SPEECH QUALITY ENHANCEMENT BASED ON SPECTRAL SUBTRACTION

Jessica C. S. Veras[1], *Thiago de M. Prego*[1,2], *Amaro A. de Lima*[1,2], *Tadeu N. Ferreira*[1,3], *and Sergio L. Netto*[1]

1. Program of Electrical Engineering, Federal University of Rio de Janeiro, Brazil.
2. Program of Electrical Engineering, Federal Center for Technological Education, Brazil.
3. Program of Telecommunication Engineering, Fluminense Federal University, Brazil.

{jessica.veras, thiago.prego, amaro.lima, tadeu.ferreira, sergioln}@smt.ufrj.br

ABSTRACT

This paper presents an algorithm for reverberant speech enhancement based on single channel blind spectral subtraction. This algorithm deals with the late components of the reverberation effect and it was optimized using 18 speech signals from the NBP database. Experimental results show that the proposed algorithm is well suited for speech enhancement in teleconference and telepresence environments and it can increase the perceptual quality by up to 31% and 62% of reverberant and noisy speech signals from databases with simulated and real reverberation and noise effects, respectively.

1. INTRODUCTION

Reverberation can strongly affect the performance of state-of-the-art systems of speech/speaker recognition and hearing-aid, motivating the use of speech enhancement techniques. Although reverberation degrades speech intelligibility and perceptual quality, in a small amount it makes speech more pleasant to common listeners [1]. The use of a microphone array is commonly associated to dereverberation techniques, but for the applications previously mentioned the use of one microphone approach is more indicated.

This paper describes the algorithm proposed in [2] for the enhancement of reverberant speech signals based on spectral subtraction. This algorithm is a modification of the two-stage algorithm for one-microphone reverberant speech enhancement [3] and it was fine tuned using the 18 signals from the NBP database [4]. The algorithm was tested in two databases composed by signals from the WSJCAM0 corpus [5] and MC-WSJ-AV corpus [6].

This paper is organized as follows: In Section 2, the spectral subtraction dereverberation algorithm is explained in details. Section 3 describes the training and test databases employed in this work. Section 4 briefly describes the REVERB Challenge 2014 [7] and its suggested quality assessment metrics. Section 5 shows the results of the training experiments used in the parameter optimizations and also the test experiments analyzing in details the results achieved with the optimized parameters. Finally, a conclusion concerning the performance increase is included in Section 6.

2. SPECTRAL SUBTRACTION ALGORITHM

The spectral subtraction algorithm [2] is based only on the spectral subtraction stage of the two-stage algorithm [3] and it aims to reduce the effect of long term reverberation components of reverberant speech signals. In this paper the reverberant signal is modeled as the convolution of the room impulse response (RIR) $h(n)$ and the anechoic (clean) speech signal $s(n)$,

$$z(n) = \sum_{l=0}^{N} h(l)s(n-l). \quad (1)$$

The Figure 1 shows the components of the spectral subtraction block, under the assumption that the early reflections and late reverberation components of the RIR are approximately uncorrelated [3].

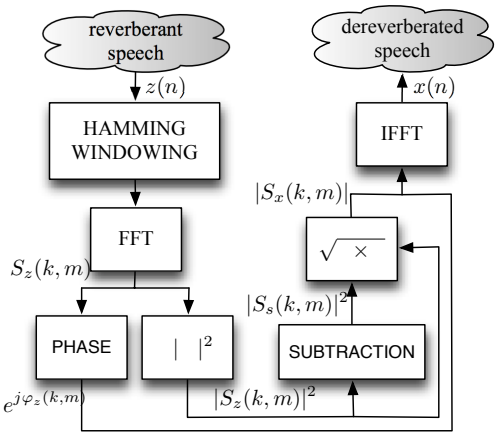

Fig. 1. Diagram of the spectral subtraction algorithm.

Let $S_z(k,m) = |S_z(k,m)|e^{j\varphi_z(k,m)}$ be the FFT of the m-th frame of the windowed version of $z(n)$, where a 32 ms hamming window of 24 ms overlap is used. Also let ρ be the length of the early reflection in frames, commonly considered

around 50 ms, γ be the scaling factor that establishes the relative strength of the late impulse components and $w(m)$ be an asymmetrical smoothing window based on the Rayleigh distribution, given by

$$\begin{cases} w(m) &= \left(\frac{m+a}{a^2}\right) e^{\left(\frac{-(m+a)^2}{2a^2}\right)}, & \text{if } m > -a \\ w(m) &= 0, & \text{otherwise} \end{cases}, \quad (2)$$

where a controls the overall spread of the function and it has to be smaller than ρ in order to provide a reasonable match to the shape of the equalized impulse response.

The model of the power spectrum of the late reverberation, in order to perform the spectral subtraction, can be described as

$$|S_l(k,m)|^2 = \sum_{\tau=-\infty}^{\infty} \gamma w(\tau-\rho)|S_z(k,m-\tau)|^2, \quad (3)$$

where k is the frequency bin and m refers to the time frame.

Considering that the early and late components are mutually uncorrelated, the power spectrum of the early impulse components can be estimated by subtracting the power spectrum of the late impulse components from the reverberant speech. The spectrum subtraction scheme performs a weighting in the power spectrum of $z(n)$, where the block SUBTRACTION is given by

$$G(k,m) = \max\left[1 - \frac{|S_l(k,m)|^2}{|S_z(k,m)|^2}, \epsilon\right], \quad (4)$$

and finally the magnitude spectrum of $x(n)$ is

$$|S_x(k,m)| = \sqrt{|S_z(k,m)|^2 \times G(k,m)}. \quad (5)$$

In order to calculate $x(n)$ the phase $\varphi_z(k,m)$ of $S_z(k,m)$ is combined to the magnitude $|S_x(k,m)|$, so

$$S_x(k,m) = |S_x(k,m)|e^{j\varphi_z(k,m)} \quad (6)$$

In order to asses the perceptual quality of a reverberant speech signal, this work uses the quality measure known as QAreverb [4], which is given by

$$Q = -\frac{\sigma^2 T_{60}}{R^\gamma}, \quad (7)$$

where σ^2 is the room spectral variance defined in [8], T_{60} is the reverberation time defined as the period of time required for the sound-pressure to decay 60 dB, in this work estimated by Karjalainen's algorithm [9], R is the direct-to-reverberant energy ratio defined in [10], with $\gamma = 0.3$. In practice, a higher T_{60} indicates a more lasting reverberation effect. σ^2, T_{60} and R are obtained directly from the RIR, $h(n)$, which is estimated from the deconvolution process between the clean and the reverberant speech signals.

The Q score is mapped to the Q_{MOS} score, which uses the MOS (mean opinion score) scale, through a third order followed by a first order polynomial functions.

The parameters scaling factor γ, attenuation limit ϵ, length of the early reflections ρ and spread control a were selected through an exhaustive search using the Q_{MOS} score as the measure to be optimized.

3. TRAINING AND TEST DATABASES

The database used to fine tune the selection for the algorithm's parameters is called New Brazilian Portuguese (NBP) database [4] and it was developed using 3 types of reverberation effect: artificial, natural and real. The 4 anechoic speech signals, 2 for male and 2 for female, were used to generate 24 degraded speech signals with artificial reverberation, 68 with real reverberation and 108 with natural reverberation effects, making a total of 204 speech signals (200 reverberant speech signals plus the 4 anechoic speech signals), all of them sampled at $F_s = 48$ kHz. The speech signals were assessed by an absolute category rate (ACR) MOS test with 30 listeners. The reverberation effect was imposed onto these 4 anechoic signals using three distinct reverberation effects, namely:

- Artificial reverberation: It is represented by 6 distinct artificially generated RIRs with source-microphone distance of 180 cm, where the early reflections were modeled via the image method, and the late reverberation used the feedback delay network method and a modified version of Gardner's method, for emulating the lower and higher reverberation times, respectively. The average measured reverberation time were $\{196, 292, 387, 469, 574, 664\}$ ms.
- Natural reverberation: This approach consists in RIRs obtained from the direct recordings of 4 different types of rooms with several source-microphone distances for each room, as detailed in [11], making a total of different 17 RIRs. The average measured reverberation time for the 4 rooms are in the range of $\{120, 230, 430, 780\}$ ms. The source-microphone distances range from 50 cm to 1020 cm.
- Real reverberation: It is the only case where the degraded signals were directly played/recorded in the rooms, without using the convolution operation between the RIRs and the anechoic speech signals. The 7 rooms used in the recordings presented different room dimensions and employed at least 3 different source-microphone distances emulating a total of 27 RIRs with average measured reverberation time in the range of $\{140, 390, 570, 650, 700, 890, 920\}$ ms. The source-microphone distances range from 50 cm to 400 cm.

The training database is composed of 18 reverberant speech signals, one for each environment (anechoic, 6 ar-

tificial RIRs, 4 natural rooms, 6 real rooms). The overall best performance was $Q_{\mathrm{MOS}} = 3.78$ for $\{\gamma = 0.35, \epsilon = 10^{-3}, \rho = 7, a = 6\}$, against $Q_{\mathrm{MOS}} = 3.46$ for the unprocessed training database.

The test database was suggested by the REVERB Challenge described in Section 4 and it is divided in development database and evaluation database, each of these databases is also divided in two datasets:

- SimData: speech signals from the WSJCAM0 database [5] convolved with RIRs measured in three different rooms with two different source-microphone distances and background noise was added to each signal at fixed signal-to-noise ratio (SNR). The reverberation time for these rooms are about $\{250, 500, 700\}$ ms and the source-microphone distances are $\{50, 200\}$ cm.
- RealData: speech signals recorded in a reverberant and noisy room from the MC-WSJ-AV database [6] with two different source-microphone distances. The reverberation time for this rooms is about 700 ms and the source-microphone distance are $\{50, 250\}$ cm.

The development database is composed by 1484 signals from SimData and 179 signals from RealData. The evaluation database is composed by 2176 from SimData and 372 from RealData.

4. REVERB CHALLENGE 2014

The REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge 2014 [7] consists of two parts: speech enhancement task and automatic speech recognition task. The speech enhancement task, which is the task in which the algorithm described in this paper is participating, consists of mitigating the effects of noise and reverberation of speech signals from the development and evaluation databases described in Section 3. The performance of the algorithms participating in the speech enhancement task is assessed by four mandatory and three optional measures. The mandatory are:

- Cepstral distance (CD) [12]: measures the discrepancy between degraded and clean signals. Can only be measured in SimData as it needs the clean signal.
- Log-likelihood ratio (LLR) [12]: is a measure of the discrepancy between degraded and clean signals. Can only be measured in SimData as it needs the clean signal.
- Frequency-weighted segmental SNR (FWSS) [12]: measures the discrepancy between degraded and clean signals. Can only be measured in SimData as it needs the clean signal.
- Speech-to-reverberation modulation energy ratio (SRMR) [13]: measures the perceptual quality of a speech signal degraded by noise and reverberation. Can be used for both SimData and RealData quality assessment.

The optional are:

- Computational cost: measures the how long (in seconds) the algorithm (ATime) took to process a given dataset. As this is strongly dependent on the platform configuration, the computational cost (RTime) of the given reference code is also computed for each dataset. The algorithms were used in MATLAB Version 7.12.0.635 (R2011a) 64-bit in a computing environment with Windows 7 64-bit operating system, AMD Vision Dual Core E-350 1.60 GHz processor and 4 GB RAM.
- Word error rate (WER) [14]: common metric to measure performance of speech recognition systems. WER is measured after the dataset is processed by the speech enhancement algorithm and the reference algorithm for automatic speech recognition (which uses HTK [15]) given by the REVERB Challenge. The automatic speech recognition algorithm was used in a linux ubuntu 12.04 virtual machine in a MAC OS-X 10.8 64-bits, with a 2.3 GHz intel quadcore i7 processor and 8GB RAM.
- Perceptual Evaluation of Speech Quality (PESQ) [16]: ITU-T standard for evaluate the perceptual quality of speech coders. As the publishing of PESQ results demands the purchase of a license, the authors of this paper did not used it in the REVERB Challenge.

5. EXPERIMENTAL RESULTS

Tables 1 to 6 show the results for quality assessment metrics cepstral distance (CD), log-likelihood ratio (LLR), Frequency-weighted segmental SNR (FWSS), Speech-to-reverberation modulation energy ratio (SRMR), Q_{MOS} (a blind version of Q_{MOS} was used in RealData), Word Error Rate (WER) and computational cost (ATime and RTime) for each subset of the development and evaluation databases. It can be observed that all averages of the quality assessment measures for the processed databases are greater then its unprocessed counterpart. Some partial results give the false impression of a worse performance for the processed compared to the unprocessed databases (WER is not the case), but it can be stated that the difference between the values are lesser then the expected estimation error. The results for SimData Room 1 show a worse performance for the automatic speech recognition algorithm for the processed speech signals because the perceptual quality of the signals were already high, i.e., the MOS [17] estimated by Q_{MOS} was around 4, which means it is labeled as GOOD.

Regarding the development database the objective metrics CD, LLR and FWSS plus the perceptual metrics SRMR and

Table 1. Quality measures results using single channel data full batch for unprocessed development SimData dataset.

Measure	Room 1		Room 2		Room 3		Avg.
	Near	Far	Near	Far	Near	Far	-
CD	1.96	2.65	4.58	5.08	4.2	4.82	3.88
LLR	0.34	0.38	0.51	0.77	0.65	0.85	0.58
FWSS	8.1	6.75	3.07	0.53	2.32	0.14	3.49
SRMR	4.37	4.63	3.67	2.94	3.66	2.76	3.67
Q_{MOS}	4.23	3.87	3.35	1.52	3.27	2.35	3.10
WER (%)	15.3	25.3	43.9	85.8	52.0	88.9	51.8

Table 2. Quality measures results using single channel data full batch for processed development SimData dataset.

Measure	Room 1		Room 2		Room 3		Avg.
	Near	Far	Near	Far	Near	Far	-
CD	3.46	3.46	4.64	4.78	4.27	4.44	4.17
LLR	0.51	0.52	0.51	0.69	0.64	0.77	0.61
FWSS	8.07	7.56	5.39	2.55	4.19	1.96	4.96
SRMR	5.06	5.68	4.71	4.32	4.74	4.13	4.77
Q_{MOS}	4.21	3.96	3.81	2.42	3.69	2.85	3.49
WER (%)	36.5	46.0	34.6	63.2	45.3	64.5	48.3
ATime	1167	1200	1185	1667	1067	1206	1249
RTime	181	164	189	199	181	192	184

Q_{MOS} and WER presented an improvement of 7%, 5%, 42%, 30%, 13% and 3.5%, for SimData and SRMR and Q_{MOS} increased by 62% and 51% and WER decreased by 22.5%, for RealData. Regarding the evaluation database the objective metrics CD, LLR and FWSS plus the perceptual metrics SRMR and Q_{MOS} and WER presented an improvement of 6%, 2%, 43%, 31%, 15% and 0.9%, for SimData and SRMR and Q_{MOS} increased by 60% and 50% and WER decreased by 14.6%, for RealData.

It is important to note that the training database used to fine tune the algorithm parameters belongs to a different corpus than both development and evaluation databases (these two databases have signals from the same corpus). In spite of that, the three databases (training, development and evaluation) share common reverberation times and source-microphone distances, what explains the performance of the algorithm proposed in [2] and shows that the algorithm is well suited for being used for speech enhancement in teleconference and telepresence environments.

6. CONCLUSION

This work analyzes the performance of the algorithm for single channel speech enhancement proposed in [2] with respect to perceptual quality metrics for reverberant and noisy speech signals.

The algorithm is based in spectral subtraction and four of

Table 3. Quality measures results using single channel data full batch for development RealData dataset.

Measure	Unprocessed dataset			Processed dataset		
	Near	Far	Avg.	Near	Far	Avg.
SRMR	4.06	3.52	3.79	6.51	5.74	6.13
Q_{MOS}	2.45	2.41	2.43	3.72	3.64	3.68
WER (%)	88.7	88.3	88.5	69.0	62.9	66.0
ATime	-	-	-	340	329	335
RTime	-	-	-	56	53	55

Table 4. Quality measures results using single channel data full batch for unprocessed evaluation SimData dataset.

Measure	Room 1		Room 2		Room 3		Avg.
	Near	Far	Near	Far	Near	Far	-
CD	1.99	2.67	4.63	5.21	4.38	4.96	3.97
LLR	0.35	0.38	0.49	0.75	0.65	0.84	0.58
FWSS	8.12	6.68	3.35	1.04	2.27	0.24	3.62
SRMR	4.5	4.58	3.74	2.97	3.57	2.73	3.68
Q_{MOS}	4.24	3.96	3.61	2.37	3.2	2.4	3.30
WER (%)	18.1	25.4	43.0	82.2	53.5	88.0	51.7

Table 5. Quality measures results using single channel data full batch for processed evaluation SimData dataset.

Measure	Room 1		Room 2		Room 3		Avg.
	Near	Far	Near	Far	Near	Far	-
CD	3.49	3.53	4.62	4.86	4.29	4.55	4.22
LLR	0.53	0.53	0.48	0.65	0.62	0.74	0.59
FWSS	7.97	7.65	5.85	3.14	4.3	2.03	5.16
SRMR	5.21	5.55	4.9	4.35	4.8	4.1	4.82
Q_{MOS}	4.22	4.02	3.99	2.87	3.73	3.88	3.79
WER (%)	47.5	52.5	38.4	57.1	43.4	66.2	50.8
ATime	1661	2028	1754	1834	1760	1709	1791
RTime	331	247	290	328	278	307	297

Table 6. Quality measure results using single channel data full batch for evaluation RealData dataset.

Measure	Unprocessed dataset			Processed dataset		
	Near	Far	Avg.	Near	Far	Avg.
SRMR	3.17	3.19	3.18	5.08	5.12	5.10
Q_{MOS}	2.51	2.57	2.54	3.79	3.8	3.80
WER (%)	89.7	87.3	88.5	76.3	71.5	73.9
ATime	-	-	-	736	622	679
RTime	-	-	-	138	126	132

its parameters (scaling factor γ, attenuation limit ϵ, frame size of the early reflection ρ and spreading factor a of the Rayleigh distribution) were fine tuned using a 18 reverberant speech signals from the NBP database [4]. The development and evaluation databases used in the REVERB Challenge 2014 were not used to train the algorithm in any way, both being used as test databases.

The metrics suggested by the REVERB Challenge 2014 (CD, LLR, FWSS and SRMR) plus the Q_{MOS} metric show that the algorithm proposed in [2] is well suited for speech enhancement for teleconference and telepresence environments. Regarding the evaliation database, the algorithm increases the estimated metrics for quality assessment CD, LLR, FWSS, SRMR and Q_{MOS} by 6%, 2%, 43%, 31% and 15% and decreases WER by 0.9%, for SimData and SRMR and Q_{MOS} is increased by 60% and 50% and WER is decreased by 14.6%, for RealData.

7. ACKNOWLEDGEMENTS

The authors would like to thank Prof. M. Karjalainen, for providing the T_{60} estimation algorithm, and Prof. D. Wang, for providing the original two-stage algorithm for reverberant speech enhancement.

8. REFERENCES

[1] R. Appel and J. Beerends, "On the Quality of Hearing One's Own Voice," *J. Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, Apr. 2002.

[2] T. de M. Prego, A. A. de Lima and S. L. Netto, "Perceptual Improvement of a Two-Stage Algorithm for Speech Dereverberation," *Proc. InterSpeech*, Lyon, France, pp. 1360–1364, Sep. 2013.

[3] M. Wu and D. Wang, "A Two-Stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, pp. 774–784, May 2006.

[4] Amaro A. de Lima, Thiago de M. Prego, Sergio L. Netto, Bowon Lee, Amir Said, Ronald W. Schafer, Ton Kalker and Majid Fozunbal, "On the quality-assessment of reverberated speech," *Speech Communication*, pp. 393–401, Mar. 2012.

[5] T. Robinson, J. Fransen, D. P ye, J. Foote, and S. Renals, "Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition," *Proc. ICASSP 95*, pp. 81–84, 1995.

[6] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357–362, 2005.

[7] K. Kinoshita, et al., "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, pp. 1–4, 2013.

[8] J. J. Jetz, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoustic. Soc. Am.*, vol. 65, pp. 1204–1211, May 1979.

[9] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy reponse measurements," *Proc. Conv. Audio Engineering Society*, Amsterdam, Netherlands, pp. 867–878, May 2001.

[10] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *J. Acoustic. Soc. Am.*, vol. 124, pp. 982–993, Aug. 2008.

[11] M. Jeub, M. Schäfer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms," *Proc. 16th Int. Conf. on Digital Signal Processing*, Santorini, Greece, pp. 550–554, 2009.

[12] Hu, and Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE T-ASLP*, 16(1), 229-238, 2008.

[13] Falk, et al., "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE T-ASLP*, 18(7), 1766-1774, 2010.

[14] D. Klakow, and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, pp. 19–28, Sep. 2002.

[15] S. Young, et al., "HTK Hidden Markov Model Toolkit," Mar. 2009. http://htk.eng.cam.ac.uk/.

[16] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," 2001.

[17] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," 1996.