On Human Capability and Acoustic Cues for Discriminating Singing and Speaking Voices

Yasunori Ohishi \(^1\)  
Katunobu Itou \(^2\)  
Masataka Goto \(^3\)  
Kazuya Takeda \(^1\)

\(^1\) Graduate School of Information Science, Nagoya University, Japan  
\(^2\) Faculty of Computer and Information Sciences, Hosei University, Japan  
\(^3\) National Institute of Advanced Industrial Science and Technology
Let’s do the Quiz

Can you discriminate between **Singing** and **Speaking** voices?
(Japanese voices)

<table>
<thead>
<tr>
<th>Question</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q.1. Can you do it?</td>
<td>(2 s long)</td>
</tr>
<tr>
<td>Q.2. Can you do it?</td>
<td>(500 ms long)</td>
</tr>
<tr>
<td>Q.3. Can you do it?</td>
<td>(200 ms long)</td>
</tr>
</tbody>
</table>
Investigation of signal length necessary for discrimination

<table>
<thead>
<tr>
<th>Signal length [ms]</th>
<th>Correct rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-s voice signals</td>
<td></td>
</tr>
<tr>
<td>500-ms voice signals</td>
<td></td>
</tr>
<tr>
<td>200-ms voice signals</td>
<td></td>
</tr>
</tbody>
</table>

- **Singing performance**
- **Speaking performance**
- **Total performance**
Investigation of signal length necessary for discrimination

![Graph showing correct rates for different signal lengths.]

- 1-s voice signals
- 500-ms voice signals
- 200-ms voice signals

Not only **temporal characteristics** but also such **short-term features** carry discriminative cues.
The goal of this study

Subjective experiments
Investigation of acoustic cues necessary for discrimination between singing and speaking voices

Based on knowledge obtained by subjective experiments

Automatic vocal style discriminator
• Spectral feature measure
• F0 derivative measure
Introduction of the voice database

- AIST humming database
  - 75 Japanese subjects (37 males, 38 females)
  - **Sing a chorus and verse A sections**
    at an arbitrary tempo, without musical accompaniment
    (25 Japanese songs selected from “RWC Music Database: Popular Music”)

- **Read the lyrics of chorus and verse A sections**
  Most of these subjects haven’t had the special musical training
The goal of this study

Subjective experiments
Investigation of acoustic cues necessary for discrimination between singing and speaking voices

Based on knowledge obtained by subjective experiments

Automatic vocal style discriminator
• Short-term spectral feature measure
• F0 derivative measure
To compare the importance of temporal and spectral cues for discrimination, voice quality and prosody are modified by using signal processing techniques.

Investigation of acoustic cues necessary for discrimination

Temporal structure of signal is modified, short-time spectral features are maintained

Random splicing technique

Randomly concatenating pieces

Let’s do the quiz

Q.1 (250 ms)  Q.2 (200 ms)  Q.3 (125 ms)
Investigation of acoustic cues necessary for discrimination

Temporal structure of signal is maintained, short-time spectral features are modified

Low-pass filtering technique

Eliminating frequency component higher than 800 Hz

Let’s do the quiz

Q.1  Q.2  Q.3
Investigation of acoustic cues necessary for discrimination

Singing voice correct rate [%]

- Original voice: 99.3%
- Low-pass Filtering: 86.9%
- Random Splicing (250 ms): 84.3%
- Random Splicing (200 ms): 76.9%
- Random Splicing (125 ms): 70.6%

Speaking voice correct rate [%]

- Original voice: 100%
- Low-pass Filtering: 98.9%
- Random Splicing (250 ms): 94.9%
- Random Splicing (200 ms): 90.0%
- Random Splicing (125 ms): 95.0%

Stimuli
Discussion

- Correct rate of singing voices declined

Random splicing technique

- Temporal structure of the original voices (rhythm and melody pattern) has been modified

before

after

- Prolonged vowels of singing voices has been divided into small pieces

before ch i r i b a

after a ch i a i b i r i

Low-pass filtering technique

- Frequency components higher than 800 Hz have been eliminated

Important acoustic cues for discrimination ??
The goal of this study

Subjective experiments
- Short-term spectral feature
- Temporal structure

Based on knowledge obtained by subjective experiments

Automatic vocal style discriminator
- Spectral feature measure
- F0 derivative measure
Automatic discrimination measure

- **Spectral feature measure**
  - Difference in spectral envelopes and vowel durations
  - Mel-Frequency Cepstrum Coefficients (MFCC)
  - $\Delta$MFCC (5-frame regression)

- **F0 derivative measure**
  - Difference in dynamics of prosody
  - $\Delta$F0 (5-frame regression)
  - F0 Extraction ($PreFEst$, Goto1999)

![Graphs showing spectral envelopes and F0 variations for singing and speaking voices](image-url)
Training the discriminative model

- Gaussian mixture models (16-mixture GMM)
  - e.g. Discrimination using $\Delta F0$

**F0 extraction and $\Delta F0$ calculation**

- for $\Delta F0$ of each frame

**Likelihood comparison**

- Input signal

**GMMs**

- Singing voice GMM
  - Relative Frequency
  - $\Delta F0$ [cent/10ms]

- Speaking voice GMM
  - Relative Frequency
  - $\Delta F0$ [cent/10ms]
Automatic discrimination results

<table>
<thead>
<tr>
<th>Input signal length [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct rate [%]</td>
</tr>
<tr>
<td>Total performance of MFCC+ΔMFCC+ΔF0</td>
</tr>
<tr>
<td>Total performance of ΔF0</td>
</tr>
<tr>
<td>Total performance of MFCC+ΔMFCC</td>
</tr>
</tbody>
</table>

87.6% Human performance
Summary and future work

- Investigation of signal length necessary
  - Not only *temporal characteristics* but also *short-time spectral feature* can be a cue for the discrimination

- Investigation of acoustic cues necessary
  - *The relative importance of the temporal structure* is found for singing and speaking voice discrimination

- Automatic vocal style discriminator
  - Feature vector (MFCC+ΔMFCC+ΔF0)
  - For 2-s signals, the correct rate is 87.6%

- Plan to propose *new measures* to improve the automatic discrimination performance