

*Separation of Singing Voice from  
Music Accompaniment for  
Monaural Recordings*

Li and Wang (2006)

Assaf Solomovitch

# Outline

- Introduction – music segregation
- What is ASA
- Comparison to known segregation methods
- System description
  - Voice detector
  - Predominant pitch detection
  - Singing voice separation
    - ASA
      - Segmentation
      - Grouping
- Results
- Extensions

# Introduction

- Many uses for music separation systems:
  - Automatic lyrics recognition & alignment
  - Singer identification
  - Music information retrieval
  - Automatic music transcription
- Speech separation has been studied thoroughly. Can we apply the same methodology to music separation ?

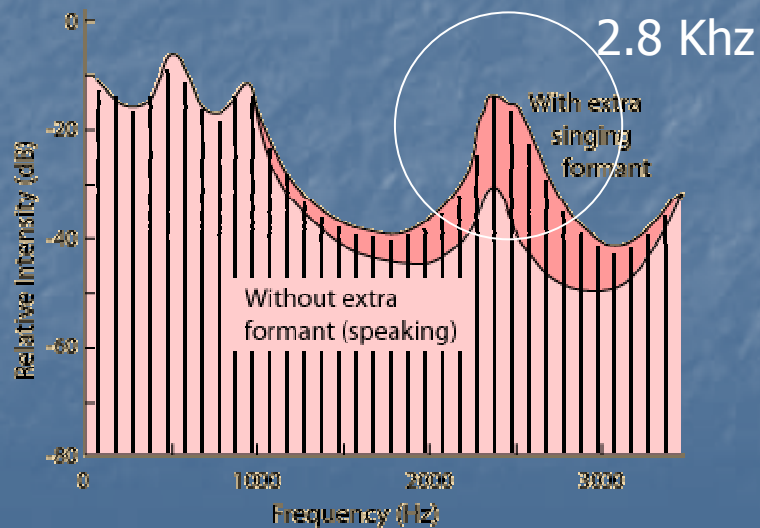
# Music Vs. Speech

## ■ Similarities:

- Both are generated in the human vocal tract
- Have voiced and unvoiced sounds

## ■ Differences

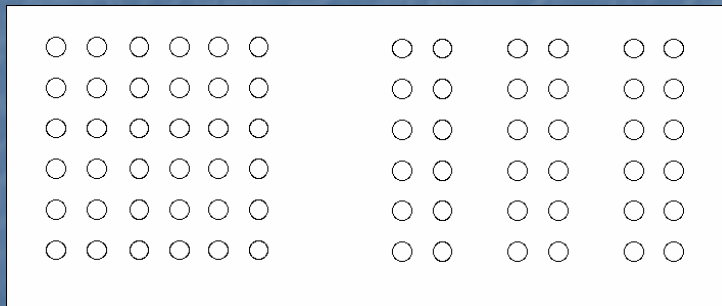
- The presence of the “singing formant” in the range [2-3] KHz – but only in some kinds of music.
- The “interference”, the music, is:
  - Broadband
  - Correlated
  - Harmonic
- About 90% of all sounds are voiced
- A wider pitch range [80-1400] Hz
- Different pitch dynamics



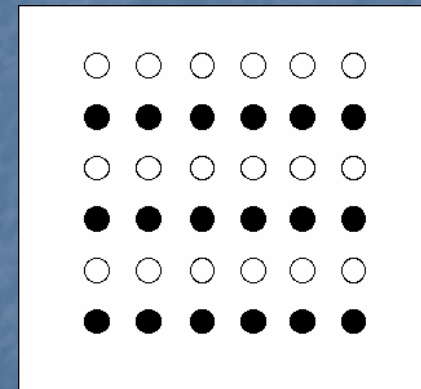
After Sundberg, *The Acoustics of the Singing Voice*

# Auditory scene analysis (Bregman'90)

- Based on the Gestalt psychological theory
- Many similarities between audition and vision (scene analysis)



Principle of proximity



Principle of similarity

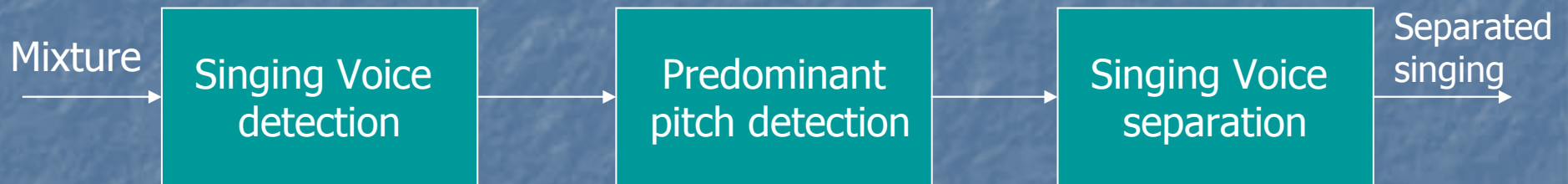
# Auditory scene analysis (Bregman'90)

- Listeners are able to parse the complex mixture of sounds arriving at the ears in order to retrieve a mental representation of each sound source
  - Ball-room problem, Helmholtz (1863)
    - “complicated beyond conception”
  - Cocktail-party problem, Cherry (1953)
- Two conceptual processes of auditory scene analysis (ASA):
  - Segmentation: Decompose the acoustic mixture into sensory elements - a collection of local time-frequency regions (segments)
  - Grouping: Combine segments into streams, so that segments in the same group are likely to originate from the same acoustic source
  - Primitive grouping cues: proximity, periodicity, onset/offset, common spatial location, smooth transition

# Other Sound separation approaches Vs CASA

- Beamforming
  - Disadvantages:
    - Configuration stationarity (if target moves)
    - Poor performance if multiple sounds come from direction near the target
- BSS using ICA:
  - Disadvantages:
    - Assumptions about the mixing process
    - The mixing matrix  $A$  needs to be stationary for a period of time in order to allow estimation of a large number of parameters (like configuration stationarity in beamforming)
    - Poor performance if multiple sounds come from direction near the target
- Speech enhancement techniques
  - Irrelevant to our case – we need separation not enhancement. Also it deals with the narrower perspective of speech + noise.

# System Description



- Singing voice detection is a CASA based separation system, previously used to separate speech.
- Pitch detection accuracy is critical for correct group segments. An algorithm for pitch detection is presented here, modified for singing voice.
- The system is able to separate voiced sounds only
- We need to add a VAD first to filter out portions without vocals

# Singing voice detection

- Goal: partition the input into vocal and nonvocal portions
- Problem 1: partitioning
  - For each frame calculate short-time features
- Problem 2: classification
  - MFCC proved fit as for use as features.
  - Chosen classifier: GMM

# Singing voice detection (cont.)

Article proposes a novel method for detection:

- Take advantage of rhythm (beats) in music:
  - Beats tend to introduce strong spectral changes (percussions)
  - Partition input according to these strong spectral changes
- Within each portion, decide for each frame according to maximum likelihood
- Classify the portion into the class with the larger overall likelihood

# Predominant pitch detection

“Detecting Pitch Of Singing Voice In Polyphonic Audio” / Li & Wang 2005

- Step 1: auditory periphery model: filtering with 128-channel gammatone filterbanks
- Step 2: for each channel compute a normalized correlogram

$$A(c, m, \tau) = \frac{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}} r(c, mT + n) r(c, mT + n + \tau)}{\sqrt{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}} r^2(c, mT + n)} \sqrt{\sum_{n=-\frac{N}{2}}^{\frac{N}{2}} r^2(c, mT + n + \tau)}}$$

- $r$  – filter output for low frequency channel
- $m$  - frame index
- $n$  - time step index

# Predominant pitch detection

"Detecting Pitch Of Singing Voice In Polyphonic Audio" / Li & Wang 2005

- Step 3: Different methods for high/low frequency channels:
  - Low frequencies - channel selection
  - High frequencies - peak selection
- Step 4: Evaluation of pitch hypotheses probability
- Step 5: Pitch tracking by HMM
- Step 6: The Viterby algorithm is used to decode the most likely sequence of pitch hypotheses

# Singing voice separation

- Step 1: auditory periphery model
- Step 2: feature extraction for each TF unit:
  - Energy
  - Autocorrelation
  - cross-channel correlation
  - cross-channel envelope correlation
- Step 3: segmentation
  - Segments are formed by merging contiguous TF units based on temporal continuity and cross-channel correlation

# Segmentation based on temporal continuity and cross-channel correlation

- $U_{c,m}$  = T-F unit of frequency channel  $c$  and time frame  $m$
- $\tilde{A}(c,m,\zeta)$  – normalized correlogram response
- $\zeta$  – the autocorrelation time lag
- $L$  – maximum lag of the correlogram in sampling steps
- Cross channel correlation between  $U_{c,m}$  and  $U_{c+1,m}$  :

$$C(c,m) = \frac{1}{L} \sum_{\tau=0}^{L-1} A(c,m,\tau)A(c+1,m,\tau)$$

- Algorithm:
  1. Select  $U_{c,m}$  If  $\tilde{A}(c,m,0)$  exceeds a certain threshold
  2. Iteratively expand from a selected T-F unit to its selected neighbours in time, and its selected neighbours in frequency if the cross-channel correlation exceeds a certain threshold. This gives one segment
  3. Repeat step 2 until all selected units have been considered

# Singing voice separation

- Step 4: labeling T-F units based on detected predominant pitches.
- Step 5: grouping - If a segment has more than half of its frames marked as singing dominant, then the entire segment is labeled as singing voice dominant. All the singing dominant segments are grouped to form the foreground stream – the segregated singing.

# T-F unit labeling

- In the low-frequency range:
  - A time-frequency (T-F) unit is labeled by comparing the periodicity of its autocorrelation with the estimated target pitch
- In the high-frequency range:
  - Due to their wide bandwidths, high-frequency filters respond to multiple harmonics. These HF responses are amplitude modulated, and their envelopes fluctuate at the freq' corresponding to the F0.
  - A T-F unit in the high-frequency range is labeled by comparing its AM repetition rate with the estimated target pitch

# Resolved and unresolved harmonics

- For voiced speech, lower harmonics are resolved while higher harmonics are not
- For unresolved harmonics, the envelopes of filter responses fluctuate at the fundamental frequency of speech
- Hence we apply different grouping mechanisms for low-frequency and high-frequency signals:
  - Low-frequency signals are grouped based on periodicity and temporal continuity
  - High-frequency signals are grouped based on amplitude modulation (AM) and temporal continuity

# Results

	Rock Music	Country Music
10db	<ul style="list-style-type: none"><li>■ Mixture: </li><li>■ vocals: </li></ul>	<ul style="list-style-type: none"><li>■ Mixture: </li><li>■ vocals: </li></ul>
-5db	<ul style="list-style-type: none"><li>■ Mixture: </li><li>■ vocals: </li></ul>	<ul style="list-style-type: none"><li>■ Mixture: </li><li>■ vocals: </li></ul>

# Possible extensions

- use ASA cues other than pitch (onset/offset, common frequency modulation) to organize the scene – thus enabling unvoiced segregation
- “Auditory Segmentation Based on Onset and Offset Analysis” / Hu & Wang 2007

# Finding onset/offset

- Convert the Canny Edge Detector to audio processing to find onset/offset
  - Convolve the intensity with a derivative of a Gaussian function
  - Identify the peaks and valleys
  - Mark those peaks that are above a certain threshold as onsets and those valleys that are below a certain threshold as offsets

# Possible extensions

- The pitch detection stage is based on autocorrelation function. Hence, the frequency resolution in the high-frequency range is limited. As a result, the system cannot segregate high-pitched singing voice, e.g. operatic voice.
- Since the pitch accuracy is acute, use pitch detection algorithms especially tuned for music
  - Using "Transcription of the singing melody in polyphonic audio" / Ryyanen & Klapuri 2006
  - Using "Efficient Pitch Detection Techniques For Interactive Music" / Cuadra, Master & Sapp 2001

The end

Questions?