Natural **transformation** of type and nature of the **voice** for extending vocal repertoire in **high-fidelity applications**

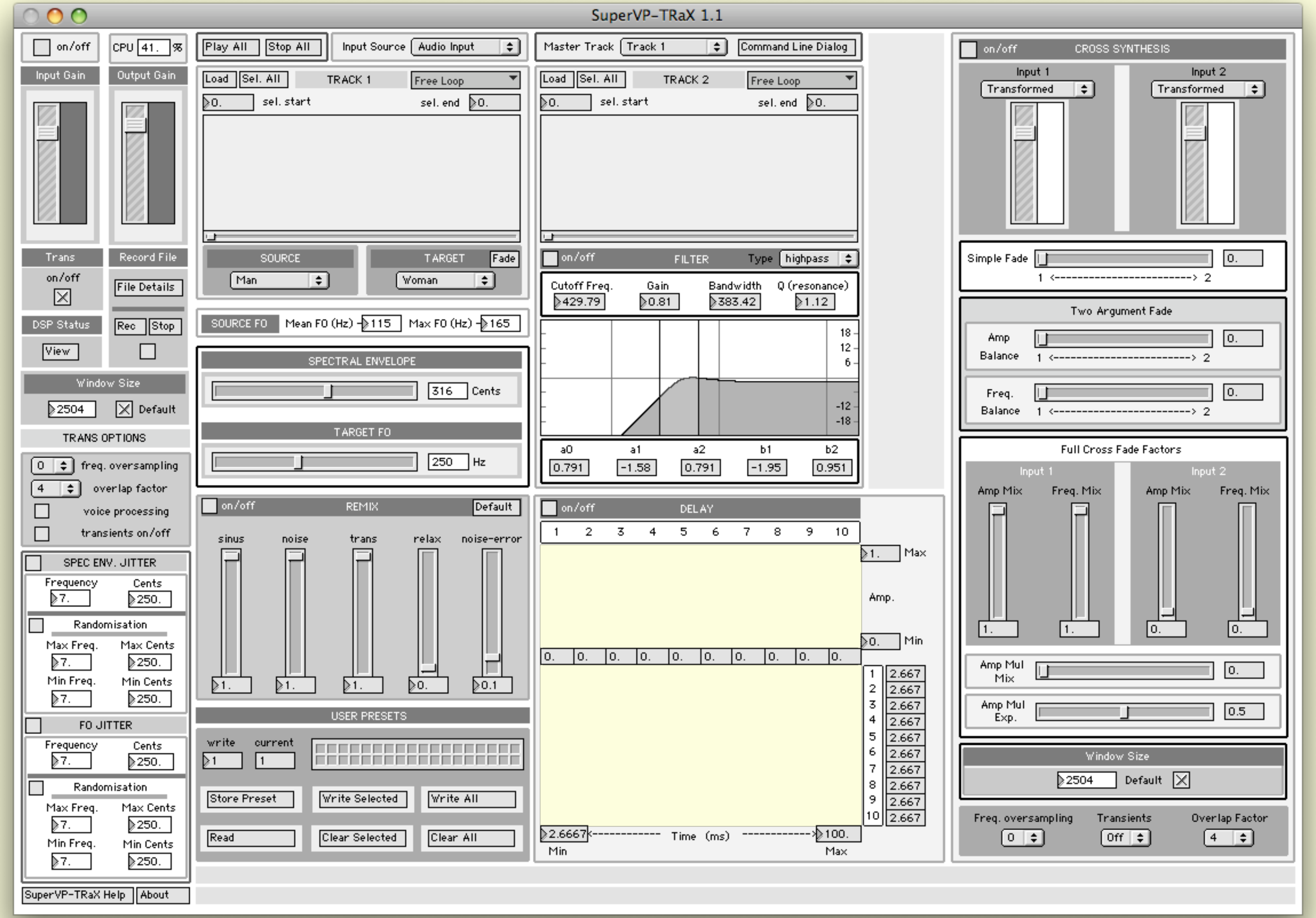
Snorre Farner,
Axel Röbel and Xavier Rodet

ircam
Centre Pompidou

Analysis/Synthesis Group
Paris, France

Funding:
AGENCE NATIONALE DE LA RECHERCHE
ANR
Affective Avatars

Demo: Real-time voice transformation


SuperVP-TRaX 1.1

□ on/off | CPU 41. % | Play All | Stop All | Input Source: Audio Input | Master Track: Track 1 | Command Line Dialog

Input Gain | Output Gain

TRACK 1: Load | Sel. All | Free Loop | sel. start 0. | sel. end 0.

TRACK 2: Load | Sel. All | Free Loop | sel. start 0. | sel. end 0.

Trans on/off ☒ | Record File | File Details | DSP Status | View | Rec | Stop □

SOURCE: Man | TARGET: Woman | Fade

SOURCE F0: Mean F0 (Hz) 115 | Max F0 (Hz) 165

SPECTRAL ENVELOPE: 316 Cents

TARGET F0: 250 Hz

Window Size: 2504 | ☒ Default

TRANS OPTIONS: 0 freq. oversampling | 4 overlap factor | □ voice processing | □ transients on/off

□ SPEC ENV. JITTER: Frequency 7. | Cents 250.

□ Randomisation: Max Freq. 7. | Max Cents 250. | Min Freq. 7. | Min Cents 250.

□ F0 JITTER: Frequency 7. | Cents 250.

□ Randomisation: Max Freq. 7. | Max Cents 250. | Min Freq. 7. | Min Cents 250.

□ on/off REMIX | Default: sinus 1. | noise 1. | trans 1. | relax 0. | noise-error 0.1

USER PRESETS: write 1 | current 1 | Store Preset | Write Selected | Write All | Read | Clear Selected | Clear All

□ on/off FILTER | Type highpass

Cutoff Freq.	Gain	Bandwidth	Q (resonance)
429.79	0.81	383.42	1.12

a0	a1	a2	b1	b2
0.791	-1.58	0.791	-1.95	0.951

□ on/off DELAY: 1 2 3 4 5 6 7 8 9 10 | Amp. Max 1. | Min 0. | 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.

1	2.667
2	2.667
3	2.667
4	2.667
5	2.667
6	2.667
7	2.667
8	2.667
9	2.667
10	2.667

Time (ms): Min 2.6667 | Max 100.

□ on/off CROSS SYNTHESIS: Input 1 Transformed | Input 2 Transformed

Simple Fade 0. (1 <------------------------> 2)

Two Argument Fade: Amp Balance 0. (1 <------------------------> 2) | Freq. Balance 0. (1 <------------------------> 2)

Full Cross Fade Factors: Input 1 Amp Mix 1. | Freq. Mix 1. | Input 2 Amp Mix 0. | Freq. Mix 0.

Amp Mul Mix 0. | Amp Mul Exp. 0.5

Window Size 2504 | Default ☒

Freq. oversampling 0 | Transients Off | Overlap Factor 4

SuperVP-TRaX Help | About

AES 35, London 2009 - Farner, Röbel, Rodet: Natural voice transformations

Overview

- Real-time demonstration
- Introduction:
  - motivation, applications, background
- Transformation of the voice:
  - gender and age, voice quality, expressivity
- Perceptive evaluation
  of transformation of gender and age
- Conclusion

Introduction 1

- Why transform the voice in games?
  - speech synthesis: avoid prerecorded utterances
  - voice transformation: avoid databases of many actors
  - enrichen voice repertoire for narrators and NPCs
  - design the voice of a game character based on the player’s voice in multiplayer role-playing games

- Other applications:
  - educational games, e-learning, “serious games”
  - music, multimedia, audiobooks, story telling,...
  - films, dubbing, cartoon characters,...

Introduction 2

- Ircam's objectives: artistic applications
  - music composition and composition tools

    => speech processing

    => voice and instrument transformation

- Requirements:
  - very high sound quality
  - very high degree of naturalness
  - automatic solution
  - real-time user control

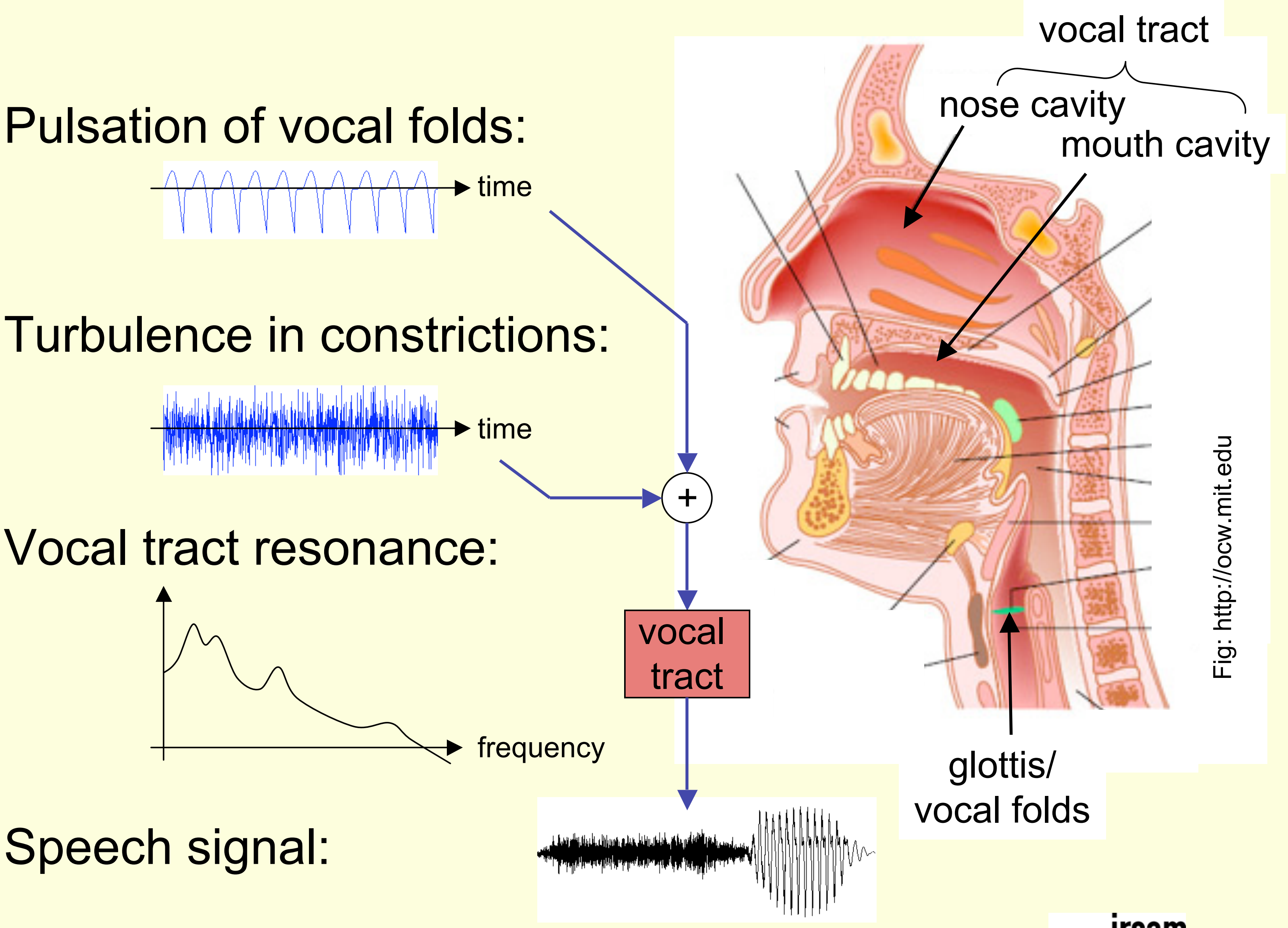
Voice transformation today

Two basic concepts:

- Voice conversion: from voice A to voice B
  - often need parallel recordings of A and B
  - learning of differences between A's and B's voices
  - a new phrase of A can then be converted to B's voice
  - artifacts such as non-uniform vocal timbre

- Voice transformation: modification of general acoustic properties of the voice to transform
  - gender and age
  - voice quality: breathy voice, whispering, more or less timbred,...
  - expressivity,...

The voice



Pulsation of vocal folds:

Turbulence in constrictions:

Vocal tract resonance:

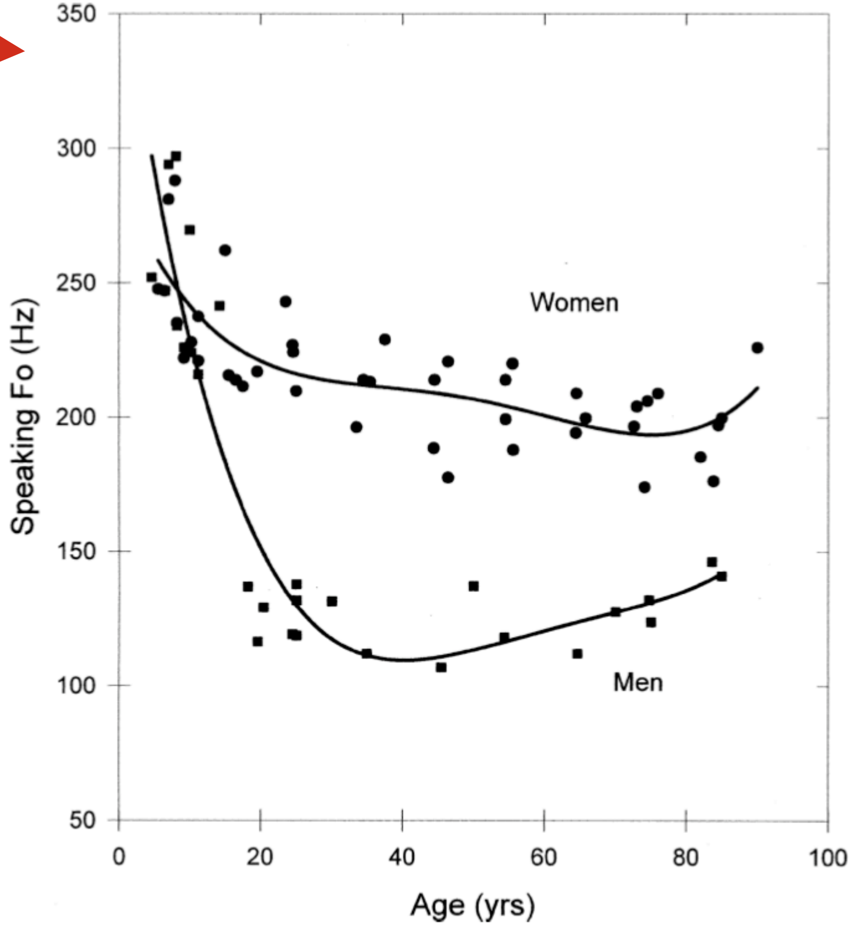
Speech signal:

Dependencies on gender and age

- General acoustic properties:

IrcamVoiceTrans

  - average F0
  - vocal tract; formants
  - pitch stability
  - ambitus ($F0_{max}$ - $F0_{min}$)
  - breathiness
  - speech rate
- prosody
- vocabulary
- linguistics

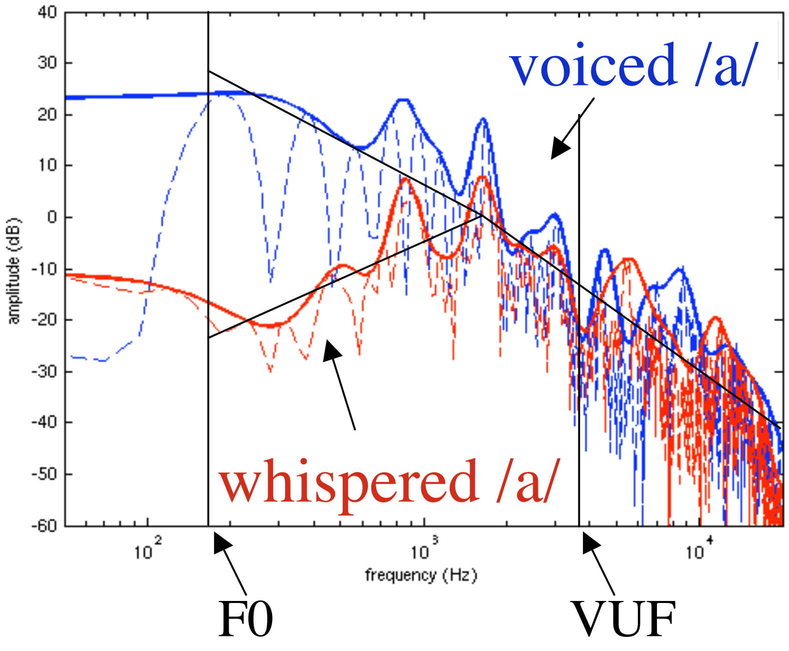


[Baken, J. Voice 2005]

Whispering and breathy voice

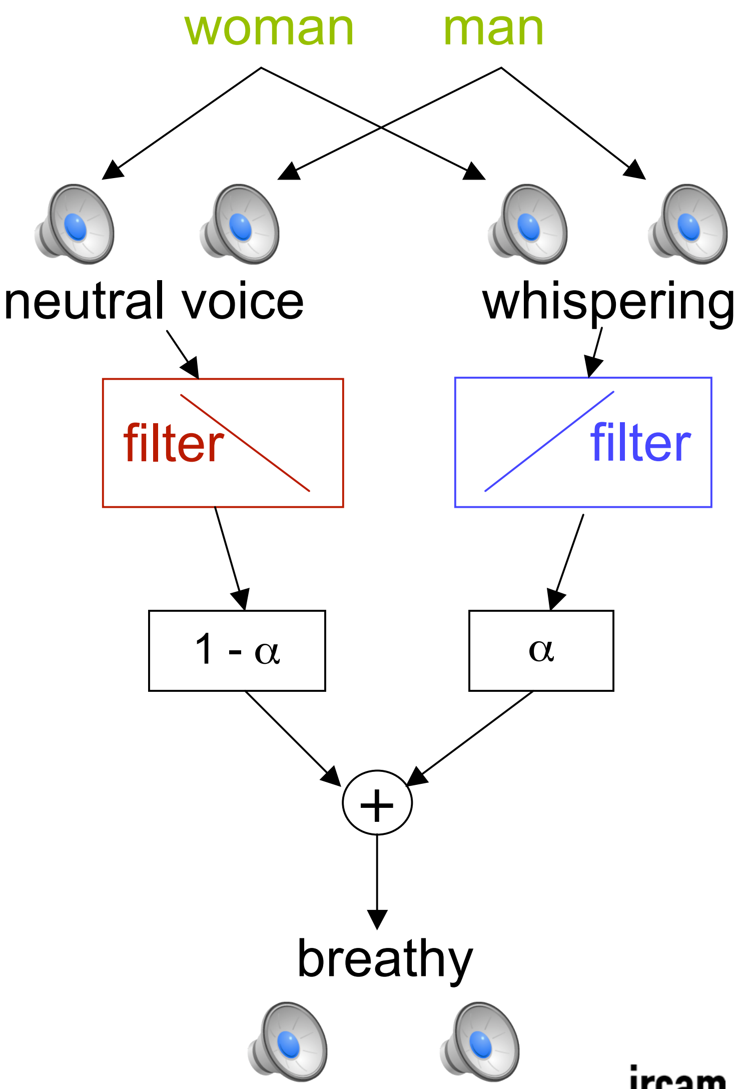
Whispering:

- filter noise by spectral envelope
- attenuate frequency bands that are voiced in original



Ex.: original: whisper:

Breathy voice:



woman man

neutral voice whispering

filter filter

1 - α α

+

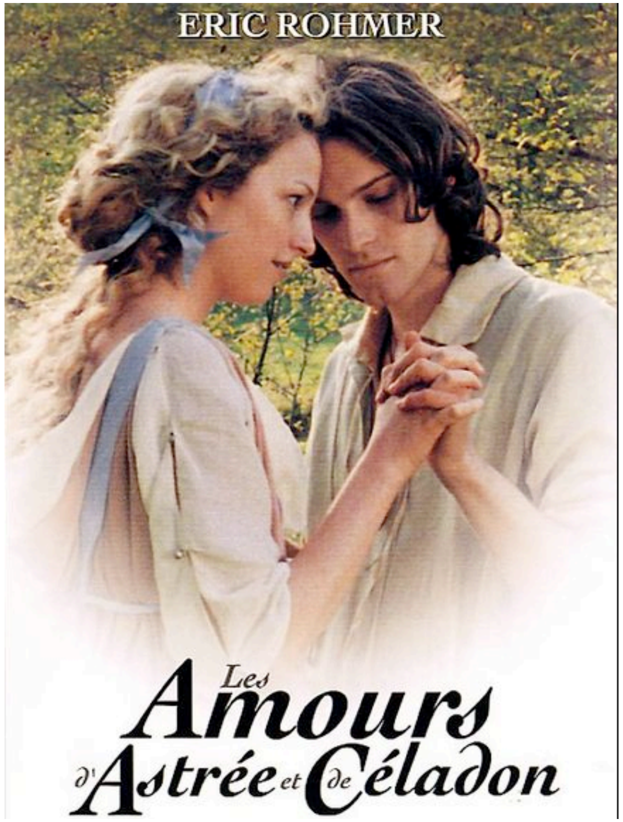
breathy

Transformation of identity

- Disguising man to woman:
  - ...also the voice: 🔈 → 🔈
  - Céladon 🔈 → Alexie 🔈

- Monologue → dialog
  🔈 🔈

- One actor to 12 persons:
  - 🔈→🔈 5th Blind (woman)
  - 🔈→🔈 Oldest Blind Woman
  - 🔈→🔈 Oldest Blind Man
  - 🔈→🔈 3rd Blind (man)



« Deux Songes
de Maeterlinck d'après Brueghel »
by J. B. Barrière, 2007

Examples: transformation of voice quality

- original: 🔊

- breathy: 🔊
- whispering: 🔊
- creaky: 🔊

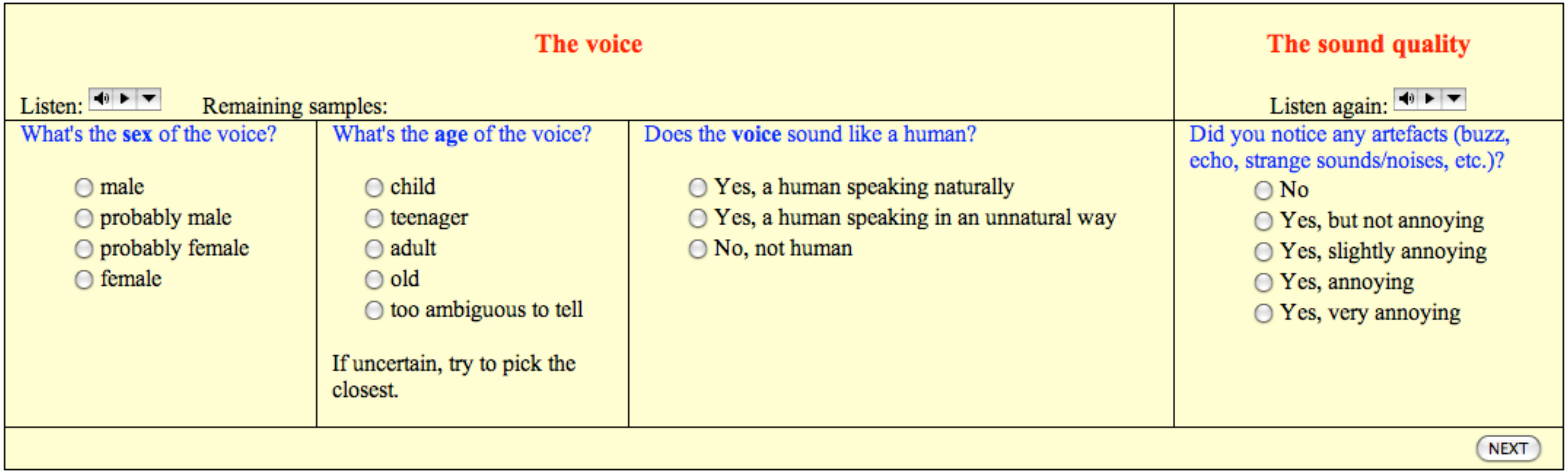
- trembling: 🔊
- pitch ambitus: greater: 🔊 smaller: 🔊 zero (robot): 🔊
- speech rate: faster: 🔊 slower: 🔊 faster vowels: 🔊

combinations

- dull: 🔊 excited: 🔊
- drunk: 🔊

Perceptual evaluation

- 13 voices (5 women, 6 men, 1 girl and 1 boy)
- 2 sentences of 2 to 3 seconds
- 7 transformations (male/female x 4 ages) +original
- 31 subjects listening to each sentence once (26)



The voice

Listen: Remaining samples:

What's the **sex** of the voice?

○ male
○ probably male
○ probably female
○ female

What's the **age** of the voice?

○ child
○ teenager
○ adult
○ old
○ too ambiguous to tell

If uncertain, try to pick the closest.

Does the **voice** sound like a human?

○ Yes, a human speaking naturally
○ Yes, a human speaking in an unnatural way
○ No, not human

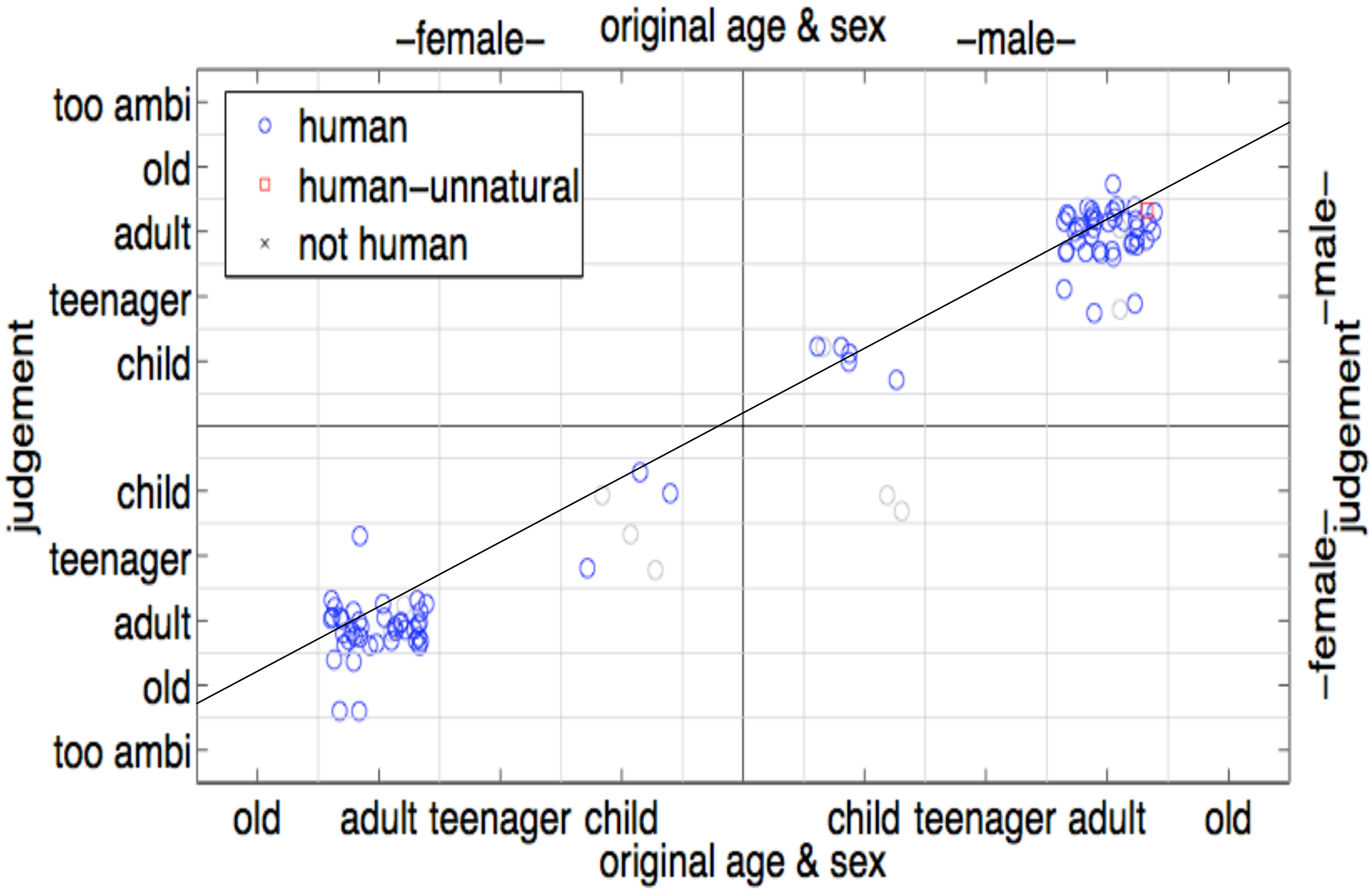
The sound quality

Listen again:

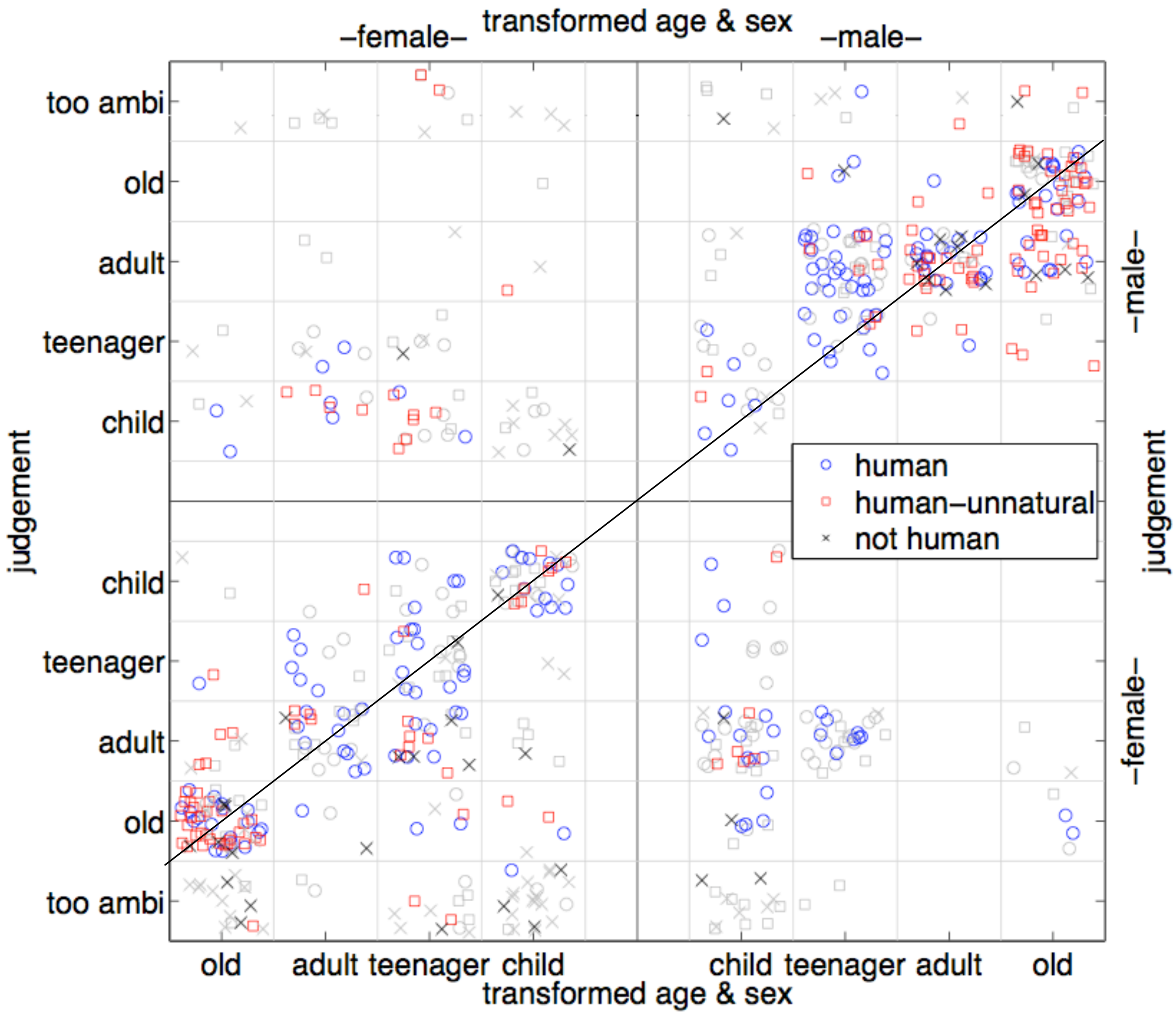
Did you notice any artefacts (buzz, echo, strange sounds/noises, etc.)?

○ No
○ Yes, but not annoying
○ Yes, slightly annoying
○ Yes, annoying
○ Yes, very annoying

Evaluation: original voices



Evaluation: transformed voices



Examples of transformation of gender and age

	source voice	
target voice	woman	man
original	. 🔊	. 🔊
little girl	. 🔊	. 🔊
teenage girl	. 🔊	🔊
woman		. 🔊
aged woman	🔊	🔊
boy	🔊	🔊
teenage boy	. 🔊	. 🔊
man	. 🔊	
aged man	. 🔊	🔊

Conclusion

- Perceptual evaluation:
  - listening test focuses on artifacts
  - in real world, attention is distracted by background sounds, visual input, story line, etc.
  - transformation of pitch and timbre not enough, e.g., girl 🔊 → man 🔊 : a girl's way of speaking
- Voice transformation is already used in high-fidelity applications: music, film, theatre
- IrcamVoiceForger:
  - C++ library
  - real-time

Demo: one actor → 4 characters

[Characters and animation by Cantoche]

Original: 🔊