Voice anonymization in urban sound recording

Alice Cohen-Hadria[1], Mark Cartwright[2], Brian McFee[2] and Juan Pablo Bello[2]

[1] Institute of Research and Coordination of Acoustic and Music (IRCAM), Sorbonne Université, Paris

[2] Music and Audio Research Laboratory, New York University

Context and Goals

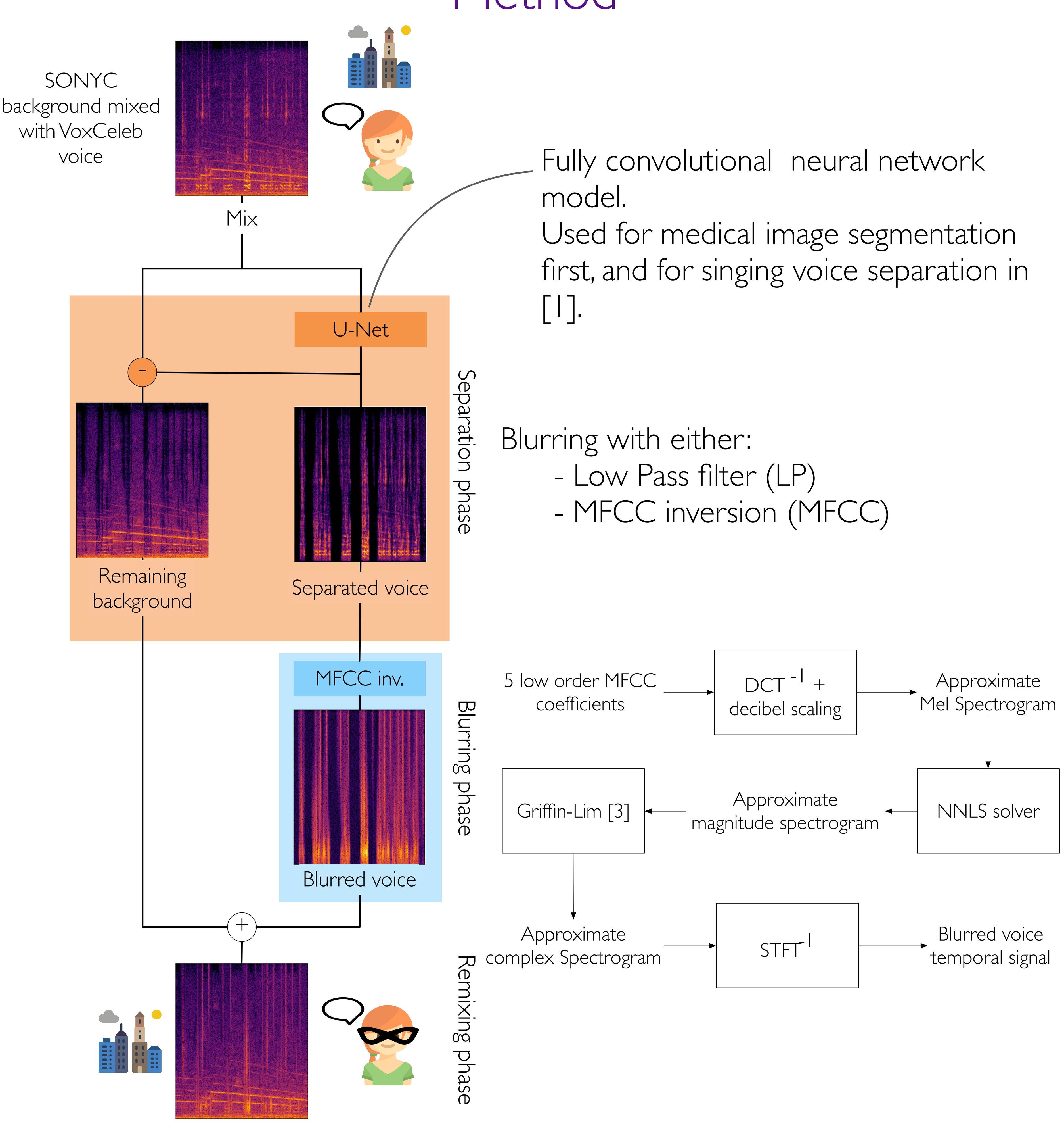
- Urban acoustic sensors may pick up human voice in some recordings.
- To maintain privacy, we need to ensure that conversations are not discernible and voices are not identifiable, we need anonymization methods.
- Need for anonymization methods which :

1. Obfuscate the content
2. Mask the speaker identity
3. Preserve the acoustic scene

Datasets

- **Backgrounds: SONYC-UST** dataset. 1744 urban recordings (without human voice) from New York City. Labeled in 7 coarse classes: engine, machinery impact, non-machinery impact, powered saw, alert signal, music.
- **Voices:**
  - **VoxCeleb** Recordings of celebrities (1211 speakers for training ans 40 for testing) labeled in speaker. Used for training the separation system and evaluating speaker identity masking.
  - **LibriSpeech** English speakers reading book extracts (1000 hours long). Used to evaluate content obfuscation.
- **Mixing**
  - **Training** Mix 2<N<5 excerpts of voice from VoxCeleb with backgrounds from SONYC
  - **Testing** 2 voice to background ratio conditions **Low** ($\alpha \in [0{:}1; 0{:}4]$),) and **High** ($\alpha \in [0{:}5; 0{:}7]$) mix = α background + $(1-\alpha)$ voice

Method



spectrogram for visualisations purpose only, some operations are performed in temporal domain

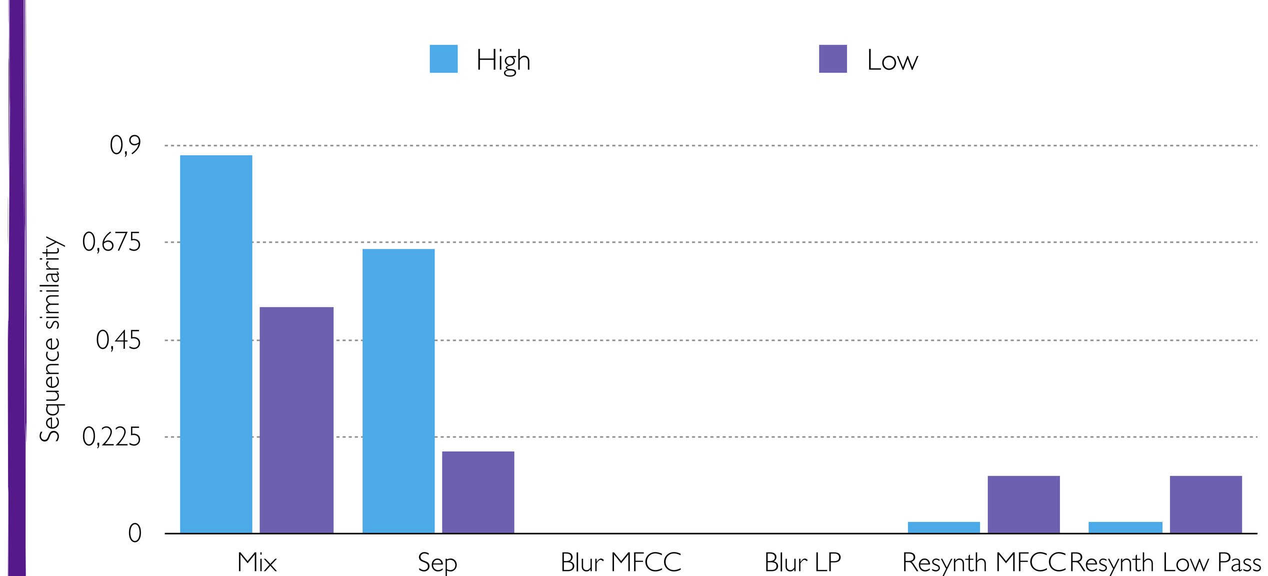
Experiments

Source separation results

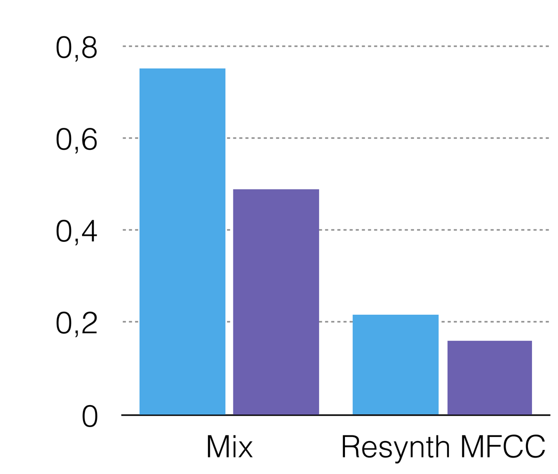
SNR	Model	SDR (std)	SIR (std)	SAR (std)
Low	U-Net	8.20 (4.9)	13.01 (5.4)	10.98 (4.5)
	IBM	12.39 (4.3)	19.21 (4.3)	13.62 (4.3)
High	U-Net	12.31 (4.0)	16.91 (4.0)	12.31 (3.3)
	IBM	17.98 (3.0)	21.26 (3.3)	16.19 (2.98)

- Source separation metrics [2]
  IBM : Ideal Binary Masks
- Better separation in High SNR setting.
- The quality of the separation impact the quality of the blurring.

Content obfuscation



Human experiment



- Use of LibriSpeech for easier transcription
- Blur separated version are never transcripted.
- Only resynthesis does not fully obfuscate the content -> due to the quality of the separation
- Trends in ASR experiment replicated in human listening test

Speaker identity masking

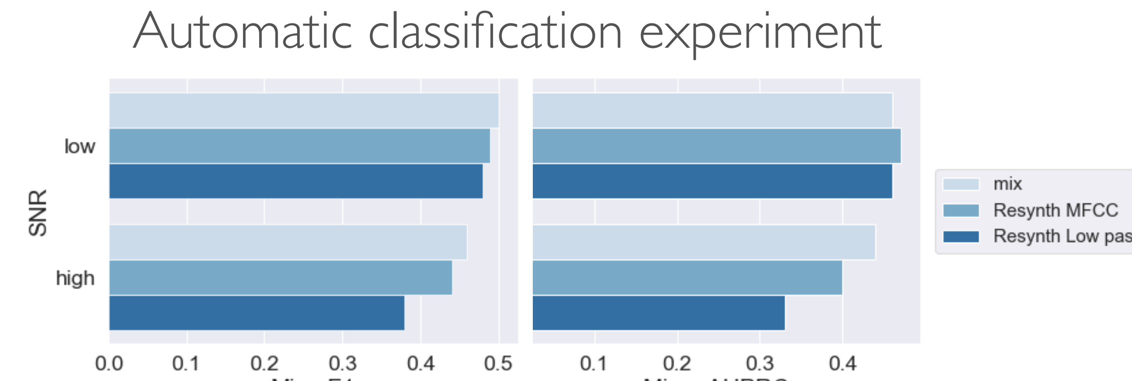
SNR	Audio	% correct Identification
High	Mix	83
	Low Pass filter	43
	MFCC inversion	43
Low	Mix	43
	Low Pass filter	29
	MFCC inversion	29

- VoxCeleb's VggVox model for speaker identification
- Both in High and Low, our blurring method decrease the identification
- Need for human evaluation, but necessitate training

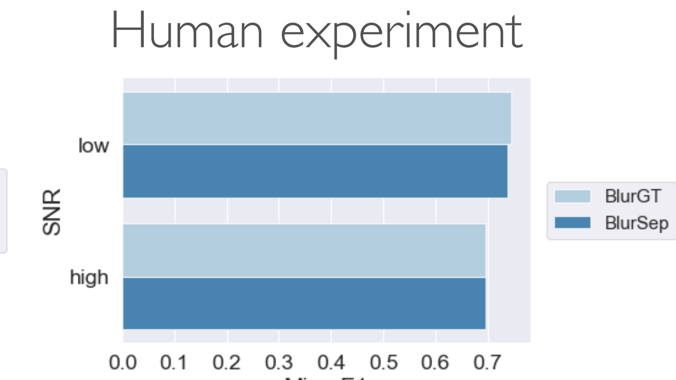
Scene preservation

- Baseline for the Urban Sound Tagging challenge for DCASE 2019, for automatic experiment. Use of VGGish features.
- 8 coarse classes.
- Classify mixes and blurred versions to assess how much of the scene is preserved
- Our blurring method preserve the acoustic scene.
- Trends in classification experiment replicated in human listening test

Automatic classification experiment



Human experiment



References

[1] Jansson et al, *Singing Voice Separation with Deep U-Net Convolutional Networks*, In Proc. of ISMIR, 2017.

[2] Vincent et al, «*Performance Measurement in Blind Audio Source Separation*, In IEEE Transactions on Audio, Speech and Language Processing, 2006.

[3] Daniel Griffin and Jae Lim, "Signal estimation from modified short time Fourier transform," IEEE Transactions on Acoustics, Speech, and Signal Processing

NEW YORK UNIVERSITY

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

ircam Centre Pompidou

MARL