Supervised Monaural Speech Separation using Computational Auditory Scene Analysis

Fazal Rehman
Deptment of Electrical Engineering
Gomal University
D.I.Khan, Pakistan
nasirsaleem@gu.edu.pk

Nasir Saleem
Deptment of Electrical Engineering
Gomal University
D.I.Khan, Pakistan
nasirsaleem@gu.edu.pk

Muhammad Irfan Khattak
Deptment of Electrical Engineering, University of Engineering & Technology
Kohat Campus, Pakistan
m.i.khattak@uetpeshawar.edu.pk

Sheeraz Ahmad
Department of Computer Science,
Iqra University
Peshawar, Pakistan
sheerazahmad306@gmail.com

***Abstract*—Monaural speech separation based on computational auditory scene analysis (CASA) is studied in many speech processing applications to separate a speech of the interest from the interfering backgrounds. Customarily, this task of separating a target speech from the background is considered as a signal processing problem. Recently speech separation approach is formulated as a supervised learning problem and discriminative patterns of speech, speakers, and background noises are learned from some input training data. This paper presents a monaural supervised speech separation method that is based on the principle of ideal binary mask (IBM) computation. In presented method, the system is trained with a set clean speech utterances and during the separation the SNR is estimated in time-frequency units and compared to a threshold. The time-frequency units satisfying the threshold are retained while time-frequency units violating the threshold are discarded to construct a binary mask. The estimated mask is than applied to mixture speech spectra to acquire the target speech by using the phase of mixture speech. Our experimental results across three mixtures situations: two-talkers, three-talkers, and four-talkers at SNRs: -5dB, 0dB, 5dB and 10dB reported that presented CASA based supervised mask outperformed the competing approaches: Nonnegative matrix factorization (NMF), Non-negative dynamical system (NNDS) and log minimum mean square error (LMMSE) estimation in terms of short time objective intelligibility (STOI) and Segmental SNR and Perceptual evaluation of speech quality (PESQ) metrics.**

Keywords— Deep neural networks, Intelligibility, Machine learning, speech enhancement, time-frequency masking,

I. Introduction

Usually speech communication takes place concurrently with acoustic interferences, 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 [1] 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 [2] and phase-based decomposition [3], 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 [4]. Research in ASA has stirred extensive work to the computational auditory scene analysis (CASA) systems for speech separation [5-8]. 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 [6], 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 [9-12].

In this paper, we have presented a supervised monaural speech separation method which can deal with speech utterances originated from different number of speakers/sources (two-four-talkers) in high-frequency range. The remaining paper is organized as: The general idea of the monaural speech separation method is presented in section II and experiments and setup is presented in section III, whereas the results and discussion are presented in the section IV. The conclusion is extracted and presented in section V.

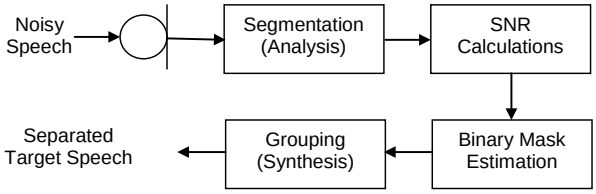

Figure 1: Block diagram

II. CASA Based Monaural Speech Seperation

To illuminate the procedure of supervised monaural speech separation problem, we describe the common problem of the monaural speech separation. This problem of speech separation is defined by estimating *M* speakers $Sp_1(t)$, $Sp_2(t)$, …., $Sp_N(t)$ for a given mixture *Mix(t)*, given by mathematical formula as:

$$Mix(t) = \sum_{k=1}^{M} Sp_k(t) \tag{1}$$

In time-frequency representation, the magnitude spectra of mixture signals $M_{IX}(\omega,t)$ is the sum of magnitude spectra of all sources by taking STFT and is given by equation as:

$$M_{IX}(\omega,t) = \sum_{k=1}^{M} Sp_k(\omega,t) \tag{2}$$

The real-valued magnitude spectra are used as supervised input to the speech separation system and time-frequency mask for the target source is computed. The magnitude spectra $|M_{IX}(\omega,t)|$ stand for the feature vectors, where ω and τ denotes frequency and time indexes, respectively. The estimated magnitude spectra are given by equation as:

$$M_{IX}(\omega,t) = M_{IX} \square\ Sp_k, \quad for \quad \sum_{k=1}^{M} Sp_K \tag{3}$$

Where, Θ indicates an element-wise multiplication. According to presented method, the target sources magnitude spectrum is obtained by subtracting the clean speech utterance from the mixture magnitude spectrum. The SNR in each time-frequency unit is estimated from the ratio of the clean speech magnitude spectrum and subtracted target speech magnitude spectrum, given by equation as:

$$SNR(\omega,t) = T_{ARG}(\omega,t) / C_{LEA}(\omega,t) \tag{4}$$

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

$$IBM(\omega,t) = \begin{cases} 1, & if \quad SNR(\omega,t) \geq LC \\ 0, & if \quad SNR(\omega,t) \ < \ LC \end{cases} \tag{5}$$

Where *t* denotes time, ω denotes frequency and *LC* is the local criterion. The estimated target speech magnitude spectra are achieved by multiplying the estimated binary mask with mixture magnitude spectra. The final time-domain signals are reconstructed by computing the inverse STFT of the estimated magnitude spectrograms using phase of the mixture magnitude spectra. The block diagram for the presented method is given in Fig.1. The presented method is abbreviated by CASA-IBM.

III. Experiments and Setup

We have considered 30 clean speech utterances from Noizeus database [13] 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) [14] to assess the quality of separated speech, Short-Time Objective Intelligibility (STOI) [15] to assess the recognition capability of the separated speech and segmental SNR (SegSNR) [13] 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. Three competing methods including Nonnegative matrix factorization (NMF), Non-negative dynamical system (NNDS) and log minimum mean square error (LMMSE) are used.

IV. Results and Discussions

Assessments of the methods including NMF, NNDS, LMMSE and CASA-IBM for speech separation are made by calculating objective speech quality, speech intelligibility and interfering source reduction measures, PESQ, STOI and SegSNR, explained in section III. Table-I shows the PESQ and SegSNR scores for one of the three situations: two-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 two-talkers situation at -5dB. Among other methods, NMF achieved better results in terms of PESQ and SegSNR. Table-II shows the PESQ and SegSNR scores for three-talker situation at -5dB, 0dB, 5dB and 10dB SNRs. 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

Table I: 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 II: 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 III: 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

situation at -5dB. The high 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. Fig. 2 demonstrates the STOI scores achieved by the NMF, NNDS, LMMSE and CASA-IBM methods in three situations at four input SNR levels. It is obvious from the Fig. 2 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, ΔSegSNR, and ΔSTQI, endorsed by Table I-III. The improvements (Δ) are computed for all three situations. It is obvious from Table III-V that CASA-IBM has achieved the highest PESQ, SegSNR and STOI improvements as compared to NMF, NNDS and LMMSE 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

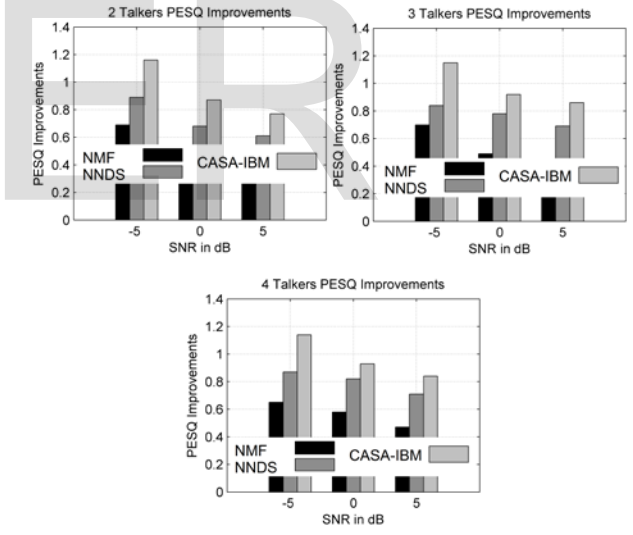

Figure 2: speech intelligibility analysis

are portrayed in Fig. 3. The spectra of the competing methods have vanished valuable speech contents, consequently granted a reduced amount of speech intelligibility which is endorsed by Fig. 2. 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.

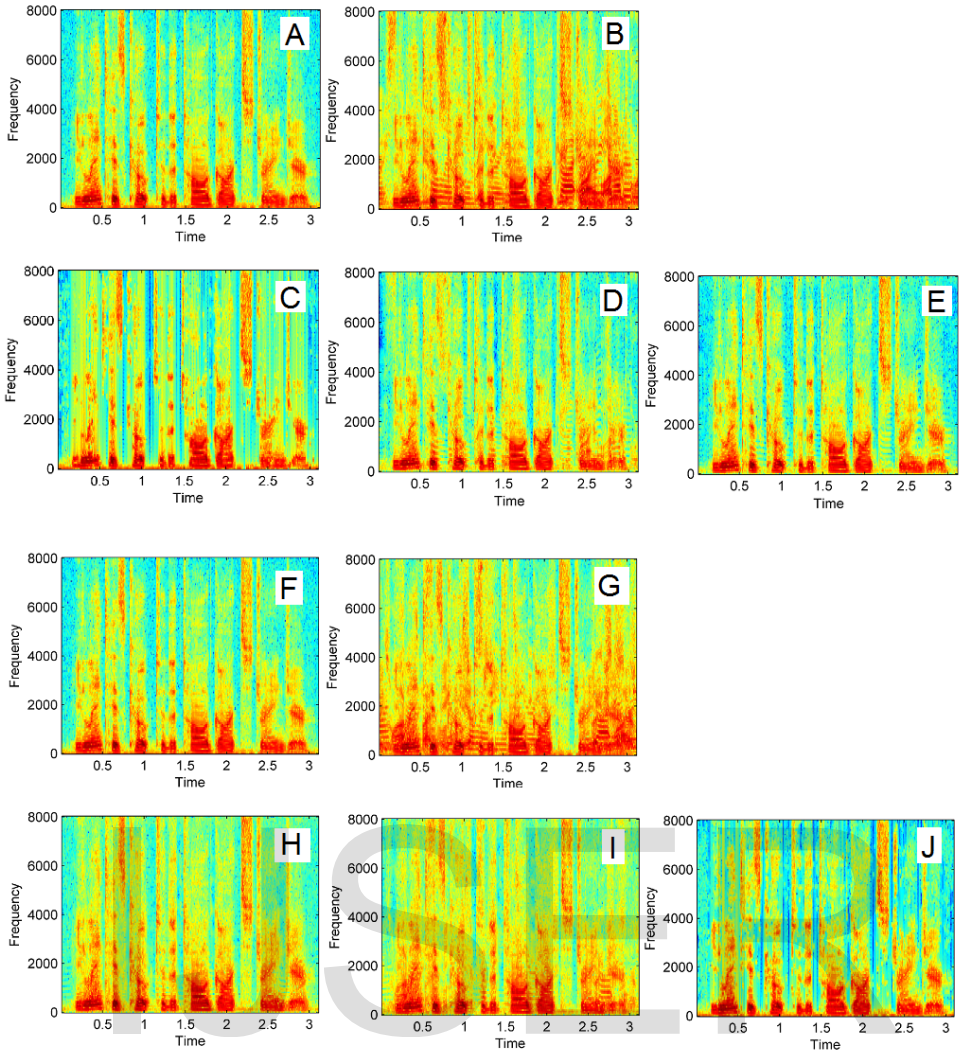

Figure 3: Spectral Analysis. (A and F) Cleans speech, (B) two-talker mixture, (C) NNDS, (D) NMF, (CASA-IBM), (G) Four-talker mixture, (H) NNDS, (I) NMF, and (J) CASA-IBM

V. 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

[1]. O'shaughnessy, D. (1987). Speech communication: human and machine. Universities press.

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

[3]. 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.

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

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

[6]. 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.

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

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

[9]. 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.

[10]. 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.

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

[12]. 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.

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

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

[15]. 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