Soft-CASA system for Single Channel Speech Separation

Belhedi Wiem[#1], Ben Messaoud Mohamed anouar[#2], Bouzid Aicha[#3]

[#] *Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis*
Laboratoire de Signal, Images et Technologies de l'Information, LR11ES17, 1002 Tunis, Tunisie

[1]bel.hedi@gmail.com

[2]anouar.benmessaoud@yahoo.fr

[3] bouzidacha@yahoo.fr

***Abstract*— In this paper we study the masking effect on Computational Auditory Scene Analysis (CASA) based systems for single channel speech separation (SCSS). In this study, we focus on the benchmark masks of the literature that are namely: the ideal binary mask (IBM), the binary mask (BM) and soft mask. Each system is evaluated objectively and subjectively in order to highlight the effect of each mask on the intelligibility and the quality of the separated speech. Based on this study we develop a new system, that we call Soft-CASA for SCSS that outperforms the original one. The proposed system achieves 28.84% improvement in the Short-time Objective Intelligibility (STOI) parameter, 7.1% improvement in SNR_{loss} and 92% improvement in overall perceptual score (OPS), compared to the original system.**

***Keywords*— SCSS, CASA, Soft mask, binary mask**

I. INTRODUCTION

The unconventional methods of separation of the signal and which are a part of the register of the Auditory Scene Analysis (ASA) [1] and which exploit properties of the speech signal, make the object of increasingly active researches.

Extensive researches in ASA have led to Computational ASA (CASA) which is the study of how a computational system can organize sound into perceptually meaningful elements.

In CASA-based techniques, the goal is to imitate the process of the human auditory system using signal processing approaches [2]. It seeks discriminative features in the composite signal in order to separate the speech signals .CASA process runs in five essential steps. The first step consists to feature analysis. To do so, front-end transform such as Gammatone filter bank or the short time Fourier transform (STFT), are applied. The second step is the segmentation of the mixture into time-frequency cells. The segmentation is based on some criteria which are: fundamental frequency, onset, offset, amplitude modulation, continuity and position. The third step is pitch tracking and time frequency (T-F) labelling. After that, final segregation is done using T-F masking. And the last step is the grouping in which the cells that are supposed to belong to one source are grouped together. The CASA process is as illustrated in Figure 1.

CASA-based approaches have proven their effectiveness in speech separation. However, they have some limitations. For example, as the grouping stage is based on periodicity; hence, it could only be applied only to the voiced frames of speech. The unvoiced frames are confused with interference because they lack harmonic structure and are characterized with weaker energy. However, in [3] an hybrid system composed CASA and spectral subtraction approach was proposed for separating speech mixture composed of both voiced and unvoiced sounds. In addition, in CASA systems, the pitch information is required to be estimated directly from the mixed signal. They often use the estimated pitch trajectories by applying a multi-pitch estimator [6] [16] [17]. Hence, the quality of the separated speech signal is limited by the accuracy of the multi-pitch estimator. One of the major limitations is that the output signals produced by the CASA systems often lack perceptual quality due to the severe cross-talk problem. In general, applying binary masks inevitably cause cross-talk and artifacts in the separated signals. In order to overcome this limitation of CASA, we extend the original CASA system.

The CASA system employed in this study is Hu and Wang model described in [3].

The new system, that we call Soft-CASA, uses a soft mask instead the hard mask that used to be employed. This greatly improves the separated signal's qualities. In fact, we prove that replacing the hard mask by a soft mask, in the CASA system provides better separation quality. The proposed system achieves a 28.84% improvement in the Short-time Objective Intelligibility (STOI) parameter, 7.1% improvement in SNR_{loss} and 92% improvement in the overall perceptual score (OPS), compared to the original system. The rest of the paper is organized as follows; Section II gives an overview of the hard and soft masks. Evaluation and comparison are made in section III, and the discussion is made in section IV. Finally Section V concludes the work.

II. MASKS STUDIED

In sound sources, energy is often concentrated in relatively small areas in the T-F cells. This sparsity allows, to some degree, the selection of T-F areas dominated

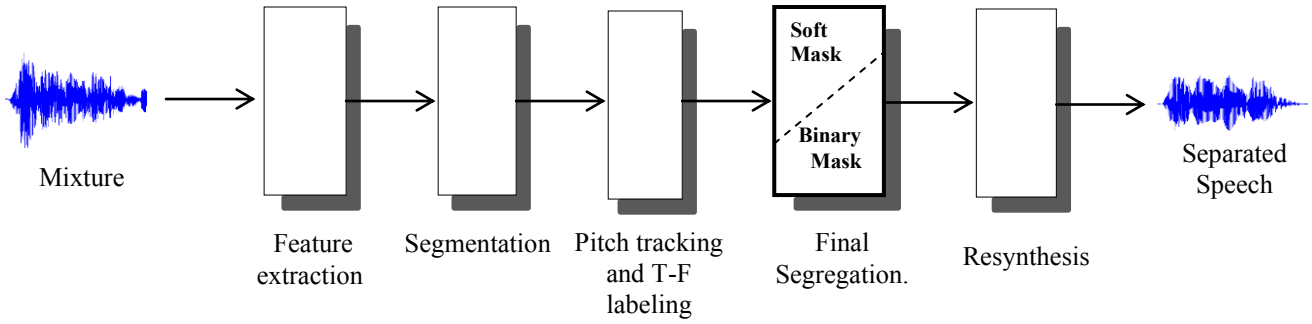

Fig.1. CASA process. In the original CASA system the Final segregation is done using a binary mask. However in the proposed approach the final segregation is done using a soft mask.

by the desired source and setting other cells to zero. This selectivity could be done in different ways. According to the value assigned, different mask types are obtained. If the value assigned is either 0 or 1 the mask is binary. However, when the mask value is in the range between 0 and 1 then the mask is called Soft.

In this paper we study the effect of using soft instead of binary mask in CASA system.

A. Binary mask and Ideal Binary mask

A binary mask is a matrix of binary numbers, defined in the T-F domain. The numbers are assigned by comparing the energy of the target speaker against the concurrent speaker. If the target energy, in a T-F cell, is greater than the one of the concurrent then the value assigned is 1. Otherwise, the value assigned to 0. This technique has been of great effectiveness in Computational Auditory Scene Analysis (CASA) when used as output representation to label the origins of the mixed speech [4].

When both target and concurrent signals are known, the ideal BM (IBM) could be determined. IBM is constructed by comparing their powers in a T-F cells against a local criterion (LC). We assume that a mixture x(t,f) can expressed as the summation of the target $s_1(t,f)$ and the masker $s_2(t,f)$, as expressed in equation below:

$$x(t,f) = s_1(t,f) + s_2(t,f) \quad (1)$$

After decomposing the signals using either a STFT, IBM is determined using Equation 2:

$$IBM(t,f) = \begin{cases} 1 & \text{if } s_1(t,f) - s_2(t,f) > LC \\ 0 & otherwise \end{cases} \quad (2)$$

such as t is the time index, LC is the local criterion and f is the frequency index.

B. Soft mask

In soft mask the probability that the segment is dominant or dominated are calculated and attributed probability to each frame of the input mixed signal [5]. Thus, unlike the BM that attributes a binary value (either 0 or 1), soft mask attributes a value in the range between 0 and 1.

The estimation of soft mask can be made using several methods; among them the statistical computation [4] or fuzzy logic [5]. In statistical computation, the probability calculation is adopted to estimate weighting factors needed for the decision of maintaining or rejecting a part of a signal.

We assume that mixing signal x of two speakers s_1 and s_2 can be expressed as in the Equation 1.

Let S_1 and S_2 be the log power spectral vectors of speakers s_1 and s_2, respectively. The distributions of S_1 and S_2 for each speaker are described as follows:

$$P(s_j) = \sum_{Nj=1}^{Nj} P_j(Nj) \prod_{d=1}^{D} N(j_d; \mu_{Nj,d}^{j}; \sigma_{Nj,d}^{j}) \quad (3)$$

such as j is equal to 1 or 2, Nj is the number of Gaussians in the mixture. In addition, $P_j(N_j)$ is a prior probability of the N_j^{th} Gaussian that was determined from training phase. D is the power spectral vector dimensionality. Also, j_d represents the d^{th} dimension of j. $\sigma_{Nj,d}^{j}$ and $\mu_{Nj,d}^{j}$ are respectively the mean variance of the d^{th} dimension of the N_j^{th} Gaussian in the mixture.

The mixture signal x satisfies the log-max approximation. This could be expressed using Equation 4:

$$x(\omega) = \max(s_1(\omega), s_2(\omega)) \quad (4)$$

Thus the probability that the segment of x belongs to s_1 is the probability that s_1 is greater than s_2 in the same segment which can be expressed by the next equation:

$$P(s_{1d} = x_d \mid x) = P(s_{1d} \rangle s_{2d} \mid x) \quad (5)$$

Then the soft mask is expressed as follows:

$$Soft = \begin{cases} \dfrac{P_{x_1}(s_1 \mid N_{s_1}) P_{s_2}(x \mid N_{s_2})}{P(x \mid N_{s_2}, N_{s_1})} + \dfrac{P_{s_1}(x \mid N_{s_1}) C_{s_2}(x \mid K_{s_2}) \delta(s_1 - x)}{P(x \mid N_{s_1}, N_{s_2})} & \text{if } s_1 \le x \\ 0 \quad otherwise \end{cases} \quad (6)$$

such as $\delta(s_1-x)$ is a Dirac delta function of s_1 centred at x and C_{s2} is described in the equation below:

$$C_{s_2}(\omega \mid N_{s_2}) = \int_{-\infty}^{\omega} \mathrm{N}(\mathrm{x}; \mu^{x}_{N_{s_2}}; \sigma^{x}_{N_{s_2}})dx \quad (7)$$

III. Evaluation and Comparison

The proposed SCSS system is evaluated on the speech separation GRID database provided in [7]. In order to study the performance of the proposed SCSS system, we consider four different s clips of same gender (SG), different gender (DG) and same talker (ST). The performance of the proposed system (Sys 2) is summarized in Table 1. In addition, we conduct an evaluation with the original system (Sys 1) and the Ideal one (Sys 3). The evaluation is conduct in terms of Short-time Objective Intelligibility (STOI) [8], Perceptual Evaluation of Speech Quality (PESQ) [9], Signal-to-Noise loss ratio (SNRloss) [10], Signal-to-interference ratio (SIR) [11], Signal-to-Artifact Ratio(SAR) [11], Signal-to-Distortion ratio(SDR) [11] provided by the Bss-Eval toolbox [11] and overall perceptual score (OPS) [12], target-related perceptual score (TPS) [12], interference-related perceptual score (IPS) [12] and artifacts-related perceptual score (APS) that are provided by the PEASS [12].

The following observations are made:

1) STOI [8]: In most cases, the proposed system achieves better performance compared to the original and ideal systems. This shows that the proposed system outperforms the other ones in terms of speech intelligibility.

2) PESQ [9]: The proposed system improves perceived speech quality over the competitive ones in most scenarios.

3) SNR_{loss} [10]: This metric predicts the speech intelligibility in noisy condition. The results show that the proposed approach outperforms the other ones.

4) BSS Eval toolkit [11]: The toolkit evaluates speech separation in terms of SIR, SAR and SDR. The following results are made:

- The proposed system outperforms the other systems in terms of SDR. Thus it gives the better overall quality of the proposed speech separation method.
- The proposed system gives better SIR than both systems when testing on Clip 1 and Clip 3. However, when testing on Clip 2, System 1 gives better SIR and when testing on Clip 3, System 3 gives better results.
- Same observation goes to SAR. In fact, the proposed system gives better SAR results only when testing on Clips 2 and 3. However in Clip 1 and 4 the results are comparable to the other systems.

5) PEASS [12]: We also report the evaluation results of the prosed system in using the perceptual evaluation methods for audio source separation (PEASS). The metrics reveal that:

- The proposed approach achieves the highest OPS score over the competitive ones. This proves that our approach gives separated signals that are the closest to the clean signal. The spectrograms in Figure 2 confirm that.
- The proposed system achieves the highest APS scores. This means that, compared to the competitive approaches, our approach performs better separation without producing artifacts. The APS scores are in line with those of SAR.
- In most scenarios, the proposed system gives the highest TPS scores over the competitive ones.
- The same observation goes to the IPS. Thus we can conclude that the proposed system permits

TABLE 1: Evaluation of the proposed system. The proposed system (Sys 2) is compared to the original system (Sys 1) and the Ideal one (Sys 3). The systems are tested on Four Scenarios which are mixtures of two speakers of: same gender (SG), different gender (DG) and same talker (ST). Evaluation and comparison are conduct in terms of STOI[8], PESQ[9], SNR_{LOSS}[10], SIR[11],SAR[11], SDR[11], OPS[12], TPS[12], IPS[12] and APS[12]. Best results are indicated in colour in each Sub-Column.

		Clip 1(SG 3dB)			Clip 2(DG 0dB)			Clip 3(ST -6dB)			Clip 4(ST 0dB)		
Criterian		Sys1	Sys2	Sys3	Sys1	Sys2	Sys3	Sys1	Sys2	Sys3	Sys1	Sys2	Sys3
	STOI	0,665	0,668	0,637	0,118	0,755	0,001	0,732	0,939	0,860	0,652	0,425	0,594
	PESQ	2,222	3,305	2,620	1,807	0,869	0,656	1,502	3,039	2,206	0,741	2,379	0,327
	SNRloss	0,994	0,929	0,976	0,997	0,954	0,977	0,994	0,852	0,899	0,984	0,953	0,956
Bss Eval	SIR	-11,186	5.898	4.952	-6,328	-4,612	-4,782	-5,212	7.006	-0,377	-7,781	-3,733	-0,272
	SAR	13.862	11.048	11.462	13.762	13.940	11.312	11.560	15.324	10.955	12.206	4.990	6.507
	SDR	-7,77	4.073	-3,311	-10,55	-1,665	-9,558	-9,458	8.716	-5,191	-9,552	-0,002	-3,39
PEASS	OPS	10.727	30.060	11.903	6.672	6.789	3.202	6.864	17.266	2.660	6.211	4.528	1.989
	TPS	52.087	62.198	44.154	53.531	10.350	13.588	51.395	52.857	19.098	40.50	48.274	19.670
	IPS	14.248	63.361	30.458	11.418	0.924	2.930	15.533	48.668	6.299	18.235	18.310	6.053
	APS	33.621	33.724	30.451	29.638	52.807	49.174	93.151	24.370	100	100	100	100

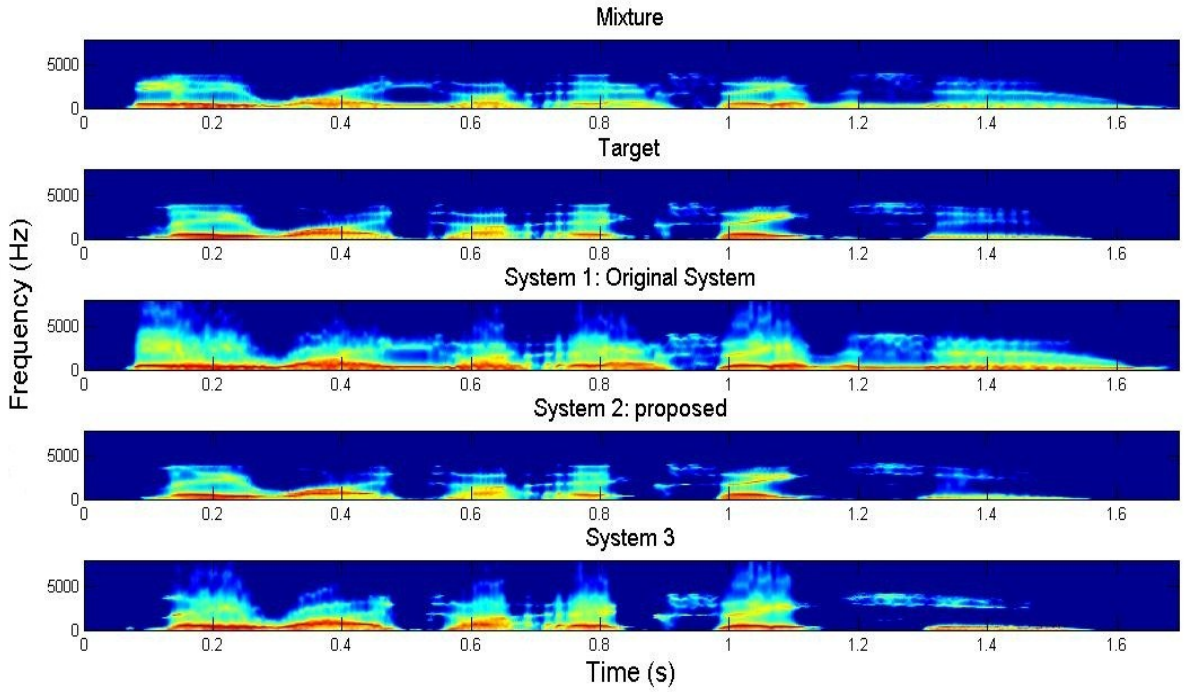

Fig.2. Spectrograms of the original signal (Mixture), Target signal, output of the original system (System 1), output of the proposed system (System 2), and of the Ideal system (System 3). This example shows that the proposed approach outperforms the other one.

interference cancellation in the separated signal.

IV. Discussion

The separated signals using CASA-based techniques often miss perceptual quality due to the severe cross-talk problem [13]. In general, applying hard masks inevitably cause artifacts and severe cross-talk in the output signals, as proved in [15]. The studies in [3] [14] reported the crosstalk percentage of the CASA-based approaches. In this work we proved that replacing the hard masking in a CASA system by a soft one considerably improves the system performance and the quality of the produced signals. The original CASA system can effectively separate speakers that are fragmented and overlapping and represented by a compact area in the time-frequency. However, because of hard masking, it produces artifacts and additional noise which decreases speech intelligibly. As soft masks attribute a probability varying between 0 and 1 instead of a binary value, it provides a better flexibility therefore a better separation quality when integrated in a CASA system.

In most of the test scenarios used here, Soft-CASA techniques outperform CASA by some degree. In fact, Soft-CASA achieves a 28.84% improvement in STOI, 7.1% improvement in SNRloss, 92% improvement in OPS and 53% PESQ.

From the performance profile of the Soft-CASA we conclude that this system is promoting and effective even when the two speakers are represented in very compact area such as mixtures of talkers of SG and ST which was the most challenging test scenarios.

In order to improve the original CASA system, another alternative technique is to train a set of classifiers to decide if each T-F cell in the mixed signal belongs to a particular source in the mixture. In [15], a similar approach was developed to identify speech-dominated regions in a mixture. However this technique can only be used for simple separation tasks. Also, the classifiers are only applied for separating mixtures composed of speakers similar to those on which they were trained. As a consequence, the separation quality is poor if the target signal is different from the trained data. The classifier method is unable to achieve SCSS unless the training data belongs to the same set of speakers as those in the mixture signal. The most serious problem with the ordinary CASA is of hard masking is that masks are binary that means that a component either belongs to the desired signal or not; this cannot be decided with certainty in most of the cases.

V. Conclusion

In this paper we proposed a new system for SCSS that we called Soft-CASA. In fact Soft-CASA is an extension of the ordinary CASA system. Almost all the previous CASA-based systems use the binary mask as principal tool required for separation, this decreases the quality of the produced signal. To solve this problem, we have used a soft mask. Soft-CASA thus benefits from the advantages of an ordinary CASA

system as well as soft masking. The results of the evaluation of the proposed system prove its efficiency and the robustness over the original system. As future work we plan to implement our algorithms in order to do real-time SCSS.

REFERENCES

[1] SHAMMA, Shihab A., ELHILALI, Mounya, et MICHEYL, Christophe. Temporal coherence and attention in auditory scene analysis. *Trends in neurosciences*, 2011, vol. 34, no 3, p. 114-123.

[2] WANG, DeLiang et BROWN, Guy J. Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE Press, 2006.

[3] HU, Guoning et WANG, DeLiang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 2004, vol. 15, no 5, p. 1135-1150.

[4] RADFAR, Mohammad H. et DANSEREAU, Richard M. Single-channel speech separation using soft mask filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, vol. 15, no 8, p. 2299-2310.

[5] VAN SEGBROECK, Maarten, *et al.* Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks. In : *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008. p. 4393-4396.

[6] MESSAOUD, Mohamed Anouar Ben; BOUZID, Aïcha. Pitch estimation of speech and music sound based on multi-scale product with auditory feature extraction. International Journal of Speech Technology, 2016, 19.1: 65-73.

[7] M. Cooke, J.R. Hershey, and S.J. Rennie, "Monaural speech separation and recognition challenge," Elsevier Computer Speech and Language, vol. 24, no. 1, pp. 1–15, 2010.

[8] TAAL, Cees H., HENDRIKS, Richard C., HEUSDENS, Richard, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In : ICASSP. 2010. p. 4214-4217.

[9] RIX, Antony W., BEERENDS, John G., HOLLIER, Michael P., et al. 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. IEEE, 2001. p. 749-752.

[10] MA, Jianfen et LOIZOU, Philipos C. SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication*, 2011, vol. 53, no 3, p. 340-354.

[11] VINCENT, Emmanuel, GRIBONVAL, Rémi, et FÉVOTTE, Cédric. Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing, 2006, vol. 14, no 4, p. 1462-1469.

[12] EMIYA, Valentin, VINCENT, Emmanuel, HARLANDER, Niklas, et al. Subjective and objective quality assessment of audio source separation. IEEE Transactions on Audio, Speech, and Language Processing, 2011, vol. 19, no 7, p. 2046-2057.

[13] WIEM, Belhedi, MESSAOUD, Mohamed Anouar Ben, et AÝCHA, Bouzid. Single channel speech separation based on sinusoidal modeling. In : Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on. IEEE, 2016. p. 672-676.

[14] MOWLAEE, Pejman, SAYADIYAN, Abolghasem, et SHEIKHZADEH, Hamid. Evaluating single-channel speech separation performance in transform-domain. Journal of Zhejiang University SCIENCE C, 2010, vol. 11, no 3, p. 160-174.

[15] M. Seltzer, B. Raj, and R. Stern, " A bayesian classifier for spectrographic mask estimation for missing feature speech recognition, " Speech Communication, vol. 43, no. 4, pp. 379 – 393, 2004.

[16] MESSAOUD, Mohamed Anouar Ben; BOUZID, Aïcha; ELLOUZE, Noureddine. An efficient method for fundamental frequency determination of noisy speech. In: International Conference on Nonlinear Speech Processing. Springer Berlin Heidelberg, 2013. p. 33-41.

[17] MESSAOUD, Mohamed Anouar Ben; BOUZID, Aïcha; ELLOUZE, Noureddine. Estimation du pitch et décision de voisement par compression spectrale de l'autocorrélation du produit multi-échelle.