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]

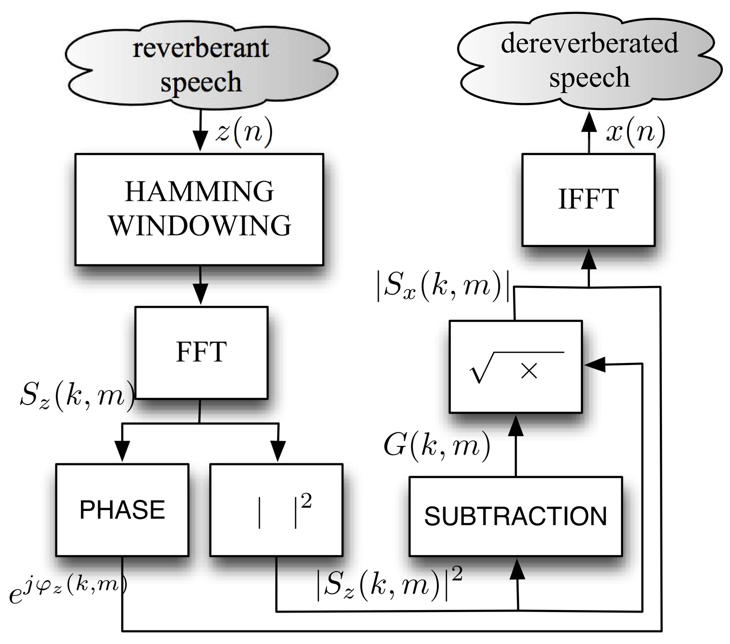
[1]Program of Electrical Engineering, Federal University of Rio de Janeiro, Brazil
[2]Program of Electrical Engineering, Federal Center for Technological Education, Brazil
[3] Program of Telecommunication Engineering, Fluminense Federal University, Brazil

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.

Spectral Subtraction Algorithm

Block Diagram



Description

- $S_z(k,m) = |S_z(k,m)|e^{j\varphi_z(k,m)}$ is the FFT of the m-th frame of the windowed version of $z(n)$.
- $w(m,a)$ is smoothing window based on the Rayleigh distribution.
- a controls the overall function time spread ($a < \rho$).
- ρ is the length of early reflections.
- Power spectrum model of the late reverberation is $|S_l(k,m)|^2 = \gamma w(m-\rho,a) * |S_z(k,m)|^2$, with γ scaling factor.
- SUBTRACTION block is $G(k,m) = \max\left[1 - \frac{|S_l(k,m)|^2}{|S_z(k,m)|^2}, \epsilon\right]$.
- $|S_x(k,m) = (\sqrt{|S_z(k,m)|^2 \times |S_s(k,m)|^2})e^{j\varphi_z(k,m)}$.
- Practical values: $\{\gamma,\epsilon,\rho,a\} = \{0.35, 10^{-3}, 7, 6\}$.

Training and Test databases

Training database

- The new Brazilian-Portuguese (NBP) database.
- F_s = 48-kHz sampling frequency.
- 4 anechoic speech signals (2 male and 2 female) were used to generate reverberant speech following three different frameworks:
  - Artificial reverberation: 6 distinct artificially generated RIRs. Source-microphone (d) distance of 180 cm and reverberation time (T_{60}) $\{196, 292, 387, 469, 574, 664\}$ ms.
  - Natural reverberation: 17 different RIRs obtained from the direct recordings. $T_{60} = \{120, 230, 430, 780\}$ ms and $d = [50, 1020]$ cm.
  - Real reverberation: 27 RIRs obtained from signals directly played/recorded in different rooms. $d = [50, 400]$ and $T_{60} = \{140, 390, 570, 650, 700, 890, 920\}$ ms.
  - Total of 204 signals (4 anechoic, 24 artificial, 68 natural, and 108 real).
- All NBP signals were assessed through ACR MOS test with 30 non-trained listeners for each signal.
- The database is available upon request by e-mail to the authors.
- In this work, the training database is composed of 18 signals from NBP, one for each environment (anechoic, 6 artificial RIRs, 4 natural rooms, 6 real rooms).

Test database

- F_s = 16-kHz sampling frequency.
- Composed of signals from two databases:
  - SimData: speech signals from the WSJCAM0 convolved with measured RIRs and background noise was added to each signal. $T_{60} = \{250, 500, 700\}$ ms and $d = \{50, 200\}$ cm.
  - RealData: speech signals from the MC-WSJ-AV database were played and recorded in a reverberant and noisy room. $T_{60} = 700$ ms and $d = \{50, 250\}$ cm.
- Two databases were suggested:
  - Development database: 1484 signals from SimData and 179 from RealData.
  - Evaluation database: 2176 signals from SimData and 372 from RealData.
- Both development and evaluation databases were used as test databases in this work.

Quality measures

REVERB Challenge measures

- The performance of the algorithms participating in the enhancement task is assessed by 4 mandatory and 3 optional measures:
  - Cepstral distance (CD): measures the discrepancy between degraded and clean signals. Can only be measured in SimData as it needs the clean signal.
  - Log-likelihood ratio (LLR): 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): 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): measures the perceptual quality of a speech signal degraded by noise and reverberation. Can be used for both SimData and RealData quality assessment.
  - 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.
  - Word error rate (WER): 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 automatic speech recognition given by the REVERB Challenge. 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.
  - Perceptual Evaluation of Speech Quality (PESQ): 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.

Aditional perceptual quality measure

- In order to evaluate the perceptual quality of a reverberant speech signal, this work employs the QAreverb measure $Q = -\frac{T_{60}\sigma_r^2}{R^\xi}$.
  - T_{60} is the reverberation time.
  - σ_r^2 is the room spectral variance (RSV).
  - R is the direct-to-reverberant energy ratio (DRR) with $\xi = 0.3$.
  - Q_{MOS} is the Q score mapped into MOS scale.

Experimental Results

Table 1 : Orig. development SimData.

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 4 : Orig. evaluation SimData.

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 2 : Proc. development SimData.

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

Table 5 : Proc. evaluation SimData.

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 3 : Development RealData.

Measure	Original 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 6 : Evaluation RealData.

Measure	Original 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

Conclusions

Advantages of the proposed approach

- Dereverberation algorithm fine tuned w/ perceptual measure.
- Improvements for the development database:
  - SimData: CD (7%), LLR (5%), FWSS (42%), SRMR (30%) and Q_{MOS} (13%) and WER (3.5%)
  - RealData: SRMR (62%), Q_{MOS} (51%) and WER (22.5%)
- Improvements for the evaluation database:
  - SimData: CD (6%), LLR (2%), FWSS (43%), SRMR (31%) and Q_{MOS} (15%) and WER (0.9%)
  - RealData: SRMR (60%), Q_{MOS} (50%) and WER (14.6%)

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