On the Relative Importance of the Short-Time Magnitude and Phase Spectra Towards Speaker Dependent Information

Kamil K. Wójcicki and Kuldip K. Paliwal

Signal Processing Laboratory
Griffith School of Engineering
Griffith University, Nathan, Australia

ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery
Department of Electronic Systems
Aalborg University, Aalborg, Denmark
Presentation Outline

1. Introduction, Background and Motivation
2. Human Listening Experiment
3. Results and Discussion
Introduction

⇒ Speech Signal Processing, Representations, Problem Definition

- Short-time Fourier analysis
- Two representations: short-time magnitude and phase spectra
- Is only the short-time magnitude spectrum useful? In applications such as: automatic speech recognition, speech enhancement, automatic speaker recognition, etc.
- Relative importance of these representations for speech intelligibility [Alsteris and Paliwal(2007), Wojcicki and Paliwal(2007)]
- What about relative importance of the short-time magnitude and phase spectra towards speaker dependent information?
Proposed Approach

⇒ Approach and Study Aims

- Conduct human speaker verification experiment to determine:
  - the relative importance of the short-time magnitude spectrum and the short-time phase spectrum for speaker verification;
  - the effect of the dynamic range of an analysis window on speaker dependent content of the short-time magnitude spectrum and the short-time phase spectrum.
Analysis-Modification-Synthesis (AMS) ⇒ Procedure Stages

1. **Analysis Stage**
   - Framing
   - Windowing
   - Fourier analysis
   \[ S(t, f) = \int_{-\infty}^{\infty} s(\tau)w(t-\tau)e^{-j2\pi f \tau} \, d\tau \]

2. **Modification Stage**
   \[ S(t, f) = |S(t, f)| e^{j\angle S(t, f)} \]
   \[ \hat{S}(t, f) = \ldots \]

3. **Synthesis Stage**
   - Inverse Fourier analysis
   - Windowing
   - Overlap add (OLA)
Analysis-Modification-Synthesis (AMS) ⇒ Modification Stage

- **Magnitude-Only (MO)**
  \[ \hat{S}(t, f) = |S(t, f)|e^{j\phi(t,f)} \]

- **Phase-Only (PO)**
  \[ \hat{S}(t, f) = e^{j\angle S(t,f)} \]

- **Original**
  \[ \hat{S}(t, f) = |S(t, f)|e^{j\angle S(t,f)} = S(t, f) \]
Experimental Setup

Speech Corpus
- Gender-balanced speaker-dependent speech corpus
  (5 sentence types, 12 speakers, 6 repetitions)
- Used 2 repetitions from each speaker of the utterance:
  “three nine zero two six seven”

Speech Stimuli
- AMS settings
  (32 ms frame duration, 4 ms frame shift, 1024 point FFT)
- 5 analysis windows
  (Chebyshev 10 dB, 35 dB, 60 dB; rectangular; Hamming)
- Stimuli types: MO, PO and original
  (132 processed stimuli)
Analysis window functions: (a) Chebyshev 10 dB; (b) Chebyshev 35 dB; (c) Chebyshev 60 dB; (d) rectangular; and (e) Hamming.
**Stimuli Spectrograms**

⇒ *Magnitude-Only (MO) and Phase-Only (PO) Comparison*

Spectrograms of the utterance: “*three nine zero two six seven*” belonging to a female speaker: (a) and (b) originals; (c) MO Chebyshev 10 dB; (d) PO Chebyshev 10 dB; (e) MO Chebyshev 35 dB; (f) PO Chebyshev 35 dB; (g) MO Chebyshev 60 dB; (h) PO Chebyshev 60 dB; (i) MO rectangular; (j) PO rectangular; (k) MO Hamming; and (l) PO Hamming.
Stimuli Spectrograms

⇒ Magnitude-Only (MO) and Phase-Only (PO) Comparison

Spectrograms of the utterance: “three nine zero two six seven” belonging to a male speaker: (a) and (b) originals; (c) MO Chebyshev 10 dB; (d) PO Chebyshev 10 dB; (e) MO Chebyshev 35 dB; (f) PO Chebyshev 35 dB; (g) MO Chebyshev 60 dB; (h) PO Chebyshev 60 dB; (i) MO rectangular; (j) PO rectangular; (k) MO Hamming; and (l) PO Hamming.
Experimental Procedure

→ Human Speaker Verification Tests

- Task: verify if two utterances, played one after another, belong to the same speaker
- 1584 speaker verification tests
- Yes or no decision via keyboard
- No feedback
- Stimuli presented over headphones in random order
- 1 listener (pilot study)
Results

⇒ False Rejection Rate (FRR) – Chebyshev Analysis Window Function

- MO stimuli contain speaker dependent information for the Chebyshev analysis window with the large dynamic range.
- PO Stimuli contain speaker information predominantly for the low dynamic range Chebyshev analysis windows.
Results

⇒ False Rejection Rate (FRR) – Rectangular and Hamming Analysis Window Function

<table>
<thead>
<tr>
<th>Modification</th>
<th>Window</th>
<th>FRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>MO</td>
<td>Rectangular</td>
<td>0.00</td>
</tr>
<tr>
<td>MO</td>
<td>Hamming</td>
<td>0.00</td>
</tr>
<tr>
<td>PO</td>
<td>Rectangular</td>
<td>0.00</td>
</tr>
<tr>
<td>PO</td>
<td>Hamming</td>
<td>33.33</td>
</tr>
</tbody>
</table>

- MO stimuli contain speaker dependent information for the rectangular and the Hamming analysis windows
- PO Stimuli contain speaker information for the rectangular window, while the speaker dependent content is reduced for the Hamming analysis window
Discussion

⇒ Relative Content of $|S(t, f)|$ and $\angle S(t, f)$ Versus Overall Content of $S(t, f)$

- The overall information content of $S(t, f)$ is same.
- If $|S(t, f)|$ contains more information then $\angle S(t, f)$ has to contain less.
- Similarly, if $|S(t, f)|$ contains less information then $\angle S(t, f)$ has to contain more.
- We conjecture that it is the properties of the analysis window function that control how the information is distributed between $|S(t, f)|$ and $\angle S(t, f)$. 
Computation of the short-time phase spectrum via a four quadrant version of the arctangent function:
\[ \angle S(t, f) = \arctan\left(\frac{\text{Im}\{S(t,f)\}}{\text{Re}\{S(t,f)\}}\right). \]

Use of large dynamic range analysis windows means that \( S(t, f) \) will have some very small (close to zero) values.

Such small values produce less reliable \( \angle S(t, f) \) estimates.

Note that in PO stimuli construction all \( \angle S(t, f) \) values are given equal importance through unity magnitude.
In our work the dynamic range of the Chebyshev analysis window was chosen as the tunable parameter under investigation.

However, with the change in dynamic range the width of the main-lobe also changes.

The dynamic range is certainly important for the short-time magnitude spectrum estimation as it affects spectral leakage.

That may not be the case, however, for the short-time phase spectrum estimation. The main-lobe width may be of more importance. Further investigation is needed.
The short-time magnitude spectrum contains speaker dependent information for the Chebyshev analysis window with large dynamic range.

The short-time phase spectrum contains speaker dependent information for the Chebyshev analysis windows with low dynamic range.

Further research into derivation of speaker discriminative features from the short-time phase spectrum is warranted.
L. Alsteris and K. Paliwal.
Short-time phase spectrum in speech processing: A review and some experimental results.

K. Wojcicki and K. Paliwal.
Importance of the dynamic range of an analysis window function for phase-only and magnitude-only reconstruction of speech.