*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Evaluating Audiovisual Source Separation in the Context of Video Conferencing

*Berkay İnan*[12], *Milos Cernak*[2], *Helmut Grabner*[23],
*Helena Peic Tukuljac*[1], *Rodrigo C. G. Pena*[1], *Benjamin Ricaud*[1]

[1]LTS2, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Logitech Europe S.A., Lausanne, Switzerland
[3]Zurich University of Applied Sciences, Zurich, Switzerland

berkayinan@gmail.com, milos.cernak@ieee.org, helmut.grabner@zhaw.ch
helena.peictukuljac@epfl.ch, rodrigo.pena@epfl.ch, benjamin.ricaud@epfl.ch

## Abstract

Source separation involving mono-channel audio is a challenging problem, in particular for speech separation where source contributions overlap both in time and frequency. This task is of high interest for applications such as video conferencing. Recent progress in machine learning has shown that the combination of visual cues, coming from the video, can increase the source separation performance. Starting from a recently designed deep neural network, we assess its ability and robustness to separate the visible speakers' speech from other interfering speeches or signals. We test it for different configuration of video recordings where the speaker's face may not be fully visible. We also asses the performance of the network with respect to different sets of visual features from the speakers' faces.

**Index Terms**: speech enhancement, source separation, multimodal, audiovisual

## 1. Introduction

Video conferencing is a convenient way of replacing physical meetings and is a central tool for exchanging collaborative information worldwide. However, it is not yet as natural as in-person meetings, and it faces several challenges. For example, it is impossible for listeners to understand when two people are speaking at the same time. Often due to the mediocre recording conditions, insignificant noise such as keyboard typing or finger tapping can quickly become a source of annoyance and distraction. Audio recordings are usually transmitted via a single audio channel, and source separation is hardly possible.

Recently, the increasing computing power has made it possible to handle audio and video data and combine them in order to separate speech and noise, even on a single audio channel [1, 2, 3]. This might allow improved experiences with video conferencing as well, where the speech of a single person could be separated from other secondary sources of noise. We hypothesize that the visual information, within videos, might give clues that help source separation. Indeed, in video conferencing the speaker appears in the video with his mouth clearly moving as (s)he speaks, providing important information on its speech.

The topic of audiovisual source separation received a lot of attention in recent works. Google proposes a recurrent neural network to predict complex ratio masks for spectrograms to separate target speakers' audio signals from the spectrogram of mixture signal [3]. They use face recognition networks to extract embeddings for the faces of the speakers in order to capture visual information. The system works with the raw complex representation for short time Fourier transform of the audio signal to avoid magnitude and phase separation. It also proposes ways to process multiple speakers to leverage joint information. In [1] the authors follow a similar idea. They propose a residual network that first predicts the magnitude spectrogram and uses this prediction along with the input's phase spectrogram to predict phase spectrogram of the output. For visual embedding extraction, they use a temporal extension of *Resnet* [4], which is trained for lip reading task, similar to [5]. These lip reading embeddings are shown to be able to capture the audiovisual correlations, which allow them to achieve source separation even in the presence of high number of concurrent speakers. The approach used in [2] is a different way to look into the problem. They have built a network to detect whether the audio is synchronized with the video or not. They show that the audiovisual representation learned by this network can be used to learn new tasks such as audio source localization, action recognition, and off-screen/on-screen source separation.

These promising audio-visual approaches have not yet be evaluated in the context of video conferencing. Their performance on different types and intensity of audio interference and their robustness to face occlusion is still largely unknown. The ability to make this process real-time is also an important question. In this work, we thus focus on building audiovisual source separation for video conferencing, inspired by [3], assessing it, and exploring its efficiency in different real-world conditions:

1. We show that the deep neural network combining audio and visual information is able to efficiently extract a speech associated to a video from i) a second unrelated speech with different intensities, including when the interfering speech is much louder, ii) additive noise such as keyboard noise or impact noise with various intensities.
2. We demonstrate that the extraction of speech against noise is robust and can still be separated in the absence of video.
3. We provide an efficiency limit for the speech-speech extraction in the case of face occlusion. We show that the performance degrades when the faces are occluded for a segment longer than 20% of the duration of the video.
4. We provide evidence that the face recognition embeddings can be advantageously replaced by lip reading embeddings, leading to more accurate results, with less training of the network.

Our experiments were performed using standard research lab resources, showing the possibility for implementing this solution with reasonable amount of computer power and making a step forward to application in video conferencing. We provide the source code for our implementations at https://github.com/berkayinan/audiovisual-separation-for-vc.



http://dx.doi.org/10.21437/Interspeech.2019-2671

# 2. Experimental Setup

## 2.1. Datasets

**AVSpeech**. For training and evaluation, we choose to use the AVSpeech dataset [3]. It is an open dataset that consists of YouTube videos with a single speaker whose face is visible through most of the video. The duration of clips is between 3 to 10 seconds. There is no restriction on languages. Given limited processing power and storage constraints, we use a random subset of AVSpeech with videos that have 25 frames-per-second. This subset contains 37737 clips from 26518 different videos. We use $80\%$ of these clips as the training set and the rest as the evaluation set. We ensure that clips from the same videos are not shared between training and evaluation splits.

**AudioSet**. We use AudioSet as the dataset of non-speech audio [6]. We selected a subset of AudioSet which contains some of the possible noise sources that could appear during video conferences such as keyboard noise, and from domestic places, like baby cry and laugh, dog barking or glass breaking. This subset contains 10006 clips. Similar to AVSpeech, we make an $80/20$ split for training and evaluation sets. Note that AudioSet is weakly labeled, the precise time location is not available, and it may not be the only audio noise found in the clip. Often human speech may appear concurrently with the noise.

## 2.2. Training process

During training, we randomly pick two videos with 3-second duration from the training dataset. We choose one of them to be the target video and the other to be the interfering one. In the speech-speech separation task, we pick both videos from AVSpeech; while in speech-noise separation, we pick the target video from AVSpeech and the interfering video from AudioSet. The pre-computed video embeddings of the target video are given as the input of the visual part. The mixture of the audio from both video samples is the input of the audio part. The mixtures' mean signal to noise ratio (SNR) is around 0 dB, with a standard deviation of $\sigma = 8.4$ dB, i.e. the interfering audio signal has on average an intensity close to the one of the target signal, with some large variations allowed. This variety of different SNR mixtures makes the network robust to changes in input SNR.

A neural network is trained to separate the speech signal of a target speaker from other interfering audio sources. The idea is that the neural network will be able to extract the necessary information about how to isolate the target speech signal by making use of the mixture signal and the visual clues from target speaker's face. The output is a spectral mask that can be applied to input mixture signal to recover the target speech signal. Mean square error between the ground-truth spectrogram and the output spectrogram is chosen as the loss function.

Our network architecture is based on the network proposed in [3]. The network is made of 3 parts: i) visual feature extractor, ii) auditory feature extractor and iii) fusion network. Visual and auditory feature extractors are convolutional networks, whereas the fusion network is a combination of a Long Short-Term Memory (LSTM) block and fully connected layers that are able to make use of the sequential nature of the signals.

## 2.3. Visual Representation

Using raw image data from videos to learn temporal correlations is challenging, even with the large amount of resources [7]. In order to make the training process more tractable, we extract low dimensional embeddings from the video frames. These embeddings are extracted using the output of the lower layers of the networks that are trained for i) face recognition tasks and ii) lip reading.

**Face recognition embeddings**. We use face recognition embeddings as described in [3]. To extract these embeddings from videos, we apply face recognition on each frame. We discard the video if the number of frames with faces detected is less than 75% of the whole video. Otherwise, we put a vector of zeroes in place of the embedding for the missing face frame. We then apply face rectification based on the facial landmarks needed by the face recognition networks.

We use *Openface* [8] to obtain the face recognition embeddings. Face images are reshaped into size of $96 \times 96$ using *dlib*'s face alignment. Using the *nn4.small2* model, and flattening output of *AvgPool2d* layer as the embedding, we obtained a 736-dimensional embedding vector per frame.

**Lip reading embeddings.** In contrast to face recognition embeddings, the lip reading is directly trained on a speech-related task with a more limited region of interest. These embeddings should be able to capture the speech related information more accurately than the face embeddings. We obtained those embeddings by following the work of [1].

We pre-process the video frames using *dlib*'s face detection [9] and its face rectification, leading to $256 \times 256$ images. We then crop the window with coordinates $126-221, 80-175$ (width,height), focused on the mouth area, and we convert it to gray scale with range $[0-1]$.

The lip reading model used in our work is the pre-trained model of [5]. The implementation is the courtesy of [10]. The output of the first part of the network, the *ResNet34*, is used as our embeddings, which is 512-dimensional. Similar to face recognition embeddings, we only keep the videos if the number of frames with faces detected is more than 85% of the whole video, and missing embeddings are replaced with vector of zeros.

## 2.4. Audio Representation

Our audio preprocessing is the same as in [3]. On the 3-second audio clips from the videos, we apply a 512-point short-time Fourier transform on 25 ms long frames and 10 ms frame steps, using a Hann window. To reduce the impact of louder components, we also apply a power-law compression of 0.3 on the real and imaginary part. The network input signal is a tensor made of the real and imaginary time-frequency representations.

# 3. Results

## 3.1. Speech-speech separation

For evaluation, similarly to the training process, we mix two speech signals from an independent evaluation subset of AVSpeech to obtain our input audio-visual mixed clip. Note that the evaluation set does not contain any samples from the videos used in the training set. Thus, the speakers in the evaluation set are unseen to the algorithm. Lip reading embeddings from the frames are used as the visual input.

We measure the algorithm's performance under different levels of interference. We run evaluation tests on mixtures with different levels of average SNR. For each input SNR, we report different metrics scores for speech quality, the SNR, signal to distortion ratio (SDR), Short Term Objective Intelligibility (STOI) [11] and Perceptual Evaluation of Speech Quality (PESQ) [12]. Figure 1 shows the results for a network trained on 0 dB speech mixtures.

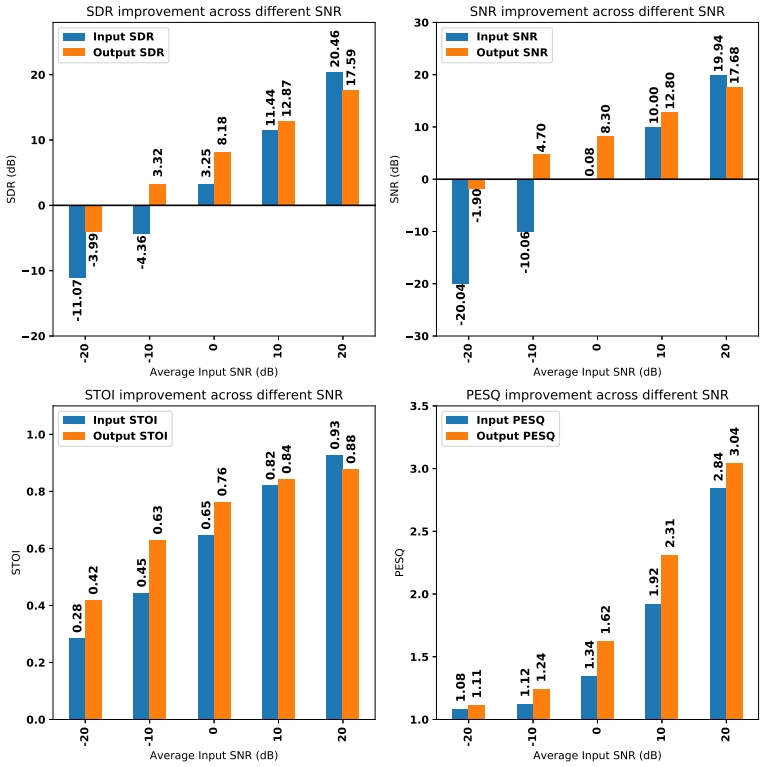


Figure 1: *Improvement for speech-speech separation task using lip reading embeddings, at different noise levels. Blue bars show the values for the unprocessed noisy input signal. Orange bars show the values after the audiovisual source separation.*

Our results suggest that the algorithm improves the intelligibility and quality of the input signal at every SNR level, except at very high SNR (20 dB). In this latter case, the decrease in quality is due to the time-frequency masking process which, while reducing the noise, distort the original signal. At this high level of the target signal, any distortion of it has a large impact.

### 3.2. Speech-noise separation

For this task, videos from AVSpeech dataset were mixed with audio samples from AudioSet. Lip reading embeddings are used as visual input. In order to see the possible advantages of visual data, we also trained an audio-only version of the network as the baseline. Results can be seen in Figure 2.

Our results show that, for both networks, the quality of the output signal is improved in all of the evaluation metrics. Listening to a few examples, we observed that the noise was efficiently removed, however, there were also small amounts of distortion and suppression on the target speech signal.

The results from audiovisual and audio-only methods are very close, with slightly better quality for the audiovisual one. In most of the cases, the noisy signal has a time-frequency signature that largely differs from speech, and this makes it easier for the network to separate it based on the spectrogram only. This also explains why the performance is better compared to speech-speech separation task, as speech signals overlap more and have similar spectral signatures.

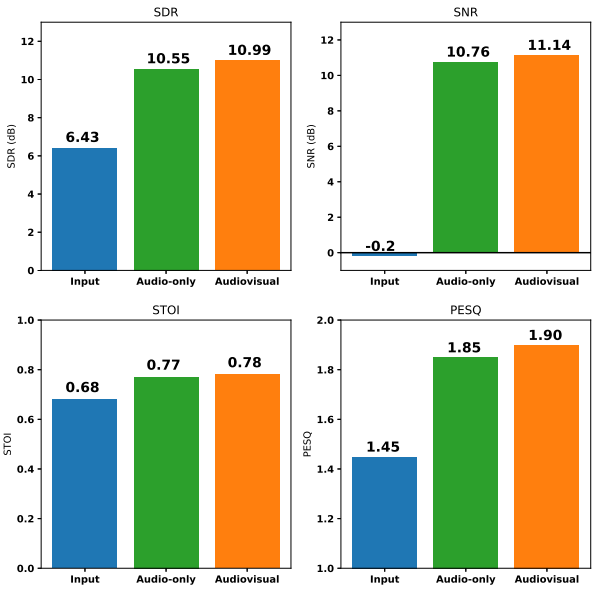


Figure 2: *Comparison of audio-only and audiovisual models on speech - noise separation, evaluated on mixtures with average SNR of 0 dB*

### 3.3. Robustness against occlusion of the face

In video conferencing, there may be cases where speakers' face and mouth will not be visible all the time. Also, given that

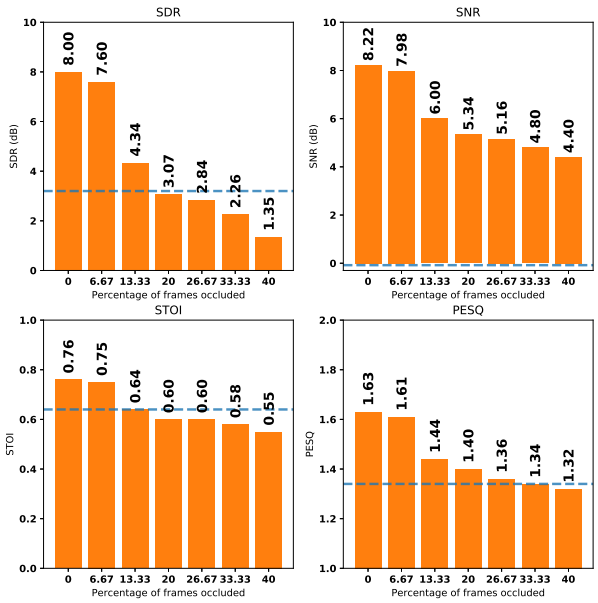


Figure 3: *Performance of speech-speech separation when face detection fails to detect the face during a variable period of time. The dashed horizontal lines show the evaluation of the input mixture without applying any separation. Beyond 20% of occluded frames, the quality of the output becomes worse than the unprocessed input signal.*

face detectors do not have perfect precision, there may be cases where the algorithm will not have a visual input during certain periods of time. Thus, we tested the robustness of the algorithm against occlusions, when the face detection fails.

During the training, we included samples that contain missing frames because of the imperfection of the face detection algorithm. These samples should help the network learn how to handle missing frames.

For the evaluation, in each video sample, we occluded a randomly placed, continuous segment of the video with different durations. Figure 3 shows the performance with respect to occlusions in speech-speech separation task at 0 dB SNR.

The results show that dropping a few frames does not affect the performance significantly. However, beyond 20% of occluded frames (0.6 seconds for a 25 frame-per-second video), the network does not improve the quality and even deteriorates it compared to the unprocessed input. It shows the limits in the robustness of the network.

### 3.4. Comparison of visual embeddings

In order to better understand the difference between face recognition and lip reading embeddings and their impact on the final speech separation, we have performed some further analysis. We created a subset of video frames from 2 different speakers, with labels indicating if the speaker was talking or silent. For each frame, we apply two dimensionality reduction algorithms, t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA), to the embeddings in order to visualize their distribution, and their ability to be separated. The results are shown on Fig.4. The lip reading embeddings provide a larger separation between talk and silence than between speakers, as opposed to face recognition embeddings. This is

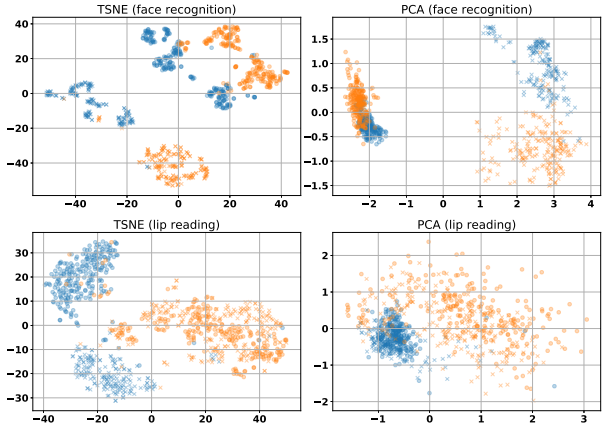


Figure 4: *Projections of face embeddings (top) and lip reading embeddings (bottom) for two speakers. The 2 speakers are distinguished by the marker's shape (circles and crosses) and the silent (in blue) / talking (orange) state is associated with the marker color. Face recognition emphasizes the distinction between speakers while lip reading separates more the speaker state.*

expected as face recognition is supposed to be invariant to face poses and focused on the differences of the faces between individuals. This explains why our network performs better with lip reading embeddings. It is much easier to detect and characterize lip movement and relate it to speech with this latter configuration. Still, it is possible to extract lip movement information from face recognition embeddings but it requires a much finer training of the visual part of the network, with more data and more epochs, as done in [3], and not (on purpose) in our study.

## 4. Conclusion

Our study confirms that the audiovisual separation network is able to make use of the visual clues of the target speaker to perform the audio source separation, especially when separating speech signal from another speech signal. It is robust and still works when interfering signals have a relatively high intensity. However, our assessment shows that the separation quality degrades when faces are not visible even for a short period of time.

The masking of audio spectrograms brings artifacts and distortion in the output audio signal. It is audible, sometimes not pleasant and may impair the intelligibility of the speech. New solutions should focus on alternative treatments of the audio signal to overcome this limitation.

The network is complex and requires a large amount of resources for training. Using lip reading embeddings, from a pretrained network, greatly reduces this need and enables experiments in a standard research lab environment. This makes a step forward to video conferencing applications.

## 5. Acknowledgements

We would like to thank Pingchuan Ma for providing the implementation for the lip reading network.

# 6. References


[1] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," *arXiv:1804.04121 [cs]*, Apr. 2018, arXiv: 1804.04121. [Online]. Available: http://arxiv.org/abs/1804.04121

[2] A. Owens and A. A. Efros, "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features," *arXiv:1804.03641 [cs, eess]*, Apr. 2018, arXiv: 1804.03641. [Online]. Available: http://arxiv.org/abs/1804.03641

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Jul. 2018, arXiv: 1804.03619. [Online]. Available: http://arxiv.org/abs/1804.03619

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[5] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," *arXiv:1703.04105 [cs]*, Mar. 2017, arXiv: 1703.04105. [Online]. Available: http://arxiv.org/abs/1703.04105

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 776–780. [Online]. Available: http://ieeexplore.ieee.org/document/7952261/

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, Jun. 2014, pp. 1725–1732.

[8] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[9] D. E. King, "Dlib-ml: A Machine Learning Toolkit," p. 4.

[10] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end Audiovisual Speech Recognition," *arXiv:1802.06424 [cs]*, Feb. 2018, arXiv: 1802.06424. [Online]. Available: http://arxiv.org/abs/1802.06424

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.