Introduction to real-time speech processing

DPlug meeting

23/08/2022
Principles of speech articulation
Anatomy of speech organs

adapted from MadBeppo.com, 2021

from National Cancer Institute
Phonation

Voiced

- modal
- breathy
- creaky

'vain', 'zen', 'game'

Voiceless

'fame', 'sane', 'came'
Places of articulation

Labials
- 'pie', 'buy', 'my'
- 'fee', 'vie'

Coronals
- 'thigh', 'thy'
- 'tie', 'die', 'nigh', 'sigh', 'zeal', 'lie'
- 'shy', 'she'

Dorsals
- 'hack', 'hang', 'hag'
Nasality

Nasal  ‘ram’, ‘ran’, ‘rang’

Nasalized  ‘on’, ‘in’ in French
<table>
<thead>
<tr>
<th>Manners of articulation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosives</td>
<td>'pie', 'buy'</td>
</tr>
<tr>
<td></td>
<td>'tie', 'die'</td>
</tr>
<tr>
<td></td>
<td>'kye', 'guy'</td>
</tr>
<tr>
<td>Fricatives</td>
<td>'fie', 'vie'</td>
</tr>
<tr>
<td></td>
<td>'thigh', 'thy'</td>
</tr>
<tr>
<td></td>
<td>'sigh', 'zeal'</td>
</tr>
<tr>
<td></td>
<td>'shy'</td>
</tr>
<tr>
<td>Approximants</td>
<td>'lie'</td>
</tr>
<tr>
<td></td>
<td>'we'</td>
</tr>
</tbody>
</table>
Vowels

MRI of 4 New Zealand English vowels  
*from Watson et al., 2009*
Speech acoustics
Example spectrogram of speech

they don't understand that
Phonation - F0

they don't understand a sad cat

Voiceless
Phonation - F0

they dono't understand a ndth a t

Voiced
Phonation - F0

don't understand that
Mel cepstrum

don't understand that
Mel cepstrum

they do not understand that

25 Mel cepstral coefficients
adapted from Bocquelet et al., 2016c
Formants F1-F2

they don't understand that
Formants F1-F2

they don't understand that
Formants F1-F2

F1 F2 diagram of some IPA vowels (Hitch, 2017)
Speech synthesis
Source-filter model of speech production

\[ x(t) = s(t) \ast h(t) \]
Analysis

\[ s(t) \]

\[ \begin{align*}
F0 \\
\text{Mel Cepstrum} \\
\text{Formants}
\end{align*} \]

\[ h(t) \]
Synthesis

Discrete

Continuous

Acoustic features

Speech units

Concatenation / parametric synthesis

Vocoder (MLSA, formants)
Synthesis

Impulse train generator (voiced)

White noise generator (voiced)

Voiced/unvoiced switch

Mel cepstrum

Synthesis filter

Speech

MLSA filter

F0

Impulse train generator

Synthesis filter

Speech (vowels)

Formants (F1,F2)

Klatt synthesizer
(simplified for vowels)
Thank you for your attention
Appendix
Cepstral analysis

\[ C(\tau) \triangleq \left| \mathcal{F}^{-1} \left\{ \log \left( |\mathcal{F}\{x(t)\}|^2 \right) \right\} \right|^2 \]
Source-filter separation with the power cepstrum

\[ C(\tau) = \left| \mathcal{F}^{-1} \left\{ \log \left( |\mathcal{F} \{ s(t) \ast h(t) \}|^2 \right) \right\} \right|^2 \]

\[ = \left| \mathcal{F}^{-1} \left\{ \log \left( |S(f)|^2 \cdot |H(f)|^2 \right) \right\} \right|^2 \]

\[ = \left| \mathcal{F}^{-1} \left\{ \log \left( |S(f)|^2 \right) + \log \left( |H(f)|^2 \right) \right\} \right|^2 \]

\[ = \left| \mathcal{F}^{-1} \left\{ \log \left( |S(f)|^2 \right) \right\} + \mathcal{F}^{-1} \left\{ \log \left( |H(f)|^2 \right) \right\} \right|^2 \]
Cepstral analysis

\[ C(\tau) \triangleq \left| \mathcal{F}^{-1} \left\{ \log \left( |\mathcal{F} \{ x(t) \}|^2 \right) \right\} \right|^2 \]

\[ C_c(\tau) \triangleq \mathcal{F}^{-1} \left\{ \log \left( \mathcal{F} \{ x(t) \} \right) \right\} \]
Cepstral analysis

Power Cepstrum

Complex Cepstrum

\[ C(\tau) \triangleq \left| \mathcal{F}^{-1} \left\{ \log \left( |\mathcal{F} \{ x(t) \} |^2 \right) \right\} \right|^2 \]

\[ C_c(\tau) \triangleq \mathcal{F}^{-1} \{ \log (\mathcal{F} \{ x(t) \}) \} \]

Keeps phase information for synthesis
Cepstral analysis

Power Cepstrum

Complex Cepstrum

Mel Cepstrum

Complex cepstrum with warping of the spectrum on the mel scale

\[ C(\tau) = \left| \mathcal{F}^{-1} \left\{ \log \left( |\mathcal{F} \{x(t)\}|^2 \right) \right\} \right|^2 \]

\[ C_c(\tau) = \mathcal{F}^{-1} \{ \log (\mathcal{F} \{x(t)\}) \} \]}
Mel scale

Original mel scale.
Subjective relation between pitch and frequency as measured by Stevens et al., 1937