CASA BASED SUPERVISED SINGLE CHANNEL SPEAKER INDEPENDENT SPEECH SEPARATION

M. Fazal Ur Rehman, Nasir Saleem, Asif Nawaz, Sadeeq Jan, Zeeshan Najam, M. Irfan Khattak, Sheeraz Ahmed

1,2Department of Electrical Engg, Gomal University, D.I.Khan, Pakistan
3Faculty of Engineering (ETS), Dubai Women College, HCT, Dubai, U.A.E
4Department of CSIT, Uni of Engg & Tech, Peshawar, Pakistan
5Department of Electrical Engg, MNS, UET, Multan, Pakistan
6Department of Electrical Engg, Uni of Engg & Tech, Kohat Campus, Pakistan
7Department of Computer Science, Iqra National University, Peshawar, Pakistan

1engrfrehman@gmail.com, 2nasirsaleem@gu.edu.pk, 3anawaz@hct.ac.ae
4sadeeqjan@uetpeshawar.edu.pk, 5shaani75@hotmail.com
6m.i.khattak@uetpeshawar.edu.pk, 7sheerazahmad306@gmail.com,

Corresponding Author: Sheeraz Ahmed
email: sheeraz.ahmad@inu.edu.pk

https://doi.org/10.26782/jmcms.2019.12.00074

Abstract

Computational auditory scene analysis (CASA) based speech separation is widely considered in a number speech processing applications and is used to separate a target speech from target-interference mixtures and usually the task of target separation is considered as a signal processing problem. However, target speech separation is formulated as a supervised learning problem and discriminative patterns of speech, speakers and background noises are learned from input training data. In this paper, we present a single channel supervised speech separation approach based on the ideal binary mask (IBM) estimation. In proposed approach, speaker independent speech separation system is trained with sets of the clean speech magnitudes and during separation; SNR is estimated in time-frequency (TF) channels using clean magnitudes and compared to a pre-defined threshold. The TF channels satisfying threshold are held while TF channels violating the threshold are discarded to construct an IBM. The estimated mask is then applied to the mixtures to reconstruct the target speech, using phase of the mixture speech. The experiments are conducted in three speaker independent mixture’s scenarios: termed as 2-talkers, 3-talkers and 4-talkers mixtures at four input SNRs: -5dB, 0dB, 5dB and 10dB. The experimental outcomes reported that proposed CASA based supervised speaker independent mask estimation outperformed the competing approaches: Nonnegative matrix factorization (NMF), Nonnegative dynamical system (NNDS) and log minimum mean square error (LMMSE) estimation in terms of PESQ, SegSNR, LLR, WSS, SIG, BAK and STOI objective measures.
Keywords: CASA, IBM, intelligibility, time-frequency masking, supervised speech separation, quality

I. Introduction

Since ancient times, speech is one of the most vital communication methods of humanity. In earlier times, the conversation was possible merely face-to-face, but development of the telephone was a revolution into age of the telecommunication. Since that time, individuals are capable to make the long-distance calls and to connect with other individuals from all over the world. With development in the technology and the growing demand for the long-lasting reachability and connectivity, the sharing of the information by means of speech is possible nowadays from any place and any time. The use of cell phones has turned out to be an integral part of everyone’s life. In order to guarantee an excellent quality transmission in the mobile and wireless networks, digital signal processing plays important role. The growing computational performance of technical platforms permits the insight of more and more sophisticated and complex algorithms in cell phones. A generalized single channel speech enhancement mechanism is provided in Fig. 1 for reference.

II. Literature Review

Usually speech communication takes place concurrently with acoustic intrusions, including sounds from the surroundings and competing-talkers. An effective monaural speech separation system attenuates acoustic interferences and greatly facilitates a number of speech applications, including: speech/speaker identification systems, hearing aids and automatic speech recognition (ASR). In various circumstances, such as telecommunications and audio retrieval systems, a monaural (single microphone) way out is essential, where fundamental attributes of speech and background interferences must be considered. Several methods for monaural speech separation [VI] have been...
proposed in the literature. All such methods presume definite properties of background interferences and generally have difficulty in dealing with background competing acoustic interfering sources. Monaural speech separation is also studied based on the statistical learning [X] and phase-based decomposition [I], however, with very limited evaluations. While speech separation is a challenge task, the auditory systems demonstrated an outstanding capacity for this task.

The auditory systems segregate acoustic input signals that correspond to different sources into streams and this segregation/separation is based on the auditory scene analysis (ASA) principle [VI]. Research in ASA has stirred extensive work to the computational auditory scene analysis (CASA) systems for speech separation [II], [III], [XVI]. These systems usually consider speech separation into two key steps: segmentation, known as analysis and the grouping, known as synthesis. During segmentation process, acoustic inputs are decomposed into sensory parts; each part is expected to originate from a one source. During grouping step, these parts which likely to be come from the same source are grouped together and this grouping is based mostly on the periodicity.

Recently, the CASA model proposed in [XVI], the sensory parts are formed, based on the cross-channel correlation (correlation between contiguous filter responses) and temporal permanence, whereas grouping between the sensory parts is executed according to the global pitch extraction inside every time-frame. In the majority of circumstances, the proposed model was capable of removing the interruptions and recovered the low-frequency energy of target speech utterance. Conversely, the proposed model was not able to handle high-frequency speech signals (above 1 kHz) effectively, and much of target speech was lost in the high-frequency range. In fact, the failure to deal with speech utterances in high-frequency range is a common problem of the CASA based speech separation systems. Many IBM based methods have been proposed in literature for speech separation/enhancement [XI-XIV].

In [VII] a two-talker speech separation system based on CASA for speaker recognition was proposed to separate a target speech from other interfering speech. A tandem method was used to classify voiced segments, and gammatone frequency cepstral coefficients (GFCCs) clustering based object function is launched to recognize a particular speaker. For sequential organization of voiced segments, the finest group is realized via exhaustive or beam search. The unvoiced segments are produced by estimating onset/offset and then unvoiced–voiced and unvoiced–unvoiced segments are divided in that order. The unvoiced–voiced segments are controlled using the binary mask of separated voiced segments, whereas the unvoiced–unvoiced segments are separated consistently. Robustness is one of the most major aspects of practical applications of speech separation and CASA based single channel speech separation suggests a way out to this problem.

A speech separation system is proposed to separate speech utterance of two talkers. Gaussian mixture models (GMMs) and vector quantization (VQ) is used to find out the grouping cues on the individual clean data for all speakers. For a given speech utterance, speakers are identified when two
speakers are presented in the utterance, followed by the factorial-max vector quantization (MAXVQ) model is considered to estimate the masks and lastly the utterances of target speakers are resynthesized in framework of CASA. In order to solve problem of speech separation, a method is proposed [VIII] that combined CASA with objective quality assessment of speech (OQAS). During grouping procedure of CASA, the OQAS is used as a steer to initiate the CASA structure. This combination of VASA and OQAS improved the performance both in terms of SNR and MOS. In order to develop robustness in ASR, a novel method is proposed based on auditory features and speaker recognition using a front-end which is based on CASA. The auditory features are made robust by investigating diverse feature dimensions and incorporated dynamic features. The features are evaluated and robustness in speaker identification is evaluated in noisy conditions.

In this paper, we have proposed a supervised single channel speaker independent speech separation method which can deal with speech utterances originated from different number of speakers/sources in high-frequency range. In this paper 2-talker, 3-talker and 4-talker scenarios are considered to test our method. The remaining paper is organized as: The single channel speech separation based on CASA method is presented in section 3, experiments and setup is presented in section 4, whereas the results and discussion are presented in section 5. The final conclusions are extracted and presented in section 6.

III. CASA Based Single Channel Speech Separation

![Fig 2: Block diagram of proposed method](image)

To explain the process of supervised single channel speaker independent speech separation, the general problem of speech separation is illustrated. The problem of the single channel speaker independent speech separation is defined by estimating \( N \) speakers \( SP_1(t), SP_2(t), \ldots, SP_N(t) \) for a given mixture \( M(t) \), expressed by mathematical formula as;

\[
\text{SNR} = \frac{\text{Signal Power}}{\text{Noise Power}}
\]

\[
\text{SNR} = \frac{\sum_{i=1}^{N} \| SP_i(t) \|^2}{\sum_{i=1}^{N} \| \text{Noise}(t) \|^2}
\]

\[
\text{SNR} = \frac{\| SP(t) \|^2}{\| M(t) - SP(t) \|^2}
\]

\[
\text{SNR} = \frac{\| SP(t) \|^2}{\| M(t) \|^2}
\]

\[
\text{SNR} = \frac{\| SP(t) \|^2}{\| M(t) - SP(t) \|^2}
\]

\[
\text{SNR} = \frac{\| SP(t) \|^2}{\| M(t) \|^2}
\]
In time-frequency (TF) version, the magnitude spectrums of the mixture signals \( M(\omega, k) \) is expressed as the sum of magnitude spectrums of all \( N \) speakers. The frequency domain depiction is obtained by using a set of gammatone filterbank instead of the STFT and is given by equation as:

\[
M(\omega, k) = \sum_{j=1}^{N} S_{j}(\omega, k)
\]  

In the proposed approach, real-valued magnitude spectrums of clean speech utterances are used to train the system in supervised manner, and time-frequency (TF) mask for the target source is estimated. The magnitude spectrum \( |M(\omega, t)| \) stands for the feature vectors, where \( \omega \) and \( t \) denote frequency and time indexes. The estimated magnitude spectrums are given by equation as:

\[
M(\omega, k) = M \odot S_{j}, \text{ for } \sum_{j=1}^{N} S_{j}
\]  

Where, \( \odot \) denotes element-wise multiplication. According to the proposed approach, the target sources magnitude spectrum is achieved by subtracting the magnitude spectrum of the clean speech utterance from the mixture magnitude spectrum as:

\[
T(\omega, k) = C(\omega, k) - M(\omega, k)
\]

The overall SNR in all time-frequency channel is estimated from the ratio of the clean speech magnitude spectrum to the subtracted target speech magnitude spectrum \( T(\omega, t) \), and is given by equation as:

\[
\text{SNR}(\omega, k) = \frac{C(\omega, k)}{T(\omega, k)}
\]

In order to estimate ideal binary time-frequency mask, the estimated SNR is compared to a predefined threshold, known as local criterion (LC) usually selected as 0dB. All the time-frequency channels satisfying the threshold are retained and all other channels are discarded. The construction of the ideal binary mask is given by equation as:

\[
\text{IBM}(\omega, k) = \begin{cases} 
1, & \text{if } \text{SNR}(\omega, k) \geq \text{LC} \\
0, & \text{if } \text{SNR}(\omega, k) < \text{LC}
\end{cases}
\]

Where \( t \) denotes time, \( \omega \) denotes frequency and \( \text{LC} \) is the local criterion. The estimated target speech magnitude spectrum is achieved by multiplying the estimated binary mask with the mixture magnitude spectrum. The final time-domain signals are reconstructed by using set of gammatone filterbank of estimated magnitude spectrogram using phase of the mixture signals. The block diagram for the proposed approach is demonstrated in Fig. 2, showing.
different steps mentioned above. The proposed method is abbreviated as CASA-IBM.

IV. Experiments and Setup

We have considered 30 clean speech utterances from Noizeus database [IV] during training the system. We have used three mixture situations to test our presented system, including two-talkers, three-talkers and four-talkers. All the mixture sources are purely nonstationary in nature. The duration of each mixture is approximately 50 seconds. The mixtures are mixed with training utterances at -5dB, 0dB, 5dB and 10dB SNRs. To conduct objective evaluations of the source separation method, we have considered Perceptual evaluation of speech quality (PESQ) [IX] to assess the quality of the separated speech, Short-Time Objective Intelligibility (STOI) [XVI] to assess the recognition capability of the separated speech and segmental SNR (SegSNR) [V] to assess the suppression of sources from the separated speech. STOI means a correlation between the clean and separated speech utterance and has been shown a high correlation to human capacity of understanding a spoken item. SegSNR provides correlation that how much competing source is removed by a separating method where a high SegSNR value shows that separated speech has low residuals. The higher values of the PESQ demonstrate that the separated target source has pleasant effect on human ears. LLR and WSS are used to measure the distance between clean utterance and the utterance processed by the speech separation algorithms. The smaller the distance, better is the performance of algorithm. Moreover, composite measures are used to measure the speech distortion (SIG) and the residual noise distortion (BAK). The higher values of composite measures indicate a better performance. Three competing methods including Nonnegative matrix factorization (NMF), Non-negative dynamical system (NNDS) and log minimum mean square error (LMMSE) are used for comparison purpose.

V. Results and Discussions
Assessments of the methods including NMF, NNDS, LMMSE and CASA-IBM for speech separation are made by calculating the objective speech quality, speech intelligibility and interfering source reduction measures, PESQ, STOI, SegSNR, LLR, WSS, BAK and SIG, respectively which are explained in section III. Table-1-3 shows the PESQ and SegSNR scores for the three situations: 2-talkers, 3-talker and 4-talkers at -5dB, 0dB, 5dB and 10dB SNRs. The results shown high performance of the CASA-IBM as compared to other state-of-the-art classical methods: NMF, NNDS, and LMMSE. The CASA-IBM method achieved remarkable PESQ and SegSNR output scores in 2-talkers, 3-talker and 4-talkers situations at -5dB. Among other methods, NMF achieved better results in terms of PESQ and SegSNR.

![Fig 3: PESQ Improvements](image-url)

**Table 1: PESQ and SNRSeg scores for 2-talkers condition**

| Processing Methods | PESQ | SegSNR |
|--------------------|------|--------|
|                    | -5dB | 0dB | 5dB | 10dB | -5dB | 0dB | 5dB | 10dB |
| Noisy              | 1.39 | 1.91 | 2.21 | 2.39 | 1.76 | 2.52 | 3.02 | 4.45 |
| LMMSE              | 2.01 | 2.24 | 2.51 | 2.68 | 2.79 | 4.01 | 5.17 | 7.21 |
| NNDS               | 1.95 | 2.19 | 2.62 | 2.75 | 3.42 | 4.12 | 5.27 | 7.34 |
| NMF                | 2.13 | 2.39 | 2.72 | 2.88 | 4.12 | 4.63 | 5.41 | 7.56 |
| CASA-IBM           | 2.39 | 2.72 | 2.93 | 3.12 | 4.98 | 5.02 | 6.15 | 8.32 |

**Table 2: PESQ and SNRSeg scores for 3-talkers condition**

| Processing Methods | PESQ | SegSNR |
|--------------------|------|--------|
|                    | -5dB | 0dB | 5dB | 10dB | -5dB | 0dB | 5dB | 10dB |
| Noisy              | 1.25 | 1.65 | 2.12 | 2.32 | 1.71 | 2.21 | 2.73 | 4.23 |
| LMMSE              | 1.98 | 2.28 | 2.56 | 2.76 | 2.67 | 3.91 | 5.02 | 6.97 |
| NNDS               | 1.82 | 2.07 | 2.52 | 2.65 | 3.32 | 4.01 | 5.13 | 6.86 |
| NMF                | 2.01 | 2.14 | 2.45 | 2.55 | 3.81 | 4.43 | 5.01 | 7.01 |
| CASA-IBM           | 2.29 | 2.55 | 2.87 | 2.93 | 4.74 | 4.93 | 5.97 | 8.76 |

**Table 3: PESQ and SNRSeg scores for 4-talkers condition**

| Processing Methods | PESQ | SegSNR |
|--------------------|------|--------|
|                    | -5dB | 0dB | 5dB | 10dB | -5dB | 0dB | 5dB | 10dB |
| Noisy              | 1.19 | 1.48 | 1.92 | 2.22 | 1.67 | 1.98 | 2.51 | 4.13 |
| LMMSE              | 1.77 | 1.91 | 2.43 | 2.66 | 2.57 | 3.71 | 4.93 | 6.73 |
| NNDS               | 1.73 | 2.01 | 2.42 | 2.52 | 3.12 | 3.91 | 4.77 | 6.76 |
| NMF                | 2.19 | 2.22 | 2.39 | 2.45 | 3.65 | 4.11 | 4.73 | 6.87 |
| CASA-IBM           | 2.18 | 2.43 | 2.74 | 2.88 | 4.54 | 4.76 | 5.72 | 8.43 |

Copyright reserved © J. Mech. Cont. & Math. Sci.
M. Fazal Ur Rehman et al 979
The achieved scores have shown high performance of the CASA-IBM as compared to other methods. The CASA-IBM method achieved notable PESQ and SegSNR output scores in three-talker situation at -5dB. The high

Copyright reserved © J. Mech. Cont.& Math. Sci.
M. Fazal Ur Rehman et al
PESQ and SegSNR scores indicate that the presented method has a high capacity to reduce the interfering sources from the input mixtures and provide a high quality of speech. Among other methods, again NMF achieved better results in terms of the PESQ and SegSNR. The PESQ improvements are shown in Fig. 3 Table-4-6 shows the Normalized WSS and LLR scores for three situations: 2-talkers, 3-talker and 4-talkers at -5dB, 0dB, 5dB and 10dB SNRs. The obtained results showed a high performance of the CASA-IBM as compared to other classical methods: NMF, NNDS, and LMMSE. The CASA-IBM method achieved notable Normalized WSS and LLR scores in 2-talkers, 3-talker and 4-talkers situations at -5dB. The achieved scores have shown high performance of the CASA-IBM as compared to other methods.

CASA-IBM method achieved notable Normalized WSS and LLR output scores in 4-talker situation at -5dB. The low Normalized WSS and LLR scores indicate that the presented method has shown a high capacity to reproduce a close replica of original speech utterance. Table-7-9 shows the SIG and BAK scores for three situations: 2-talkers, 3-talker and 4-talkers at -5dB, 0dB, 5dB and 10dB SNRs. The obtained SIG and BAK results showed a high performance of the CASA-IBM as compared to other classical methods: NMF, NNDS, and LMMSE. The CASA-IBM achieved notable SIG and BAK scores in 2-talkers, 3-talker and 4-talkers situations at -5dB. The achieved scores have shown that CASA-IBM as compared to other methods comparatively reduced background noise and offered less speech distortion. The CASA-IBM method achieved notable Normalized WSS and LLR output scores in 4-talker situation at -5dB. The high SIG and BAK scores indicate that the presented method has shown a high capacity to reduce the background noise with an acceptable speech distortion. The CASA-IBM offers high ability to understand the spoken items (intelligibility). ΔPESQ and ΔSegSNR endorsed by Table I-III. Improvements (Δ) are computed for all input SNR levels. It is obvious from the Fig. 4 that the STOI based speech intelligibility achieved by the CASA-IBM is consistently higher than three competing methods and unprocessed mixtures. The high STOI scores suggest that the separated speech utterances provide by the CASA-IBM offers high ability to understand the spoken items (intelligibility). ΔPESQ and ΔSegSNR endorsed by Table I-III. Improvements (Δ) are computed for all

Fig 4: Intelligibility Analysis
three situations. It is obvious from Table IV-IX that CASA-IBM has achieved the highest Normalized WSS, LLR, SIG and BAK improvements as compared to NMF, NNDS and LMMSSE competing methods.

In a speech separation system, speech distortion plays a vital role and defines the intelligibility of separated speech. A separated speech with great distortion may lose much main speech contents; hence the understanding potential of the separated speech (intelligibility) is lowered. Consequently, it is imperative to carry out speech separation in such a way that competing-talker/talkers is/are eliminated but not at the cost of speech intelligibility. Similarly, the separation task needs effective removal of the competing sources from mixtures so that listener may have clear understanding of speech signals. To assess these capabilities of our presented system, we have evaluated spectra of the separated speech by conducting time-varying spectral analysis. The spectra of the separated speech are portrayed in Fig. 5. The spectra of the competing methods have vanished valuable speech contents, consequently granted a reduced amount of speech intelligibility which is endorsed by Fig. 4. In contrast, if we examine the spectrogram of the speech separated by CASA-IBM, much closed copy of the clean speech spectrogram is achieved. The essential speech contents are well potted by the CASA-IBM.
Moreover, the residuals are evident in the spectrograms of NMF, NNDS and LMMSE which is greatly reduced by the CASA-IBM.

VI. Conclusions

In this paper we studied the speech intelligibility potential of single-microphone speech enhancement based on DNNs which belongs to machine learning family. We have shown that when DNNs are trained purposely to handle different noise types and SNRs, these networks have great potential of achieving considerable improvements in speech intelligibility. We have trained the DNNs to learn mapping from the noisy spectrums and estimated the time-frequency masks. The masks are than applied to noisy speech spectrums to obtain a speech with enhanced intelligibility using phase of the noisy speech. Our experimental results across six noisy situations: airport, babble, car, coffee shop, and exhibition hall and five SNRs: -10dB, -5dB, 0dB, 5dB and 10dB report that DNN based estimated ratio mask outperforms the competing approaches: NMF and LMMSE in terms of the STOI and NSEC objective speech intelligibility measures.

References

I. Cauwenberghs, G. (1999). Monaural separation of independent acoustical components. In Circuits and Systems, 1999. ISCAS’99. Proceedings of the 1999 IEEE International Symposium on (Vol. 5, pp. 62-65). IEEE.

II. Darwin, C. J. (1997). Auditory grouping. Trends in cognitive sciences, 1(9), 327-333.

III. Ellis, D.P.W. (1996). Prediction-driven computational auditory scene analysis (Doctoral dissertation, Massachusetts Institute of Technology).

IV. Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. Speech communication, 49(7-8), 588-601.

V. Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). Independent component analysis. In Natural Image Statistics (pp. 151-175). Springer, London.

VI. Li, H., Wang, Y., Zhao, R., & Zhang, X. (2018). An Unsupervised Two-Talker Speech Separation System Based on CASA. International Journal of Pattern Recognition and Artificial Intelligence, 32(07), 1858002.

VII. Li, P., Guan, Y., Xu, B., & Liu, W. (2006). Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. IEEE Transactions on Audio, Speech, and Language Processing, 14(6), 2014-2023.

VIII. Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on (Vol. 2, pp. 749-752). IEEE.

Copyright reserved © J. Mech. Cont.& Math. Sci.
M. Fazal Ur Rehman et al

983
IX. O'shaughnessy, D. (1987). Speech communication: human and machine. Universities press.

X. Roweis, S. T. (2001). One microphone source separation. In Advances in neural information processing systems (pp. 793-799).

XI. Saleem, N., Shafi, M., Mustafa, E., & Nawaz, A. (2015). A novel binary mask estimation based on spectral subtraction gain-induced distortions for improved speech intelligibility and quality. University of Engineering and Technology Taxila. Technical Journal, 20(4), 36.

XII. Saleem, N., & Irfan, M. (2018). Noise reduction based on soft masks by incorporating SNR uncertainty in frequency domain. Circuits, Systems, and Signal Processing, 37(6), 2591-2612.

XIII. Saleem, N. (2017). Single channel noise reduction system in low SNR. International Journal of Speech Technology, 20(1), 89-98.

XIV. Saleem, N., Mustafa, E., Nawaz, A., & Khan, A. (2015). Ideal binary masking for reducing convolutive noise. International Journal of Speech Technology, 18(4), 547-554.

XV. Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 4214-4217). IEEE

XVI. Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. IEEE transactions on neural networks, 10(3), 684-697.

XVII. Weintraub, M. (1985). A theory and computational model of auditory monaural sound separation (Doctoral dissertation, Stanford University).