Why singing is interesting

Simon Dixon    Queen Mary University of London, UK
Masataka Goto   AIST, Japan
Matthias Mauch Queen Mary University of London, UK

ISMIR 2015 Tutorial

2015/10/26
Why Singing is Interesting

- All popular music cultures around the world use singing
- The singing voice is the most expressive of all musical instruments
- “Of all musical instruments the human voice is the most worthy because it produces both sound and words, while the others are of use only for sound” (Summa Musice, 13th century)
- Our representations (e.g. MIDI, Western notation) are inadequate for expressive singing
- Knowledge about singing from other disciplines (e.g. physiology, psychology, pedagogy) is rarely exploited in MIR
- Many MIR tasks involving singing have never been attempted
Our Plan

What we said we’d do
... introduce to the ISMIR community the exciting world of singing styles, the mechanisms of the singing voice and provide a guide to representations, engineering tools and methods for analysing and leveraging it.

Our aim
... for music information retrieval specialists to walk away with a newly sparked passion for singing and ideas of how to use our knowledge of singing, and singing information processing, to create new, exciting research.
Overview

10:00-10:05  Overview of this tutorial, brief introduction of three speakers

10:05-10:50  **Part 1: Singing Styles and Psychology of Singing** (45 min)
             by Simon Dixon
             questions (10 min)

11:00-12:15  **Part 2: Practical Guide to Singing Information Research** (45 min)
             by Matthias Mauch
             (11:30-12:00:break)
             questions (10 min)

             by Masataka Goto
             questions (10 min)

13:20-13:30  Conclusions
Part 1: Simon Dixon

- **Queen Mary University of London** (2006-)
  - Reader
  - Deputy Director of the Centre for Digital Music

- **Working on music informatics since 1996**
  - Mainly music signal analysis
  - E.g. automatic transcription, beat tracking, audio alignment

- **President of ISMIR** (2014-2015)
Part 2: Matthias Mauch

- Senior Applied Researcher in industry
- Visiting Lecturer at Queen Mary University of London
- Working on music informatics since 2006
- Passionate choir singer and pop singer
Part 3: Masataka Goto

- Prime Senior Researcher / Leader of the Media Interaction Group, AIST (National Institute of Advanced Industrial Science and Technology)
- Working on music information research since 1992
- General chair of ISMIR 2009/2014

No. 5

PreFEst http://songle.jp
Musicream
http://songrium.jp
SmartMusicKIOSK
Robot singer
http://songrium.jp
ISMIR 2015 Tutorial: Why singing is interesting

Part 1: Singing Styles and Psychology of Singing

Centre for Digital Music, Queen Mary University of London

Simon Dixon

2015/10/26
Part 1: Singing Styles and Psychology of Singing

- Singing Styles and Vocal Expression
- Physiology of the Singing Voice
- Intonation, Accuracy, Drift, Poor Singing
- MIR and Singing: Open Problems
Singing Styles and Vocal Expression
Singing Styles

- The voice is a versatile instrument
- It is universal: everyone has one, can use it, and it is suitable for music of all cultures
- It is portable, affordable and expressive
- Use cases: entertainment, art, worship, communication, social
- We observe a great diversity of styles of singing\(^1\)
- Aesthetics (taste, appreciation of beauty) vary by style, and sometimes within styles

Aesthetics: Natural or Artificial?

- Natural
  - Authenticity of expression (e.g. rock, pop, folk styles)
  - Speech-like quality (e.g. Broadway), directness
  - Clarity of lyrics: rap (lyrics foremost) vs opera (intelligibility sacrificed for volume)
  - Amplification destroyed the effort/reward equation

- Artificial
  - Purity of tone, effortless (e.g. Western classical: “objectifying control”)
  - Training, discipline (“high” vs “low” culture)
  - Technical prowess (e.g. classical, jazz)
  - Performance, acting (e.g. rock, opera, musicals)
  - Microphone technique
  - Audio effects

- Exceptions to the general patterns disprove any simplistic view
Aesthetics: Other Factors

- Entertainment vs artistic or intellectual traditions
- Individuality
  - Choral: aim to act as one, breathing and articulating together; accurate but not expressive; no vanity
  - Pop: centrality of the star
- Historical shifts in Western art music
  - The “perfect voice is thus high, sweet and clear” (Isadore of Seville, d. 636)
  - “not effeminate, nasal, forced, strained, nor animal-like” (Scientiae artis musicae, Salomon, 1274)
  - Renaissance: small ranges; change music rather than register
  - Baroque: register switch (use of falsetto); throat articulation; don't move any part of body except glottis
  - 18th-19th century: smoothness, little/no vibrato, portamento, imperceptible register switch, no force, precise intonation
  - Garcia (1840): scientific approach: begin notes forcefully
  - Modern: power and unity of timbre across the range
Baroque Chorale: J.S. Bach
Broadway Belt: from "Oklahoma"
Western Opera: Diane Damrau
Beijing Opera
European Art Song: Ian Bostridge
Dutch Folk
Balkan: Neli Andreeva & Philip Kutev Choir

the Bulgarian National Ensemble & Choir "Filip Kutev"
Malka Moma- Nely Andreeva - Live
Inuit Throat Singing
Tuvan Overtone Singing
Pakistani Qawwali: Nusrat Fateh Ali Khan
Indian Filmi: Lata Mangeshkar
Indian “Beatboxing”: Sheila Chandra
South Africa: Ladysmith Black Mambazo
Sacred Harp (Shape Note Singing)
Jazz Scat: Ella Fitzgerald
Jazz Vocalese: Lambert, Hendricks and Ross
Jazz Acapella: Take 6
Vocal Acrobatics: Bobby McFerrin
American Soul: James Brown
Early Rap: Sugar Hill Gang
Rap: Eminem
Heavy Rock meets MIR

Highway to Hell (AC/DC)
Uri Nieto - March 28th, 2015 - New York City
Physiology of the Singing Voice
How the Voice Works

- Respiratory system: compresses lungs to create airflow
- Vocal folds: chop airstream into a periodic pulsation
- Vocal tract: filters source waveform according to resonances (formants)²

Breathing

- Controlled by rib muscles diaphragm, abdominal wall
- Pressure at glottis determines loudness and affects pitch
- Lung pressure 0.4-1.5 kPa gives 65-87 dB SPL at 0.5m
- Air flow: alternating open phase (triangular pulse) and closed (or almost closed) phase, resulting in 12 dB/octave rolloff
- Slope of closing of glottis varies with loudness
Vocal Folds (Cords)

- Run from front (Adam’s apple) to back (arytenoid cartilage), with an opening called the \textit{glottis}
- Abducted (spread) and adducted (brought together) by laryngeal muscles operating on the arytenoid cartilages
- This determines the tension in the vocal cords
- Myoelastic theory: explains cyclic opening and closing of glottis
  - the vocal cords are initially closed
  - breath pressure is applied from beneath (\textit{subglottic pressure})
  - cords remain closed until sufficient pressure builds up to push them apart
  - air then escapes and the pressure drops
  - muscle tension brings the folds back together
- The rate of repetition of this cycle determines the pitch
Phonation Modes

Continuum of tension in vocal cords:

- Completely relaxed (open): cords do not vibrate (voiceless phonation)
- Partially lax: high air flow, no closed phase (breathy phonation)
- Moderate tension: “sweet spot” of maximum vibration, normal state for spoken vowels (flow phonation)
- High tension: low air flow, long closed phase (pressed phonation)
- Pressed together (closed): vocal cords block airstream (glottal stop)

---

Vocal Fold Oscillation Modes

- **Vocal fry**
  - Folds thick and relaxed
  - Multiple air bursts followed by a long closed phase
  - Two folds vibrate asynchronously
  - Occurs typically at the end of spoken phrases

- **Modal (chest voice)**
  - Symmetrical vibration
  - Open phase at least 50% of cycle
  - Sole register of classical tenor, baritone and bass

- **Falsetto (head voice)**
  - Folds thin and stretched
  - Symmetrical vibration
  - Almost no closed phase
  - Register of countertenor (with closed phase)
Pitch

- Singer’s pitch range is determined by length and mass of vocal folds
- Classical voices
  - Soprano: 260-1050 Hz (C4-C6)
  - Alto: 175-700 Hz (F3-F5)
  - Tenor: 130-520 Hz (C3-C5)
  - Bass: 80-330 Hz (E2-E4)
- Vibrato (classical)
  - Pitch modulation via pulsations in cricothyroid muscle
  - Rate: 5-7 Hz
  - Depth (pitch variation): $\pm 0.5$-1.5 semitones
- Vibrato (pop)
  - Amplitude modulation via variations in subglottal pressure
Vocal Tract

- Resonances occur in the vocal tract according to its configuration
- Up to 5 formants are relevant for singing
- Vowel quality: mainly determined by first 2 formants
- Voice quality: determined by individual factors (size, shape)
- Singer’s formant
  - strong peak in spectral envelope of classical singers
  - clustering of the 3rd, 4th and 5th formants
  - bass (2.2 kHz), tenor (2.9 kHz), alto (3-3.5 kHz)
  - contributes to brilliance of sound and audibility over an orchestra without excessive effort
Examples of Singing Styles and Techniques

- **Choral**
  - no singer’s formant, closer to speech than operatic singing

- **Pop and country**
  - more similar to speech (breathing patterns, lung pressure)
  - pressed phonation used for high pitches
  - (general) absence of low-larynx technique, diaphragm-oriented breathing, pure tone

- **Theatrical (belting)**
  - narrow pharynx, raised larynx, high lung pressure, long closed phase
  - loud, speech-like
  - boosts high overtones
  - extends range of chest register

- **Overtone singing (some Asian cultures)**
  - fixed F0
  - formant 2 or 3 is tuned to enhance a specific partial, sometimes stronger than F0
  - results in a new (additional) pitch
Intonation, Accuracy, Drift and Poor Singing
Poor Singers

- Reveal relationship between perception, memory and production; could identify interventions to help people sing
- Pfordresher compared imitation and discrimination task results to isolate causes of poor singing in non-musicians\(^4\)
- Possible models:
  - perceptual deficit: would predict production covarying with perception, small intervals harder to reproduce than large, and little impact of auditory feedback (masking, augmenting)
  - motor deficit: predicting random direction of errors, large intervals harder than small, gravitation towards a “comfortable” pitch, no correlation with discrimination
- “Poor-pitch singing results from mismapping of pitch onto action, rather than problems specific to perceptual, motor, or memory systems.”

Poor Singers

- The majority of occasional singers can carry a tune\(^5\)
- For a well-known tune at a slow tempo, nonmusicians are as proficient as professional singers
- Various categories of poor singers exist, mostly in the pitch domain, but sometimes in timing (selective impairment)
- Not normally the result of impoverished perception
- Absolute and/or relative accuracy in pitch and tempo suggest a multicomponent system underlying proficient singing
- Pitch accuracy (lack of bias) and precision (lack of spread) in singing familiar and unfamiliar melodies were investigated\(^6\)
- Most participants had low systematic bias, but many had a large spread of results for each pitch class (i.e. were imprecise)

---


Pitch, Intervals and Temperament

- Octaves are divided into 12 (equal?) semitones
- For convenience: \( p = 69 + 12 \log_2 \left( \frac{f}{440} \right) \)
- Musical intervals correspond to fundamental frequency (F0) ratios between constituent tones
- Consonant intervals correspond to simple whole-number ratios
  - 2:1 octave (12 semitones)
  - 3:2 perfect fifth (7 semitones)
- Problem: \( 2^7 \neq \left( \frac{3}{2} \right)^{12} \)
- For fixed-pitch instruments, some or all fifths are adjusted (tempered) when tuning in order to find a suitable compromise
- Variable pitch instruments (e.g. voice) adjust to the context
  - Temperament is not really needed
  - Different instances of the same note can have different pitches
  - Pitch drift: lack of a fixed reference pitch
Intonation and Drift

- Intonation is the pitch accuracy of a realisation of a note.
- Assumes a reference (e.g. accompaniment or previous notes).
- Reported to be a main priority of choir rehearsals.
- Drift: cumulative pitch error observed by unaccompanied singers over tens of seconds.\(^7\)
- Harmonic progressions can induce drift\(^8\), but drift is also observed in solo singing.

---


Modelling Drift and Memory

- Solo singing has no external reference pitch; the reference must be internal, in memory
- Drift corresponds to forgetting the reference pitch
- 24 singers of varying ability sang *Happy Birthday* three times (a run) for various conditions
- Semi-automatic analysis to track and segment pitch trajectories
- Median of pitch trajectory used as note-wise pitch

Accuracy assessed in terms of:

- Interval Error: relative to the score, assuming equal temperament
- Pitch (Note) Error: relative to inferred tonic (linear fit)
- Pitch Drift: between 1st and 3rd runs

---

Example: Note Segmentation and Framewise/Notewise Pitch Estimates
Results: Interval Errors by Interval

interval error (with respect to ET)

n = 333 669 216 1100 665 860 658 223 219 210 211
Results: Interval Errors by Note Number

![Diagram showing interval errors by note number for phrases 1 to 4. The x-axis represents intervals in semitones, ranging from -5 to 24.]
A Model of Reference Pitch Memory

- Assume that intonation is based on two components: a reference pitch $r_i$ and the score information $s_i$ relative to the reference:
  \[ p_i = r_i + s_i + \epsilon_i \]

- Assume the memory of the reference $r_i$ is given by the following causal process:
  \[ r_i = \mu r_{i-1} + (1 - \mu)(p_{i-1} - s_{i-1}) \]
  where $p_{i-1} - s_{i-1}$ is a point estimate of the current reference pitch, and $\mu \in [0, 1]$ is a parameter relating to the memory of the previous reference pitch $r_{i-1}$

- Then $r_i = r_{i-1} + (1 - \mu)e_{i-1}$, i.e. the reference pitch is pulled in the direction of the observed error
  \[ e_{i-1} = (p_{i-1} - s_{i-1}) - r_{i-1} \]
Example: Local Reference Pitch and Note Errors
Estimating the Memory Parameter $\mu$

- **Boundary case 1: $\mu = 0$**
  - The previous note realisation is used for reference, with no further memory of the reference pitch
  - Errors are passed on fully, and variance increases with time; this is very different from our observed data

- **Boundary case 2: $\mu = 1$**
  - The reference pitch is maintained perfectly, unaffected by local errors
  - Variance is constant over time and there is no drift; this is again different from our observations

- **Best fit: $\mu = 0.85$ (varying with singer)**
Happy Birthday Study: Summary

- Median absolute pitch error = 19 cents; interval error = 27 cents
- Errors were correlated with choir experience and self-reported singing ability, but not with musical background
- Median absolute intonation drift = 11 cents
- Drift was significant in 22% of recordings
- Drift magnitude did not correlate with other measures of singing accuracy or singing experience
- Neither a static intonation memory model nor a memoryless interval-based intonation model account for the observations
- A simple causal tonal reference memory model provides a better fit
MIR and Singing: Open Problems
MIR Tasks Related to Singing

- Singing transcription and analysis
  - Predominant melody extraction
  - F0 estimation (monophonic, polyphonic)
  - Note segmentation 🎵
  - Representation issues
- Vocal activity detection
- Singer identification
- Singing skill evaluation
- Vocal timbre analysis
- Lyric transcription and synchronisation
- Singing synthesis
Open Problems — Challenges for MIR

- Representation of singing
  - Event based representations (scores, MIDI) are insufficient
  - Continuous pitch tracks capture detail of intonation (ornaments, glides, vibrato, kobushi)
  - but segmentation into notes is difficult
  - Integration of timbral information (phonation, spectral characteristics, phonemes/lyrics) into singing representations

- Algorithms to compare and assess pitch tracks

- Holistic similarity (or skill) estimation that includes pitch, timing and timbre

- New MIREX tasks: assess a singer’s naturalness, authenticity or purity of tone

---

Questions?
Section 1

Outline
Brief history of singing analysis tools

Pitch and note tracking state of the art

Annotation/transcription tools

Practical Intonation Analysis

Singing data resources
Section 2

Brief history of singing analysis tools
Figure: Carl Seashore and his tonoscopec

\[1\text{Carl E Seashore (1914). “The Tonoscope”. In: The Psychological Monographs 16.3, pp. 1–12.} \]
Pitch tracking using phonophotography

Figure: Example of phonophotography “score”, superposition of four separately recorded and transcribed melody lines.

---

Mid-20th century

1967 use of digital computers from the 1960s, e.g. cepstrum analysis⁴

1972 “An analyser has been developed which allows presentation of spectral analyses, amplitude and frequency vibrato on paper, without need for photography. Signal frequencies between 100 Hz and 10 kHz can be accepted [...]”⁵ (emphasis mine)

1975 “[...] the problem of tracking the frequency of a single (monophonic) periodic signal is one that has been addressed extensively by the speech community. Some groups consider this to be a solved problem.”⁶

1977 application to music archives⁷


The past 20 years

- pitch tracking becoming a commodity: RAPT,\textsuperscript{8} PRAAT,\textsuperscript{9} STRAIGHT,\textsuperscript{10} YIN,\textsuperscript{11} SRH,\textsuperscript{12} Tartini,\textsuperscript{13} pYIN\textsuperscript{14}


\textsuperscript{9}P. Boersma (2001). “Praat, a system for doing phonetics by computer”. In: Glot International 5.9/10, pp. 341–345.

\textsuperscript{10}H. Kawahara, J. Estill, and O. Fujimura (2001). “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT”. In: Proceedings of MAVEBA, pp. 59–64.


Recent large-scale research

A recent large scale study of singing (11258 children!) assessed ability aurally, i.e. without measuring even pitch.\textsuperscript{15}

\textsuperscript{15}G. F. Welch et al. (2014). “Singing and social inclusion”. In: \textit{Frontiers in psychology} 5.
Section 3

Pitch and note tracking state of the art
Usage in community

our own survey\textsuperscript{16}:

\begin{itemize}
  \item sent to ISMIR Community, Auditory and music-dsp
\end{itemize}

\begin{tabular}{lll}
\hline
Field of work & Position & \\
\hline
Music Inf./MIR & 17 (55\%) & Student 11 (35\%) \\
Musicology & 4 (13\%) & Faculty Member 10 (32\%) \\
Bioacoustics & 3 (10\%) & Post-doc 6 (19\%) \\
Speech Processing & 2 (5\%) & Industry 4 (13\%) \\
\hline
\end{tabular}

Experience

\begin{itemize}
  \item Pitch track 18\* (58\%)
  \item Note track 16\* (52\%)
  \item Both 7 (23\%)
  \item None 3 (10\%)
\end{itemize}

Usage in community

our own survey\textsuperscript{16}:

\begin{itemize}
  \item sent to \textit{ISMIR Community, Auditory} and \textit{music-dsp}
\end{itemize}

The DSP algorithms mentioned by survey participants were: YIN (5 participants), Custom-built (3), Aubio (2), and all following ones mentioned once: AMPACT, AMT, DESAM Toolbox, MELODIA, MIR Toolbox, Tartini, TuneR, SampleSumo, silbido, STRAIGHT and SWIPE.

General overview of pitch trackers

“F0 estimators often have three major components:

a) A pre-processing, or signal conditioning stage,

b) a generator of candidate estimates for the true period sought and

c) a ‘post-processing’ stage that selects the best candidate and refines the F0 estimate.”

Talkin 1995

In addition: voiced/unvoiced detection (either as part of the third step, or as a separate one)
Monophonic pitch tracking, a nearly-solved problem

survey of pitch trackers for singing till 2013 Babacan et al. 

Table 1. Error Rates Across the Whole Dataset

<table>
<thead>
<tr>
<th></th>
<th>GPE (%)</th>
<th>FPE (C)</th>
<th>VDE (%)</th>
<th>FFE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAPT</td>
<td>1.01</td>
<td>21.96</td>
<td>1.05</td>
<td>1.99</td>
</tr>
<tr>
<td>RAPT*</td>
<td>0.65</td>
<td>21.98</td>
<td>1.05</td>
<td>1.66</td>
</tr>
<tr>
<td>STRAIGHTv</td>
<td>1.26</td>
<td>17.22</td>
<td>1.05</td>
<td>2.22</td>
</tr>
<tr>
<td>STRAIGHTv*</td>
<td>1.25</td>
<td>17.22</td>
<td>1.05</td>
<td>2.21</td>
</tr>
<tr>
<td>PRAATu</td>
<td>1.47</td>
<td>21.91</td>
<td>0.81</td>
<td>2.18</td>
</tr>
<tr>
<td>PRAAT</td>
<td>1.41</td>
<td>21.93</td>
<td>0.81</td>
<td>2.15</td>
</tr>
<tr>
<td>PRAAT*</td>
<td>1.41</td>
<td>21.94</td>
<td>0.81</td>
<td>2.13</td>
</tr>
<tr>
<td>SRHv</td>
<td>1.91</td>
<td>18.99</td>
<td>1.28</td>
<td>3.08</td>
</tr>
<tr>
<td>SRH</td>
<td>1.72</td>
<td>17.33</td>
<td>1.33</td>
<td>2.95</td>
</tr>
<tr>
<td>SRH*</td>
<td>1.61</td>
<td>17.36</td>
<td>1.33</td>
<td>2.84</td>
</tr>
<tr>
<td>SSHu</td>
<td>3.51</td>
<td>19.66</td>
<td>1.27</td>
<td>4.55</td>
</tr>
<tr>
<td>SSH</td>
<td>2.40</td>
<td>19.46</td>
<td>1.39</td>
<td>3.61</td>
</tr>
<tr>
<td>SSH*</td>
<td>1.91</td>
<td>19.43</td>
<td>1.39</td>
<td>3.16</td>
</tr>
<tr>
<td>YINvu</td>
<td>2.69</td>
<td>8.38</td>
<td>1.05</td>
<td>3.56</td>
</tr>
<tr>
<td>YINv</td>
<td>2.44</td>
<td>12.79</td>
<td>1.05</td>
<td>3.32</td>
</tr>
<tr>
<td>YINv*</td>
<td>0.91</td>
<td>12.95</td>
<td>1.05</td>
<td>1.9</td>
</tr>
</tbody>
</table>

---

Partially solved problem: predominant pitch

- predominant pitch trackers (e.g. PreFEst,\textsuperscript{18} MELODIA\textsuperscript{19})
- vocal activity detection\textsuperscript{20,21}


\textsuperscript{19}J. Salamon and E. Gómez (2012). “Melody extraction from polyphonic music signals using pitch contour characteristics”. In: \textit{Audio, Speech, and Language Processing, IEEE Transactions on} 20.6, pp. 1759–1770.


Pitch tracking implementations I

- YIN Java implementation in Tarsos:
  https://github.com/JorenSix/TarsosDSP, Matlab implementation see
  https://code.soundsoftware.ac.uk/projects/pyin

- MELODIA (Vamp plugin)
  http://mtg.upf.edu/technologies/melodia

- pYIN Vamp plugin and source code:
  https://code.soundsoftware.ac.uk/projects/pyin or Python implementation:
  https://github.com/ronggong/pypYIN

- STRAIGHT (Matlab, available upon request)
  http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html
Pitch tracking implementations II

- SWIPE Matlab: http://www.cise.ufl.edu/~acamacho/publications/swipep.m
- SPTK/Python: http://pysptk.readthedocs.org/en/latest/sptk.html#f0-analysis
- Tartini for Supercollider
  http://doc.sccode.org/Classes/Tartini.html or standalone http://miracle.otago.ac.nz/tartini/
- Aubio http://aubio.org/ or in Vamp:
  http://aubio.org/vamp-aubio-plugins/
- RAPT (Matlab) http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/fxrapt.html
- Cepstral Pitch Tracker https://code.soundsoftware.ac.uk/projects/cepstral-pitchtracker
Note tracking

- much less explored (only few papers\textsuperscript{22,23,24,25,26})
- ill-defined for music that is not sung from a fixed note representation


\textsuperscript{24}E. Gómez and J. Bonada (2013). “Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing”. In: Computer Music Journal 37.2, pp. 73–90.


Note tracking — low performance

- **COnPOff (Precision)**
- **COnPOff (Recall)**
- **COnPOff (F-measure)**
- **COnP (Precision)**
- **COnP (Recall)**
- **COnP (F-measure)**
- **COn (Precision)**
- **COn (Recall)**
- **COn (F-measure)**
- **OBO (RateGT)**
- **OBP (RateGT)**
- **OBOff (RateGT)**
- **Split (RateGT)**
- **Merged (RateGT)**
- **Spurious (RateTR)**
- **Non-detected (RateTR)**

**Baseline method**
- (a) Gómez & Bonada
- (b) Ryynänen
- (c) Melotranscript
Melody note tracking implementations:

- pYIN: [https://code.soundsoftware.ac.uk/projects/pyin](https://code.soundsoftware.ac.uk/projects/pyin)
- CANTE [http://cofla-project.com/cante.html](http://cofla-project.com/cante.html)
- Cepstral Pitch Tracker [https://code.soundsoftware.ac.uk/projects/cepstral-pitchtracker](https://code.soundsoftware.ac.uk/projects/cepstral-pitchtracker)
Section 4

Annotation/transcription tools
The tools with graphical user interfaces mentioned by survey participants were: Sonic Visualiser (12 participants), Praat (11), Custom-built (3), Melodyne (3), Raven (and Canary) (3), Tony (3), WaveSurfer (3), Cubase (2), and the following mentioned once: AudioSculpt, Adobe Audition, Audacity, Logic, Sound Analysis Pro, Tartini and Transcribe!. 
Praat

AMPACT

“Automatic Music Performance Analysis and Comparison Toolkit”

- monophonic sung/MIDI alignment
- perceived pitch
- vibrato rate/depth
- note slope calculation

---

## Requirements

<table>
<thead>
<tr>
<th>Function</th>
<th>Melodyne</th>
<th>Praat</th>
<th>Sonic Visualiser</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate pitch</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>estimate notes</td>
<td>✔️</td>
<td>~ ✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>note/pitch correction</td>
<td>✔️</td>
<td>✘</td>
<td>✘</td>
</tr>
<tr>
<td>note/pitch sonification</td>
<td>✘</td>
<td>✘</td>
<td>✘</td>
</tr>
<tr>
<td>save note/pitch track</td>
<td>~ ✗</td>
<td>✘</td>
<td>✔️</td>
</tr>
<tr>
<td>load note/pitch track</td>
<td>✘</td>
<td>✘</td>
<td>✔️</td>
</tr>
</tbody>
</table>
## Requirements

<table>
<thead>
<tr>
<th></th>
<th>Melodyne</th>
<th>Praat</th>
<th>Sonic Visualiser</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>estimate pitch</strong></td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>estimate notes</strong></td>
<td>✔️</td>
<td>~✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>note/pitch correction</strong></td>
<td>✔️</td>
<td>✘</td>
<td>✘</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>note/pitch sonification</strong></td>
<td>✘</td>
<td>✘</td>
<td>✘</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>save note/pitch track</strong></td>
<td>❍</td>
<td>✘</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><strong>load note/pitch track</strong></td>
<td>✘</td>
<td>✘</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Section 5

Practical Intonation Analysis
A few possible research questions

- vibrato: vibrato depth and rate in different instruments\(^{31}\)
- scale: equal tempered vs. just intonation\(^{32}\)
- poor vs. good singing: accuracy and precision\(^{33}\)
- intonation drift: does intonation reference change over time?\(^{34}\)

---


Note pitches

Tony demo 1 (Erhu)
Intonation

*Intonation* is defined as “accuracy of pitch in playing or singing”\(^\text{36}\) or “the act of singing or playing in tune”\(^\text{37}\).

\[
p = 69 + 12 \log_2 \frac{f_0}{440}.
\] \hspace{1cm} (1)

- pitch strongly associated with fundamental frequency
- we usually measure pitch as in (1), i.e. in semitones with middle C corresponding to \(p = 60\)


Interval error
Interval error
Interval error
Interval error

sung interval
Interval error

Interval

Sung interval

Nominal interval
Interval error
Interval error stats from our paper. 

---

Intonation analysis example

Imports

```python
In [118]: # numpy and matplotlib-related
import matplotlib
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
#matplotlib inline

In [119]: # other funky imports
import bz2
from xml.etree import ElementTree
```

Constants and helper functions

```python
In [120]: fs = 44100
```
Tony demo 3 (Happy Birthday singing!) + updated
IPython Notebook

link to notebook online, don’t use in presentation :)
What can go wrong? Things to consider.

**Timing.** You’re interested in intonation, but people sing at different speeds, so you don’t know whether intonation differences are caused by the different speed.

**Solution:**
- provide click track to singers
Unintended pitch bias. You’re using Audacity’s standard clicks, but after a while you notice that they’re actually pitched, so you’re unsure if singers pick up their pitch from the clicks.

Solution:

- use un-pitched noise clicks
- eliminate other pitched sounds (fans etc.) from the environment
What can go wrong? Things to consider.
What can go wrong? Things to consider.

**Audio analysis fail.** You recorded in a reverberant or noisy room, and in the analysis stage you realise that it’s not clear what is the clean singing signal, and what is echo or noise.

**Solution:**

- use rooms with little reverb
- seek quiet rooms and CHECK FOR HUMS!
- use close mic’ing (we used a headset)
What can go wrong? Things to consider.

**Singers out of range.** You recorded singers and asked them to sing a melody, giving them a particular pitch. Some singers had to try very hard to reach the high notes (or couldn’t reach them), while for others the melody was in their comfort range.

**Solution:**
- nothing to worry about if testing for reaction to vocal strain was your aim ;)
- find pitch that singers are comfortable with
What can go wrong? Things to consider.

Data unsharable. You recorded singers but didn’t ask them whether you could (anonymously) share their singing recordings with the community.

Solution:

▷ do ask them in writing (and record the answers)
▷ by the way: also make sure you’ve got ethics approval for your experiments (this is usually very easy, at least at UK universities)
What can go wrong? Things to consider.

**Poor singers.** Some of the singers either don’t sing the right song, or sing it so badly that it’s unrecognisable.

**Solution:**
- devise a rule in advance to remove poor singers (e.g. all that have an interval error $> 1$ semitone in more than 20% of the notes.
Section 6

Singing data resources
- Meertens Tune Collections
  http://www.liederenbank.nl/mtc/
- Dawn Black’s Singing Voice Audio Dataset
  http://isophonics.net/SingingVoiceDataset
- QBSH Corpus
- IRMAS (instrument recognition, but has 778 voice samples)
  http://www.mtg.upf.edu/download/datasets/irmas
- iKala Dataset
  http://mac.citi.sinica.edu.tw/ikala/
- MTG-QBH
  http://mtg.upf.edu/download/datasets/mtg-qbh
- Molina’s evaluation framework incl. some data
  http://www.atic.uma.es/ismir2014singing/
- Jiajie Dai’s Singing Experiment data
  http://figshare.com/articles/Media_Content_for_Analysis_of_Intonation_Trajectories_in_Solo_Singing_/1482221
Polina Proutskova’s phonation modes dataset http://www.doc.gold.ac.uk/~mas02pp/phonation_modes/

MedleyDB
http://medleydb.weebly.com/downloads.html

RWC (Musical Instrument Sound) https://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i.html

TONAS Dataset
http://mtg.upf.edu/download/datasets/tonas

ccmixter corpus (separate vocal tracks)
http://www.loria.fr/~aliutkus/kam/

Jamendo corpus
http://www.mathieeuramona.com/wp/data/jamendo/

Ultrastar Database (annotations for vocals, gender, ..., but no audio for popular songs) http://openaudio.eu/

RWC Melody line annotations and more detailed vocal/instrumental activity https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/
- Tunebot Database
  http://music.cs.northwestern.edu/data/tunebot/
- MARG database http://marg.snu.ac.kr/?page_id=767
- NTENT Singing Voice Database (by Liliya Tsirulnik and Shlomo Dubnov.)
  http://liliyatsirulnik.wix.com/svdb
- Mixing Secrets Dataset https://sisec.inria.fr/professionally-produced-music-recordings/
- Cofla Flamenco Annotations
  http://cofla-project.com/corpus.html
Bibliography I


Bibliography III


Welch, G. F. et al. (2014). “Singing and social inclusion”. In: *Frontiers in psychology 5*. 
Lunch!
ISMIR 2015 Tutorial: Why singing is interesting

Part 3: Singing Information Processing Systems

AIST (National Institute of Advanced Industrial Science and Technology)

Masataka Goto

2015/10/26
“Singing information processing” [Goto, et al., 2009-]

- Music information research for singing voices
  Signal processing + machine learning + interfaces + …
- Singing is one of the most important elements of music
  Many people listen to music with a focus on singing
- Attract attention not only from a scientific point of view but also from the standpoint of industrial applications
  - Automatic singing pitch correction (e.g., “Auto-Tune”)
    Intentionally used to achieve a desired effect such as T-Pain (USA) and Perfume (Japan)

F0 of natural singing

![F0 of natural singing](image1)

After Auto-Tune

![After Auto-Tune](image2)
“Singing information processing” [Goto, et al., 2009-]

- Music information research for singing voices
  Signal processing + machine learning + interfaces + …
- Singing is one of the most important elements of music
  Many people listen to music with a focus on singing
- Attract attention not only from a scientific point of view
  but also from the standpoint of industrial applications
  - Automatic singing pitch correction (e.g., “Auto-Tune”)
  - Singing synthesis (e.g., “VOCALOID”)
  - Query-by-humming (e.g., “midomi”)
  - Singing skill evaluation for Karaoke
- The concept is broad and still emerging
Singing Information Processing Systems

- **Vocal Timbre Analysis**
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- **Lyric Transcription and Synchronization**
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- **Singing Skill Evaluation**
  - Singing skill evaluation/visualization/training

- **Singing Synthesis**
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
Singing Information Processing Systems

- **Vocal Timbre Analysis**
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- **Lyric Transcription and Synchronization**
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- **Singing Skill Evaluation**
  - Singing skill evaluation/visualization/training

- **Singing Synthesis**
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
Music information retrieval based on **singing voice timbre**

- Retrieve and discover songs that have **vocal timbre** similar to a query song

VocalFinder: MIR system based on vocal timbre similarity by Hiromasa Fujihara and Masataka Goto

Query song

Avril Lavigne: I'm With You
Britney Spears: Oops!
Celine dion: Because You Loved Me
VocalFinder: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate vocal melody from polyphonic sound mixtures by using predominant-F0 estimation method *PreFEst*

- **Vocal timbre modeling method** [Fujihara, et al., 2005-]
  - Train singer GMM for each song by using feature vectors on reliable vocal frames
Male/Female Estimation

Music browsing based on male/female estimation

- Visualize a song collection by using male/female estimation

You can try this at http://songrium.jp/sings
Male/Female Estimation: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate vocal melody from polyphonic sound mixtures by using predominant-F0 estimation method *PreFEst*

- **Male/Female modeling method**
  - Train male/female SVM classifier by using feature vectors on reliable vocal frames

![Diagram of vocal extraction and modeling process]
Singer identification (ID) for polyphonic music recordings

- Identify the name of the singer who sang the input song
  Similar to speaker recognition
- You can retrieve a song without metadata

Song: “Open Arms”
Artist: Journey
Singer name: Steve Perry

Song: “With or Without You”
Artist: U2
Singer name: Bono
Singer ID: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate **vocal melody** from polyphonic sound mixtures by using predominant-F0 estimation method *PreFEst*

- **Vocal timbre modeling method** [Fujihara, et al., 2005-]
  - Train **singer GMM** for each singer by using **feature vectors** on **reliable vocal frames**
Singing Information Processing Systems

- Vocal Timbre Analysis
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- **Lyric Transcription and Synchronization**
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- Singing Skill Evaluation
  - Singing skill evaluation/visualization/training

- Singing Synthesis
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
LyricSynchronizer

- Automatic synchronization of **lyrics** with polyphonic music recordings
  - Display _scrolling lyrics_ with the _phrase currently being sung highlighted_ during playback of a song

The current playback position

You can listen from a clicked word

[Fujihara, Goto, Okuno, 2006-]
LyricSynchronizer: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate vocal melody from polyphonic sound mixtures by using predominant-F0 estimation method \textit{PreFEst}

- **Automatic lyrics synchronization method** [Fujihara, Goto, Okuno, 2006-]
  - Detect vocal sections by using HMM
  - Locate each phoneme in resynthesized vocal melody by using the \textit{Viterbi (forced) alignment} technique
Lyric Synchronization: References

Lyric Synchronization: References


Automatic transcription of **lyrics**

- Use repeated choruses to improve **automatic lyric recognition** (solo sung voice)

---

McVicar, Ellis, Goto, 2014


Creating hyperlinks between phrases in lyrics

- Create a hyperlink from a phrase in the lyrics of a song to the same phrase in the lyrics of another song

Songs with text lyrics

Song A

... In your eyes ...

Song B

... In your eyes ...

Hyperlink

(1) Extract a key phrase

(2) Detect a key phrase

(3) Create a hyperlink

Songs without text lyrics

[Fujihara, Goto, Ogata, 2008-]
Hyperlinking Lyrics: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate vocal melody from polyphonic sound mixtures by using predominant-F0 estimation method *PreFEst*

- **Keyword spotting method** [Fujihara, Goto, Ogata, 2008-]
  - Locate each key phrase in extracted vocal melody by using the *Viterbi alignment* with phoneme HMMs

- Interactive editing of lyrics animation based on automatic video composition on the web browser
  - Reduce manual labors: **music understanding** techniques
  - Compose videos on-the-fly: **live programming** techniques

- Interactive editing of lyrics animation based on automatic video composition on the web browser
  - Reduce manual labors: music understanding techniques
  - Compose videos on-the-fly: live programming techniques

Available on http://textalive.jp
Singing Information Processing Systems

- Vocal Timbre Analysis
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- Lyric Transcription and Synchronization
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- Singing Skill Evaluation
  - Singing skill evaluation/visualization/training

- Singing Synthesis
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
Singing Skill Evaluation

Evaluate pitch interval accuracy w/o score

F0 trajectory after low-pass filtering w/o silent sections

Semitone stability

Pg(F,t) in good singing

Pg(F,t) in poor singing

[Nakano, Goto, Hiraga, 2006]
Singing Skill Evaluation [Nakano, Goto, Hiraga, 2006]

- Evaluate pitch interval accuracy w/o score

Semitone stability

F = 19 cent

F = 78 cent

\[ P_g(F,t) \]
Singing Skill Evaluation

[Nakano, Goto, Hiraga, 2006]

$g(F)$: long-term average of $P_g(F,t)$

Pitch Interval accuracy

Sharpness of $g(F)$: second moment avg. slope of $g(F)$

Vibrato

Sample

Classifier

SVM

Training dataset

Good

Poor

$P_g(F,t)$ in good singing

$P_g(F,t)$ in poor singing

$F$ [cent]

0

100

$100 0 100 0$
MiruSinger [Nakano, Goto, Hiraga, 2007-]

- **Singing skill visualization and training**
  - Help you imitate the *vocal part* of a target song
  - Analyze and visualize *vocal singing* with reference to the *vocal part* of a target song
  - Real-time feedback of F0 and *vibrato sections*

Vocal part (original singing) of a target song

User singing

Automatically detected vibrato

![Diagram](image)

F0 (log freq.)

time

User interface with graph showing F0 variations over time for vocal part and user singing with noted vibrato sections.
MiruSinger: Technology

- **Automatic vocal extraction method** [Goto, 1999-]
  - Segregate *vocal melody* from polyphonic sound mixtures by using predominant-F0 estimation method *PreFEst*

- **Automatic Vibrato Detection Method** [Nakano, Goto, Hiraga, 2006-]
  - Calculate *vibrato likeliness* by using STFT of delta F0
Singing Information Processing Systems

- **Vocal Timbre Analysis**
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- **Lyric Transcription and Synchronization**
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- **Singing Skill Evaluation**
  - Singing skill evaluation/visualization/training

- **Singing Synthesis**
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
Singing Synthesis

- **Text-to-singing synthesis**
  - Input: Note-level score information with its lyrics + Singing synthesis parameters such as pitch (F0) and dynamics (power)

- **Speech-to-singing synthesis**
  - Input: Speaking voice reading the lyrics of a song

- **Singing-to-singing synthesis**
  - Input: Singing voice singing the lyrics of a song
Singing Synthesis

- **Text-to-singing synthesis**
  - Input: Note-level score information with its lyrics + Singing synthesis parameters such as pitch (F0) and dynamics (power)

- **Speech-to-singing synthesis**
  - Input: Speaking voice reading the lyrics of a song

- **Singing-to-singing synthesis**
  - Input: Singing voice singing the lyrics of a song
Text-to-singing synthesis

- Singing synthesis engine “VOCALOID2”
  - Singing synthesis software “Hatsune Miku” was released on August 31st, 2007
  - Virtual singer was embodied (illustrated) by a cartoon girl
    This has inspired many people to create, share, and remix
Text-to-singing synthesis

- **Singing synthesis engine “VOCALOID2”**
  - Singing synthesis software “Hatsune Miku” was released on August 31st, 2007
  - Virtual singer was embodied (illustrated) by a cartoon girl. This has inspired many people to create, share, and remix.
  - Both amateur and professional musicians started using singing synthesizers as their main vocals.
  - A lot of different voices have already been on the market.
Hatsune Miku Phenomenon

- Commercial film of “Google Chrome” browser
  - 1 minute introduction of *Hatsune Miku* Phenomenon

*Bronze Lion Award, Cannes Lions International Festival of Creativity, June 2012*

Used with permission from Crypton Future Media, INC.
Hatsune Miku Phenomenon

- **Live concerts featuring** Hatsune Miku
  - Tokyo, Sapporo, Wakayama, Yokohama, Osaka, etc.
  - Los Angeles, New York, Singapore, Hong Kong, Taipei, Jakarta, Shanghai, etc.
  - Opening act for Lady Gaga’s USA concert tour in 2014
Hatsune Miku Phenomenon

- **Live concerts featuring Hatsune Miku**
  - Tokyo, Sapporo, Wakayama, Yokohama, Osaka, etc.
  - Los Angeles, New York, Singapore, Hong Kong, Taipei, Jakarta, Shanghai, etc.
  - Opening act for Lady Gaga’s USA concert tour in 2014
  - US television debut “Late Show with David Letterman” in 2014
  - Hatsune Miku’s Opera at Théâtre du Châtelet in Paris in 2013
Hatsune Miku Phenomenon

- The most surprising change

Singing synthesis breaks down the long-cherished view that “listening to a non-human singing voice is worthless”, emerging the “culture in which people actively enjoy songs with synthesized singing voices as the main vocals”
Future of Singing Synthesis

- **Music technologies** have already changed **music cultures** in the history of music
  - The piano and guitars were *brand-new technologies* when people started using them.

- **Sound synthesis**
  - *Sound synthesizers* are *widely used* and have become *indispensable* to popular music production.

- **Singing synthesis**
  - There is no reason that the same will not happen for singing.
  - It is historically inevitable that *singing synthesizers* will become *widely used worldwide* and likewise *indispensable*.
Singing Synthesis

- **Text-to-singing synthesis**
  - Input: Note-level score information with its lyrics
  - + Singing synthesis parameters
    such as pitch (F0) and dynamics (power)

- **Speech-to-singing synthesis**
  - Input: Speaking voice reading the lyrics of a song

- **Singing-to-singing synthesis**
  - Input: Singing voice singing the lyrics of a song
SingBySpeaking

- **Speech-to-singing synthesis**
  - Convert a *speaking voice* to a *singing voice* by changing F0, phoneme duration, and singing formant

[Saitou, Goto, 2007-]
**SingBySpeaking**

[Saitou, Goto, 2007-]

- **Automatic Lyrics Synchronization Method**
  - Locate each phoneme in the speaking voice by using the Viterbi alignment with phoneme HMMs

- **F0 Contour Generation Method**
  - Add four types of F0 fluctuations on musical notes

![Graph showing Melody contour with F0 fluctuations](image)
SingBySpeaking  [Saitou, Goto, 2007-]

- **Spectral Control Method**

**Before spectral control**

**After spectral control**

Original input + Singer’s formant + Vibrato & AM + All
SingBySpeaking [Saitou, Goto, 2007-]

- **Speech-to-singing synthesis**
  - Convert a *speaking voice* to a *singing voice* by changing F0, phoneme duration, and singing formant.
Singing Synthesis

- **Text-to-singing synthesis**
  - Input: Note-level score information with its lyrics + Singing synthesis parameters such as pitch (F0) and dynamics (power)

- **Speech-to-singing synthesis**
  - Input: Speaking voice reading the lyrics of a song

- **Singing-to-singing synthesis**
  - Input: Singing voice singing the lyrics of a song
Singing-to-Singing Synthesis: VocaListener

“Packaged” by VocaListener

2010/10/04 [sm12320140]

Used with permission from Packaged / kz
What is VocaListener?

- **VocaListener** synthesizes natural singing voices by analyzing and imitating human singing
  - Imitate the pitch, dynamics, phoneme timing, and breath of the singer's voice
  - Estimate parameters of singing synthesizer “VOCALOID”

Original VOCALOID  
**Game of Love**  

VOCALOID + VocaListener
VocaListener

- Generate a **musical score** by analyzing the input singing voice
- Estimate **synthesis parameters** for each virtual singer

**Input**

Singing voice
Lyrics

**VocaListener**

**Singing synthesizers**

- Hatsune Miku
- Kagamine Rin
- Gackpoid
- Megpoid

**F0**
**Power**
**Time**

Timing of phonemes

[nakano, goto, 2008-]
VocaListener

- Generate a **musical score** by analyzing the input singing voice
- Estimate **synthesis parameters** for each virtual singer

**Input**

- **Singing voice**
- **Lyrics**
  - nome to iwarete
  - youkini nonda

**VocaListener**

**Analysis**

- F0
- Power
- Timing of phonemes

**Synthesis**

**Singing synthesizers**

- Hatsune Miku
- Kagamine Rin
- Gackpoid
- Megpoid

[Nakano, Goto, 2008-]
VocaListener

- Generate a **musical score** by analyzing the input singing voice
- Estimate **synthesis parameters** for each virtual singer

---

**Input**

- **Singing voice**
- **Lyrics**

**VocaListener**

- **Analysis**
  - **F0**
  - **Power**
  - **Timing of phonemes**

**Synthesizing synthesizers**

- **VOCALOID**
  - Hatsune Miku
  - Kagamine Rin
  - Gackpoid
  - Megpoid

---

[Nakano, Goto, 2008-]
Why is this difficult?

• Synthesized results are different because of singer DBs
• We needed iterative parameter estimation
VocaListener as a Commercial Product

- 2011/06: Press release of a product version
  - As a Job Plugin of VOCALOID3 by YAMAHA Corp.
- 2012/10: The product appears on the market
  - “VOCALOID3 Job Plugin VocaListener”
- 2015/08: The upgraded version is released
  - “VOCALOID4 Job Plugin VocaListener”
Sagashituzuketa Futarinosohtono Prologue
探し続けた 二人のほんとの プロローグ
Robot Singer HRP-4C Miim
VocaListener + VocaWatcher

AIST, Japan
Shuuji Kajita, Tomoyasu Nakano,
Masataka Goto, Yosuke Matsusaka,
Shin'ichiro Nakaoka, and Kazuhito Yokoi
Humanoid Robot Singer: HRP-4C Miim

- Imitating a human singer
  - **VocaListener**
    Imitate vocal expressions to synthesize singing voices
  - **VocaWatcher**
    Imitate facial expressions to generate robot motions

[Kajita, Nakano, Goto, Matsusaka, Nakaoka, Yokoi, 2010-]
“Uncanny Valley”
Toward Future Technologies

- Let’s overcome the “Uncanny Valley”!
  - We should not be afraid of jumping into the valley
  - Otherwise, we cannot go beyond
  - Let’s challenge!

Uncanny? Creepy? Cute?
No. 60

Robot Singer HRP-4C Miim

VocaListener + VocaWatcher

Video clips are available on the web!

http://staff.aist.go.jp/t.nakano/VocaListener/

http://staff.aist.go.jp/t.nakano/VocaWatcher/
Singing Information Processing Systems

- **Vocal Timbre Analysis**
  - MIR based on vocal timbre similarity
  - Male/female estimation
  - Singer identification

- **Lyric Transcription and Synchronization**
  - Lyric synchronization/transcription
  - Lyric animation (kinetic typography)

- **Singing Skill Evaluation**
  - Singing skill evaluation/visualization/training

- **Singing Synthesis**
  - Text-to-singing synthesis
  - Speech-to-singing synthesis
  - Singing-to-singing synthesis
  - Robot singer
Summary

- Introduce singing information processing systems

Let’s work on singing information processing!

- Singing possesses aspects of both speech and music
- Many unsolved research problems
  
  Automatic recognition of singing (lyrics) is the most difficult class of speech recognition (ASR) because of loud accompaniment and large fluctuations

  Singing synthesis requires dynamic, complex, and expressive changes in the voice pitch, power, and timbre

- Research activities on speech and music will be integrated
MIR based on singing voices / Query-by-Humming (QBH)

References

Vocal Timbre Analysis / Singer Identification

References


Lyric Synchronization

Lyric Transcription


Hyperlinking Lyrics

References

Singing Skill Evaluation / Singing Training / Other topics

References

Singing Synthesis

References

References

VocaRefiner / Singing Voice Conversion


Psychology / Physiology / Vocal Pedagogy

References

Voice Percussion


The references above are shown in the following papers:

Conclusions

Simon Dixon, Queen Mary University of London, UK
Masataka Goto, AIST, Japan
Matthias Mauch, Queen Mary University of London, UK

2015/10/26
Why singing is interesting

▶ inherent reasons
  ▶ people love to sing
  ▶ people love to listen to other people singing

▶ scientific reasons
  ▶ scientific discovery in music psychology: how people sing, and how people perceive singing
  ▶ scope for historical and cultural analysis: how people’s singing differs and changes

▶ MIR reasons
  ▶ many MIR tasks relating to singing can be improved, and new ones explored!
  ▶ there’s a lot of data out there (even annotated), which we can exploit
Why singing is interesting

- inherent reasons
  - people love to sing
  - people love to listen to other people singing
  - people love to listen to computers singing

- scientific reasons
  - scientific discovery in music psychology: how people sing, and how people perceive singing
  - scope for historical and cultural analysis: how people’s singing differs and changes

- MIR reasons
  - many MIR tasks relating to singing can be improved, and new ones explored!
  - there’s a lot of data out there (even annotated), which we can exploit