INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Deep audio-visual speech separation based on facial motion

Rémi Rigal[1], *Jacques Chodorowski*[1], *Benoît Zerr*[2]

[1]Orange Labs, Lannion, France
[2]ENSTA Bretagne, Lab-STICC, Brest, France

remi.rigal@orange.com, jacques.chodorowski@orange.com, benoit.zerr@ensta-bretagne.fr

Abstract

We present a deep neural network that relies on facial motion and time-domain audio for isolating speech signals from a mixture of speeches and background noises. Recent studies in deep learning-based audio-visual speech separation and speech enhancement have proven that leveraging visual information in addition to audio can yield substantial improvement to the prediction quality and robustness. We propose to use facial motion, inferred from optical flow techniques, as a visual feature input for our model. Combined with state-of-the-art audio-only speech separation approaches, we demonstrate that facial motion significantly improves the speech quality as well as the versatility of the model. Our proposed method offers a signal-to-distortion improvement of up to 4.2 dB on two-speaker mixtures when compared to other audio-visual approaches.

Index Terms: speech separation, deep learning, computer vision

1. Introduction

Speech is the predominant modality used by humans to interact and share information. Building a system that is capable of isolating one's speech in real-world acoustic environments is challenging yet essential to improve speech processing capabilities or remote communications between humans. As so, speech separation has became a burgeoning topic in the recent years, with many studies addressing various environments and setups.

While earlier work focused on unsupervised learning or multiple microphones setups, most recent studies use deep learning techniques on single-channel audio mixtures[1, 2, 3, 4], inspired by the ability of the human brain to focus on a specific acoustic source, a popular phenomenon called the *cocktail party effect*. This domain is rapidly evolving, the approach proposed in [5] for example is the first to be evaluated on up to five speakers mixtures. However, relying solely on a single-channel audio mixture to separate multiple speeches can be limiting, especially when associating speeches with their respective speakers at the utterance level. This permutation problem has been addressed but not completely solved by using clustering techniques on trained speech embeddings[6], or specific loss function during training that accounts for the possible permutations[7, 8].

Simultaneously, other approaches use an additional visual information to perform the speech separation. The benefits of audio-visual speech separation are threefold: First, lip movements and facial motion are inextricably connected to the acoustic characteristics of speech. Second, the visual information is unaffected by acoustic interferences, meaning that it becomes a robust alternative in extreme acoustic environments. Third, the visual and acoustic information being partially redundant, using both helps in solving the permutation problem. In addition, the previously mentioned *cocktail party effect* is known to be improved by visual data[9, 10].

Most existing audio-visual speech separation solutions use the static image stream as visual features, either processed with a pre-trained face recognition neural network[11, 12] or a face landmark detection algorithm[13]. However, neuroscience researches show that the human visual cortex contains two pathways, the ventral and the dorsal stream, that performs static object recognition and motion recognition respectively[14]. In computer vision, motion is typically inferred from a visual stream using dense optical flow techniques that estimate the speed and the movement direction of each pixel between two frames. These techniques are becoming increasingly popular for facial motion related tasks such as active speaker detection[15] or facial expression recognition[16]. Both tasks are quite similar to speech separation as they heavily rely on the analysis of motion discontinuities.

Inspired by recent progress in audio-only speech separation approaches, we propose in this paper a novel architecture of neural network for audio-visual speech separation. We mainly focus on the contribution of visual information to the separation process, especially the facial motion inferred by optical flow techniques. Our paper makes three main contributions:

1. We show that a deep neural network using time-domain audio signal and facial motion estimated from a visual stream can accurately isolate the speech of one or multiple speakers in a noisy environment.
2. We demonstrate that a neural network can learn and leverage the link between facial cues and acoustic speech signal for the purpose of speech separation.
3. We show that facial motion contributes more to the audio-visual speech separation than raw static images.

2. Neural network

2.1. Model design

Our deep neural network is using the intermediate fusion paradigm, meaning that the visual and acoustic streams from the videos are processed separately with different model architectures before being fused and processed together. Once processed, visual and acoustic features are fused using a concatenation-based fusion along the temporal axis. Both features are however sampled at different rates, the visual features are then upsampled to meet the rate of the acoustic ones by using a simple nearest neighbor interpolation.

The second stage of the network is the separation stage performed onto the concatenated feature tensor. The separated signals are then converted back to waveforms in the last stage by a 1-D transposed convolution layer which serves as a decoder.

Figure 1 shows the overall architecture of the network.

http://dx.doi.org/10.21437/Interspeech.2021-1560

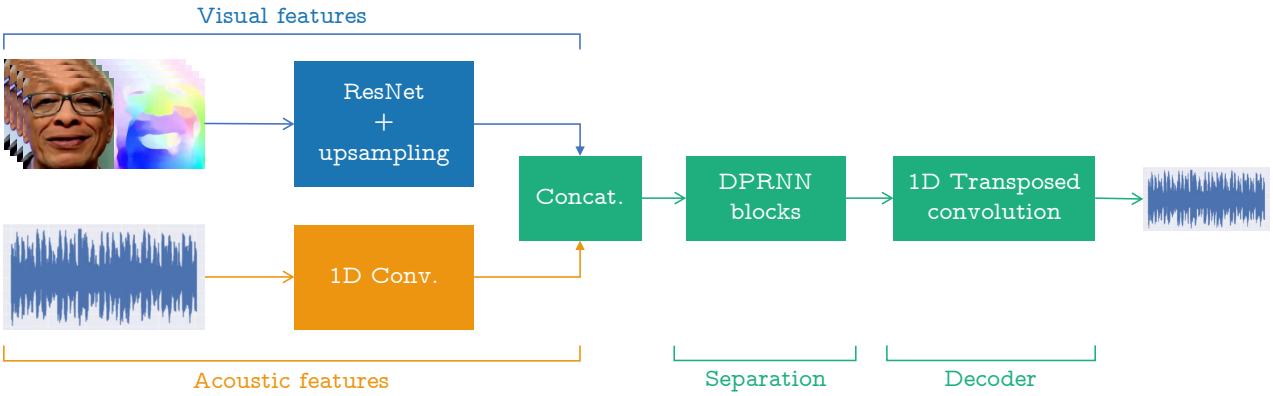

Figure 1: *Architecture of our neural network.*

2.2. Visual features

Images from the videos would be challenging to process as is since the active speaker's face may not be the only focus of the video and it is also common to have non-speaking people visible in the stream. The proposed methodology consists in detecting and isolating the active speaker's face across the frames before estimating the local facial motion with optical flow techniques.

The final visual input of the neural network is the raw images of the active speaker's face concatenated with their optical flow representations. In cases where there are multiple visible active speakers, the same process is applied independently to the face of each active speaker.

2.2.1. Face detection

We use the Face Alignment library from [17] to detect and crop the region around the active speaker's face. To account for the videos with multiple faces, we use the initial position of the active speaker given by the dataset as well as a Kalman filter to track the face across all frames. This approach has the benefit of being robust to temporary occlusions and false negatives from the face detection library. The obtained images are then aligned and reshaped to the normalized size of 96 x 96.

The face landmarks estimated by the Face Alignment library are later used for the face distortion process detailed in section 2.2.2.

2.2.2. Face distortion

The regions of the face around the lips and cheekbones participate the most to the speech separation task[11]. In order to increase the size and details of these regions, we apply a distortion to the face of the speaker. To do so we define the target face landmarks positions presented in Figure 2. This particular pattern has two benefits, on the one hand it minimizes the background scene that can commonly be seen in the two lower corners of the cropped image, on the other hand it increases the details found around the lips and cheekbones, resulting in more relevant data for the network to learn on. Using the position of the face landmarks estimated during the face detection stage, we compute a homography by matching them with the target pattern. The homography is then applied to the raw image of the face.

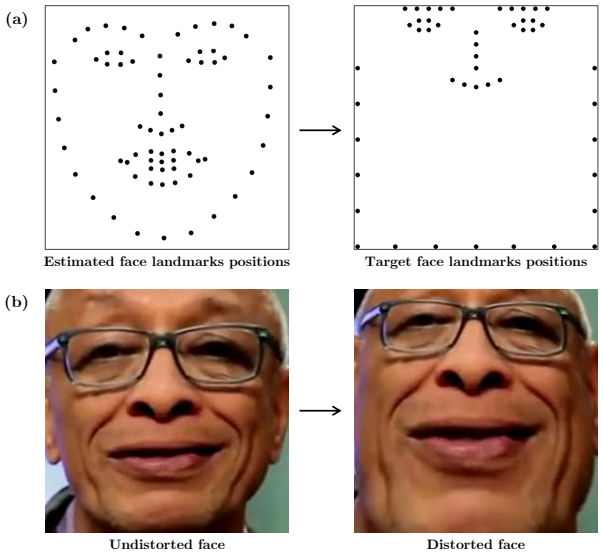

Figure 2: *Illustration of the face distortion process. We compute a homography based on estimated and target face landmarks positions (a), then we apply the distortion to the raw image (b).*

2.2.3. Optical flow

The extensive evaluation study done in [18] shows that two optical flow techniques outperforms the others on facial data: Farnebäck[19] and Flow Fields[20]. We choose to use Flow Fields to estimate the facial motion. In order to increase the displacement, the optical flow is computed between the third previous frame and the current frame. The new image of size 96 x 96 obtained for each pair of frames is concatenated with the raw image of the current frame.

If the face of the speaker is occluded during one or multiple frames, the last frame containing the face is used to compute the optical flow.

2.2.4. Visual features processing

The input visual information composed of the raw images of the active speaker's face concatenated with their optical flow representations is processed using a customized ResNet[21] architecture with only a few layers. In addition, the fully connected layer in the original ResNet is replaced with a global

average pooling layer, more robust to spatial translations of the input[22]. Our ResNet is configured to predict a feature vector of size 64 for each frame of the visual input.

2.3. Acoustic features

The audio signal from the videos is processed directly in the time domain with a linear 1-D convolution encoder. This type of encoder has proven to be particularly effective at learning a latent representation of a waveform audio signal[3, 23]. This approach replaces the time-frequency representations such as the short-time Fourier transform (STFT) typically used in audio processing tasks. Furthermore the latent space is jointly learned with the rest of the network which results in a more appropriate representation. Similarly to [3], the number of filters for the encoder is set to 64.

2.4. Separation network

The separation stage of the network is similar to the dual-path RNN (DPRNN) from [23]. We keep the three-stages architecture of DPRNN, respectively segmentation, block processing and overlap-add. The latter is used to reconstruct the separated signal representations from the overlapping segments obtained from the segmentation.

3. Experimental setup

3.1. Datasets

We choose to use the AVSpeech[11] dataset for training and evaluation, mainly because it has no restriction on languages and it contains videos recorded in a wide range of environments and conditions. The dataset consists of YouTube videos with a unique speaker facing the camera most of the time. The duration of the clips ranges from 3 to 10 seconds, with variable video resolutions and audio sample rates. Due to limited computational power only a random subset of 62253 clips from the dataset is used. 90% of the clips are taken for the training set, and the remaining 10% for the evaluation set. The clips from both sets are taken from distinct videos.

In order to simulate real world environments, the AVSpeech dataset is paired with the AudioSet[24] dataset for non-speech audio clips. The latter consists of samples from YouTube videos which are annotated with one or multiple sound classes. We excluded speech related sound classes such as *Speech* or *Singing* so that the samples don't interfere with the speeches from the AVSpeech dataset. The *Crowd* sound class was however kept as it is a frequent interference type in practical use cases.

The neural network is trained and evaluated using synthetic mixtures generated by mixing samples from both AVSpeech and AudioSet datasets in various configurations. The most versatile configuration is having a visual stream only for one speaker whereas the audio signal contains the streams of one or multiple speakers, eventually mixed with non-speech noises. This setup is tailored for use cases where the target is to focus on one specific speaker and the general context is unknown, video conferences for example. Configurations with an even number of visual and audio streams are considered too.

3.2. Dataset refinement

A dataset with important disparities such as AVSpeech is a real benefit since the network will learn an exhaustive set of configurations resulting in better real-world predictions. However, while being heterogeneous the samples have to be of utmost quality as they are used as ground truth for the loss function. We refine the dataset by discarding some of the samples based on two criteria: the visibility of the speaker and the quality of the speech segments.

The visibility of the speaker is measured with the face detection process detailed in Section 2.2.1. We use 3D face landmarks to compute the pan and tilt angles of the head of the speaker. Clips where the latter angles reach extreme values, respectively 35°and 30°, for more than half of the duration of the clip are discarded. In addition, if the face is not detected in more than 15% of the frames the clip is discarded too.

The quality of the speech segments is assessed by the Waveform Amplitude Distribution Analysis SNR (WADA-SNR) algorithm[25]. This algorithm is designed for estimating the signal-to-noise ratio of speech signals. Clips where the resulting estimated SNR is below 15 dB are discarded.

As a whole, the refinement process leads to a rejection rate of less than 2%.

3.3. Training

Considering the large size of the mixture set and its disparity, the entire network is trained for only 3 epochs. Adam[26] is used as the optimizer with an initial learning rate of $1e^{-3}$ halved after each epoch. The network is trained on samples with a duration of 3 seconds in which the audio is down-sampled to 8kHz. All the video clips are normalized to 25 frames per second, resulting in 75 images per sample for each visible speaker.

The usual metric used for speech separation tasks is the signal-to-distortion ratio improvement (SDRi)[27]. The network is thus trained with the objective of maximizing the SDRi.

4. Results and discussion

4.1. Results

Table 1: *Quantitative analysis and comparison of speech separation performance in SDRi (dB) as a function of the number of speakers (S), the number of visible speaker faces (F) and the presence of noise (N) in the audio mixture.*

	1S+1F+N	**2S+1F**	**2S+2F**	**2S+2F+N**
DPRNN	17.5	13.2	13.2	13.8
Ephrat et al.	16.0	9.9	10.3	10.6
Ours	**17.6**	**13.6**	**14.3**	**14.8**

Our network is compared with the model proposed by Ephrat et al. in the original paper that introduced the AVSpeech dataset[11]. This deep neural network uses an intermediate fusion paradigm where the visual stream is processed using a face recognition model and the time-frequency representation of the audio stream is processed with a dilated convolution network. The separation is performed with a bidirectional LSTM and fully connected layers that predict a complex time-frequency ratio mask for each speaker.

In addition, we compare our network to the dual-path RNN (DPRNN) model which serves as an audio-only baseline. We choose this particular baseline because our network is quite similar to it. The results shown for this model are obtained from our own implementation, trained and evaluated using the acoustic streams from the clips of our dataset. Due to the high disparity of recording qualities, speech languages and encodings in

our dataset, our results for this model are unsurprisingly lower than the one obtained by the authors on the cleaner WSJ0-2mix dataset.

The results shown in Table 1 demonstrate that our model outperforms both baselines in every configurations. It should however be noted that such results between our model and the audio-visual baseline were to be expected. As a matter of fact, the dual-path RNN model our network is inspired from significantly outperforms the audio-only model from [7] that inspired Ephrat et al.

In comparison with DPRNN, our model performs similarly on one-speaker mixtures but produces a gain of 0.4 dB to 1.1 dB on two-speaker mixtures. We can also see that the constraints imposed by having two visual streams yields better results.

4.2. Additional experiments

4.2.1. Speaker focused separation

In real-world applications such as video conferences or social robotics it is common that the main target is to focus on a specific speaker independently from it being the predominant voice or sound in the acoustic mixture or not. By using dual-path RNN as the audio-only speech separation baseline, we compare the ability of the networks to focus on one of the speakers, denoted *main speaker*. To this end, we generate two-speakers mixtures by mixing samples from the AVSpeech dataset at various levels, from +6 dB where the *main speaker* is the predominant voice to -9 dB where the second speaker is predominant. The models are configured to output only one clean speech. For each mixing level and for both models, we measure the number of times the speech from the *main speaker* is predicted, independently from the quality of the speech.

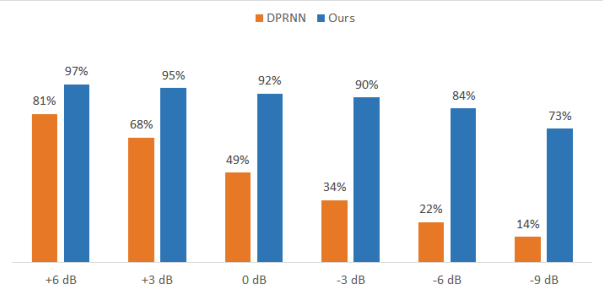

Figure 3: *Comparison of performance between audio-only and audio-visual models on speaker focused separation. Values depict the success rates obtained for various mixing levels of two-speaker mixtures.*

Figure 3 shows that the audio-only speech separation network tends to unsurprisingly favor the predominant speech. The acoustic signal given to the network has indeed no other information than speech power levels to discriminate the *main speaker* from the other one. This limitation illustrates the permutation and assignment problem inherent to audio-only approaches. Conversely, the results obtained with our audio-visual network suggest that the visual information contributes extensively to the assignment between a speech and its speaker. In fact, the success rate of our model reaches 97% at best and drops only to 73% when the voice of the *main speaker* becomes barely audible over the voice of the other speaker.

This experiment shows that our neural network has effectively learned the intricate connection between facial cues and acoustic speech signals. Ultimately it shows that if available, visual information can substantially improve the robustness and versatility of a speech separation algorithm in real-world applications.

4.2.2. Robustness to partial visual information

In order to better understand the contribution of the visual features to the separation task, we conduct an experiment in which we discard partially or completely the input visual information. The study is made on the two-speaker setup with only one visual stream, the network is trained with both modalities. Three configurations are tested and compared to the standard setup: (a) the visual information is discarded entirely, (b) only the static raw images are kept and (c) only the facial motion is kept. This experiment is conducted on the entire evaluation set, and discarded values are simply set to zero.

Table 2: *Analysis of the robustness of our trained model on partial visual information. The model is trained for two-speaker mixtures with one visual stream.*

	SDRi (dB)
Full visual information	**13.6**
No visual information	8.1
Raw images only	8.3
Facial motion only	9.8

Table 2 shows that while our trained model performs poorly when some input visual information is missing, it relies primarily on the facial motion for the visual features. We can see that the facial motion increases significantly the separation ability of the network with a gain of 1.7 dB over completely discarded visual information whereas the raw images offer only a slight gain of 0.2 dB.

In all cases, the loss induced by having partial visual information is significant, this is the consequence of using a concatenation-based fusion of the features which results in the network jointly learning the contribution of the acoustic and visual data. However, the predictions are still acceptable and the loss is balanced between all the cases, highlighting the robustness of our model.

5. Conclusion

In this paper, we introduced a novel deep neural network architecture for speaker independent audio-visual speech separation based on time-domain audio and facial motion. Our study was mainly focused on the contribution of the visual information to the separation ability. We demonstrated that our neural network is capable of accurately separating speeches from different speakers by using both visual and acoustic features jointly. Our method offered a significant gain over the state-of-the-art approaches on the AVSpeech dataset.

We conducted two experiments to better evaluate our model. First, we showed that our network learned and took advantage of the visual information to perform the speech separation task. Second, we demonstrated that facial motion contributes more to the prediction of our model than the raw images of the active speaker's face.

In future work, we plan on evaluating the relevance of the face distortion process as well as analyzing the robustness of our model to temporary facial occlusions.

6. References

[1] X. L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[2] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 8, pp. 1256–1266, sep 2018.

[4] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End Speech Separation by Speaker Clustering," *arXiv e-prints*, p. arXiv:2002.08933, Feb. 2020.

[5] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 7164–7175.

[6] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *CoRR*, vol. abs/1508.04306, 2015.

[7] D. Yu, M. Kolbaek, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 241–245.

[8] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multi-talker speech separation and tracing with permutation invariant training of deep recurrent neural networks," *CoRR*, vol. abs/1703.06284, 2017.

[9] K. L. Shapiro, J. Caldwell, and R. E. Sorensen, "Personal Names and the Attentional Blink: A Visual "Cocktail Party" Effect," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 2, pp. 504–514, 1997.

[10] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra, "Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space," *PLoS ONE*, vol. 4, no. 3, p. e4638, mar 2009.

[11] 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, apr 2018.

[12] B. İnan, M. Cernak, H. Grabner, H. P. Tukuljac, R. C. Pena, and B. Ricaud, "Evaluating audiovisual source separation in the context of video conferencing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, 2019, pp. 4579–4583.

[13] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[14] M. A. Goodale and A. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[15] C. Huang and K. Koishida, "Improved active speaker detection based on optical flow," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2020-June, 2020, pp. 4084–4090.

[16] B. Allaert, I. M. Bilasco, and C. Djeraba, "Consistent optical flow maps for full and micro facial expression recognition," *VISIGRAPP 2017 - Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, pp. 235–242, 2017.

[17] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[18] B. Allaert, I. R. Ward, I. M. Bilasco, C. Djeraba, and M. Bennamoun, "Optical flow techniques for facial expression analysis: Performance evaluation and improvements," *CoRR*, vol. abs/1904.11592, 2019. [Online]. Available: http://arxiv.org/abs/1904.11592

[19] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, J. Bigun and T. Gustavsson, Eds., vol. 2749. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.

[20] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," *CoRR*, vol. abs/1508.05151, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[22] M. Lin, Q. Chen, and S. Yan, "Network In Network," *arXiv e-prints*, p. arXiv:1312.4400, Dec. 2013.

[23] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 46–50, 2020.

[24] 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)*, 2017, pp. 776–780.

[25] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2598–2601, 2008.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.