SPEECH/MUSIC/SILENCE AND GENDER DETECTION ALGORITHM

[Hadi HARB]
outline

• Introduction
• State-of-the-art
• Our approach
• Conclusion
Video indexing

- Video data are increasing exponentially
- Video = Images + Audio
Why audio analysis

- Video
- Audio
  - Speech
    - Transcription
    - Speaker
  - Music
    - Music type

Seg.+Sem.
Indexing
Difficulties
Difficulties

- Temporal variability
  - different words
  - different speakers
- Spectral variability
  - different acoustic conditions
Difficulties

- Finding good features
  - Robust for temporal variability
  - Robust for spectral variability
State of the art

- GMM
- HMM
- Rule based
• -- data needed for learning

• ++ simple modeling
- -- data needed for learning

- ++ simple implementation if ASR exists
Rule Based [Zhang 99]

- -- Thresholding

- ++ Gives important information
What we need in video indexing

• Great variability in video types ==> few learning data
Our approach

Audio ➔ Good features
Features

- **When?**
  - Humans classify audio in relatively large windows (0.3s)
- **What?**
  - Silence Crossing Ratio SCR
  - Frequency Tracking FT
• Music spectrum = Speech Spectrum in 10ms
• Spectrum distributions in 2s are different
<table>
<thead>
<tr>
<th>Introduction</th>
<th>State-of-the-art</th>
<th>Our approach</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

FT

9/27/2001
Good features!

SF + TPPS \rightarrow Good Features
Silence

Energy

Energy *ZCR

9/27/2001
Gender detection

- ZCR
- Spectral distribution
Gender detection

Woman

Man

9/27/2001
Gender detection

0 Hz

4 KHz

Woman

Man

9/27/2001
Gender detection

\[
W = \frac{\sum_{f=1}^{5} X_f}{\sum_{f=35}^{40} X_f} \cdot \frac{1}{ZCR} \cdot G
\]

\[
G = \frac{\sum_{f} X_f \cdot f}{\sum_{f} f}
\]
Conclusion

- Speech error: 7%
- Music error: 21%
- Man/Woman: 23%
- $1/40 \times RT$
Future

- Modeling Music, Speech based on SCR, FT, others
- Demo