Modified Speech Separation Deep Learning Network Based on Hamming window

Hussein A Al-Barhan¹², Sinan M Elyass², Thamir R Saeed², Ghufran M Hatem³ and Hadi T Ziboon²

¹Kerbala University, Kerbala, Iraq
²University of Technology, Electrical Engineering Department, Iraq
³Najaf Technical College Al-Furat Al-Al-Awsat Technical University, Iraq

E-mail of corresponding author: 50257@uotechnology.edu.iq

Abstract. Speech separation is attracting widespread interest due to the sound mixing in real environments in and out door applications. Although the researchers have used many algorithms, the separation rate in the real environment is still poor. This paper presents speech separation using a modified Deep learning neural (DLN) algorithm. Interestingly, the modification has reduced the complexity of the original DLN algorithm, while, high value of separation rate has been gained caused by using Hamming instead of Hanning windows against the other algorithms. The separation rate reaches 98.6%, while, the advancement over the nearest algorithm is 2.8%.

1. Introduction

The importance of speech separation is as resulting from the nature of mixing and interference of voice sources in real environments. Thereby, the speech separation is attracting more interest due to its ability to extract a certain speech source from the mixing background sources [1]. Therefore, it has been utilized in signal processing field in many sectors; speaker recognition, hearing aid design [2], telecommunication applications [3], and determining the number of independent voice sources [4]. The major current focus on the cocktail party problem as an important application in speech processing [3]. The speech separation has been implemented by many signal processing algorithms such as: beamforming, Principal components analysis (PCA), Singular value decomposition (SVD), and Independent component analysis. (ICA), Dependent component analysis (DCA), Non-negative matrix factorization (NMF), Stationary subspace analysis (SSA), and Common spatial pattern (CSP) [5]. While, many researchers in the recent years have focused on a useful method like the human ability for the speech separation field, intelligent algorithms are attracting widespread interest in that field. Because, the intelligent algorithms always simulate the human ability, unlike other algorithms, in which, the separation has been treated as matrix algebra. Unfortunately, this treatment does not always guarantee optimal separation and the separation rate is reduced.

While, the overperform of the intelligent algorithms is caused by treats with the time, frequency, and statistical properties of the mixed and original sources [6, 7], strongly, the separation rate is improving, especially when using the DLN algorithm [8].
In [9] different DLN’s were applied, with different cost functions and masks, to maximize the estimated ratio of the sources. In this context, the deep neural network (DNN) for different noise conditions has been analyzed, while, the algorithm has been evaluated by STOI [8]. [10] Investigated his proposed with reverberant environments and for multi-channel separation based on speaker direction estimation. [11] Analyzed with the same environmental condition, but with the assumption of clean signals. [12] Presented an audio source separation based on the PGD-style generative priors to optimize the loss function using an inference-time technique with varying of known sources.

However, the use of the intelligent algorithms might not be completely succeed, practically with close feature sources and noisy environments due to the form the filter’s side-lobes interference.

This paper presents evaluation and modification of the DLN algorithm for speech separation. Taking in the account the effect of window filter on the separation process, attention is paid to the window’s filter sidelobe, so, Hamming instead of the Hanning window has been used.

2. Separation Algorithm

Deep learning neural networks (DNN) were used in this study to perform the speech separation task. Such architecture was inspired by the hierarchical structure of the brain. Deep neural networks feature a hierarchical, layer-wise arrangement of nonlinear activation functions (neurons) fed by inputs scaled by linear weights (synapses). Since the brain is adept at abstraction, it is anticipated that deep neural architecture might somehow capture abstract representations.

Time-frequency representation by short time Fourier transform (STFT) was unrestrained of time or frequency only, which allows the samples to move freely, and the signal energies as features are extracted. Where the STFT has been widely utilized for curing signals in adaptive applications. It adds a time distance by division a non-stationary signal in many frames that must include semi-stationary portions and utilize a window function into decrease the side lobes in the spectra.

Therefore, in order to analyze the DLN of speech separation, masking, and mapping- based target must be described. The masking describes the time-frequency relations between clean signals to background interference, while the mapping shows the spectral representation of clean signals. The algorithm parameters are;

2.1 Ideal Binary Mask (IBM)

The IBM is a 2-D time frequency (T-F) representation of a noisy signal, such as a cochleagram or a spectrogram as in equation (1) [13]:

\[
IBM = \begin{cases} 
1, & \text{if } \text{SNR}(t,f) > \text{LC} \\
0, & \text{otherwise} 
\end{cases}
\]

Where;
\(t, f\): denote time and frequency.
\(LC\): local criterion (LC) or threshold

2.2 Target Binary Mask (TBM)

TBM derives the label by comparing the target speech energy in each T-F unit with a fixed interference.

2.3 Ideal Ratio Mask (IRM)

Instead of a hard label on each T-F unit, the IRM can be viewed in equation (2) as a soft version of the IBM [14, 15]:

...
\[ IRM = \left( \frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2} \right)^\beta \]

Where:

(\tau, f)^2 \text{ and } N(\tau, f)^2 \text{ : speech and noise energy within a T-F unit.}

\( \beta \) \text{ : Tunable parameter, scales the mask, and is commonly chosen to 0.5.}

With the square root the IRM preserves the speech energy with each T-F unit, under the assumption that \( S(t,f) \) and \( N(t,f) \) are uncorrelated.

2.4 Spectral Magnitude Mask (SMM)

The spectral magnitude mask called fast Fourier transform (FFT)-MASK and defined by STFT (short-time Fourier transform) magnitudes of clean speech and noisy speech [3], see equation (3):

\[ SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} \]

Where:

\(|(t, f)|\) and \(|Y(t,f)|\) : spectral magnitudes of clean speech and noisy speech, respectively.

Unlike the IRM, the SMM is not upper-bounded by 1. To obtain separated speech, the SMM or its estimated is applying to the spectral magnitudes of noise speech. Then, resynthesized of separated speech with the phases of noisy speech (or an estimate of clean speech phase)

Although the usage of STFT, the window is needed to remove or reduce the effect of interference. Therefore, to localize the speech signal at the time, a specific windowing function \( w[n, \tau] \) was used to remove the interference section in the speech stream. The major windows are Hamming and Hanning. The difference between them is in the sidelobe as in the Figure below [16].

![Figure 1. (a) Hanning window. (b) Amplitude spectrum of Hanning window.](image-url)
To optimize the feature selection a Gradient descent was used, where, it is an optimization algorithm for finding a local minimum of a differentiable function which is based on Cauchy algorithm.

The Deep learning neural networks separation algorithm is based on taking the STAFF of the sound signal before and after the mixing process through 128 samples Hanning window. The features (IBM, IRM, and SMM) were extracted from the signals; which are used for separation. Then classify the separated signals according to the minimum distance between their features and unmixed one. Therefore, the current work adopts [17] algorithm, which is modified it by utilizing Hamming window instead of the Hanning window. As the Hamming has low sidelobe than the Hanning, the interference was reduced.

3. Results and Discussion

Generally, data represents the most important parameter in analysis of any system and evaluate its hypothesis. The speech separation task has been treated with the data which represents the sound of human and other sources as noise when it is mixing. Therefore, all of the participants were at the same age (adult), after that, different ages and gender were taken. Then, the human voice sources have been intersecting and mixing with the different voice sources like: water, hammer, and fan. All these sources are tested for different situations.

Two methods were used to apply the voice to the algorithm; the first was directly by the microphone, while the second was by recording the voice and then entered to the algorithm. The data was sampled at 44 kHz to cover all voice bands

The algorithm, which presented in [17] has been modified in this work. The modifications including: filtering window, Validation RMSR, losses, the number of iterations, Sequence overlap, Sequence length, Max Epochs, Mini Batch size, learning rate, as shown in Table 1 The advancement of this modification is resulting as; the processing time and rate of Sound Separation as shown in the Table 1 also. The modified DLN algorithm results have been compared with other references results as shown in Table 2.

At the beginning, the mixing happened when the microphones were used, and when these microphones have a wide pattern and posted at a closed distance.

Figure 3 represents human (male-female and mixing) voice signals, when the microphone is used. Obviously, the signal’s clarity was dropped. All signals are normalized, and their time-frequency are represented. Short time Fourier Transform (STFT) is used for time-frequency representation with Hamming window of length 128, and 96 overlap as the first step of the algorithm.
Figure 3. Voice sources with Time-Frequency representation. (a) Voice sources with mixing signal, and (b) Time-Frequency.

Afterward, to separate the man voice from the mixing (man and woman), a man Moral Time-Frequency Masks is created by multiplying the IBM with mixing time-frequency to isolate the desired (man) resource, as shown in Figure 4.

Figure 4. Moral binary mask for the man talker.

In this context, the Moral binary mask for the woman is created by using the inverse of the man's IBM. Then estimated the male and female sound signals from mixed voice signals by inverse STFT (ISTFT) after multiplying by the IBM for each source, as shown in Figure.
Next, the deep learning network is used to guess the moral soft mask for source estimation; which has properties as; Input size (1×1300), two hidden layers, each with 1300 neurons. Later on, the estimated Source signals as shown in Figure 6, while, the time-frequency using STFT of the sources binary mask is shown in Figure 7.

Figure 5. Estimated voice sources.

Figure 6. Estimated the source signals by estimated the binary mask using DNN.
The improvement of using the modified algorithm can be clearly seen in Figure 7, where, the frequency tracks appears clearly for the voice sources. Other voice sources (water and hammer) were tested for separation as a voice and noise sources as shown in Figure 8.

Figure 7. The STFT of the sources binary mask.

Figure 8. Water and hammer mixed voice separation.
The improvement is clear from the evaluation as shown in Figure 9. It can be seen that, the RMS error is reduced from 7 to 2.8 dB, and the Elapsed time is reduced from 20 min 27 sec to 6 min 24 sec. The operating system characteristics are; Intel ® Core ™ i5-4200 CPU@ 1.6 GHz 2.3 GHz.

The first set of analyses highlighted the impact of using the Hamming window, while, the overall response to this usage was surprising quite overperform. Interestingly for high values of separation rate and reduction in processing time and losses were found due to using the Hamming window. Further analysis replications with other voice sources were showed that the advancement of the modification over the original algorithm. These advancements in the results are caused by the low level of Hamming window side lobes, which was used instead of the Hanning window. Two comparisons are made, the first compares the effect of modified parameters on the results, against the original algorithm as shown in Table 1. It can be seen that processing time, loss, and RMSE for the same separation rate are reduced, these modifications in results have been occurred caused by using the Hamming window instead of Hanning. The second comparison is made with other researches results as shown in Table 2.

**Table 1.** Comparison with original algorithm.

| Sources        | algorithm | RMSE | Loss dB | Iteration | Sequence overlap | Length | Epoch | BS§ | LR¶ | PT£ (min) |
|----------------|-----------|------|---------|-----------|------------------|--------|-------|-----|-----|-----------|
| Human (MW a)   | Original [16] HANW | 6.3104 | 20      | 206       | 20               | 10     | 1     | 64  | 0.001 | 1.59      |
|                | Modified HAMW | 3.0917 | 4.9     | 594       | 5                | 3      | 1     | 74  | 0.001 | 1.05      |
|                | Original [16] HANW d | 5.9877 | 20      | 206       | 20               | 10     | 2     | 64  | 0.0009 | 2.5       |
|                | Modified HAMW | 2.9831 | 4.9     | 594       | 5                | 3      | 2     | 74  | 0.0009 | 2.21      |
|                | Original [16] HANW | 5.8025 | 19      | 206       | 20               | 10     | 3     | 64  | 0.00081 | 2.45      |
|                | Modified HAMW | 2.9025 | 4.9     | 594       | 5                | 3      | 3     | 74  | 0.00081 | 2.52      |
| Instrument (noise –) | Original [16] HANW | 6.5472 | 20      | 620       | 20               | 10     | 1     | 64  | 0.001 | 9.23      |
Evidently, the modified algorithm in which the Hamming window was used delivers better results compared to the original one where Hanning window was used. The first set of analyses highlights the impact of the modification algorithm. While the results offer powerful evidence for source signal separation, the optimal choice was the Hamming window instead of the Hanning filtering window. Where the Hanning window is a complementary window. In fact, the low level of the Hamming window side-lobes' reduces the interference. Therefore, the iterations (processing time) and the RMS errors are reduced, while, the separation ratio is same with original algorithm.

Table 2. Separation rate comparison.

| Ref. | Parameter | Separation Rate % |
|------|-----------|-------------------|
| [1]  | Low SNR   | 73                |
| [8]  | Estimated STOI | 75           |
|      | Ideal STOI | 90                |
| [10] |           | 93                |
| [11] |           | 89.5              |
| [5]  | ICA       | 96.1              |
|      | PCA       | 92.5              |
|      | SVD       | 87.4              |
| Present work | DLN | 98.6              |

The advancement of the DLN over the other algorithms is caused by its nervous systems, which are significantly different from other algorithms in several key respects. The DLN architecture was inspired by the hierarchical structure of the brain. Deep neural networks feature a hierarchical, layer-wise arrangement of nonlinear activation functions (neurons) fed by inputs scaled by linear weights (synapses). Furthermore, the DLN depends on time-frequency properties for source separation. This interpretation is different from that of other algorithms which depend on matrix algebra. Table 2 shows the overperform of modified DLN over other algorithms, where, the DLN with the modified algorithm is more accurate in tests of signal separation by at least 2.4 from the nearest ICA algorithm.
4. Conclusions
The problem of speech separation which may happen in real speech environment "cocktail party" was studied. A development of the common theoretical framework for DLN has been presented to make the required speech separation. The study has been done under different assumptions and using different speech and sound sources: Human (man and woman with different ages) and tools (fan and hammer) and nature sound like water. Many conclusions have been considered including; the masking based target is to outperform, while the masking is advancement at higher SNR, the mapping is overperforming at LOW SNR. The DLN based ratio masking is more advantageous than the other intelligent algorithms. The advancement of the presented modified work's over the original algorithm is clear from Table 1. Where, for human sound separation the RMSE was reduced by 50-60% for 1, 2, and 3 epochs, while the loss is reduced to 25% for the same epochs. In this context, for noise human mixing separation, the RMSE was reduced by 60-65% for 1, 2, 3 and 3 epochs, while the loss is reduced to 20% for the same epochs.

Various separation algorithms have been compared to the modified DLN algorithm. A surprising outcome of the comparative study was that the modified DLN is outperformed other separation algorithm by 2.4% over the nearest algorithm. To explain this advancement finding, the usage of Hamming instead of the Hanning window is the key to this advancement.

References
[1] K. Masutomi, N. Barascud, M. Kashino, J.H. McDermott, and M. Chait, "Sound segregation via embedded repetition is robust to inattention," Journal of Experimental Psychology: Human Perception and Performance, vol. 42, pp. 386-400, 2016.
[2] J. Chen, Y. Wang, and D.L. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 22, pp. 1993-2002, 2014.
[3] DeLiang Wang, Fellow, and Jitong Chen, “Supervised Speech Separation Based on Deep Learning: An Overview”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume: 26, Issue: 10, 2018.
[4] Wenmei Hu, Ruifang Liu, Xuming Lin, Yameng Li, Xin Zhou, and Xiaoxin He, “A Deep Learning Method to Estimate Independent Source Number”, The 2017 4th International Conference on Systems and Informatics (ICSAI 2017)
[5] Hussein A. Al-Barhan, Sinan M. Abdulsatat, Thamir R. Saeed, Ghufran M. Hatem, and Hadi T. Ziboon, “Blind Source Separation Based on random mixing coefficient”, accepted in 3rd International Conference of Engineering Technology and Their Applications, the Islamic University, Al-Najaf-Iraq, 6-7 Sept.2020.
[6] Hussein Ali Aldelfy, Mahmood Hamza Al-Mufraji, and Thamir R. Saeed, “ Improved Key Frame Extraction Using Discrete Wavelet Transform with Modified Threshold Factor”, TELKOMNIKA, Vol.16, No.2, pp. 567–572, April 2018.
[7] Ivan A. Hashim, Jafar W. Abdul Sadah, Thamir R. Saeed, Jawad K. Ali “Recognition of QAM Signals with Low SNR Using a Combined Threshold Algorithm”, IETE Journal of Research, Vol. 61 | No. 1 | Jan 2015.
[8] Shreya Sose, Swapnil Mali, and S.P. Mahajan, “Sound Source Separation Using Neural Network”, 10th ICCCNCT 2019 July 6-8, 2019, IIT – Kanpur Kanpur, India
[9] Emad M. Grais, Gerard Roma, Andrew J.R. Simpson, and Mark D. Plumley, “Single Channel Audio Source Separation using Deep Neural Network Ensembles”, 140th International AES Convention, Paris, June 4-7, 2016.
[10] Zhong-Qiu Wang and DeLiang Wang, Fellow, “Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation”, IEEE/ACM Transactions on audio, speech, and language processing, Volume: 27, Issue: 2, 2019.
[11] Masahito Togami, Yoshiki Masuyama, Tatsuya Komatsu, and Yu Nakagome, “Unsupervised Training for Deep Speech Sources Separation Kullback-Leibler Divergence based Probability Loss function”, arXiv:1911.04228v1 [eess.AS] 11 Nov 2019

[12] Vivek Narayanaswamy, Jayaraman J. Thiagarajan, Rushil Anirudh and Andreas Spanias, “Unsupervised Audio Source Separation using Generative Priors”, arXiv:2005.13769v1 [eess.AS] 28 May 2020.

[13] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Proc., vol. 19, pp. 2125-2136, 2011.

[14] Y. Wang, A. Narayanan, and D.L. Wang, "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Proc., vol. 22, pp. 1849-1858, 2014.

[15] C. Hummersone, T. Stokes, and T. Brooks, "On the ideal ratio mask as the goal of computational auditory scene analysis," in Blind Source Separation, G.R. Naik and W. Wang, Ed., Berlin: Springer, pp. 349-368, 2014.

[16] Katie Wilson, Stephen G. Wilson, and Ezio Biglieri, “Academic Press Library in Mobile and Wireless Communications. Transmission Techniques for Digital Communications”, Academic Press, first edition, 2016.

[17] Andrew J. R. Simpson, “Probabilistic Binary-Mask Cocktail-Party Source Separation in a Convolutional Deep Neural Network”, ArXiv, Vol. abs/1503.06962, 2015.