QUALITY ASSESSMENT OF SPEECH ENHANCED USING PARTICLE FILTERS

Frédéric Mustière, Martin Bouchard, Miodrag Bolić

School of Information Technology and Engineering, University of Ottawa
800 King Edward Ave., Ottawa, ON, Canada, K1N 6N5
Email:{mustiere,bouchard,mbolic}@site.uottawa.ca

ABSTRACT

Among the various types of existing speech enhancement algorithms, respective advantages are usually well-known and well-studied, making it possible to choose the right algorithm given a certain type of audio quality requirement (e.g. background noise intrusiveness, naturalness of the speech, etc.). Very little has however been said so far on the quality of speech enhanced by particle filters (PFs). In this paper, we show the detailed results of the analysis of PF-enhanced speech signals. This analysis is conducted using several objective measures, and is based on the comparison with other enhancement algorithms. We find that in general, PF-based algorithms yield speech signals which are among the most natural-sounding, with a residual noise that can be seen as a white noise with its variance "modulated" by the resulting speech.

Index Terms— particle filters, speech enhancement, speech quality, objective measures

1. INTRODUCTION

In the diversity of speech enhancement algorithms, a fairly recent family of methods is that of particle filters. A PF is a sequential estimation method based on Monte-Carlo simulation, which can operate on the broadest range of state-space formulated problems [1]. Using an auto-regressive (AR) speech model, researchers have been able to successfully apply PFs to speech denoising, employing various additional techniques such as Rao-Blackwellisation [2,3], smoothing [3], auxiliary PF [4], etc. Unfortunately, the quality of speech signals enhanced by particle filtering algorithms has not yet been thoroughly documented. If we consider that the assessment of speech quality necessitates the use of several objective measures, descriptions of informal listening tests and comparisons with other algorithms, then the existing literature has little to offer on the subject.

In [2], only SNR results are reported. The authors describe the result of an experiment on a spoken sentence, with a particular input SNR: the two main observations are that the residual noise is approximately white and time-varying, and that there is no musical noise. A comparison is made with another algorithm (one that is presented in [5]), although this comparison is summarized by the fact that the algorithm in [5] introduces some musical noise, while the particle filter does not. In [3], again only SNR results are reported. Among the experiments, one is conducted on a sentence at a given SNR, and the results are compared with those obtained with an extended Kalman smoother (EKS). The conclusions given are that the Rao-Blackwellised particle smoother proposed removes more noise than the EKS, but that there is a higher sibilant residual noise. In [4], no objective measures are used, and the authors note again the absence of musical noise, and some residual sibilance during unvoiced sounds.

The SNR is unfortunately a poor indicator of speech quality, for it is not well correlated with mean subjective opinions on the matter [6]. Since test results on speech signals are given for a single SNR value, what we may conclude from the sources abovementioned in terms of intelligibility and quality is therefore limited. We can however expect a PF-enhanced speech signal to contain no musical noise, and we also expect to hear some sibilance, as well as a time-varying white residual noise.

In this paper, we are interested in filling this gap by thoroughly analyzing the performance of two PF algorithms. The analysis will be mainly conducted using 4 popular objective measures: the overall SNR, the average segmental SNR, the PESQ, and the LAR, which are first described in section 2. Then, in section 3, a description of the algorithms used is given, which also include 5 non-PF algorithms, and we report some commented simulation results. The reader is welcome to reproduce some of these results using some code that is made available online. Finally, we conclude on the quality of particle filter-enhanced speech signals.

2. OBJECTIVE MEASURES OF SPEECH QUALITY

To assess the quality of speech sounds, although the systematic use of subjective tests on a large population is the ideal solution, unfortunately usually in practice the population available is limited, making the results highly variable and sometimes ambiguous. In addition, the whole process may take too long to be practical, especially for initial testing purposes. Therefore some mathematical, objective quality measures are necessary. Ideally, these measures should be as highly correlated as possible to a standard mean-opinion score (MOS) test[1], or to a subjective intelligibility measure. In other words, we would like the measures to be good predictors of average subjective preferences. We choose here to use a few different measures, each having advantages and disadvantages.

2.1. Overall SNR (OSNR)

The classic overall signal-to-noise ratio (OSNR) is simply the ratio between a (clean) signal's energy and the energy contained in the noise or error. The SNR is higher if the squared difference between the estimated speech and the original speech signal is smaller. It is considered that the SNR is not a reliable indicator of intelligibility and speech quality. Still, the overall SNR can be an interesting

This work was supported by a NSERC scholarship

[1] by *standard*, we mean "as defined by the Telecommunication Standardization Sector of the International Telecommunication Union" – abbreviated by ITU-T

"first glance" at the performance of an algorithm: a very low SNR is usually not a good sign! In addition, it is very inexpensive computationally, and thus it can be used as a simple online indicator that, for example, an algorithm is not on a divergent path.

2.2. Average segmental SNR (ASSNR)

The average segmental SNR (ASSNR) is based on the same principle as the overall SNR, except that it is computed as the average of the SNR of segments, or frames (possibly overlapping and windowed). In this paper, where we are focusing on speech sampled at 8 kHz, we use a Hanning window of width 30 milliseconds (240 samples), and 75% overlaps (that is, the window is anchored every 7.5 milliseconds or 60 samples). The ASSNR is a more insightful quality measure than the overall SNR, in the sense that it is a better MOS predictor [7, 8]. Based on the multiple experiments and informal listening tests conducted in this paper, we find that in general a higher ASSNR means less residual background noise. This finding is in accordance to a study presented in [9], in which objective measures for speech enhancement are evaluated. In [9], listeners are given an enhanced signal, and are asked to give three scores from 1 to 5. They must first rate the distortion on the speech signal itself ("SIG" score, in terms of naturalness), then they must rate the background noise (the "BAK" score, in terms of intrusiveness), and finally they must give an overall score (the "OVRL" score). Subsequently, the correlation between several objective measures (including the ASSNR) and these three scores is inferred. The ASSNR is found to be much more correlated to the "BAK" score than to the "SIG" score.

2.3. PESQ

The PESQ (Perceptual Evaluation of Speech Quality) algorithm [10, 11] is an objective method to predict the results of subjective MOS tests, designed purposely for handset telephony speech codecs. Although PESQ scores were not designed for speech enhancement algorithms evaluation, they are still found to provide a meaningful indication of performance and they are frequently used by researchers for this purpose.

The PESQ algorithm compares the original, clean speech signal to the output of the enhancement algorithm, and penalizes the final score based on measures of the distortion. The PESQ is perceptual in the sense that the amount of distortion is measured in the context of a model for the human auditory system. According to the ITU-T, PESQ scores demonstrate a good correlation with subjective test results [1].

2.4. Log Area Ratio (LAR)

The Log-Area Ratio (LAR) score is another objective speech quality measure; it is recommended in [12] for the evaluation of speech enhancement algorithms. As opposed to the previous three measures, the LAR measures a distance, and therefore it increases as distortion increases. The distance measured is based on the reflection coefficients of the corrupted (or enhanced) speech, $\{\rho_{\hat{x}}(m)\}_{m=1}^{M}$, and those of the clean speech, $\{\rho_x(m)\}_{m=1}^{M}$ (M is the order of the linear prediction analysis). For a given frame, it is computed as follows:

$$LAR_l = \left| \frac{1}{M} \sum_{m=1}^{M} \left(\log \frac{1+\rho_x(m)}{1-\rho_x(m)} - \log \frac{1+\rho_{\hat{x}}(m)}{1-\rho_{\hat{x}}(m)} \right)^2 \right|^{\frac{1}{2}} \quad (1)$$

This measure can be seen as a way to estimate efficiently the differences between the logarithms of the spectra of the clean and the corrupted speech signals, (i.e. the power spectra are directly related to the reflection coefficients). The LAR measure has been used for testing the performance of speech enhancement strategies (for example [13, 14]). The LAR is also a better indicator than the overall SNR, as it is better correlated with expected opinion scores [6]. In [6], it is observed that the LAR scores are more correlated to subjective listening test results than the Itakura-Saito measure, which is another popular spectral distortion measure based on the same linear prediction principle. In our experience, the LAR distance is much more focused on the signal's intelligibility and naturalness than on the background residual noise. This observation is in accordance with the results reported in [15], where several objective measures are assessed (including the PMF, the PSQM, the Itakura-Saito distance, the log-likelihood ratio, the ASSNR and the weighted spectral slope distance). [15] reports that the LAR has the highest correlation with speech naturalness, and even with overall subjective ratings.

3. ALGORITHMS USED, SIMULATION RESULTS AND ANALYSIS

3.1. Selected algorithms and experimental conditions

It is proposed to assess two different PF-based algorithms, which are given in Table 1 (along with other algorithms). The first one is a regular particle filtering algorithm (PF in the table), as described in [4]. We only noticed modest improvement when using an auxiliary PF, so we only used a simple PF. In addition, we use a Rao-Blackwellised particle filter (RBPF in the table), as described in [2, 3]. For both algorithms, we employed the very light technique of fixed-lag smoothing [2,16] (as opposed to more advanced, but heavy, smoothing techniques such as the one explained in [3], in order to avoid excessively long processing time).

Algorithm	Abbreviation	Reference
Regular PF	PF	[4]
Rao-Blackwellised PF	RBPF	[2, 3]
Spectral subraction	SSUB	[17]
Kalman fi lter + EM	KF+EM	[5]
MMSE short-time spectral amplitude	MMSE-STSA	[18]
Wiener Filter + a priori SNR estimator	WF+ASNRE	[19]
Dual EKF	DEKF	[20]

Table 1. Index of algorithms used

For comparisons, we also used several other speech enhancement algorithms: a basic spectral subtraction algorithm (SSUB), a Kalman filter-based algorithm using an EM algorithm to update the speech parameters (KF+EM, [5]), a method of denoising using a minimum mean-square error log-spectral amplitude estimator (MMSE-STSA, [18]), an algorithm based on a priori SNR estimation (WF+ASNRE, [19]), and the Dual extended Kalman filter (DEKF, [20]). The clean speech signal, a male voice uttering the sentence "*Primitive tribes have an upbeat attitude*", is available for download at the demonstration page:

`http://cslu.ece.ogi.edu/nsel/demos/`

[1] Even though this is generally the case, the PESQ cannot be blindly trusted. For example, in the following page: `http://microtronix.ca/pesq-disc.html`, audio examples are shown where two degraded signals obtain the same PESQ score, even though one of them is of signifi cantly lower quality.

In the following experiment, we consider that the noise variance is known in advance[1]. The PF and RBPF algorithms use Gaussian random walks on the speech parameters (the AR vector and the excitation noise log-variance), with variances according to those of [2]. The importance density used for the regular PF is the one described in Section 5.3 of [4]. The PF is set to use 2000 particles, and the RBPF 600 of them (using more particles was not found to have any significant average impact on the results). Accordingly, we simplify the KF+EM algorithm of [5] in such a way that the observation variance is not estimated, but known[2]. The EM algorithm estimating the parameters is set to iterate 20 times towards convergence, on frames of 128 samples. For the SSUB, the MMSE-STSA, and the WF+ASNRE algorithms, an estimate of the noise spectrum is computed from the first few frames, containing only noise (specifically, we use the first 1000 samples, or 0.125 seconds)[3]. For the DEKF algorithm, we directly use the two demonstrations for the AWGN subcase with known variance presented on the DEKF demonstration webpage above. We corrupt the clean signal with a computer-generated white Gaussian noise sequence to produce noisy signals with overall SNRs of approximately 0, 4, 7, and 10 dB. For 0 and 7 dB, we directly use the noisy signals available on the webpage abovementioned.

3.2. Simulation results

Type of algorithm	*Quality measure*	Input SNR (dB) 0	4	7	10
PF $L = 8$	OSNR	7.38	10.18	11.91	14.43
	ASSNR	-0.75	1.03	2.31	3.98
	PESQ	1.88	2.14	2.17	2.27
	LAR	6.15	5.68	5.14	4.71
RBPF $L = 8$	OSNR	7.50	10.32	12.36	14.60
	ASSNR	0.40	2.11	3.21	4.76
	PESQ	1.81	2.06	2.17	2.26
	LAR	5.39	5.26	4.72	4.36
SSUB	OSNR	6.60	9.00	11.38	13.12
	ASSNR	-2.38	-0.91	0.88	2.30
	PESQ	2.10	2.25	2.30	2.40
	LAR	6.27	5.79	5.34	4.71
KF+EM	OSNR	7.39	10.02	11.93	14.12
	ASSNR	-0.80	1.04	2.12	3.67
	PESQ	1.87	2.03	2.10	2.21
	LAR	5.53	5.17	5.02	4.47
MMSE-STSA	OSNR	4.44	5.78	6.49	8.01
	ASSNR	-1.96	-0.53	0.66	2.36
	PESQ	1.80	1.92	2.00	2.23
	LAR	6.40	6.18	6.00	5.79
WF+ASNRE	OSNR	3.71	5.25	5.96	7.70
	ASSNR	-0.49	0.80	1.79	3.40
	PESQ	1.57	1.65	1.97	2.16
	LAR	7.46	7.79	7.70	7.62
DEKF	OSNR	7.54	–	11.72	–
	ASSNR	0.60		2.72	
	PESQ	2.24		2.56	
	LAR	6.29		5.47	

Table 2. Comparison of PF-based algorithms to other enhancement schemes . *The input signal (a male voice) is corrupted with AWGN. The results obtained from the PF-based methods are an average over 10 experiments, and L denotes the value of the fixed lag.*

[1]Note that this is not a signifi cant advantage for PF-based algorithms, which can conveniently estimate this variance online as well. Not knowing its value may however constitute a strong penalty for other algorithms.

[2]An implementation for the RBPF and the KF+EM algorithms used can be found at `www.site.uottawa.ca/~mustiere/`

[3]An implementation for the MMSE-STSA and the WF+ASNRE algorithms used can be found at `http://dea.brunel.ac.uk/cmsp/Home_Esfandiar/`

The raw simulation results are presented in Table 2. For clarity, a rough summary of the observations made in section 2 to describe the objective measures now follows. The OSNR is merely an indicator that some enhancement (i.e. noise reduction) is taking place. A very high value may or may not indicate a very high speech quality, but a low value likely indicates a low speech quality. The ASSNR is mostly correlated with the background noise intrusiveness, or with interspeech residual noise. Next, the PESQ is well correlated with the overall speech quality. Finally, the LAR is mostly correlated with the speech naturalness.

From Table 2, we can now draw several conclusions. First, comparing the two PF-based algorithms with each other, we observe that the RBPF outperforms the regular PF for all measures used, except for the PESQ score, for which the PF is only slightly better for 0, 4, and 10 dB. Judging by the PESQ only, the PF and RBPF are therefore almost equivalent in terms of overall speech quality. However, the ASSNR measure indicates that the background noise is significantly less intrusive for the RBPF method. In addition, based on the LAR score, the RBPF-enhanced signal is more natural than the PF-enhanced signal. Informal listening clearly confirm these tendencies: there is less background noise in the output of the RBPF, and this noise is less annoying, resulting in a more natural overall signal. In [2], it was observed that the RBPF-enhanced signals contains a "time-varying" white noise. Listening to the residual noise only, and observing its waveform (see the two top graphs of Figure 1), we find that this residual noise can be further described as a white noise that is "modulated" by the speech, with an instantaneous power also depending on the input SNR. At the bottom of Figure 1, we show a segment of the absolute values of the residual error (locally time-averaged for improved visualization): the first half of the segment shown corresponds to a silence. The purpose of this graph is to support our claim that for any of the input SNR tested, the RBPF algorithm recognizes very well the absence of speech, i.e., the interspeech residual noise is brought to a very low value. In contrast, we find that the PF-enhanced signal is polluted with a "frying"-type noise, which remains more noticeable between utterances. In terms of intelligibility only, both the PF and the RBPF are found to be roughly equivalent.

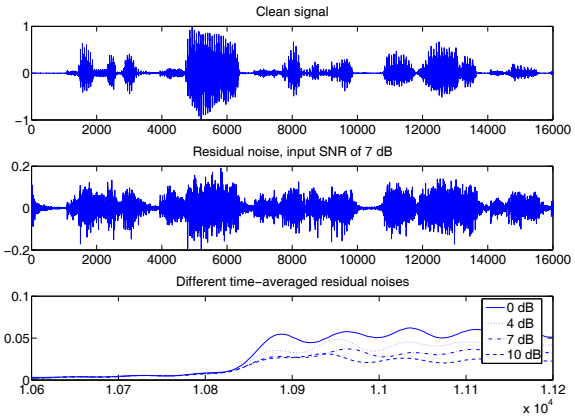

Fig. 1. Observation of the residual noise, RBPF algorithm. *The top figure shows the original speech, the middle one shows the residual noise in a RBPF-enhanced signal with 4 dB of input SNR. The bottom graph shows a smoothed segment of the residual noise in absolute value.*

From the comparison with other algorithms, we can see two important features of the RBPF-enhanced signals: in each of the conditions tested, they obtain the best overall SNR, ASSNR, and LAR scores (except in the 0 dB case, where the DEKF signal is given a better overall SNR – not significantly, and a better ASSNR). Considering PESQ scores, no algorithm is clearly standing out, except the basic spectral subtraction algorithm. The latter does obtain good PESQ scores, however the resulting speech is corrupted by very annoying and intrusive background musical noise, as indicated by a very poor ASSNR score. From informal listening, our observations are the following: in terms of listening comfort and naturalness, the RBPF produces the best signals overall. As in [4], we note a sibilance during unvoiced sounds in the resulting speech. We believe that it is in fact present within the entire enhanced speech segment, but only more perceived during these sounds. Among all other algorithms, the KF+EM is the one which outputs signals that are the closest to what can be expected from a RBPF, although with a quality that is overall inferior, as seen with the scores reported in the table, and as confirmed by informal listening tests.

According to further subjective listening tests, we find that among all the algorithms tested, the intelligibility is at best preserved, but not clearly enhanced – listening comfort is what constitutes the central difference between them. For example, although the DEKF algorithm is rated with a high PESQ score, the nature of the background noise penalizes our overall subjective perception of the speech quality.

4. CONCLUSION

From the experiments conducted, the two main conclusions are the following. First, we find that, expectedly, it is worth using a RBPF rather than a regular PF. In RBPF-enhanced signals, the background noise is less intrusive, and this noise is more "natural" (it can be identified to a white noise with power modulated by the output speech). Secondly, we find that in comparison to the other speech enhancement algorithms tested, the RBPF yields speech signals which are the most "comfortable" to listen to. Future research directions include modifications of PF-based algorithms to accomodate different types of noises, to reduce the impact of the residual noise, and to improve the intelligibility.

5. REFERENCES

[1] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: particle filters for tracking applications*, Artech House, London, February 2004.

[2] J. Vermaak, C. Andrieu, A. Doucet, and S.J. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 173–85, March 2002.

[3] W. Fong, S.J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing with application to audio signal enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 438–49, February 2002.

[4] S. J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing for nonlinear time series," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 156–168, March 2004.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[6] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[7] K. Hasan and L. Akter, "Quality improvement of enhanced speech in DCT domain using modified a priori SNR," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, Germany, 2004, pp. 733–6.

[8] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codes," in *Proceedings of ICASSP 82. IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, 1982, pp. 1000–3.

[9] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *International Conference on Spoken Language Processing, INTERSPEECH - Proceedings*, Pittsburg, PA, USA, September 2006.

[10] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. part i - time-delay compensation," *AES: Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, October 2002.

[11] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. part II - psychoacoustic model," *AES: Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, October 2002.

[12] J.H.L. Hansen and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *ICSLP, International Conference on Spoken Language Processing*, Sydney, Australia, December 1998, vol. 7, pp. 2819–2822.

[13] H. Gustafsson, S.E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, November 2001.

[14] I.A. McCowan, D.C. Moore, and S. Sridharan, "Near-field adaptive beamformer for robust speech recognition," *Digital Signal Processing*, vol. 12, no. 1, pp. 87–106, January 2002.

[15] M. Marzinzik, *Noise Reduction Schemes for Digital Hearing Aids and their Use for the Hearing Impaired*, Ph.D. thesis, Carl von Ossietzky University Oldenburg, 2000.

[16] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.

[17] T. Quatieri, *Discrete-time speech signal processing: principle and practice*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–5, April 1985.

[19] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, 1996, pp. 629–32.

[20] E. A. Wan and A. T. Nelson, "Removal of noise from speech using the dual EKF algorithm," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Seattler, WA, USA, May 1998, vol. 1, pp. 381–384.