

Singing-driven Interfaces for Sound Synthesizers

Jordi Janer i Mestres
Music Technology Group
Universitat Pompeu Fabra

Director:

Xavier Serra
Music Technology Group
Universitat Pompeu Fabra

Defense Committee:

Marcelo Bertalmio (UPF)
Philippe Depalle (McGill University)
Climent Nadeu (UPC)
Sergi Jordà (UPF)
Emilia Gomez (ESMUC)



overview

We aim to study the singing voice as a source of control, providing new interfaces for DMI's by exploring the control strategies to sound synthesizers

overview

We aim to study the singing voice as a source of control, providing new interfaces for DMI's by exploring the control strategies to sound synthesizers

Singing-driven Interfaces for Sound Synthesizers

outline

I: STUDY

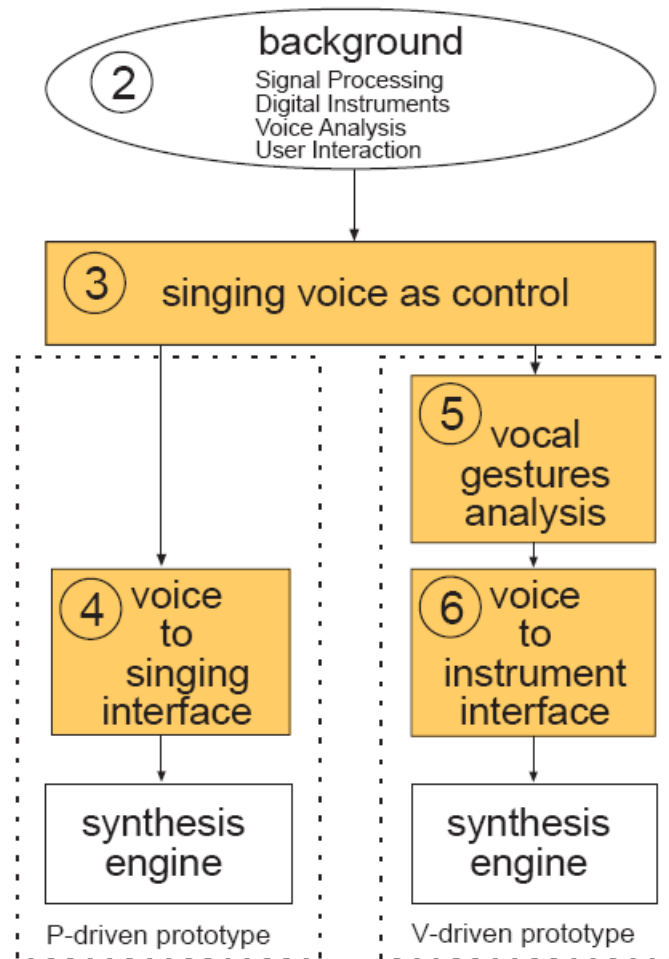
1. Presentation
2. Motivation, context and objectives
3. Singing voice as a control signal

II: EXPERIMENT

5. Controlling a singing voice synthesizer
6. Vocal gestures analysis
7. Controlling an instrumental sound synthesizer

III: RESULTS

9. Prototypes
10. Conclusions



Dissertation's chapters structure

motivation

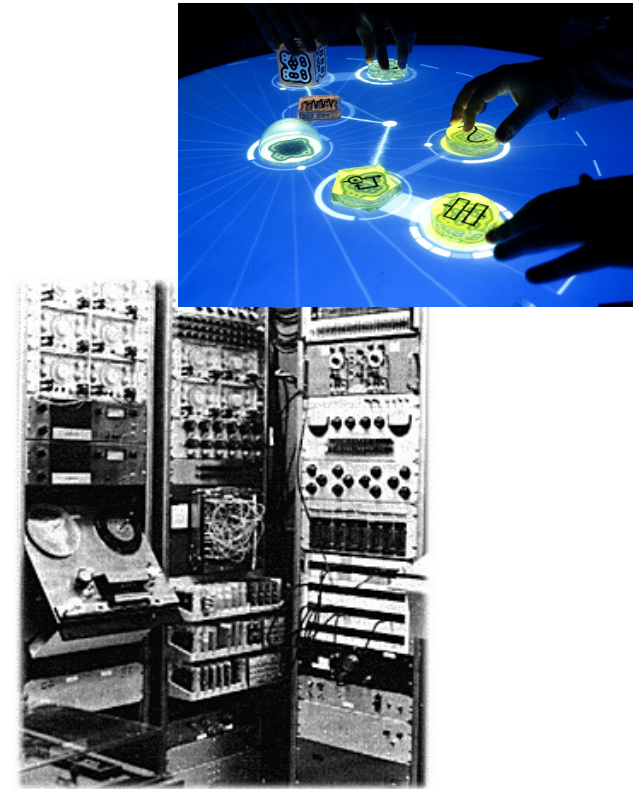
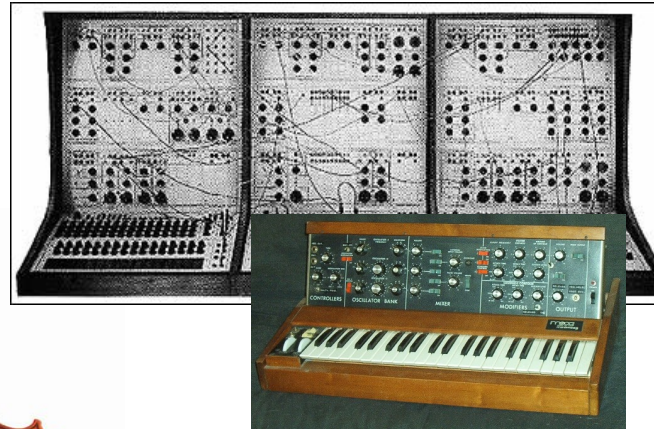
- Research at the MTG
 - voice analysis/transformation/synthesis, spectral processing, morphing examples...
- Personal background
 - Electronic Engineer, interactive installations, DSP developer...



personal activities

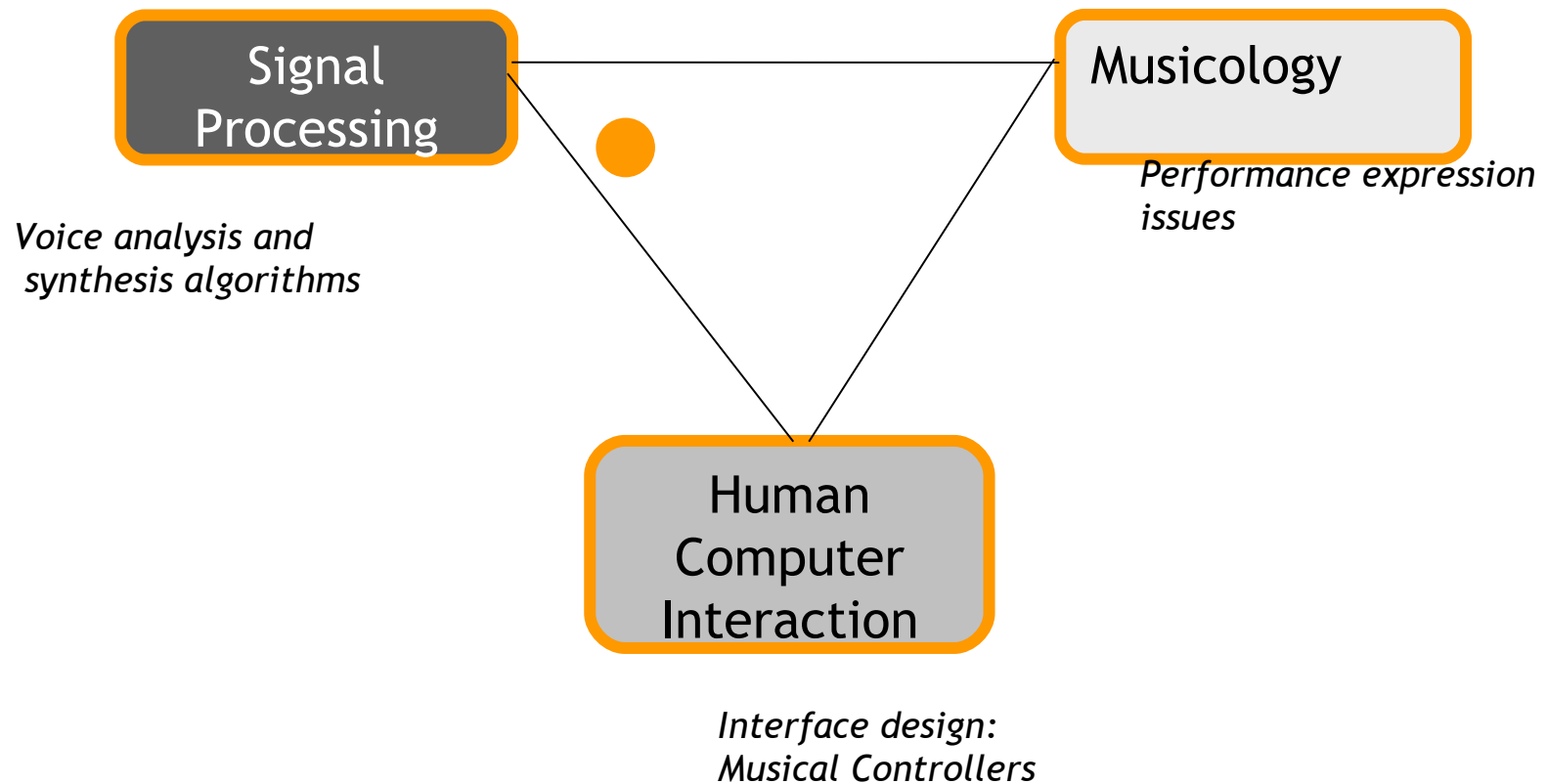
- Research activities:
 - Cuidado project (2002-2003): EU IST 20194
 - ESPRESSO project (2004-2005): Joint research with Yamaha
 - Semantic HiFI project (2004-2006): EU FP6-IST-507913
 - Short-term research stay (2005): McGill University
 - SALERO project (2006-2008): FP6 IST 027122
 - GAROTA project (2007-2008): Tech. transfer BMAT
- Academic activities:
 - Assistant Lecturer in Signal Processing at UPF (2004-2007)
 - Course on Sound Synthesis at ESMUC (2007)

Interaction in musical instruments



neanderthal flute | violin | buchla modular | minimoog | groove | reactable

disciplines

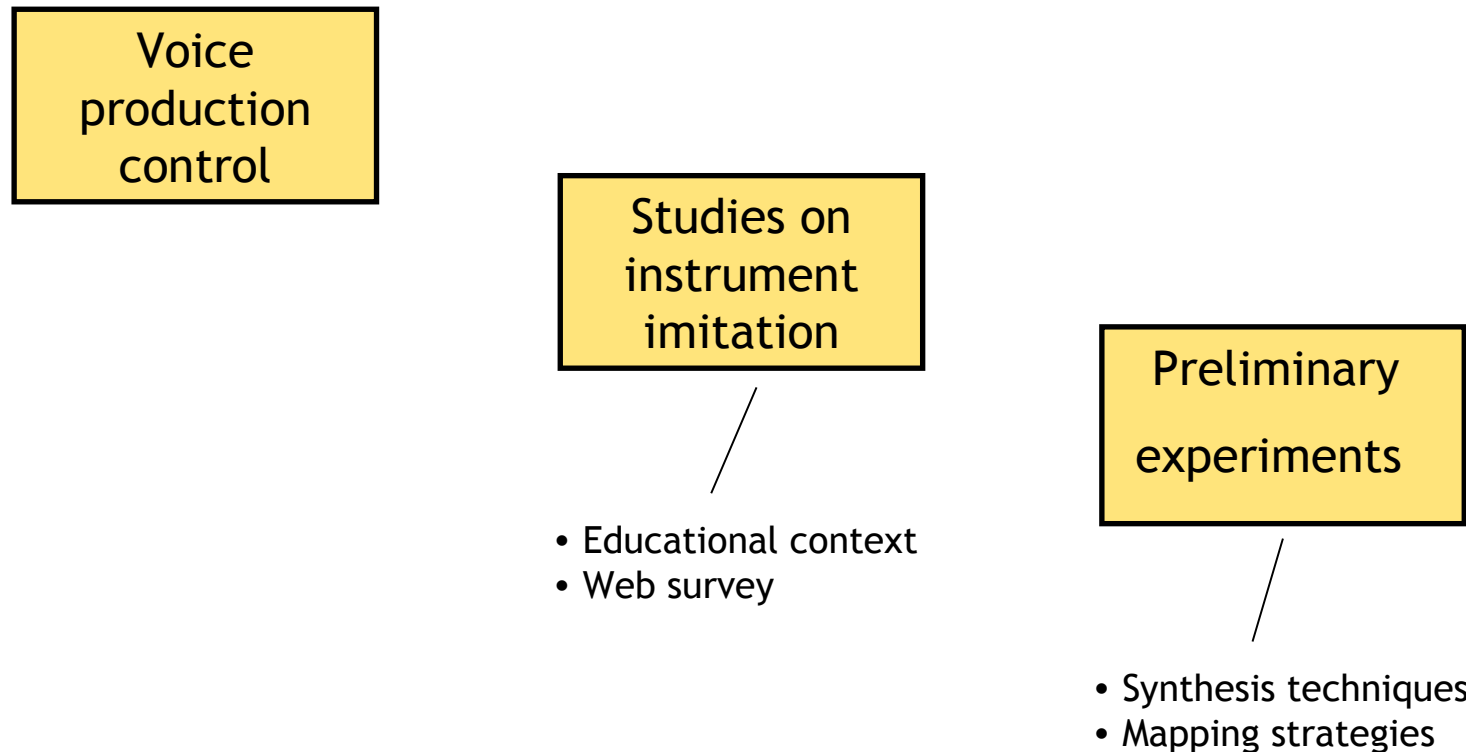


thesis objectives

- highlight the possibilities of the voice for controlling synthesis
- study how people imitate musical instruments with their voices
- design signal processing methods for extracting voice features
- design appropriate mappings to synthesizer parameters
- develop prototypes to demonstrate the results

singing voice as a control signal

singing voice as a control signal



voice production and control



voice production and control



breathing

vocal folds

vocal tract

Action

compression

phonation

articulation

Control

sub-glottic pressure

laryngeal muscles
tension

articulators position

*Voice
attribute*phonation
loudnessphonation
fundamental frequency
vocal disorders

formants

why use voice for interaction?

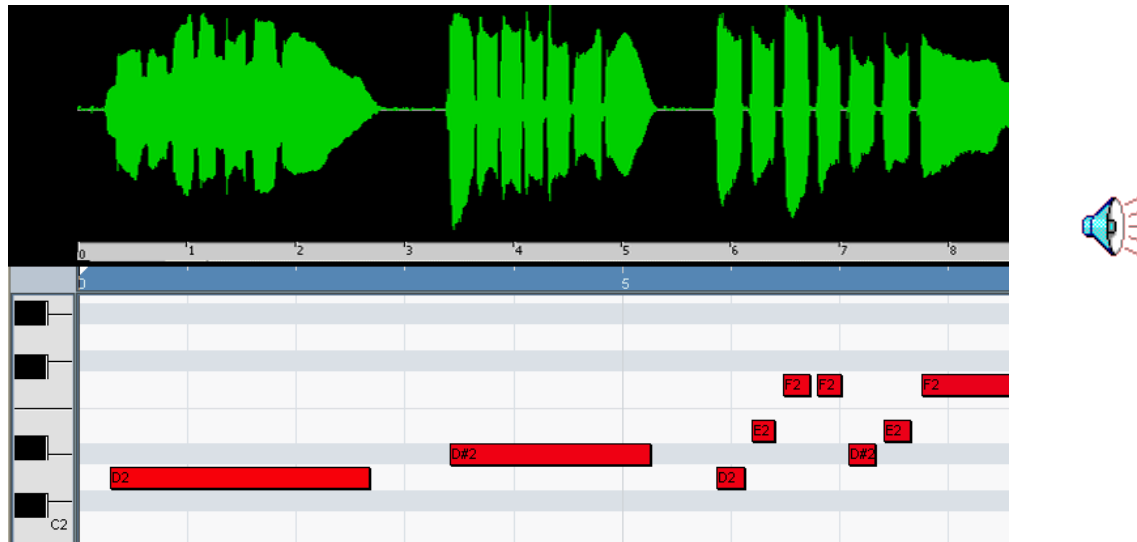


instrument imitation

why use voice for interaction?

- **expressive qualities of singing:**
 - first musical instrument and widely exploited in musical history
- **weakness of current DMI's on the control side**
 - roadmap on Sound and Music Computing Research (ed. Serra et al., 2007)
- **musical efficiency**
 - it presents at the low entry fee and the degree of complexity (Jordà, 2005)
- **ubiquity**
 - neuropsychology studies show that 90% people can sing! (Dalla Bella et al., 2007)

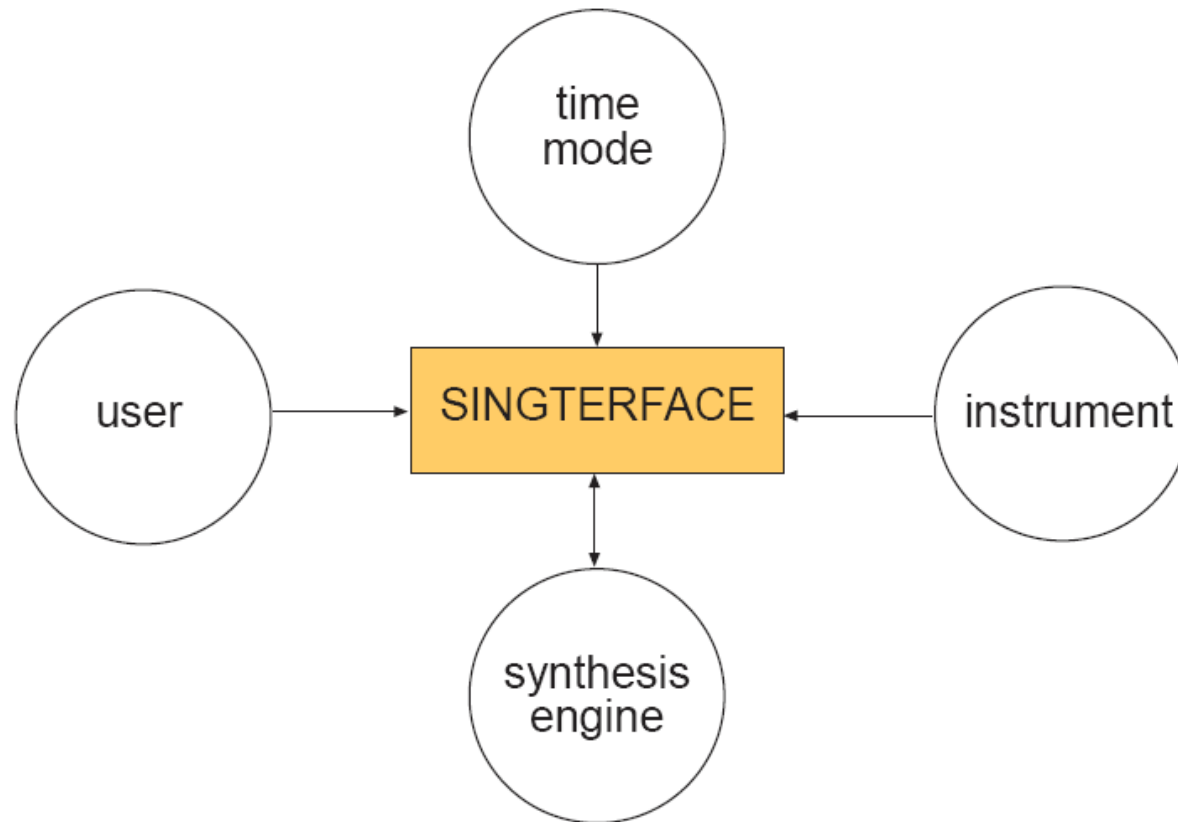
limitations of generic pitch-to-MIDI systems



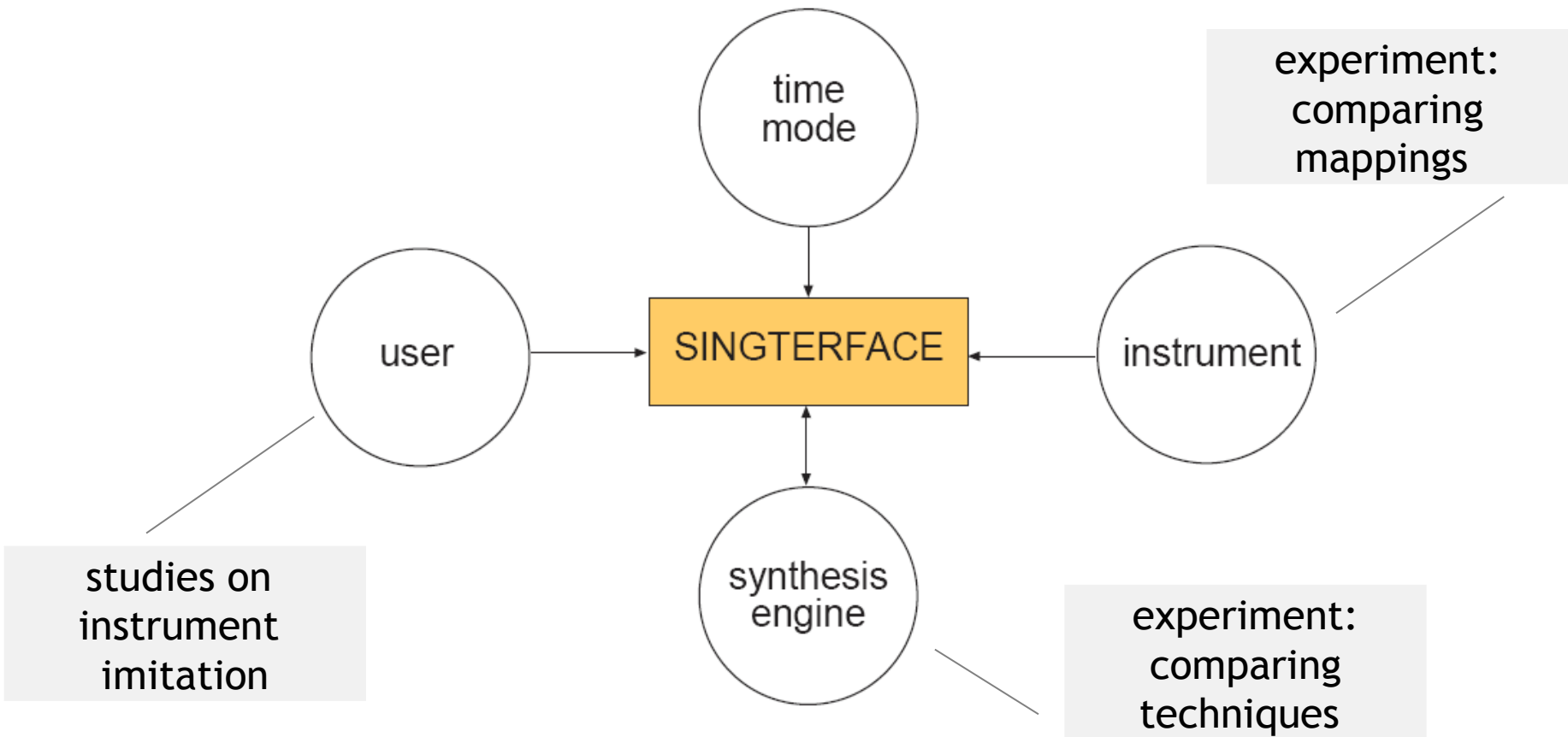
(MIDI output using DigitalEar software)

- MIDI was not designed to describe acoustic signals
- Analysis algorithms are not adapted to voice signals
- These systems are decoupled from the sound synthesis engine

design factors



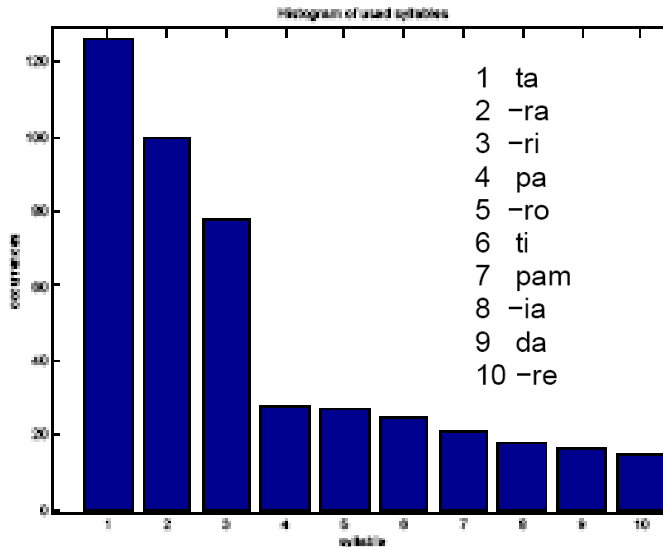
design factors



studies on instrument imitation

studies on instrument imitation

1. educational context

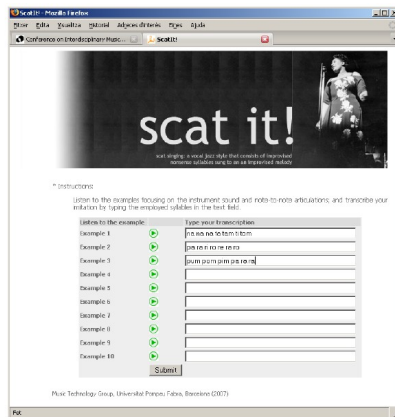


- Teachers give instructions by imitating the sound of instruments.
- Nonsense text singing is also referred as “syllabling” (Sundberg, 1994)

- **Experiment:** annotation of master classes recordings (N = 82)
- **Results:**
 - Few syllables mostly chosen
 - Influence of timbre on chosen vowel
 - Intrinsic pitch (Laver, 1994)

studies on instrument imitation

1. web survey



Examples focusing on the instrument sound and note-to-note syllables employed in the text field.

| Example | Type your transcription |
|-----------|-------------------------|
| Example 1 | na na na ta tam ti tom |
| Example 2 | pa ra ri ro re ra ro |
| Example 3 | pum pom pim pa ra ra |
| Example 4 | |
| Example 5 | |
| Example 6 | |
| Example 7 | |

Musical phrases:

instrument (3):

bass guitar, sax and violin.

articulation (3):

legato, medium and staccato.

note interval (2):

low (2) and high (7 semitone).

tempo (2):

slow and fast.

Experiment:

Number of participants: 60

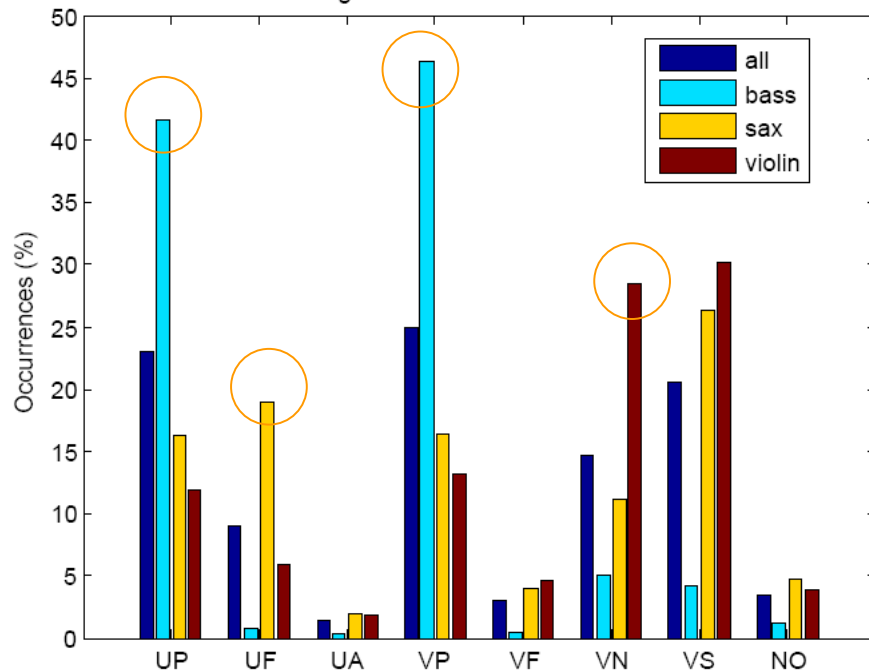
Musical background: novice (8), amateur (39), professional (13).

Language: romance (43), anglo-germanic (15) slavic (1) and other(1).

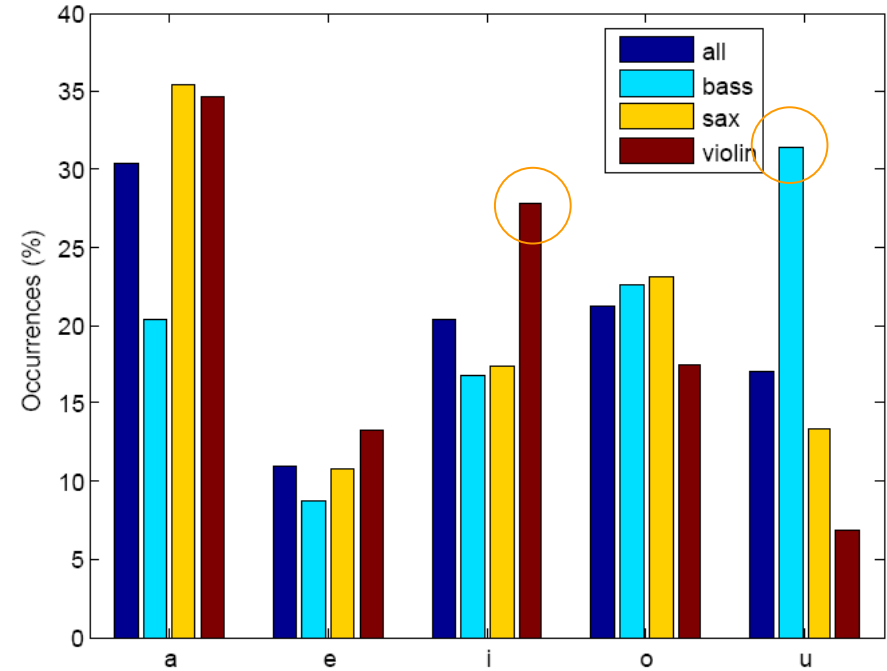
Number of transcribed phrases: 869.

studies on instrument imitation

Begin consonant - INSTRUMENT



Vowel - INSTRUMENT



Assumption:

1 syllable = consonant + vowel + consonant

UP: unvoiced plosive

UF: unvoiced fricative

UA: unvoiced affricative

NO: none

VP: voiced plosive

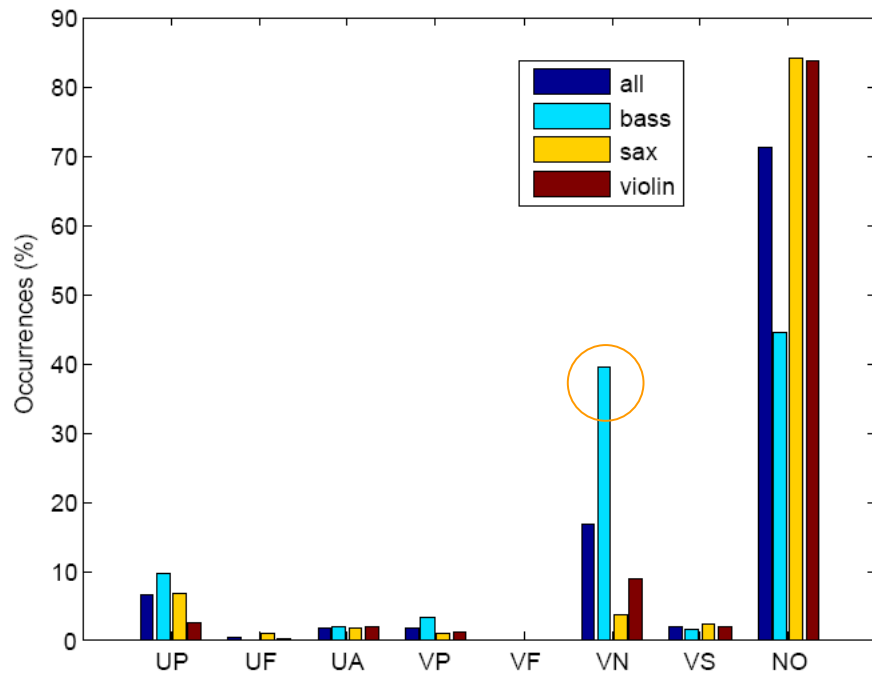
VF: voiced fricative

VN: voiced nasal

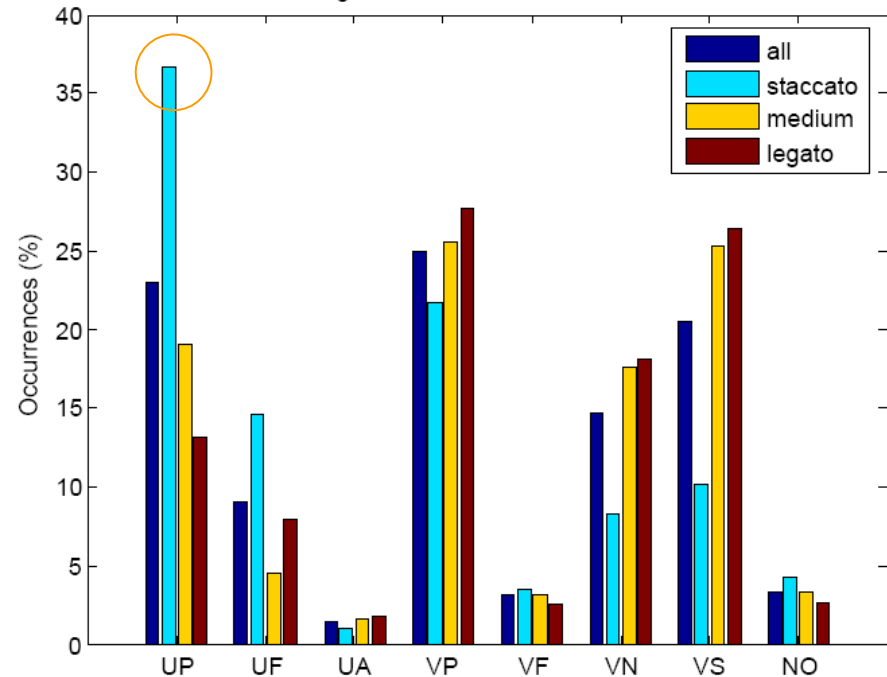
VS: voiced semivowel

studies on instrument imitation

End consonant - INSTRUMENT



Begin consonant - ARTICULATION



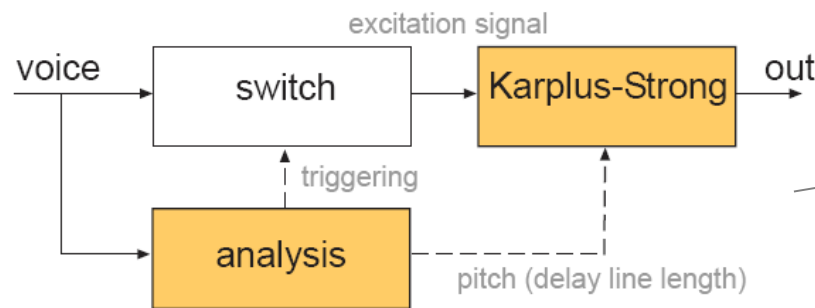
UP: unvoiced plosive
 UF: unvoiced fricative
 UA: unvoiced affricative
 NO: none

VP: voiced plosive
 VF: voiced fricative
 VN: voiced nasal
 VS: voiced semivowel

preliminary technical experiments

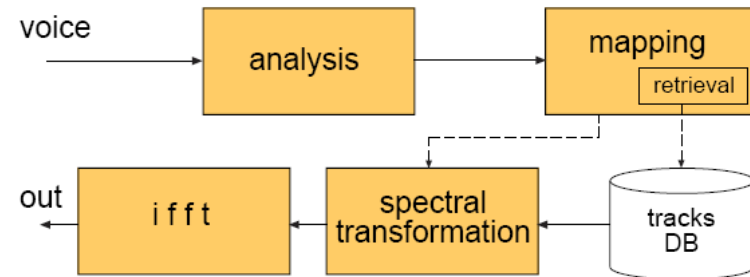
preliminary technical experiments

2. Comparing techniques for bass guitar sound synthesis



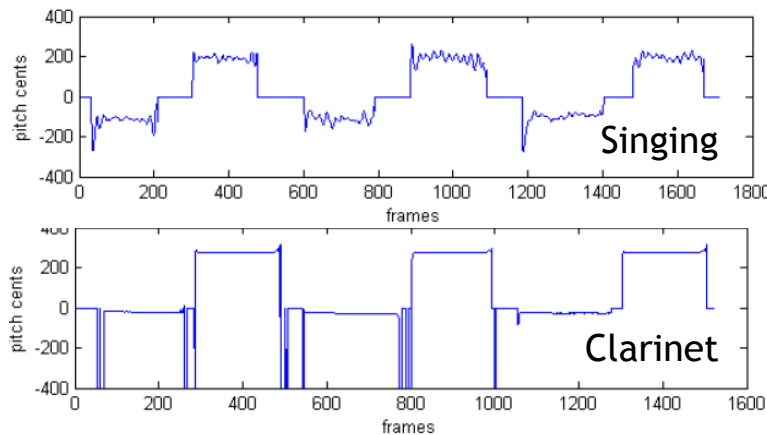
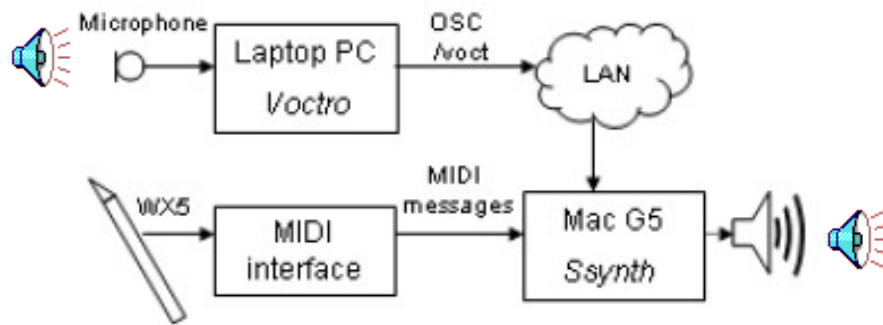
Physical modeling

*Sample-based
spectral transformation*



preliminary technical experiments

2. Comparing mappings for clarinet sound synthesis



• Experiment:

- **Task:** brightness control independently from loudness

- **Subjects:**

- singers → centroid
- saxophonists → lip press.

• Results:

- mappings are instrument specific
- task not intuitive for classical singers
- centroid is related to loudness

Collaboration with McGill University. 2005

27

Singing voice as a control signal: discussion

- Interesting path to explore for controlling DMI's
- Pitch-to-MIDI present several limitations:
 - Algorithms not designed for voice input
 - Decoupling from synthesis engine
- Sample-based with transformation is more flexible
- Phonetics have a musical function in instrument imitation

outline

I: STUDY

- Motivation, context and objectives
- Interacting with digital musical instruments
- Singing Voice as a control signal

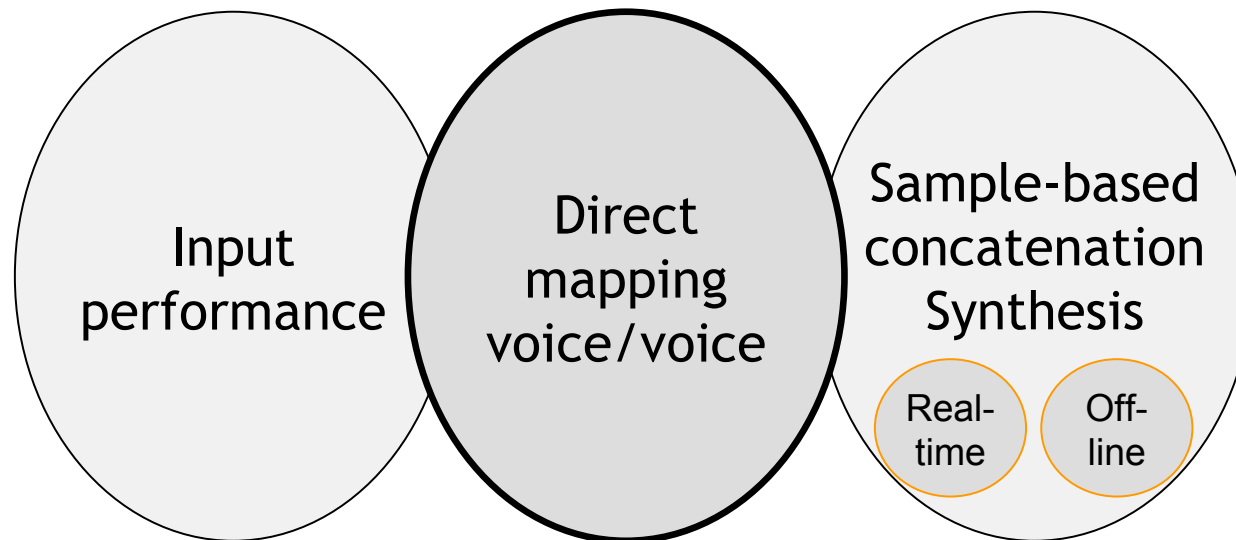
II: EXPERIMENT

- Controlling a singing voice synthesizer
- Vocal gestures analysis
- Controlling an instrumental sound synthesizer

III: RESULTS

- Prototypes
- Conclusions

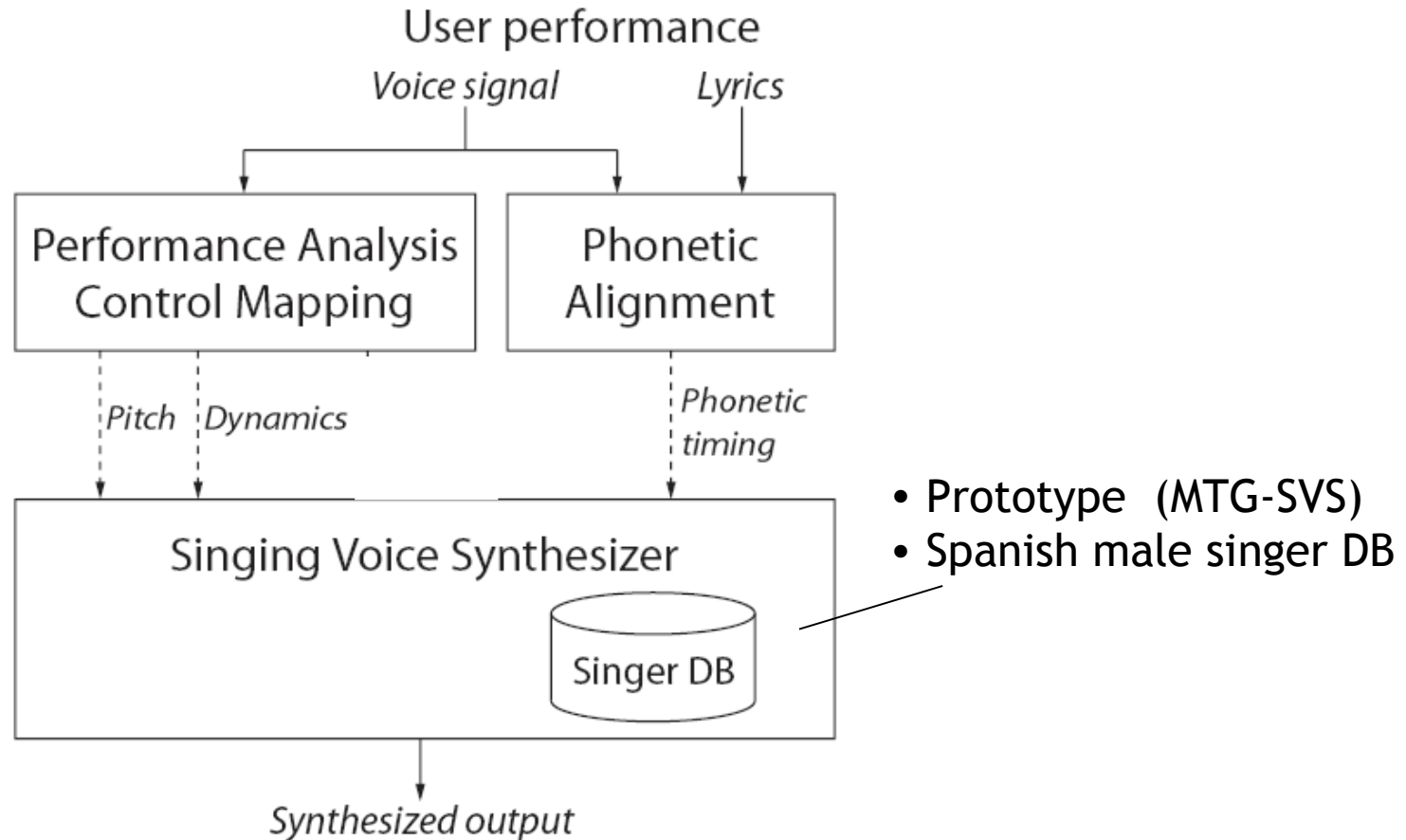
controlling a singing voice synthesizer (SVS)



- **Goals:**
 - improve the naturalness
 - facilitate the interaction/usage

- Use a SVS prototype based on performance sampling and spectral concatenation of diphones (Bonada and Serra, 2007)
- Two different approaches are proposed
 - off-line (OL) and real-time (RT)
- Presents control constraints:
 - Fixed duration of consonant phonemes → OL/RT
 - Latency in phoneme transitions due to the diphone samples → RT
 - Slow-varying control envelopes → OL/RT

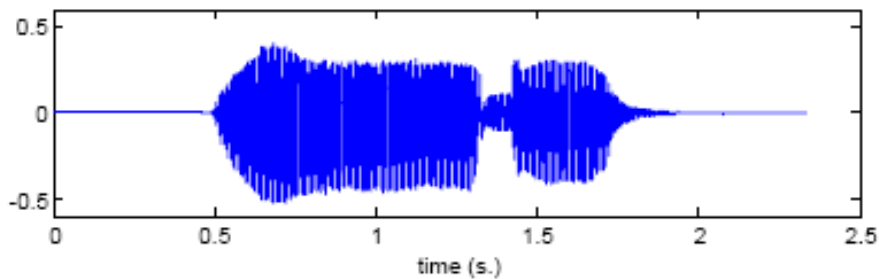
system overview



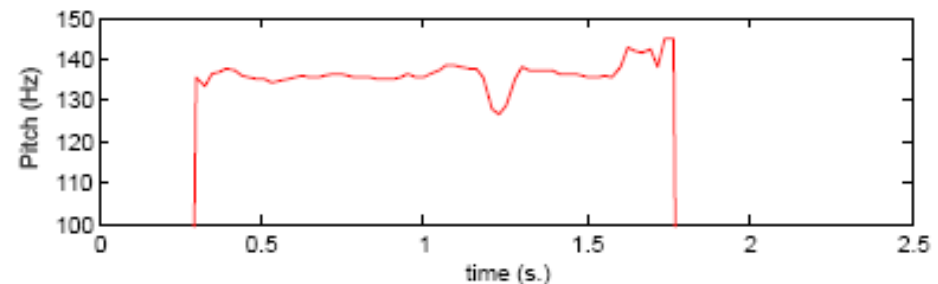
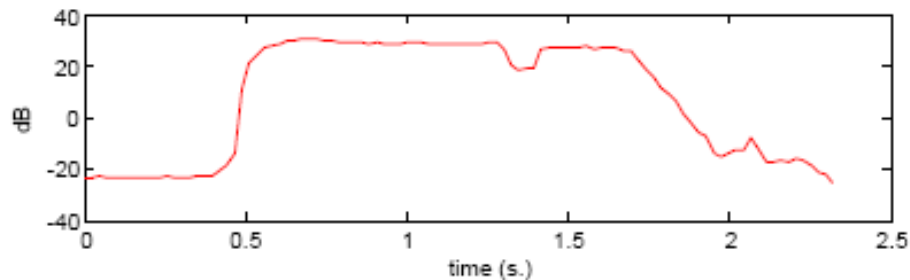
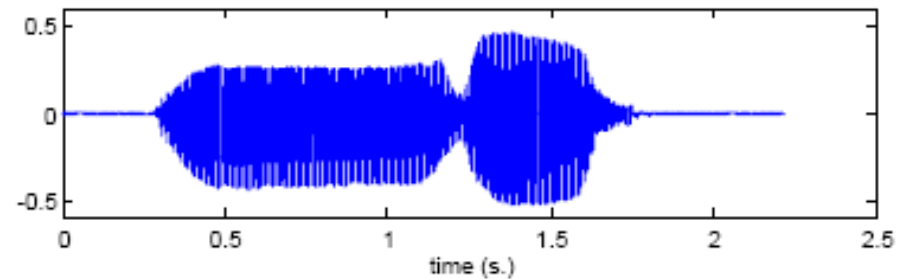
performance analysis

Constraint: variations due to phonetics already present in sample
Goal: to filter variations but keeping expression (e.g. vibrato)


Energy measurement



Pitch estimation

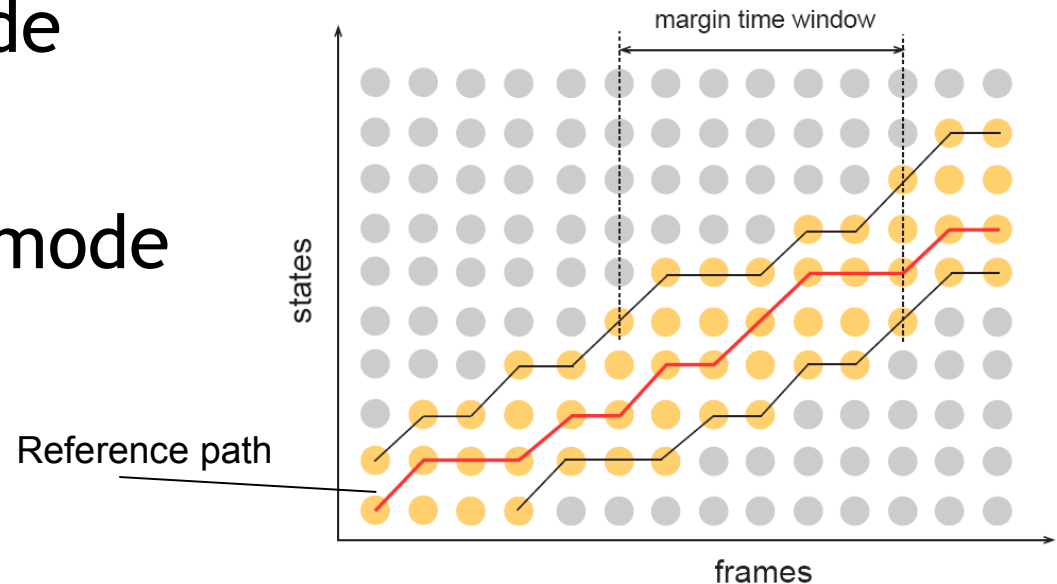
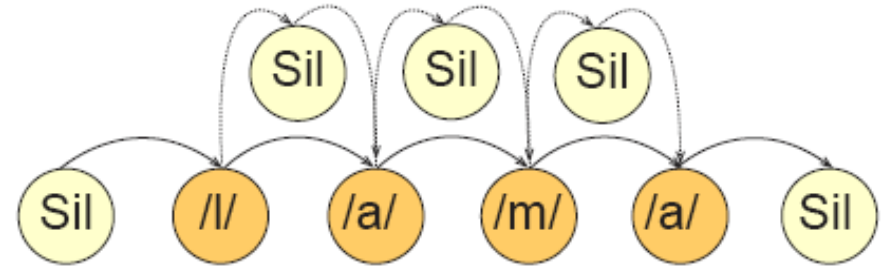


/ama/ 
 dynamics

/aga/ 
 pitch

phonetic alignment

- Modifications:
 1. insertion of silences
 3. low-delay mode
 5. synchronized mode



phonetic alignment

- Evaluation:
 1. Data set: 116 examples (~27.8 phon.)

2. Offline

| <i>Configuration</i> | <i>Accuracy (%)</i> | <i>observations</i> |
|----------------------------------|---------------------|----------------------------------|
| Viterbi(Julius) | 76.91 | Rate: 10 ms DMFCC(2 frames) |
| Viterbi (Singterface) | 73.66 | Rate: 11.6 ms DMFCC(2 frames) |

5. Real-time







| <i>Configuration</i> | <i>Overl.. dist (%)</i> | <i>observations</i> |
|----------------------|-------------------------|----------------------------------|
| LowDelay-2 | 70.34 | Rate: 11.6 ms DMFCC(2 frames) |
| LowDelay-0 | 68.36 | Rate: 11.6 ms DMFCC(0 frames) |

assessment

- Examples:

- Male singer (DB): input  synthesis 
- Female singer: input  synthesis 

- Perceptual experiment:

1. Voice input  
2. Violin input  
3. Piano input  

| N = 5 | <i>Voice</i> | <i>Violin</i> | <i>Piano</i> |
|-------------------------|--------------|---------------|--------------|
| Naturalness | 2.8 | 1.4 | 2.4 |
| Correctly ident. | 3/5 | 3/5 | 5/5 |

controlling a singing voice synthesizer: discussion

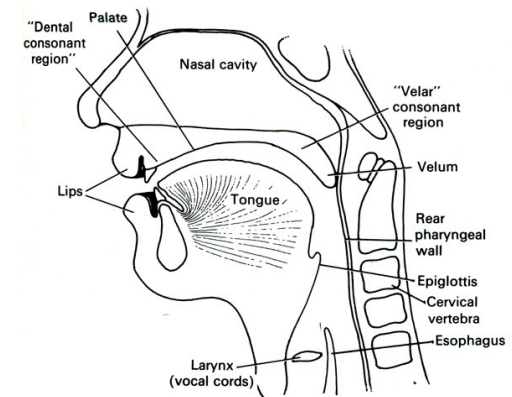
- good results in offline mode:
 - adds naturalness to SVS
 - facilitates creation process
- ✂ → improvements for real-time mode:
 - Robustness of phonetic alignment
 - Spanish acoustic models
 - Reduce the synthesis latency
 - different synthesis technique (frame based)

vocal gestures analysis

vocal gestures analysis

- In Phonetics:
 - movement of the articulators
- Instrumental gestures:
 - movement of a performer to produce sound

✂️ → describe vocal imitation of instruments



vocal gestures representation

- Classification (Cadoz, 1988):

| Instrumental Gest. | Vocal Gest. | direct acquisition |
|--------------------|--|---|
| Excitation → | loudness | airflow press. |
| Modulation → | fundamental freq. formants breathiness | vocal folds vibra. articulators pos. phonation degree |
| Selection → | phonetic class | |

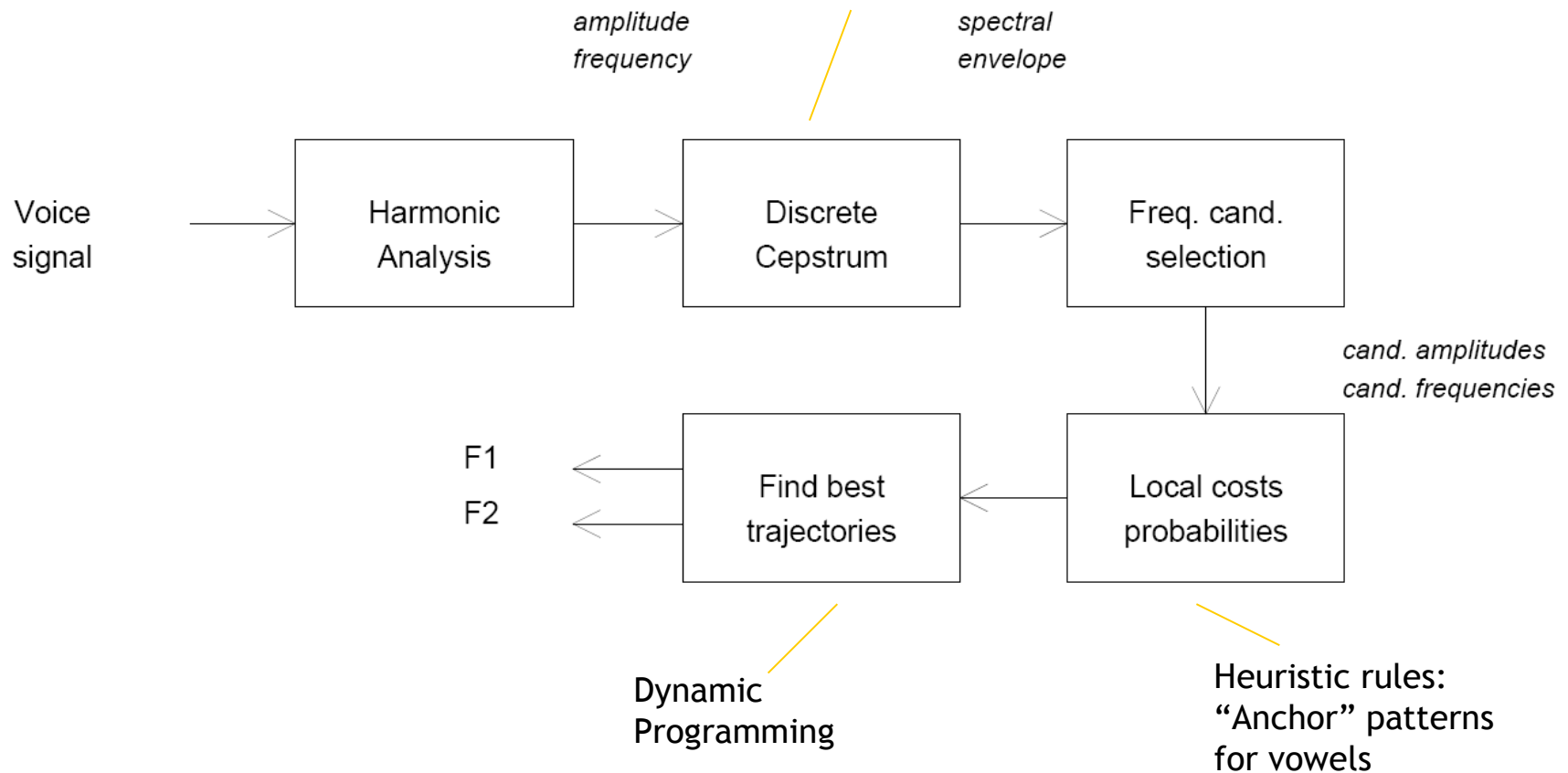
vocal gestures representation

- Classification (Cadoz):

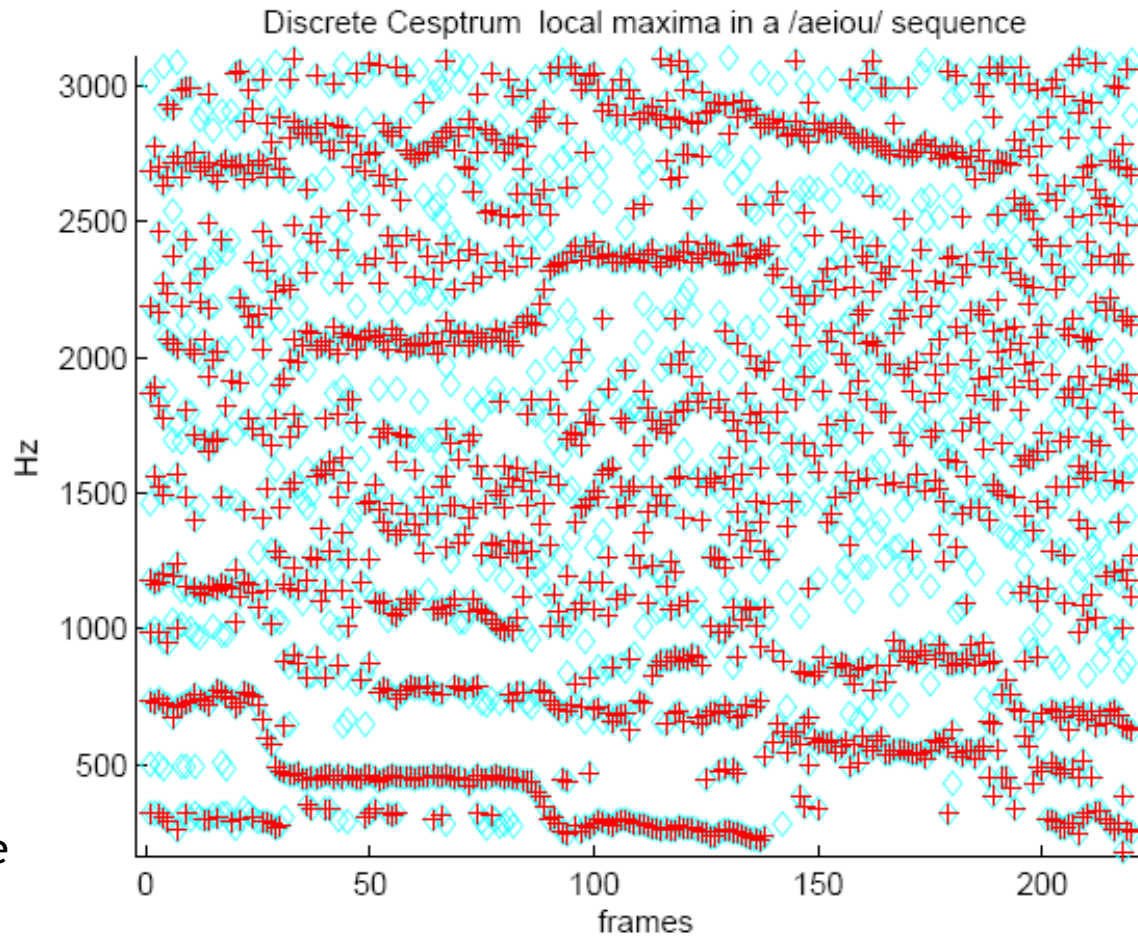
| Instrumental Gest. | | Vocal Gest. | direct acquisition |
|--------------------|---|--|---|
| Excitation | → | loudness | airflow press. |
| Modulation | → | fundamental freq. formants breathiness | vocal folds vibra. articulators pos. phonation degree |
| Selection | → | phonetic class | |

formant tracking

(Galas, Rodet, 1990)



formant tracking



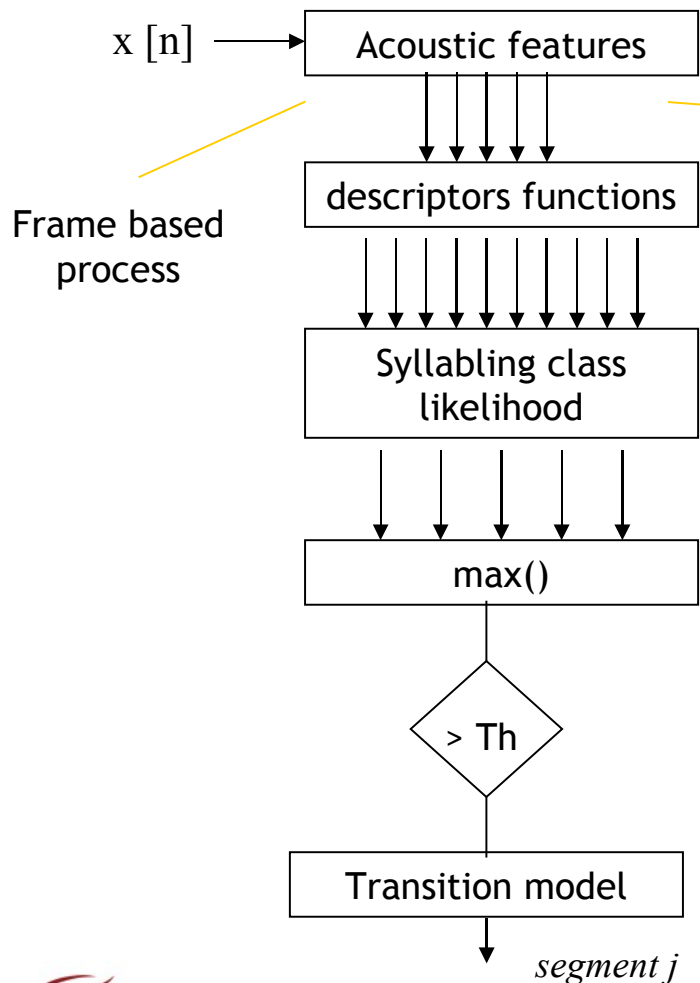
/aeiou/ example

syllabling segmentation

| CLASS | PHONEMES |
|---------------------------------|------------------------------|
| <i>Speech Phon. classes</i> | |
| Vowels | [a] , [e] , [i], [o], [u] |
| Plosive | [p], [k], [t], [b], [g], [d] |
| Liquids and glides | [l], [r], [w], [y] |
| Fricatives | [s], [x],[T], [f] |
| Nasal | [m], [n], [J] |
| <i>Syllabling Phon. classes</i> | |
| Attack | [p], [k], [t], [n], [d], [l] |
| Sustain | [a], [e], [i], [o], [u] |
| Transition | [r], [d], [l], [m], [n] |
| Release | [m], [n] |
| Special | [s],[x],[T], [f] |

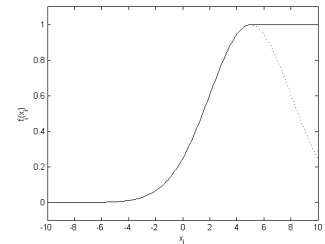
- Aim: to segment according to the musical function

syllabling segmentation

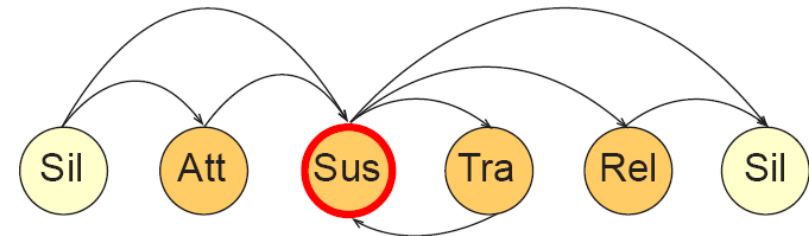


energy, MFCC, zero-cross, pitch,...

$$f_i(x_i) = \begin{cases} e^{-\frac{(x_i - \mu_i)^2}{\sigma_i^2}} & , x_i > \mu_i \\ 1 & , x_i \leq \mu_i \end{cases}$$



$$B_j = \prod_i f(x_i)_i^{\alpha_i}$$



Phonetic example /t/ /a/ /r/ /m/

syllabing segmentation

- Evaluation: annotated ground truth (94 rec.)

| SINGTERFACE | Mean | Std |
|---------------------|-------------|-------|
| Correct detect (%) | 90.68 | 15.38 |
| False positives (%) | 13.99 | 52.86 |
| Abs. Deviation (ms) | 8.93 | 9.85 |

| ESSENTIA | Mean | Std |
|---------------------|--------------|-------|
| Correct detect (%) | 89.20 | 14.56 |
| False positives (%) | 11.77 | 61.84 |
| Abs. Deviation (ms) | 22.78 | 14.04 |

| AUBIO | Mean | Std |
|---------------------|--------------|--------|
| Correct detect (%) | 96.81 | 8.67 |
| False positives (%) | 109.81 | 105.69 |
| Abs. Deviation (ms) | 13.93 | 11.83 |

| MAMI | Mean | Std |
|---------------------|-------|-------|
| Correct detect (%) | 90.59 | 15.92 |
| False positives (%) | 19.99 | 53.59 |
| Abs. Deviation (ms) | 16.56 | 8.38 |

| JULIUS | Mean | Std |
|---------------------|-------|--------|
| Correct detect (%) | 71.85 | 34.28 |
| False positives (%) | 92.38 | 163.25 |
| Abs. Deviation (ms) | 15.58 | 11.90 |

Speech recognition test
Vowels are considered onsets

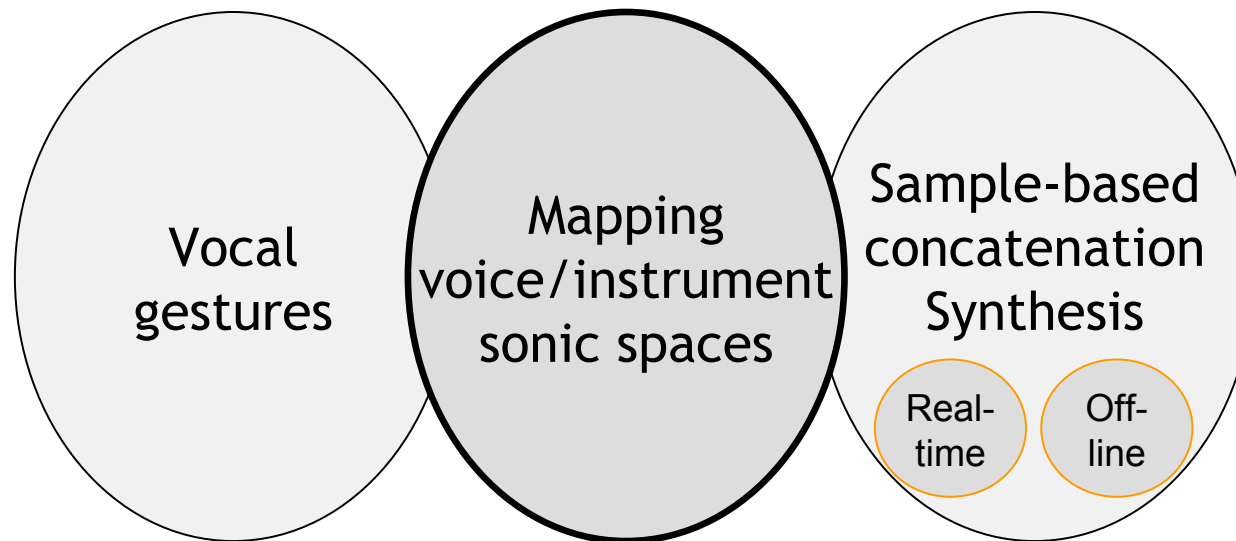
vocal gestures analysis: discussion

- vocal gestures are proposed to describe vocal imitation of instruments
- novel algorithms are presented

→ improvements:

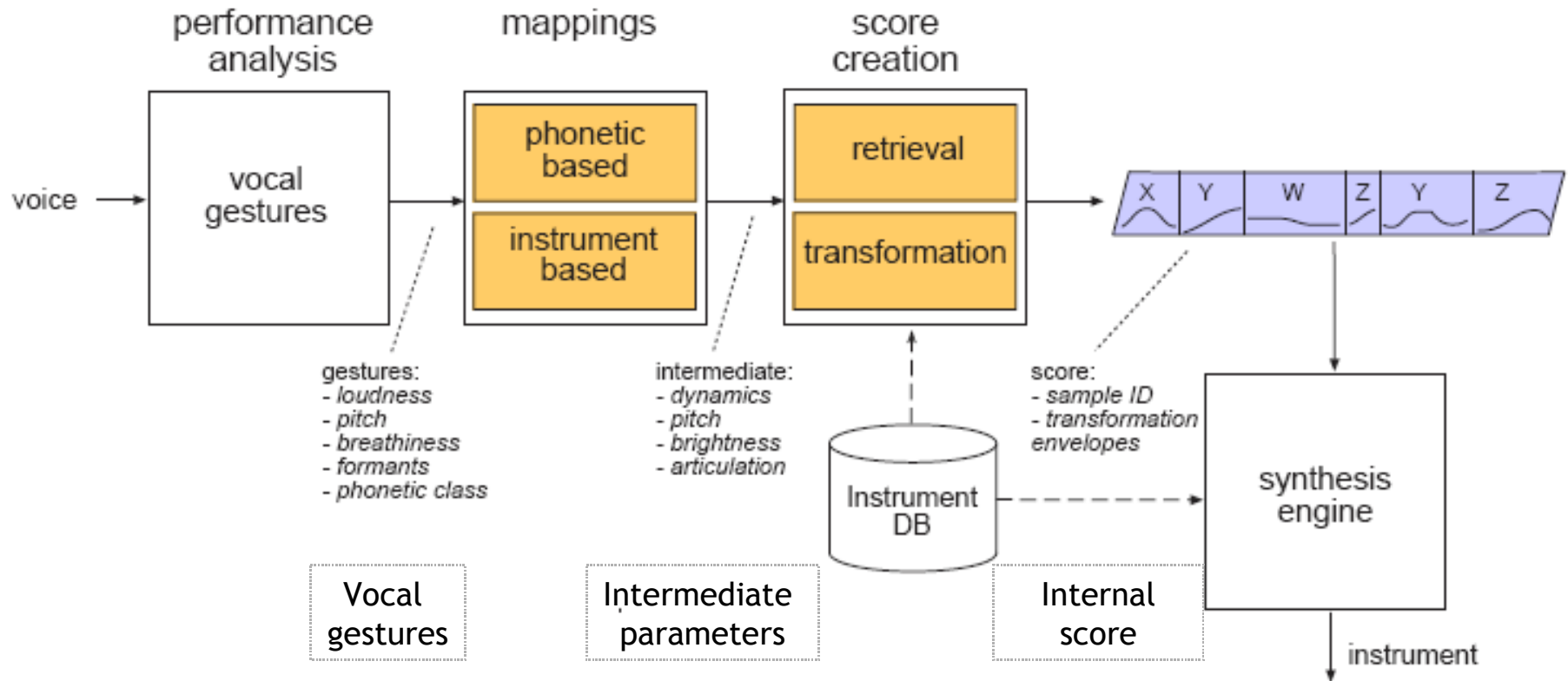
- formants:
 - Improve accuracy in high pitched voices
- segmentation:
 - Identify *Release* syllabbling class (nasals)

controlling an instrumental sound synthesizer



system overview

- Multi-layer mappings (Arfib, 2002):

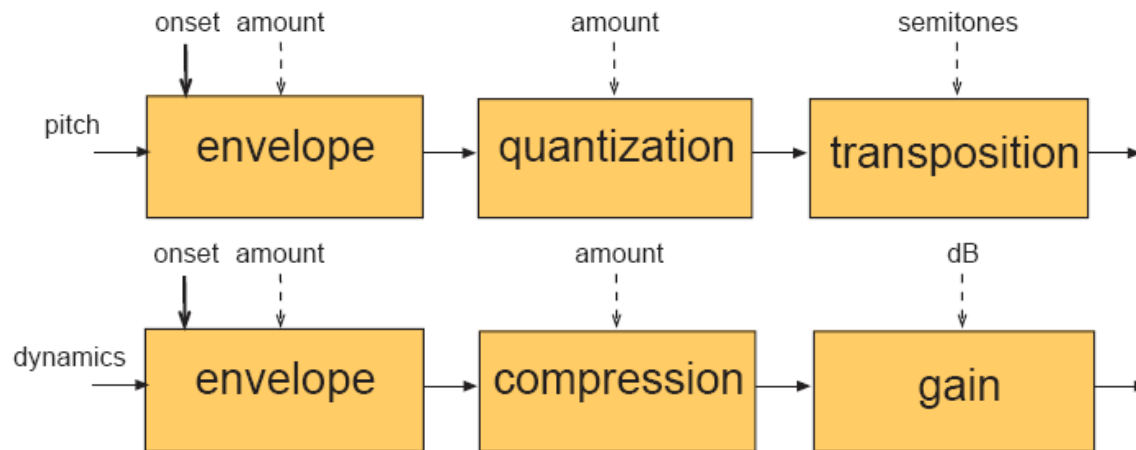


studies on mapping

- Phonetic-specific mappings:
 - intermediate parameter for articulation
 - Based on a classifier from the syllabling segmentation.
- Instrument-specific mappings:
 - Intermediate parameter for:
 - Pitch, dynamics, brightness, time-scaling
 - Set of functions: $g(x, i)$, x : parameter, i : instrument

studies on mapping

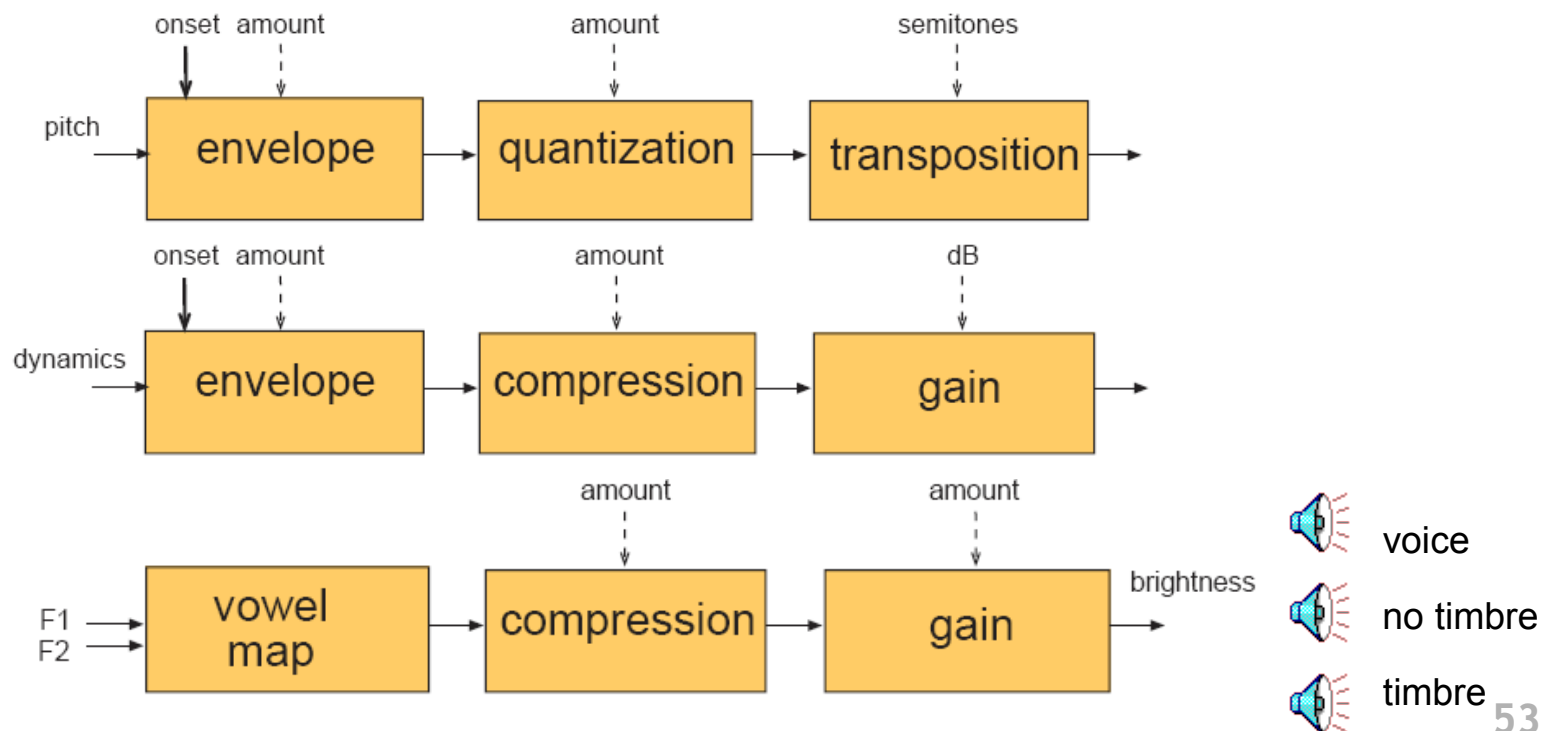
- Instrument-specific mappings:



Envelope function, $y[n] = (1 - a) \cdot y[n - 1] + a \cdot x[n]$
with $a = a(t)$


studies on mapping

- Instrument-specific mappings:



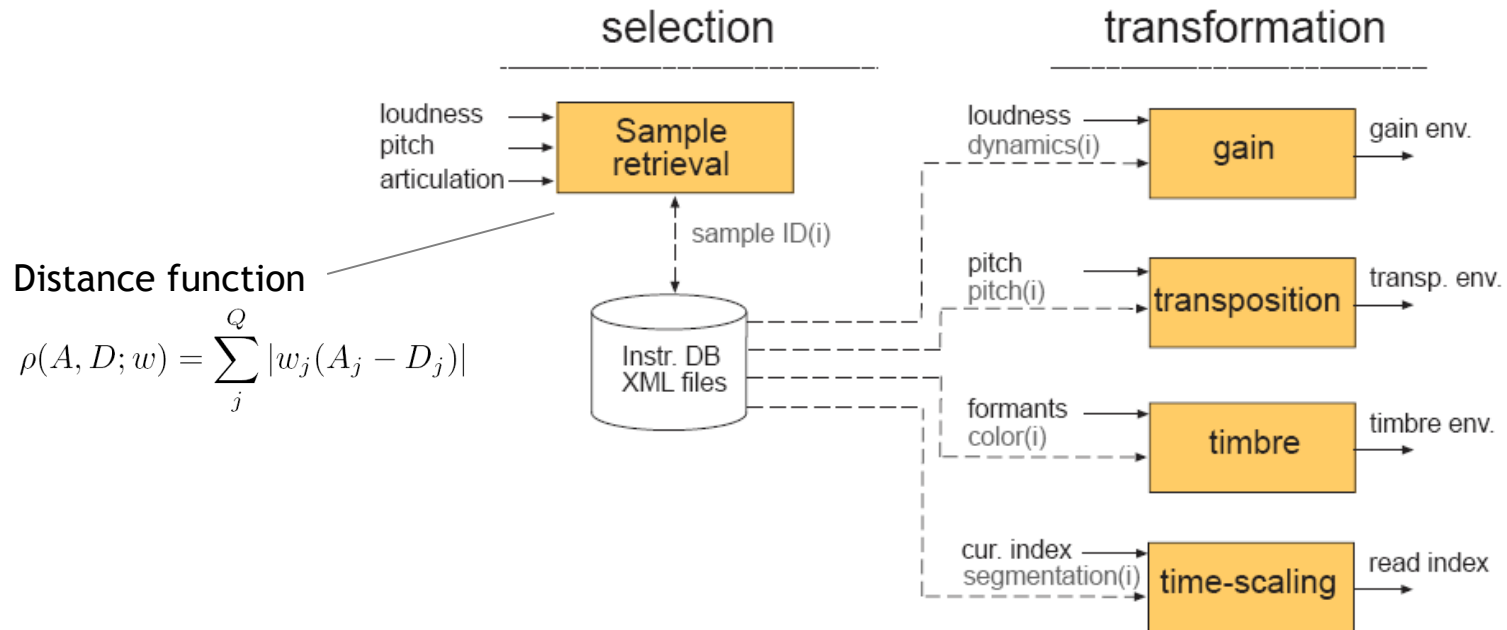
studies on mapping

- Instrument-specific mappings:
 - Time-scaling functions

|  <i>Time-scaling type</i> | |
|---|--|
| Sustain-stretch | $TS = \begin{cases} 1 & t < D_a^s \\ \frac{D_T^i - D_a^s - D_r^s}{D_s^s} & D_a^s < t < D_a^s + D_s^i \\ 1 & D_a^s + D_s^i < t < D_a^s + D_s^i + D_r^s \end{cases}$ |
| Sample-stretch | |
| Sustain-frame | |
| None | |

Real-time mode: looping strategy is employed


score creation



| <i>Intermediate Parameter</i> | <i>Type</i> | <i>Range</i> |
|-------------------------------|-------------|-----------------|
| Loudness | envelope | [0..1] |
| Pitch | envelope | [-12000..12000] |
| Brightness | envelope | [0..1] |
| Articulation | float | [0..1] |
| Duration | float | <i>seconds</i> |

| <i>Internal Score Parameter</i> | <i>Type</i> |
|---------------------------------|-------------|
| SampleID | integer |
| Gain | envelope |
| Transposition | envelope |
| Timbre mapping | envelope |
| Frame index | envelope |

assessment

- **Examples:**
 - violin synthesis 
- **Feedback from user tests:**
 - latency
 - limited synthesizer's sonic space
 - auditory feedback
 - learning curve

outline

I: STUDY

- Motivation, context and objectives
- Interacting with digital musical instruments
- Singing Voice as a control signal

II: EXPERIMENT

- Controlling a singing voice synthesizer
- Vocal gestures analysis
- Controlling an instrumental sound synthesizer

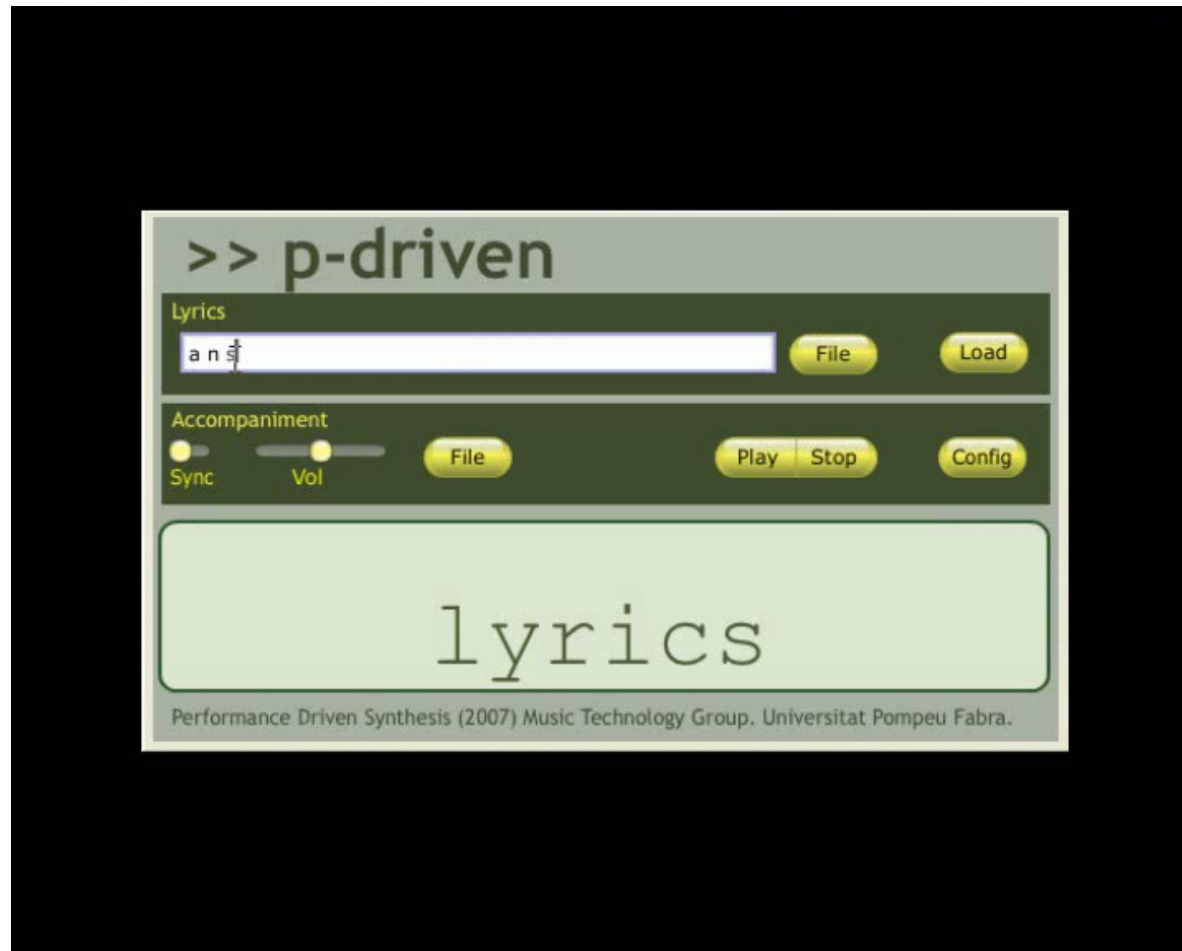
III: RESULTS

- Prototypes
- Conclusions

prototypes: *p-driven*

- performance-driven singing synthesis
 - VST plugin implemented in C++
 - Offline and real-time operation
 - User inputs lyrics
- latency:
 - Audio I/O: 11.6 ms
 - Alignment: 11.6 ms
 - Synthesis: 70 - 150 ms

prototypes: *p-driven*



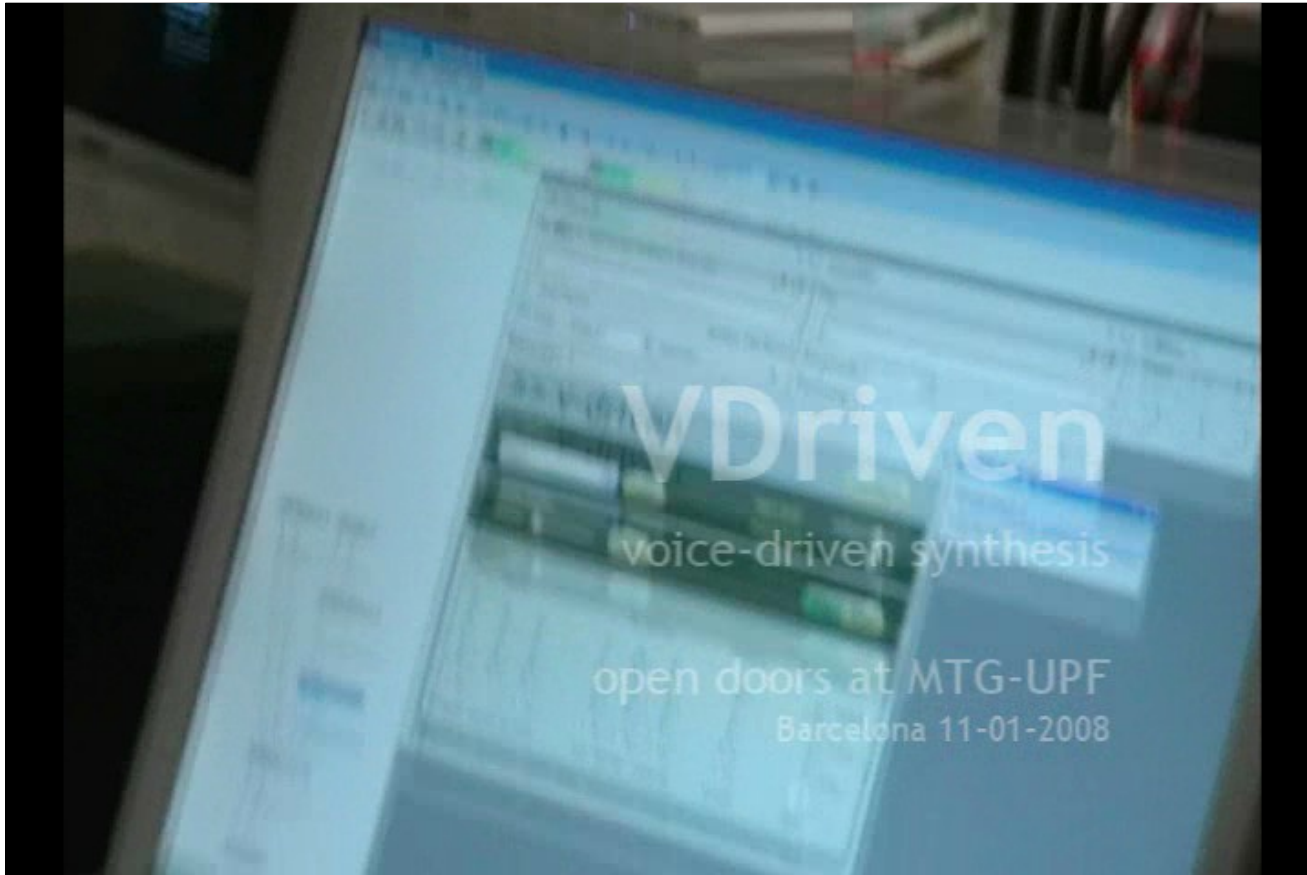
prototypes: *v-driven*

- voice-driven synthesis
 - VST plugin implemented in C++
 - Offline and real-time operation
 - User configures mappings and instrument
 - OSC output
- latency:
 - Audio I/O: 11.6 ms
 - Internal processing: 29.02 ms

v-driven performance

saxophone
sound

v-driven installation



thesis contributions

- Study of vocal imitation of musical instruments
- Low-delay Phonetic Alignment
- Formant Tracking Algorithm
- Syllabling Segmentation Algorithm
- Mapping strategies for voice to instrument
- Two prototypes

thesis discussion

The **control of DMI's with the singing voice** is a possible and interesting way of interaction


Instrument imitation is an area that should be further explored in other domains (musicology).

Vocal gestures representation permits to describe instrument imitation signals and to define structured mappings

Naturalness of singing voice synthesis can be improved when driven by an input performance

Instrumental sound synthesis requires of high-quality synthesizers to get realistic results.

future work

- Context aware voice-driven synthesis
- Timbre-oriented voice-driven synthesis
- Voice-driven sound retrieval  *
- Cross-instrumental auditory feedback
- Generalization instrument-driven synthesis

* Sound example of voice-driven audio mosaicing
(generated with BeatMash)

publications

- Janer, J. 2008. '*Analysis of vocal gestures for voice-driven synthesis*'. (in preparation)
- Janer, J. Maestre, E. 2008. '*Mapping phonetic features for voice-driven sound synthesis*', EBusiness and Telecommunication Networks - Selected papers from ICETE 2007. CCIS Series. Springer-Verlag.
- Janer, J. Maestre, E. 2007. '*Phonetic-based mappings in voice-driven sound synthesis*', Proceedings of International Conference on Signal Processing and Multimedia Applications - SIGMAP 2007; Barcelona, Spain.
- Janer, J. Peñalba, A. 2007. '*Syllabling on instrument imitation: case study and computational methods*', Proceedings of 3rd Conference on Interdisciplinary Musicology; Tallinn, Estonia.
- Janer, J. Bonada, J. Blaauw, M. 2006. '*Performance-driven control for sample-based singing voice synthesis*', Proceedings of 9th International Conference on Digital Audio Effects; Montreal, Canada.
- Janer, J. Bonada, J. Jordà, S. 2006. '*Groovator - an implementation of real-time rhythm transformations*', Proceedings of 121st Convention of the Audio Engineering Society; San Francisco, CA, USA.

publications

- Janer, J. 2005. '*Feature Extraction for Voice-driven Synthesis*', Proceedings of 118th Audio Engineering Society Convention; Barcelona.
- Janer, J. 2005. '*Voice-controlled plucked bass guitar through two synthesis techniques*', Proceedings of 2005 International Conference on New Interfaces for Musical Expression; Vancouver, Canada.
- Janer, J. Loscos, A. 2005. '*Morphing techniques for enhanced scat singing*', Proceedings of 8th Intl. Conference on Digital Audio Effects; Madrid, Spain.
- Fabig, L. Janer, J. 2004. '*Transforming Singing Voice Expression - The Sweetness Effect*', Proceedings of 7th International Conference on Digital Audio Effects; Naples, Italy.
- Janer, J. 2004. '*Voice as a musical controller for real-time synthesis*', Doctoral Pre-Thesis Work. UPF. Barcelona.

- Website: <http://www.mtg.upf.edu/~jjaner/phd>

thanks to:

I wish to thank specially my colleagues Jordi Bonada, Esteban Maestre, Merlijn Blaauw, Alex Loscos, Maarten de Boer, Alfonso Perez, Sergi Jordà, Lars Fabig, Oscar Mayor and Perfecto Herrera. Thanks to Graham Coleman, Greg Kellum and Paul Brossier for reviewing the written document. Thanks to Jose Lozano, Oscar Celma, Inês Salselas, Lucas Vallejo, Amaury Hazan, Ricard Marxer and Jose Pedro García-Mahedero for the recording sessions.

Jordi Janer, 2008

The dissertation is published under



end