ISSN 1870-4069

Analysis of Speech Separation Methods based on Deep Learning

Jessica Rincón-Trujillo, Diana Margarita Córdova-Esparza

Universidad Autónoma de Querétaro, Facultad de Informática, Querétaro, Mexico
{jessssy_1993,diana_mce}@hotmail.com

Abstract. In this paper, we perform an analysis of speech separation methods using deep learning. We provide a review of the literature, discussing the main features of the works based on audio and audiovisual processing, as well as the deep learning method. For the analysis of the articles, we provide a description where we identify the category, characteristics, methodology, results, and application. According to the study, we observed that the acoustic-based algorithms require audio portions of voices with external interferences to improve the intelligibility and the quality of the voice signals. Audio-visual-based methods use thousands of hours of segments of noisy videos to obtain stable performances and enhancing the quality of the separation. This latter category, also known as the cocktail party problem represents an ongoing open problem in the deep learning community.

Keywords: Deep Learning, Speech Separation, Computer Vision.

1 Introduction

According to Wang & Chen [23], speech separation is the task of dividing the target voice from background interference. Speech separation from multiple sources of sound is defined by Cherry [3] as cocktail party problem. It can be said that the separation of speech isolates the source of the sound. For this, the separation corresponds to the segregation of the auditory current. Speech separation can be used for assistive technologies such as hearing aids and cochlear implants, automatic captioning, dictation systems, and solving multi-speaker simultaneous speech [5, 4].

The first systematic study on flows segregation carried out by Miller [18] establishes that listeners divide a signal with two alternating sinusoidal wave tones in two flows. He also reviewed human intelligibility scores when he interfered with a variety of tones, broadband noise, and other voices.

In the 1950s, Rosenblatt created some useful brain analogs for analytic tasks [20]. His research began the learning of machines, developing techniques that allow computers to learn and classify as human beings do. Having as main objective *the ability to recognize and process complex patterns of information related from all dimensions, such as the human brain.* Rosenblatt was the one who invented the Perceptron, an artificial neuron or basic unit of inference, on which the

multi-layered learning networks are based, which are the basis of what is known as Deep Learning [24]. In the context of artificial intelligence (AI), deep learning according to Gómez, et al. [12] refers to the automatic activity of knowledge acquisition, through the use of machines that use several levels for extraction.

Nowadays, due to massive amounts of data, deep Learning-based methods have pushed the state-of-the-art beyond what was possible with traditional techniques either using audio-only sources or audiovisual involving audio signal processing and computer vision [19].

1.1 Audio-based Methods

Below we discuss articles that address this topic in the acoustics category and the results accomplished.

The work developed by Simpson, et al. [21] proposes a statistical analysis based on the hypothesis of the separation of sources, making use of twelve models already tested on voice separation and with fifty pieces of music which were produced professionally. They used non-parametric statistics, establishing reliable evidence for significant conclusions about the performance of the various models. They conclude with the design of a basic procedure based on the hypothesis for the non-rigorous statistical analysis of the results of the source separation model. They obtained reliable evidence in the study. However, they found no evidence of any significant difference between their two best models based on Deep Neural Networks (DNN).

The proposal implemented by Gao, et al. [11] describes a unified voice improvement framework for jointly managing both background noise and interfering speech in a speaker-dependent scenario based on deep neural networks (DNN). They explore the speech improvement that depends on the speaker optimizing the performance of the system in comparison with the independent speakers; they considered the interference as a type of noise and demonstrated that their system could achieve performance comparable to specific systems where only noise or voice interference is present. They then use a framework based on joint learning to further improve system performance in low signal-to-noise (SNR) environments.

In the present year, Dadvar et al. [7] describe a robust binaural voice separation system based on a deep neural network (DNN). The system is based on three main stages of processing. The first stage deals with the spectral processing from multiresolution cochlea characteristic extracted from the signal beamforming. The second stage comprises spatial processing, based on a spatial feature of a softly-masked binaural cue (smITD + smILD) obtained by soft masking missing data from binaural signals. The third stage contains a deep neural network with combined spectral and spatial characteristics, designed for noisy and reverberant conditions. The proposed system is evaluated and compared with the two recent binaural voice separation systems as baselines in various noisy

and reverberant conditions. They show that their system exceeds current baseline systems because improving the intelligibility and quality of separate speech signals in reverberant and noisy conditions. The results confirm the efficiency of each component of the system, especially in highly reverberant scenarios.

1.2 Audiovisual-based Methods

Next, we present the articles in the audio-visual category and the results they have obtained.

Khan, et al. [15] proposes a study into whether visual speech information can be used to assist in the estimation of the mask for the separation of audio speakers to improve speech quality and intelligibility. They developed two visual methods for the separation of the speakers, which use a deep neural network that assigns the visual characteristics of the speech to a space of audio attributes from which the visually derived binary masks and the visually derived relationship masks are estimated before the application to the mix of voices. The other method entails masking in relation to the audio, forming a line focus for speaker separation. The collection of the audio mix from the speakers is made from a single microphone, and the visual characteristics of the speech are extracted from the mouth of each speaker in the combination. The audio speech characteristics are subsequently estimated for each speaker from the corresponding visual components and then used within the proposed mask estimation methods. Obtaining good results in situations in which the two speakers are equidistant from the microphone and have a similar volume. However, the accuracy of the mask reduces when the audio power of the two speakers differs.

In the same year Afouras, et al. [1] propose an audiovisual neural network that can isolate the voice of a speaker from other people, obtained from a noisy audio signal and the corresponding speaker video, producing an improved audio signal containing only the speaker's voice with the rest of the speakers and the background noise suppressed. The proposed model is evaluated for five simultaneous voices, demonstrating both qualitative and quantitative reliable performance. The performance of the model was put to the test in open environments. It consists of two modules: a magnitude sub-network and a phase sub-network. The first sub-network receives the magnitude spectrograms of the noisy signal and the video of the speaker as inputs and generates a smooth mask. Then, the input magnitudes are multiplied elementally with the mask to produce a spectrogram of filtered scale. The prediction of the scale, together with the phase spectrogram obtained from the noisy signal, is introduced into the second subnet, which produces a residual phase. The residue is added to the noisy stage, generating the improved phase spectrograms. Finally, the enhanced magnitude and phase spectra are transformed back into the time domain, producing the improved signal.

Another work, developed by Ephrat, et al. [9] establishes an audiovisual model to isolate a single voice signal from a mix of sounds such as speakers and background noise. The input is a video with one or more people speaking, the speech of interest interfered with by other speakers and background noise.

Both auditory and visual characteristics are extracted and incorporated into the joint model for audiovisual separation. The output is a decomposition of the audio track input into clean voice tracks, one for each person detected in the video. Videos are obtained in which the speech of specific people is improved while the rest of the sound is suppressed. The model is based on a neural network that incorporates visual and auditory signals. They introduced an audiovisual dataset called AVSpeech composed of thousands of hours of video segments from the web. Demonstrating the applicability of their method to classical tasks of separation of voices, as well as to real situations. Their demo only requires the user to specify the face of the person in the video whose speech they wish to isolate. They make two main contributions: A model of audiovisual separation by voice that surpasses the audiovisual and unique models in the classic tasks of voice separation. A new set of audiovisual data on a large scale, AVSpeech, composed of segments of videos in which the audible sound belongs to a single person, visible on the video, and without audio background interference.

The purpose of this work is to provide an overview of the state-of-the-art techniques for speech separation using either audio-only or audiovisual input sources.

2 Methodology

In this section, we present the analysis of multiple deep learning-based speech separation studies using acoustic and audiovisual sources.

For our analysis, we considered the following: category, features, method, results, application, advantages, and disadvantages.

2.1 Acoustic Category

Simpson et al. [21] analyzed separation results on twelve different models, as data they used 50 pieces of recorded music. They applied statistical analysis of the results for audio-based speech separation discovering that the highest performance models were the ones that used deep learning.

Chen et al. [2] used the ideal Radio Mask (IRM) as the target of supervised learning. The IRM is estimated from the 64-channel cochleagrams of a combination of plain speech and noise. According to [22], the cochleagram is a time-frequency representation of an acoustic signal. They showed that random frequency noise perturbations on the spectrogram gives the best speech separation results in classification accuracy. They found that when training a DNN, the quality of the relationship mask improves as the classifier has been exposed to more noise interference scenarios. A disadvantage is that when a training set is created from limited speech and noise resources, a classifier will likely adapt to the training set and make wrong predictions in a test set, especially when the background noise it is highly non-stationary, it is suggested to expand the noise resources.

Gao et al. [11] used clear speech and speech interference followed by log-power spectral (LPS) features [8] of both speeches to train an ensemble of deep neural networks with dropout. Their system were able to decrease background noise and speech interference in speaker-dependent situations. They found that a speaker-dependent system is much more robust than one independent of the speaker, can unify speech improvement and speech separation and that it is possible to achieve better performance for mixed conditions of noise and voice interference. However, improvements can still be made for environments with high SRN and speech improvement architecture.

Liu and Wang [16] inspired by auditory scene analysis [22], they decompose the task into two stages, a frequency-domain concurrent grouping step, and a time-domain grouping step. The two stages are trained combined and using recurrent neural networks achieving state-of-the-art performance using permutation invariant training (PIT) and deep clustering (DC). The experiments show that the proposed system improves on the best-reported results of PIT and DC. The training of the first two stages can be done together and sequentially (1st phase of simultaneous grouping to separate two speakers in the frame level, 2nd sequential grouping to transmit spectra).The proposed approach takes advantage of both variable permutation training and deep grouping and has been shown to produce better results than both methods. The disadvantage is that the algorithm does not perform in real-time.

Dadvar and Geravanchizadeh, [7] used Multiresolution cochleagram (MRCG) features extracted from the beamformed signal and spatial features of softly-masked binaural cues (smITD + smILD) to train a supervised deep neural network with mixed spectral and spatial features to determine an Ideal Radio Mask (IRM) based on the signal-to-reverberant noise ratio.The results revealed that MRCG is the best spectral feature in the binaural speech separation task in noisy and reverberant environments. The application of their system is the binaural speech separation task in noisy and reverberant environments. They show that the proposed system exceeds baseline systems in a variety of simulated conditions, especially in low SNR and high RT scenarios. They compared their system proposed with two binaural separation systems and RT demonstrate the superiority of the proposed system in terms of the gains of STOI, ESTOI and PESQ. As a disadvantage, the system needs to be extended to be applicable in real environments.

2.2 Audiovisual Category

Khan et al. [15] used visual features taken from Active Appearance Models (AAM) from each speaker in the mixture. Audio speech features are afterward estimated for each speaker from the equal visual features and used within the suggested mask estimation methods. They trained a deep neural network to map a stack input spectral features to a ratio mask. The mask is later applied to the noisy mix to determine the target speaker spectral features. They obtained the

highest performance when merging audio and visual information to create the masks. Applications of their system entail situations where the two speakers are at the same distance from the microphone and have comparable loudnesses. They found that the audiovisual masking that combines visual and audio masks offers higher performance in all the tests carried out as well as in all the SNR. As disadvantage, the resulting voice of the resulting masks was of lower quality and sometimes intelligible compared to the result of the relationship masks. - When the audio power of the two speakers differs the accuracy of the mask is reduced.

Afouras et al. [1] used visual features extracted from images with a spatio-temporal residual network. Acoustic features are derived from the audio waveforms using Short Time Fourier Transform (STFT) with a Hann window function to generate spectrograms. They trained a Convolutional Neural Network (CNN) capable of producing clear speech from noisy audio segments recorded in real environments. They used two databases: LRS2 and VoxCeleb2. As a disadvantages, the LRS2 database contains information only in English and the method can fail in conditions where there is a lot of noise.

Ephrat et al. [9] used the AVSPEECH Dataset containing around 4700 hours of video fragments with nearly 150,000 distinct speakers to train a multi-stream architecture that takes visual data from faces and noisy audio to generate spectrogram masks for each face detected in the video. The spectrograms are then used to obtain isolated speech signals for each speaker suppressing other interfering signals. They obtained state-of-the-art results on speech separation as well as a potential application to video captioning and speech recognition.The proposed model was tested in different videos which contained various types of noise (in a bar, restaurant, debates, etc.). The model works well in situations with a lot of background noise or several people talking. They carried out several tests, that allowed them to observe the results of the proposed model in different scenarios. They used 3 Mandarin databases: TCD-TIMIT and CUAVE databases. The only current disadvantage is that the method does not work in real time without a powerful GPU due to faces movement.

Lu et al. [17] used the WSJ0 Dataset (audio-only) [14] that contains 30 hours of training data as well as the GRID Dataset [6] containing 34 speakers each with 1000 frontal video recordings. They trained an Audiovisual matching network obtaining improvements on speech separation quality over the state of the art of audio-only speech source separation. They found that When the audio is not separated correctly, the audiovisual focus remains stable and in some cases, the correct audio separation is not achieved.

Gabbay et al. [10] used the GRID Dataset [6] and the TCD-TIMIT dataset [13] consisting of 60 speakers with about 200 videos each to train a Convolutional neural network that takes the frames of silent video as input, and predicts sound features that are converted into intelligible speech. Compared to audio-only techniques, their method is not influenced by similar speech vocal components generally observed in same-gender speech separation. Their system has no problems caused by the gesticulation. As disadvantages, they did not perform testing with

multiple people talking and according to the tests carried out when working with TCD-TIMIT vid2speech generates unintelligible content.

3 Results

According to our analysis, we observe that in the case of the acoustic category the deep learning methods that use pieces of music or audios voices with external interferences (noise) improve the intelligibility and quality of voice signals.

The articles presented in this category use deep neural networks, based on non-parametric supervised recurrent neural networks characterized by having the required information to ensure adequate data sets to improve the intelligibility and the quality of signals for speech separation in reverberant situations. In the case of the Audio-Visual category, we observed that its main feature is that it requires high amounts of training data, comprising thousands of hours of noisy video segments.

Some of the advantages found in the combination of audio and visual sources are stable performances improving the quality of speech separation, therefore associating the separate voice tracks with the speakers visible in the video.

4 Conclusions

In this article, we carried out an analysis of deep learning-based speech separation methods. Compared to traditional techniques, deep learning techniques have had a breakthrough using acoustic-only sources, improving the clarity and quality of voice signals.

For audiovisual sources, there have been significant advances, obtaining stable performances, and enhancing the quality of the separation. However, there is still active work in progress to solve the so-called called cocktail party problem on unconstrained environments.

Acknowledgements. The authors wish to acknowledge the financial support for this work by the Consejo Nacional de Ciencia y Tecnología (CONACYT) through financial support scholarship number (CVU): 925734. We also want to thank Universidad Autónoma de Querétaro (UAQ) through project number FIF-2018-06.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121 (2018)
2. Chen, J., Wang, Y., Wang, D.: Noise perturbation for supervised speech separation. Speech communication 78, 1–10 (2016)
3. Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America 25(5), 975–979 (1953)

4. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3444–3453. IEEE (2017)
5. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision. pp. 87–103. Springer (2016)
6. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120(5), 2421–2424 (2006)
7. Dadvar, P., Geravanchizadeh, M.: Robust binaural speech separation in adverse conditions based on deep neural network with modified spatial features and training target. Speech Communication 108, 41–52 (2019)
8. Du, J., Huo, Q.: A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In: Ninth Annual Conference of the International Speech Communication Association (2008)
9. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018)
10. Gabbay, A., Ephrat, A., Halperin, T., Peleg, S.: Seeing through noise: Visually driven speaker separation and enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3051–3055. IEEE (2018)
11. Gao, T., Du, J., Dai, L.R., Lee, C.H.: A unified dnn approach to speaker-dependent simultaneous speech enhancement and speech separation in low snr environments. Speech Communication 95, 28–39 (2017)
12. Gómez Gil, M.: Aprendizaje profundo, el poder del aprendizaje automático unido al poder de cálculo de las computadoras actuales (2016)
13. Harte, N., Gillen, E.: Tcd-timit: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia 17(5), 603–615 (2015)
14. Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 31–35. IEEE (2016)
15. Khan, F.U., Milner, B.P., Le Cornu, T.: Using visual speech information in masking methods for audio speaker separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10), 1742–1754 (2018)
16. Liu, Y., Wang, D.: A casa approach to deep learning based speaker-independent co-channel speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5399–5403. IEEE (2018)
17. Lu, R., Duan, Z., Zhang, C.: Listen and look: Audio–visual matching assisted speech source separation. IEEE Signal Processing Letters 25(9), 1315–1319 (2018)
18. Miller, G.A., Heise, G.A.: The trill threshold. The Journal of the Acoustical Society of America 22(5), 637–638 (1950)
19. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S.: A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR) 51(5), 92 (2018)
20. Rosenblatt, F.: The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory (1957)
21. Simpson, A.J., Roma, G., Grais, E.M., Mason, R.D., Hummersone, C., Liutkus, A., Plumbley, M.D.: Evaluation of audio source separation models using hypothesis-

driven non-parametric statistical methods. In: 2016 24th European Signal Processing Conference (EUSIPCO). pp. 1763–1767. IEEE (2016)
22. Wang, D., Brown, G.J.: Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press (2006)
23. Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10), 1702–1726 (2018)
24. Wason, R.: Deep learning: Evolution and expansion. Cognitive Systems Research 52, 701–708 (2018)