2019 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 13–16, 2019, PITTSBURGH, PA, USA

VOICE ANONYMIZATION IN URBAN SOUND RECORDINGS

Alice Cohen-Hadria,[1] *Mark Cartwright,*[2,3] *Brian McFee,*[2,4] *Juan Pablo Bello*[2,3,4]

[1] UMR STMS 9912, Sorbonne Université, IRCAM, CNRS, France, cohenhadria@ircam.fr
[2] Center for Urban Science and Progress, New York University, USA
[3] Music and Audio Research Lab, New York University, NY, USA
[4] Center for Data Science, New York University, NY, USA
{mark.cartwright, brian.mcfee, jpbello}@nyu.edu

ABSTRACT

Monitoring health and noise pollution in urban environments often entails deploying acoustic sensor networks to passively collect data in public spaces. Although spaces are technically public, people in the environment may not fully realize the degree to which they may be recorded by the sensor network, which may be perceived as a violation of expected privacy. Therefore, we propose a method to anonymize and *blur* the voices of people recorded in public spaces—a novel, yet increasingly important task as acoustic sensing becomes ubiquitous in sensor-equipped smart cities. This method is analogous to Google's face blurring in its Street View photographs, which arose from similar concerns in the visual domain. The proposed blurring method aims to anonymize voices by removing both the linguistic content and personal identity from voices, while preserving the rest of the acoustic scene.

The method consists of a three-step process. First, voices are separated from non-voice content by a deep U-Net source separation model. Second, we evaluate two approaches to obscure the identity and intelligibility of the extracted voices: a low pass filter to remove most of the formants in the voices, and an inversion of Mel-Frequency Cepstral Coefficients (MFCC). Finally, the blurred vocal content is mixed with the separated non-vocal signal to reconstruct the acoustic scene. Using background recordings from a real urban acoustic sensor network in New York City, we present a complete evaluation of our method, with automatic speech recognition, speaker identification, sound event detection, and human perceptual evaluation.

Index Terms— voice anonymization, source separation, urban recordings, privacy

1. INTRODUCTION

Sensor networks are increasingly used to monitor urban environments and optimize the use of municipal resources and assets. In this context, large-scale sensing in public spaces is becoming increasingly prevalent. The data captured by these sensors can contain identifiers or other information that people may expect or prefer to be kept private. For example, video monitoring may capture an individual's face or gait [1], which could be de-anonymized to reveal their location and activity. In the case of acoustic sensing, if proper measures are not implemented, this can lead to the recording of conversations, with similar implications for privacy. In this work, we explore new methods for vocal anonymization in urban sound recordings, with the goal of obscuring speech content and de-identifying the speaker, while preserving the remainder of the acoustic scene. This work is conducted in the context of the Sounds of New York City (SONYC) [2] project, which is aimed at monitoring, analyzing, and mitigating urban noise pollution.

Vocal anonymization is a challenging task that requires the segregation and modification of sounds in complex acoustic scenes, and for which only partial solutions exist. For example, Ribaric et al. [3] reviewed voice de-identification methods which were limited to masking the identity of the speaker, but did not directly address speech content. Qian et al. [4] proposed a method to anonymize both content and speaker, but it was developed for recordings of clear voices which do not have backgrounds needing preservation.

Our approach is inspired by the process of face blurring in images: we separate the original audio signal into voice and background components, then selectively distort ("blur") the voice signal before mixing back with the background. By using source separation, processing and remixing, our method is the first to achieve the three objectives of speaker de-identification, content obfuscation and scene preservation in environmental sound recordings. Our work further contributes new datasets for the training and testing of our models, as well as a novel evaluation framework.

2. APPROACH

The proposed anonymization process is depicted in fig. 1. It consists of a three-step process:

1. We extract the *voices* from the *mix* of voice and background using a deep neural network called U-Net [5, 6], described in Section 2.1. This step estimates two separated audio signals: the *voice* and the residual *background*.
2. The separated voice is *blurred* to remove identifiable information. We use two different blurring methods in Section 2.2.
3. The blurred voice is recombined with the background signal, resulting in the anonymized *resynthesis*.

2.1. U-Net source separation

The U-Net model is a fully convolutional neural network, originally designed for cell segmentation [5]. Jansson et al. [6] successfully applied this architecture to spectrogram representations to isolate singing voice from non-vocal instrumentation.

In our case, the input $X \in \mathbb{R}_{+}^{T \times F}$ is a magnitude spectrogram of *mix*, an urban sound recording containing both a *background* acoustic scene we wish to preserve and voices we wish to anonymize.

978-1-7281-0824-7/19/$31.00

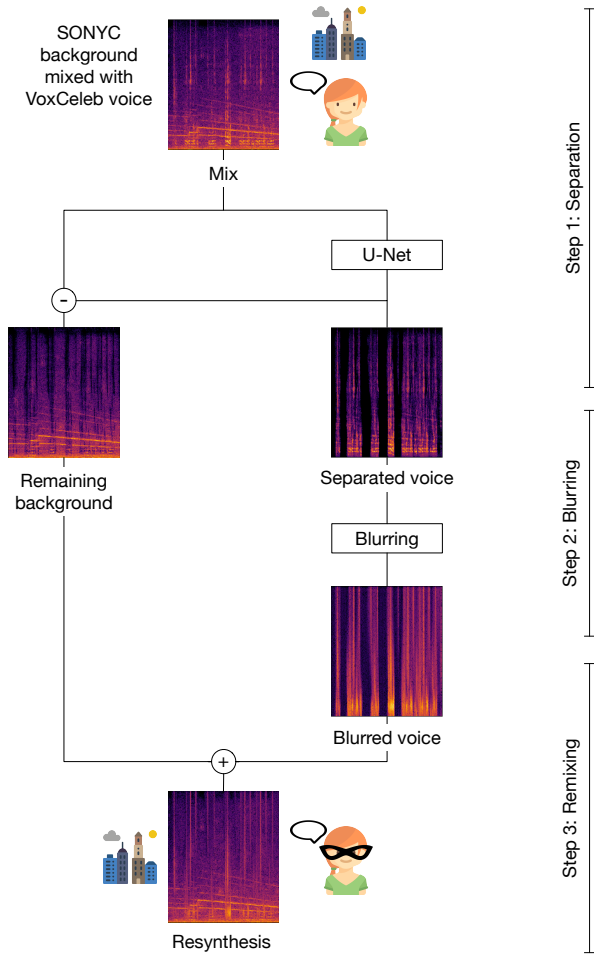


Fig. 1: Schematic diagram of the proposed anonymization method. Note that the spectrogram visualizations are for illustrative purposes only. Some steps are actually performed in the time-domain. See section 2 for details.

From X, the U-Net learns a compressed, encoded representation, which is decoded to reconstruct a target magnitude spectrogram, which in our case, is derived from an isolated vocal signal provided as supervision. More precisely, the U-Net with parameters θ computes a continuous (soft) mask $f_\theta : \mathbb{R}_+^{T\times F} \rightarrow [0,1]^{T\times F}$, and the separated magnitude spectrogram is estimated by the element-wise product of X with the mask: $X \otimes f_\theta(X)$. To train the U-Net, we minimize the mean absolute error between the estimation and the ideal isolated voice magnitude spectrogram $Y \in \mathbb{R}_+^{T\times F}$:

$$\mathcal{L}(\mathcal{X}, \mathcal{Y}; \theta) = \|X \otimes f_\theta(X) - Y\|_1.$$

To preserve details lost in the encoding stage, skip connections are added between corresponding layers of the encoder and decoder subnetworks. After estimating the vocal magnitude spectrogram, we combine it with the phase of the input spectrogram to reconstruct, with an ISTFT, an temporal signal of the separated voice.

2.2. Voice blurring

We present here two blurring methods, intended to conceal both the identity and content of the separated vocal signal.

2.2.1. *Low-pass filtering*

Because most of the voice content is localized to the high frequencies, for this first blurring method, we chose to simply low-pass filter the separated voice at 250 Hz, by sub-sampling the signal at 500 Hz and up-sampling the results back to 16 kHz. This was done using the sample-rate conversion utilities provided by `librosa` [7].

2.2.2. *MFCC inversion*

To further remove the identity of the speaker, we also propose another method of blurring using approximate MFCC inversion, like in [8]. MFCCs represent the spectral envelope, or timbre, of an audio signal [9], and the low-order coefficients have been used in speech recognition systems because they broadly retain phonetic information while discarding much of the speaker's identifiable characteristics. In this work, we compute the MFCC coefficient and choose to retain only the first 5 MFCC coefficients for inversion, which are insufficient for speech recognition, but still capture the general spectral envelope.

Given a time-series of the first 5 MFCCs, we apply the inverse discrete cosine transform and decibel-scaling, resulting in an approximate mel power spectrogram. We then use a Non-Negative Least Squares (NNLS) solver to convert the mel spectrogram into a linear frequency power spectrogram. Finally, the Griffin-Lim algorithm [10] is used to estimate phase, and the resulting complex spectrogram is transformed to the time domain using the inverse short-time Fourier transform. We stress that the reconstruction here need not be perfect, since the goal is to produce an unintelligible signal. Griffin-Lim is used here primarily to reduce obvious phasing artifacts which might otherwise distract from the overall acoustic scene. This components of this process have been implemented in `librosa` 0.7 [7].

3. EVALUATION

This section presents the methods we used to evaluate our anonymization process. The first experiment (Section 3.2) evaluates the separation step, and uses standard source separation metrics. The next three experiments address each of our goals:

1. *Content obfuscation*, evaluated by a speech recognition experiment (automatic and human);
2. *Speaker anonymization*, evaluated by a speaker identification experiment (automatic); and
3. *Scene preservation*, evaluated by an urban sound tagging experiment (automatic and human).

3.1. Datasets

The U-Net model is trained and evaluated on pairs of *mix* and corresponding separated *voice*. We constructed our own separated datasets using SONYC [2] recordings of New York City as the background, and two sources of foreground voice recordings: VoxCeleb [11] celebrity voices recordings and LibriSpeech audio book speech recordings [12]. VoxCeleb will be used for speaker recognition, and LibriSpeech for content obfuscation.

3.1.1. SONYC dataset

The SONYC dataset [13] has been released for the Urban Sound Tagging Challenge for DCASE 2019.The SONYC data consist of 10-second recordings, labeled for 8 coarse classes: *engine, machinery impact, non-machinery impact, powered saw, alert signal, music, human voice* and *dog*. These labels are weak, i.e. there is only one label for all 10 seconds, and multi label, i.e. one example can be labeled with one or several classes. It is split into a training set of 2351 recordings and a validation set of 443 recordings, as only the development set has currently been released. Using the labels, we selected the recordings without human voices in them, resulting in 1500 background recordings for training, and 244 backgrounds for validation. The training set was used to fit the U-Net separation model, and validation set is used as test data in our four experiments. We detail the mixing procedure of our synthetic datasets in section 3.1.4

3.1.2. VoxCeleb dataset

The VoxCeleb dataset contains recordings of celebrity voices curated from YouTube and was originally designed for speaker recognition and verification. We chose the VoxCeleb dataset due to its similarity to real-world conversations and speaker identity annotations. VoxCeleb is partitioned into training and test sets, consisting of 1211 and 40 distinct speakers respectively. We used the VoxCeleb dataset for training and testing the U-Net model. We chose one random voice from the testing VoxCeleb dataset for each SONYC *background* of our test set, resulting in 244 *mix* test signals.

3.1.3. LibriSpeech dataset

LibriSpeech is a dataset designed for automatic speech recognition (ASR) [12]. It contains 1000 hours of recordings of people reading English texts for audio books. Because these recordings are clear and easy to transcribe, we chose this dataset for evaluating the system's ability to obfuscate speech content. Similarly to the VoxCeleb dataset, we choose one random voice from the LibriSpeech dataset for each background signal, resulting in 244 *mix* test signals.

3.1.4. Mixing voices and backgrounds

To generate training data, we randomly added $N \in \{3, 4, 5\}$ 1 second segments of voice from VoxCeleb training set to our SONYC backgrounds. Both the *backgrounds* and *voices* were normalized using the root mean square (RMS) of the signal.

Test signals were generated in two different conditions: High-SNR and Low-SNR, intended to simulate scenarios in which the vocal signal is or is not prominent in the acoustic scene. Signals were combined using two ranges of mixing coefficient α: $\text{mix} = \alpha \cdot \text{voice} + (1 - \alpha) \cdot \text{background}$. In the Low-SNR range ($\alpha \in [0.1, 0.4]$), we mixed the voices lower than the background to resemble the current recordings in the SONYC dataset that contain voice. In the High-SNR range ($\alpha \in [0.5, 0.7]$), we mixed the voices higher to mimic other urban scenarios (for example crowdsensing with mobile phones) in which the voice may be more prominent. In both settings, α is drawn uniformly at random from the given range. For each experiment, we randomly selected 244 voices from either VoxCeleb testing set or LibriSpeech as the *voices*, mixed with the 244 backgrounds of the evaluation set of SONYC dataset, resulting in evaluation sets of 244 mixes.

3.2. Experiment 1: Source separation quality

To evaluate the source separation quality of the model, we use the traditional metrics of this field — *Source-to-Artifact Ratio (SAR)*, *Source-to-Interference Ratio (SIR)* and *Source-to-Distortion Ratio (SDR)* [14] — as implemented in the `mir_eval` toolbox [15]. For a reference point, we also computed the results of the *ideal binary mask (IBM)*, an oracle method using the ground-truth separated signals. To compute the IBM, we set the time-frequency mask to 1 whenever the target source is stronger than the acoustic background, and 0 otherwise. The IBM gives an upper bound for the performance our two test conditions (Low-SNR and High-SNR). We use an evaluation set composed of the evaluation backgrounds of SONYC and voices from VoxCeleb testing set. While this experiment does not directly address our anonymization goals, we report on this experiment to inform the analysis of our other experiments and to provide a reference point to relate our results to other source separation algorithms.

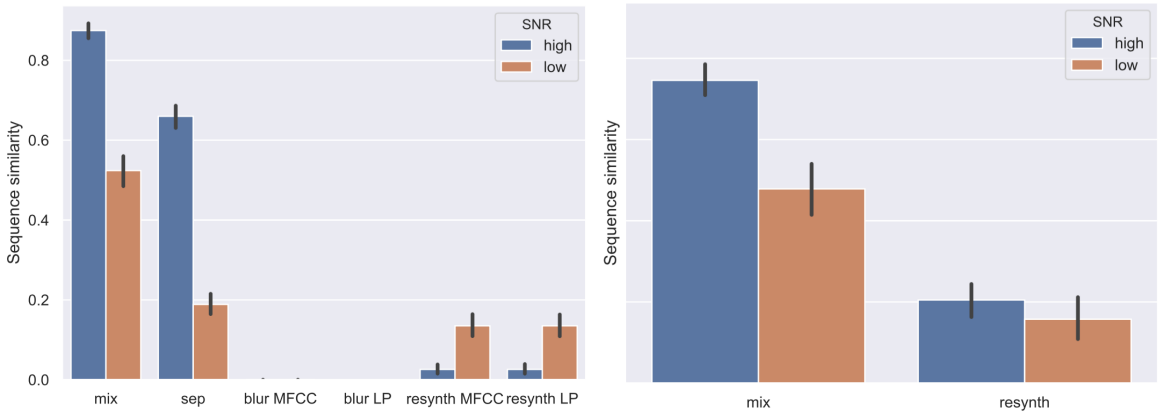
3.3. Experiment 2: Content obfuscation

To demonstrate that our method successfully obfuscates speech content, we ran an ASR system using the Google API [16] on the output of our processing pipeline: for the *mix*, *separated voice*, *blurred voice* and *resynthesis*. We evaluate this using the LibriSpeech set, which includes text transcriptions of the spoken content. We evaluate the performance using Ratcliff-Obershelf algorithm [17] as implemented in Python's `difflib.SequenceMatcher`. This metric measures the similarity between two words. Using this similarity to the ground truth, we compare the results for the *mix*, the *separated voice*, *separated background*, *blurred voices*, and *resynthesis*. Evaluating on each step of the process, as well as the estimated background, allows us to quantify how much intelligible speech is retained by each stage.

Since ASR may not perform well on noisy signals, we also ran a listening test in which we asked 10 participants to transcribe the *mix* and the MFCC-inversion blurred signals. The 10 participants collectively transcribed 30 source recordings (15 High-SNR and 15 Low-SNR), the *mix* and *resynthesis* (MFCC-inversion) of each. To prevent participants from transcribing both the *mix* and *resynthesis* version of the same source recording, the recordings were split into two sets, each containing 30 recordings with High/Low-SNR *voice*, *mix*, and *resynthesis* versions evenly distributed. The set assignment and presentation order of recordings were both randomized.

3.4. Experiment 3: Speaker anonymization

To measure the ability of the proposed system to obscure speaker identities, we ran an automatic speaker identification system, trained on the VoxCeleb dataset. We used the `VggVox` model [18], which is a convolutional neural network trained to produce a speaker-oriented embedding from an input spectrogram. First, we used the model compute the embeddings of a pool of speakers composed of the ground-truth voice recordings. Then, to identify the speaker in a test signal, we simply compute the embedding of the signal and compare it to all of the embeddings of the pool using cosine distance. The nearest neighbor estimates the speaker of our test signal. We used this method to identify the speakers in the *mixes*, the *separated voices*, and the different blurring approaches.



(a) ASR with LibriSpeech data, MFCC and Low-Pass blurring. (b) Human transcription with LibriSpeech, MFCC blurring

Fig. 2: Content obfuscation results. Mean inverse sequence similarity to the ground-truth word sequence. Bars are the 95% CIs.

Table 1: Example of transcription with ASR and human transcription. The bolded words are the correctly transcribed portions.

Audio	Method	Transcription
—	Ground truth	sort of magic touchstone by which they are saved the labour of investigation but there is no such thing as a single
Mix	ASR	**sort of Magic touch stone by which they are saved the labour of Investigation but there is no such thing as a single**
	Human	**Sort of magic** touchdown **by which they are saved the** role **of** their imagination **but there is no such thing as a single**
Resynthesis	ASR	take me to PE15 **there's no such thing as single**
	Human	**but there** was **no such**

3.5. Experiment 4: Acoustic scene preservation

Lastly, to evaluate the preservation of the acoustic scene, we ran an urban sound tagging model trained on the SONYC dataset. This model was released as a baseline for the Urban Sound Tagging challenge for DCASE 2019.The model uses logistic regression with VGGish input features [19] to estimate the 8 urban sound classes of the SONYC dataset. The evaluation set is computed with voices from the VoxCeleb dataset. We evaluated tagging performance with the micro-averaged F1 score (with a 0.5 threshold) and the micro-averaged and macro-averaged area under the precision-recall curve (AUPRC), [20]. Here we focus the comparison on the tagging performance of the MFCC-inversion *blurred voice* and the original *mix* signals.

The same 10 participants from the earlier task were also asked to perform multi-label annotation of 30 excerpts of 10 seconds (15 High-SNR, 15 Low-SNR). Two different audio sets were annotated using the same set of classes as the automatic classifier. The first set contains the output of our anonymization system (this set will be called `BlurSep` in the remaining sections). The other set is the equivalent ground-truth anonymization. The ground truth voice from VoxCeleb is blurred by MFCC-inversion and mixed with a background from SONYC dataset (this set will be called `BlurGT`).

3.6. Model training

We used the 1500 *mixes* from the train set of SONYC mixed with VoxCeleb voices for training as specified in section 3.1.4. Input spectrograms consist of 128 frames of 16 kHz monophonic audio, which have been transformed by 1024-point Hann-windowed STFT with hop size of 256 samples. The U-Net output activations are sigmoid to limit values between 0 and 1 for the output mask. We used the ADAM optimizer with default parameters for 5 epochs. The model architecture follows Janson et. al. [6]. Each layer of encoding is a 2D convolution with 5 by 5 filters and 2 by 2 strides, batch normalization and leaky ReLU activations. Each layer of the decoder is a strided deconvolution layer, with stride 2 and filters 5 by 5, batch normalization and ReLU activations.

4. RESULTS

4.1. Experiment 1: Source separation quality

The results for the source separation step are presented in table 2. The IBM rows indicates an upper bound for the source separation problem with our data. As expected, the separation is better in the High-SNR context. In that data subset, the voice is rarely masked by the background, making it easier to extract with our deep neural network. Since we are using the estimated separated voice signal to also compute the remaining background, the voice separation performance indicates how much of the voice we are blurring and how much is retained in the estimated background signal. Because of this, we can see that the quality of the separation impacts the quality of the content anonymization in section 4.2.

Table 2: Results of the source separation step, mean and standard deviation (IBM: ideal binary masks)

Model	SDR (std)	SIR (std)	SAR (std)
U-Net Low-SNR	8.20 (4.91)	13.01 (5.36)	10.98 (4.49)
U-Net High-SNR	12.31 (4.03)	16.91 (4.03)	12.31 (3.13)
IBM Low-SNR	12.39 (4.27)	19.21 (4.27)	13.62 (4.27)
IBM High-SNR	17.98 (3.02)	21.26 (3.31)	16.19 (2.98)

4.2. Experiment 2: Content obfuscation

Figure 2 presents the results of the ASR and the human evaluation experiments for the obscuring speech content. The *mix* results show that background noise already affects the ability of the ASR system to transcribe, and that the effect is larger for the Low-SNR dataset.

Separating the voices from the *mix* does not improve the results of the ASR: we see substantial decrease between `mix` and `sep` in both the High-SNR and Low-SNR context. This indicates that the ASR system is sensitive to the artifacts created by the separation step. The blur results (`blur MFCC` bar for MFCC blurring and `blur Low Pass` for low pass blurring) in fig. 2a) are the results of both blurring methods: MFCC and Low Pass. Our two blurring methods fully obscure the content of the voices, at least from the perspective of ASR. All of the remaining content we can observe in the *resynthesis* data comes from the fact that the separation is not perfect, and some un-blurred vocal content remains in the residual *background* after separation and *resynthesis*. This effect is even more noticeable in the case of the Low-SNR data, when the separation is less accurate, leaving more vocal signal in the estimated background. Overall, both blurring methods successfully obscure speech content, but the efficacy of the entire process is a function of the quality of the separation.

Because manual text transcription is a long and tedious task, we limited the human transcription experiments to the *mix* (input) and *resynthesis* (output) signals. MFCC blurring was used for the *resynthesis*. Similar to the ASR results, we again find the Low-SNR dataset more difficult to transcribe than High-SNR. The transcription score after *resynthesis* tracks is substantially lower than on the original signals, confirming that our blurring method does obscure the content, even with human listeners. Noise in the *mix* hampers the human subjects' ability to transcribe more than the ASR system in high SNR conditions. Notably, Low-SNR results are very similar between humans and the ASR system.

An example of human and automatic transcription is provided in Table 1, with a High-SNR signal. While the original *mix* can be perfectly interpreted by ASR, and mostly interpreted by the human, the obfuscated signal destroys most of the speech content. Although a few words are correctly transcribed, the results suggest that the overall content of the utterance is successfully obscured.

4.3. Experiment 3: Speaker anonymization

The results of the speaker verification experiment are presented in Table 3, with the percentage of correctly identified speakers from 244 test examples. The Low-SNR dataset remains the hardest to identify the speaker, even for the unmodified *mix*. Only 43% of the tracks are correctly assigned to the right speaker for the Low-SNR dataset. Finally, our blurring method increases the anonymity of speakers, reducing the percentage of correct identifications from 43% for the *mix* to 29%% for the MFCC inversion (Low-SNR dataset), and from 83 to 43% (High-SNR dataset).

Table 3: Results of speaker verification, Low-SNR and High-SNR.

	Data	% correct identification
Low-SNR	Mix	43%
	Low-pass filter	29 %
	MFCC inversion	29 %
High-SNR	Mix	83 %
	Low-pass filter	43 %
	MFCC inversion	43 %

4.4. Experiment 4: Acoustic scene preservation

The results of the scene preservation task are presented in Figure 3. From the results, we can see that the urban sound tagging is a hard task: what matters most here is the difference in accuracy obtained by the classifier in each test condition (original mix, MFCC-inversion, and low-pass filtering).

For the *mix*, the micro-averaged F1 score is 0.46 for the Low-SNR dataset. For the *resynthesis* with MFCC blurring signal, the F1 score is 0.44. Since the voice is not masking the background as in the High-SNR dataset, the classification results are better for the Low-SNR dataset, for all metrics (Figures 3a to 3c). For the Low-SNR dataset, our blurring method does not decrease the performance of the urban sound classification. For the High-SNR dataset, since the voices are louder, the results decrease between *mix* and the *resynthesis*.

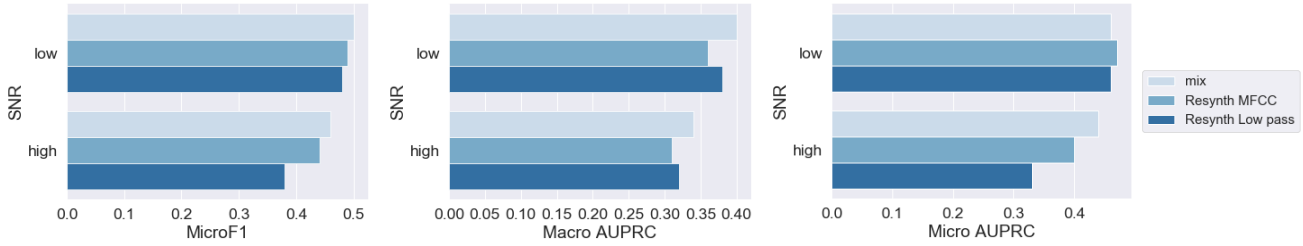
Finally, among our two blurring methods, we can see that the MFCC-inversion method is the one that best preserves the scene.

The results of the human evaluation are in Figure 4. For both the Low-SNR and High-SNR variants, the performance of the participants on the smaller human-based evaluation was higher than the machine's performance on the automatic evaluation task. In agreement with the automatic classification results, humans were better at classifying the Low-SNR dataset compared to the High-SNR, probably because the mixed vocals did not obscure as much of the overall acoustic scene.

The `BlurGT` can be seen as the best classification score we would achieve given our blurring process. We can see that the results of the `BlurGT` and the `BlurSep` are comparable in term of F1 score. Therefore, while these scores are below the published F1-score agreement between novice annotators and ground-truth annotations on this task using the full dataset (0.86) [21], this is likely due to additional masking caused by the blurring, rather than the separation performance. But this may also be due to the particular small subset of SONYC-UST chosen for this experiment. Thus, human evaluation assesses that our anonymization method does not completely preserve the acoustic scene, but this is not due to the separation quality. Therefore, this could possibly be improved with a different blurring method or by simply mixing the blurred voice at a lower level to reduce masking.

5. CONCLUSION AND FUTURE WORKS.

We presented a new task and a new proof of concept method for preserving individual privacy in urban sound recordings. The evaluations demonstrated that our method successfully anonymized speakers, obscured speech content, and generally preserved the acoustic



(a) Micro F1 score (b) Macro AUPRC (c) Micro AUPRC

Fig. 3: Results of the urban sound classification experiments.

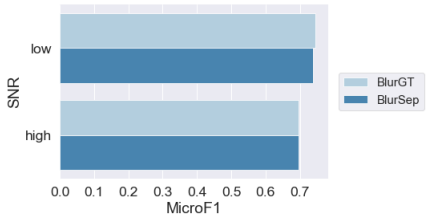


Fig. 4: Human evaluation for acoustic scene preservation, with MFCC blurring

background of the scene. The proposed blurring methods were intentionally simple, but show promise. However, more advanced blurring methods could be considered to improve performance, which will be the subject of future research. The results of the speech recognition experiments indicate that source separation is a crucial step in the overall system, and improvements to the separation model could be readily integrated into the proposed framework.

6. REFERENCES

[1] A. Kale, A. Sundaresan, A. Rajagopalan, N. Cuntoor, A. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on image processing*, 2004.

[2] Juan Pablo Bello, Cláudio T. Silva, Oded Nov, R. Luke DuBois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy, "Sonyc: A system for the monitoring, analysis and mitigation of urban noise pollution," *CoRR*, vol. abs/1805.00889, 2018.

[3] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, 2016.

[4] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, Xiang-Yang Li, Yu Wang, and Yanbo Deng, "Voicemask: Anonymize and sanitize voice input on mobile devices," 2017.

[5] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. 2015, LNCS, Springer.

[6] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. of ISMIR*, 2017.

[7] Brian McFee, Matt McVicar, Stefan Balke, Vincent Lostanlen, Carl Thom, Colin Raffel, Dana Lee, Kyungyun Lee, Oriol Nieto, Frank Zalkow, Dan Ellis, Eric Battenberg, Ryuichi Yamamoto, Josh Moore, Ziyao Wei, Rachel Bittner, Keunwoo Choi, nullmightybofo, Pius Friesch, Fabian-Robert Stter, Thassilo, Matt Vollrath, Siddhartha Kumar Golu, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis Hawthorne, and CJ Carr, "librosa/librosa: 0.6.3," Feb. 2019.

[8] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in *ICASSP*, 2000.

[9] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs: Prentice Hall, 1978.

[10] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, 1984.

[11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *CoRR*, 2017.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

[13] Mark Cartwright, Ana Elisa Mendez Mendez, Graham Dove, Jason Cramer, Vincent Lostanlen, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Pablo Bello, "SONYC Urban Sound Tagging (SONYC-UST): a multilabel dataset from an urban acoustic sensor network," Mar. 2019, This work is supported by National Science Foundation award 1544753.

[14] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2006.

[15] Colin A. Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis, "Mir_eval: A transparent implementation of common mir metrics," in *ISMIR*, 2014.

[16] Zhang, A., "Speech recognition," .

[17] John W. Ratcliff and David Metzener, "Pattern matching: The gestalt approach," *Dr. Dobb's Journal*, p. 46, 1988.

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *CoRR*, 2017.

[19] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *ICASSP*. 2017.

[20] Vincent Van Asch, "Macro-and micro-averaged evaluation measures," *Belgium: CLiPS*, 2013.

[21] Mark Cartwright, Ana Elisa Mendez Mendez, Graham Dove, Jason Cramer, Vincent Lostanlen, Ho-Hsiang Wu, Justin Salamon, Oded Nov, and Juan Pablo bello, "Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network," in *Preparation*, 2019.