

Objective Comparison of Speech Enhancement Algorithms under real world conditions

Stavros Ntalampiras
Department of Electrical
and Computer
Engineering,
University of Patras
26500, Rion, Patras,
Greece
+30 2610 969806
sntalampiras@upatras.gr

Todor Ganchev
Department of Electrical
and Computer
Engineering,
University of Patras
26500, Rion, Patras,
Greece
+30 2610 96 9808
tganchev@ieee.org

Ilyas Potamitis
Department of Music
Technology & Acoustics,
Technological Educational
Institute of Crete
74100, Rethymno, Crete,
Greece
+30 28310 21911
potamitis@stef.teicrete.gr

Nikos Fakotakis
Department of Electrical
and Computer
Engineering,
University of Patras
26500, Rion, Patras,
Greece
+30 2610 996 216
fakotaki@wcl.ee.upatras.gr

ABSTRACT

Over the past decades the problem of one channel, speech enhancement has been addressed by a great deal of researchers. In this work selected methods belonging to a variety of categories are applied to denoise speech signals corrupted by non-stationary urban noise. The performance of spectral subtraction, signal subspace, model-based and Kalman filtering approaches is evaluated. Several objective measures which are designed to predict human listening tests are employed in order to reach accurate conclusions. Two series of experiments were carried out while multiband spectral subtraction along with a short-time spectral amplitude (STSA) estimator based on the minimization of the mean square error (MSE) of the log-spectra are shown to outperform the rest of the algorithms.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: *Speech recognition and synthesis*

General Terms

Algorithms, Performance

Keywords

Speech Enhancement, Spectral Subtraction, Signal Subspace, Model-based Enhancement, Kalman Filtering

1. INTRODUCTION

The primary objective of noise compensation methods as applied in the context of speech processing is to reduce the effect of any signal that is alien to and disruptive of the message conveyed among participants in a communicative event (whether humans or ASR machines). Depending on the application, speech enhancement methods aim at speech quality improvement and/or signal preprocessing for speech or speaker recognition. The key

Permission to make digital or hard copies of part or all of this work or personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

PETRA'08, July 15-19, 2008, Athens, Greece.

Copyright 2008 ACM 978-1-60558-067-8... \$5.00

difference is that in the latter case, the complexity of the problem to be solved by the recognizer is relaxed by a pre-processing transformation from the time domain to a domain with more desirable properties as regards the recognition process. When speech quality and intelligibility is the issue, it is essential that we respect the specific idiosyncrasies of human speech hearing and, therefore, reconstruct the time-domain signal. In brief, a speech enhancement algorithm aims at one or more of the following goals:

- The improvement of speech quality by reduction of effort, fatigue and original message ambiguity.
- The reduction of noise-induced stress that could probably effect the articulation of speech at low SNRs – well-known as the Lombard effect.
- The elimination of speech-coding inconsistencies.
- Robust Automatic Speech/Speaker Recognition. Although ASR has come to a point that it enables the launch of commercial products, operational systems still face the problem of maintaining high recognition performance in adverse environments due to the mismatch between training and operational acoustical characteristics.

Due to the polymorphic manifestations and detrimental effect of noise, speech enhancement remains an open challenge. Comprehensive assessments of noise compensation methods that belong to different speech processing strategies can be found in [1].

After more than three decades of advances on the one-channel, speech enhancement problem, to our opinion, four distinct families of algorithms seem to have predominated in the literature, namely: a) the *spectral subtractive algorithms* [2], b) the *statistical model-based approaches* [3, 4, 5], c) the *signal subspace approaches* [6, 7] and d) the enhancement approaches based on a *special type of filtering* [8].

In this work, eight speech enhancement methods are evaluated on a real-world database recorded for the needs of speech recognition in motorcycle environment. This database consists of speech recordings coherent with the communication protocol of UK-police force, and especially with the motorcycle policing units. Thus, the performance of the speech enhancement algorithms under consideration is evaluated in conditions characterized with highly non-stationary noise. Specifically, we perform an objective

comparison by utilizing four distortion measures: SIG, BAK, OVL and PESQ [9].

The remaining of the present contribution is organized as follows: A brief description of the selected methodologies is provided in section 2. Section 3 offers a brief outline of the application scenario. The evaluation procedure followed in this study is outlined in Sections 4. Section 5 analyzes the distortion measures and the experimental set-up, respectively. Finally, Section 6 concludes this work.

2. ALGORITHMS

In our the present work, we consider eight speech enhancement methodologies belonging to the four categories specified in Section 1: spectral subtractive, statistical-model based, subspace and Kalman based filtering. All parameters of the algorithms are kept to the defaults, as originally published in the corresponding reference otherwise the changes will be mentioned explicitly.

The first algorithm in Tab. 1 is magnitude spectral subtraction [2], where noise is estimated during non-speech periods. It takes into account that colored noise varies over different frequency bands. Subsequently we tested the spectral subtraction with Martin’s [11] noise estimation algorithm, which derives an unbiased noise estimator based on the optimally smoothed power spectral density (psd) estimate as well as the analysis of spectral minima.

The next three speech enhancement methods belong to the statistical model based category. This category models the noise signal during a non-speech period while their main objective is to minimize a specific distortion measure in order to compute the clean signal. Ephraim and Malah in [3] proposed the minimum mean square error of the log-spectra resulting to a short-time spectral amplitude estimator. Next, we considered an algorithm which utilizes a Bayesian estimator of the short-time spectral amplitude of speech minimizing the weighted cosh distortion measure (a symmetric quantity derived by the two forms of the IS measure) [4]. The *musical noise* effect that many methods are introducing to the signal during the enhancement procedure is partially due to the large variance estimates of the spectra. A method trying to solve this issue and has been examined during our experiments is based on wavelet thresholding the multitaper spectrum [5].

Subspace approaches project the noisy signal onto two subspaces: signal plus noise subspace and the noise subspace using singular value decomposition (SVD) or the eigenvalue decomposition (EVD). Therefore, the clean signal can be obtained by nulling the components of the signal in the noise subspace, considering only the components in the signal subspace weighted accordingly to each methodology. In the present work, we consider two subspace algorithms: the first one has built-in pre-whitening, so it can be used in general for colored noise [6]. The second incorporates a human hearing model (MPEG-1 psychoacoustic model) to deal with the annoying effect of musical noise while it doesn’t employ prewhitening [7].

The last algorithm considered in this comparison is Kalman filter-based approach with the integration of the estimate-maximize (EM) method to iteratively estimate the spectral parameters of the speech signal. The clean signal is obtained as a byproduct of the parameter estimation algorithm [8].

No.	Algorithm	Ref
1	Multiband Spectral Subtraction	[2]
2	Spectral Subtraction with noise estimation	[10]
3	MMSE log spectral amplitude estimator	[3]
4	Bayesian estimator based on weighted cosh distortion measure	[4]
5	Wavelet thresholding the multitaper spectrum	[5]
6	Subspace algorithm with embedded pre-whitening	[6]
7	Perceptually motivated subspace	[7]
8	Kalman filter based	[8]

Table 1. Catalog of the Algorithms

Voice activity detection (VAD) is utilized by the majority of these algorithms for noise spectrum estimation. The subspace category uses a different approach as well as threshold value. In total, seven of the algorithms represent the aforementioned two groups, and one [10] utilizes a noise estimation-based methodology with spectral subtraction.

In section 5, the performance of these eight speech enhancement algorithms is evaluated on recordings made in conditions representing the specific real-world application of interest.

3. THE MOVEON SCENARIO

The MoveOn project aims at the creation of a multi-modal and multi-sensor, zero-distraction interface for motorcyclists. In the considered demonstration scenario, this interface provides the means for hands-free operation of a command and control system that enables for information support of police officers on the move. Due to the specifics of the MoveOn application, involving hands-busy and eyes-busy motorcyclists, speech is the dominating modality. The noisy motorcycle environment requires proper measures for addressing the speech enhancement task, in order to guarantee reliable speech recognition performance.

For the purpose of research and technology development a dedicated speech database was recorded in the motorcycle environment [11]. Specifically, a group of 30 professional motorcyclists, members of the operational police force of UK, was recruited. Each participant was asked to repeat a number of domain-specific commands and expressions, or to provide a spontaneous answer to questions related to time, current location, speed, etc. The prompt sheets (each one containing 302 prompts) were implemented as audio sequences that are played to the motorcyclists via earplug.

In total, the speech corpus consists of about 40 hours of recordings, obtained in 40 recording sessions by 29 police officers riding UK-police motorbikes. Different motorbikes and helmets were used, and the trace of road differed among sessions. Specifically, each session included in city driving, highway, tunnels, suburbs, etc. In addition, there are 5 recording sessions with the same hardware but in studio environment. Every session of the database consists of 4 channels recorded simultaneously: two from omni-directional microphones (AKG C 417”) placed within the helmet – 10 cm one from another – at the two sides of the mouth; one channel from a throat microphone (Alan AE 38), and finally one channel that mixes the first of the in helmet microphones with the audio prompts that were played to the speaker (see Fig. 1). This 4th channel served for synchronization

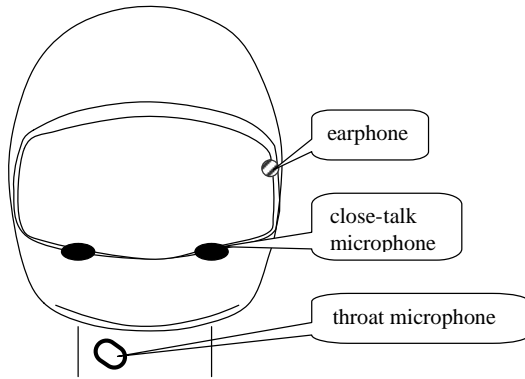


Figure 1. Recording setup: positions of the microphones and the earphone within the helmet, and the throat microphone.

purposes during annotation. The language of all recordings is British English spoken by native speakers.

All recordings were annotated in a multi-tier scheme. The annotations include different tiers for speech transcriptions, emotional tags, and various noise tags, such as: background noise, transient interferences (air-wind noise, engine noise, other noise and sound events). The transient noises are labeled by their position and estimated magnitude. One additional tier indicates when the helmet visor is open or closed, since this condition affects significantly the amount and the shaping of noise. The MoveON corpus offers a profile of the difficulties of speech communication on noisy motorcycle environment, as well as a test bed for the evaluation of various noise reduction techniques.

4. EVALUATION PROCEDURE

The evaluation procedure is organized in two setups:

Setup 1: Firstly, we studied the case when characteristics noise samples from the MoveOn database are added to the clean-speech sessions recorded in studio environment. This set-up provided controlled experimental conditions, which were utilized in preliminary study of the performance of the speech enhancement methods of interest.

Setup 2: In the second case, we relied entirely on the recordings

Rating	SIG	BAK	OVL
1	Very Unnatural, very degraded	Very conspicuous, very intrusive	Bad
2	Fairly unnatural, fairly degraded	Fairly conspicuous, somewhat intrusive	Poor
3	Somewhat natural, somewhat degraded	Noticeable but not intrusive	Fair
4	Fairly natural, little degradation	Somewhat noticeable	Good
5	Very natural, no degradation	Not noticeable	Excellent

Table 2. Explanation of the distortion measures

made in the motorcycle environment, while on the move. In this second scenario, the clean-speech reference signal is obtained through the throat microphone, which is less susceptible to the environment acoustic noise than the close-talking microphones. However, it is noteworthy mentioning that the speech signal captured by the throat microphone has reduced bandwidth, and thus, has a reduced intelligibility. Thus, one should keep in mind that there is no exact match between the spectral content provided by the throat microphone and the close-talking microphones.

The ITU-T recommendation P.385 was followed during the testing procedure. This strategy was organized in order to narrow listener’s uncertainty in a subjective test and make them evaluate three components of the signal separately, i.e., the speech signal, the background noise and both. Consequently it consists of three ratings for the enhanced sample which are the following:

- 1) the speech signal alone using a five-point scale of signal distortion (SIG),
- 2) the background noise alone using a five-point scale of background intrusiveness (BAK)
- 3) the overall effect using the scale of the Mean Opinion Score (OVRL)

The last distortion measure was Perceptual Evaluation of Speech Quality (PESQ - ITU-T Rec. P.862). This measure is designed to predict the results of subjective listening tests. It incorporates a psychoacoustic model to compare the original signal with the enhanced one ($-0.5 < PESQ < 4.5$).

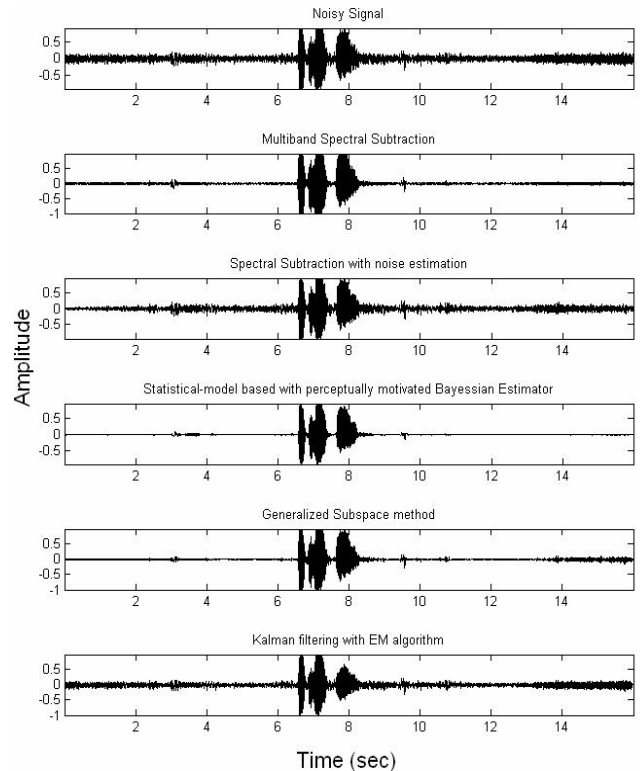


Figure 2. Top panel shows the noisy signal retrieved from the MoveON database. The remaining panels show the enhanced signals with respect to the algorithm used.

5. EXPERIMENTAL RESULTS

The eight speech enhancement methods were evaluated in common experimental setup following the evaluation procedure defined in Section 4.

Starting with Setup 1, in Figure 2, we depict the waveforms of the noisy signal in contrast to a characteristic output of each class. As it can be seen in the figure, the multiband spectral subtraction outperforms the other methods. The log MMSE is the second-best method.

The SIG, BACK, OVRL, PESQ distortion measures were computed for the enhanced speech signals produced by all algorithms. In Table 3, we summarize the results obtained from this experimental setup along with the noisy signal for reference purposes. As it can be seen from these results, the approach that distinguishes the spectrum of the noisy signal into bands and executes magnitude subtraction [2] demonstrates the best performance considering all the distortion measures. Real world noise does not affect in a spectrally balanced way the speech signal. Therefore, this approach refines each frequency portion utilizing a band-specific noise function. Spectral subtraction with

<i>Algorithm</i>	<i>SIG</i>	<i>BAK</i>	<i>OVRL</i>	<i>PESQ</i>
<i>Noisy signal</i>	2.0	1.9	2.4	3.0
Multiband Spectral Subtraction	1.9	2.2	2.3	3.0
Spectral Subtraction with noise estimation	1.7	2.0	2.2	2.9
MMSE log spectral amplitude estimator	1.5	2.6	2.0	2.9
Bayesian estimator based on weighted cosh distortion measure	1.0	2.1	1.3	2.8
Wavelet thresholding the multitaper spectrum	1.0	1.9	1.0	2.5
Subspace algorithm with embedded pre-whitening	1.0	1.85	1.2	2.5
Perceptually motivated subspace	1.6	1.9	1.8	2.4
Kalman filter based	2.0	1.7	2.0	2.6

Table 3. Experimental results for Setup 1

<i>Algorithm</i>	<i>SIG</i>	<i>BAK</i>	<i>OVRL</i>	<i>PESQ</i>
<i>Noisy signal</i>	2.6	1.8	2.3	2.2
Multiband Spectral Subtraction	2.8	2.0	2.6	2.4
Spectral Subtraction with noise estimation	2.8	1.8	2.4	2.2
MMSE log spectral amplitude estimator	3.0	2.1	2.6	2.5
Bayesian estimator based on weighted cosh distortion measure	2.7	2.0	2.4	2.5
Wavelet thresholding the multitaper spectrum	2.5	1.9	2.2	2.1
Subspace algorithm with embedded pre-whitening	2.2	2.0	2.2	2.4
Perceptually motivated subspace	2.4	1.9	2.2	2.2
Kalman filter based	1.8	1.9	2.0	2.3

Table 4. Experimental results for Setup 2

noise estimation [10] demonstrated a slightly worse performance than multiband subtraction. This is due to the absence of VAD while it depends on tracking the spectral minima in each frequency band to approximate the noise.

Of the two subspace algorithms, the generalized approach [6] showed the worst outcome according to all distortion measures but PESQ. This degradation underlines the usage of MPEG-1 psychoacoustic model and stresses the importance of including such a masking procedure. Of the three statistical-model based methods, the log MMSE performed the best followed by Bayesian estimation based on weighted cosh distortion measure. Despite its simplicity, minimizing the mean square error of the log-spectra provides better speech quality than minimizing cosh distortion measure. Nonetheless both outperformed the STSA estimator based on wavelet-thresholding the multitaper spectra which was unable to achieve adequate speech quality. Finally, the evaluation of Kalman-filter based methodology with EM algorithm indicated fair results while it has the most computational needs.

In Table 4, we present the experimental results for Setup 2. Division of the frequency spectrum into different bands in combination with the application of a different kind of spectral subtraction in each one displayed again very good performance, while the MMSE approach provided the best results across all distortion measures. As we can see the outcomes of the second setup are similar to the ones of the first, having the perceptually motivated subspace approach to outperform the generalized method. Furthermore the algorithm which minimizes the weighted cosh distortion measure demonstrated better performance than the third statistical model-based methodology.

6. CONCLUSIONS

In this work, we investigated the performance of eight speech enhancement techniques, which represent all four major categories of speech enhancement algorithms, in a real-world scenario. The database used in the experimentations is representative for the police communication protocol used by motorcycle policing units in UK. The main challenge for the speech enhancement algorithms under investigation comes from the non-stationary noise of the real world environment, and the significant variation in the environmental conditions inherent for the application. In the comparative evaluation, we relied on four subjective distortion measures. In our experimentations, the multiband spectral subtraction approach was found to offer the best performance, providing the speech with the highest perceived quality.

7. ACKNOWLEDGEMENT

This work was supported by the MoveOn project (IST-2005-034753).

8. REFERENCES

- [1] Yariv Ephraim, Hanoach Lev-Ari and William J.J. Roberts. A brief survey of speech enhancement. The Electronic Handbook, CRC Press, April 2005.
- [2] Kamath, S., Loizou, P. 2002. A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise. Proceedings of ICASSP-2002, Orlando, FL, May 2002.

- [3] Ephraim, Y., Malah, D. 1985. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. ASSP-33, pp. 443-445, Apr. 1985.
- [4] Loizou, P. 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857-869, Sept. 2005.
- [5] Hu, Y., Loizou, P. 2004. Speech enhancement by wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59-67, Jan. 2004.
- [6] Hu, Y., Loizou, P. 2003. A generalized subspace approach for enhancing speech corrupted with colored noise. *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334-341, July 2003.
- [7] Jabloun, F., Champagne, B. 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700-708, Nov 2003.
- [8] Gannot, S., Burshtein, D., Weinstein, E. 1998. 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.
- [9] ITU, ITU-T Rec. P. 862. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, 2000.
- [10] Martin, R. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.
- [11] Winkler, T., Kostoulas, T., Adderley, R., Bonkowski, C., Ganchev, T. Köhler, J., Fakotakis, N. 2008. The MoveOn Motorcycle Speech Corpus. Submitted to LREC 2008.
- [12] www.m0ve0n.net, Multi-modal and multi-sensor zero-distraction interface for two wheeled vehicles ON the move.