2010]

[Nakamura et al., 2009–2010]
Silent EL-to-Speech

- **$F_0$ estimation process**
  - Similar to EL-to-Speech, *i.e.*, converted to normal speech or whisper

- **Design of external excitation signals**
  - Accepting various power ranges over power of silence parts
  - Accepting various spectral envelops

Silent EL-to-Speech with EVC

![Silent EL-to-Speech Waveform](image1)

**Waveform**

- **$F_0$**
- **Aperiodic factors**
- **Spectral envelope**
Effectiveness of Proposed Methods

- **Naturalness** and **intelligibility**
  - Enhanced EL
  - Enhanced ES
  - Silent EL-to-Speech
  - ES-to-Speech
  - EL-to-Speech

- **Speaker individuality**
  - Controlled using only one arbitrary utterance of target voice
Conclusions

• Proposing alaryngeal speech enhancement based on statistical voice conversion techniques

1. Esophageal speech (ES)
   - Very good: no device
   - Bad: hard to learn
   - Effective for telecommunication

2. Electrolaryngeal speech (EL)
   - Bad: noisy excitation
   - Very good: easy to learn
   - Effective for telecommunication

3. Body-conducted silent EL
   - Very good: silent excitation
   - Good: relatively easy to learn
   - Effective for face-to-face communication

   NEW
   - Very good: natural and intelligible
   - Good: controllable voice quality

   NEW
   - Very good: natural and intelligible
   - Good: controllable voice quality

   NEW
   - Good: Relatively natural and intelligible
   - Good: controllable voice quality
Future Work

• Improving controllability of voice quality
• Development of a real-time conversion system
• Implementing online adaptation process
• Improving silent sound source unit and NAM microphone
• Evaluation under real environments
Acknowledgement:

• **Keigo Nakamura** (JSPS postdoc, NAIST)
• **Hironori Doi** (Ph.D. student, NAIST)