International Journal of Speech Technology (2021) 24:687–699
https://doi.org/10.1007/s10772-021-09829-9

An improved MMSE estimator based modified group delay spectrum for Forensic Automatic Speaker Recognition

Salim Djeghiour[1] · **Mhania Guerti**[1]

Received: 31 March 2020 / Accepted: 1 March 2021 / Published online: 15 March 2021

Abstract
This paper presents an improved speech enhancement algorithm called Minimum Mean Square Error (MMSE), based on the MODified Group Delay spectrum (MODGD), for Forensic Automatic Speaker Recognition (FASR) under noisy environments. This algorithm uses the MODGD instead of the amplitude spectrum, to compute the power spectrum of the noise-corrupt signal. In the proposed estimator, the MODGD retains most of the formants information. Therefore, it enhances the noisy speech signal with high quality even at extremely low Signal-to-Noise Ratio (SNR) levels. The evaluation of the improved algorithm in simulated FASR scenarios was performed by adding different noise levels, extracted from the NOISEX-92 database to the clean NIST2000-traces. The results obtained show that the proposed MMSE–MODGD estimator provides greater suppression of noise components in regions of low SNR than the MMSE estimator. In addition, there is a drastic reduction in Equal Proportion Probability (EPP) (the improvements are 1.84% for babble noise and 1.25% for factory and white noises), combining FASR techniques with the proposed MMSE–MODGD estimator than with the conventional estimator.

Keywords Forensic Automatic Speaker Recognition · MMSE estimator · MODGD spectrum · Noisy environments · Equal proportion probability

1 Introduction

In some criminal cases, the voice recorded during a telephone call is the only clue available to investigators. There is therefore a very pressing and fully justified demand from the judicial police and magistrates, to use these recordings to guide the investigation, and to establish the guilt of a suspect or prove his/her innocence. Hence, speaker recognition techniques provide a valuable contribution to the Forensic Speaker Recognition System.

To this end, forensic speaker recognition system is considered as one of the disciplines of Speaker Recognition (SR) for both identification and verification. Although several forensic applications for SR have been developed in recent years, they have not been successful due to the high complexity and variability of the speech signal, and the mismatch between modeling and testing conditions especially in real life (Deshpande & Holambe, 2011). The latest can be caused by various sources, such as reverberation, compressed audio, degraded channels, and environmental noise that degrade the performance of the forensic system (Scheffer et al., 2013). Thus, the challenging task for forensic experts is to find effective algorithms for speech enhancement in highly degraded environments, such as additive noise (Zhang & Abdulla, 2007).

Several speech enhancement algorithms, which are based on the magnitude spectrum of the speech signal, have been developed to overcome this challenge, namely: Spectral Subtraction Method (SS) (Gustafsson et al., 2004), Spectral Subtraction with Over subtraction Model (SSOM) (Dixit & Mulge, 2014), Non-Linear Spectral Subtraction (NSS) (Verschuur et al., 2006), Adaptive Noise Cancellation (ANC) (Kwatra et al., 2017) and the Minimum Mean Square Error (MMSE) estimators (Lu & Loizou, 2011).

This study proposes a modification of the MMSE estimators, by replacing the magnitude spectrum estimated using a Fourier Transform (FT), by the MODGD spectrum (Asbai &

✉ Salim Djeghiour
d.salim3@yahoo.fr

Mhania Guerti
mhania.guerti@g.enp.edu.dz

[1] Signal and Communications Laboratory, Ecole Nationale Polytechnique, Algiers, Algeria

Amrouche, 2017; Parthasarathi et al., 2011). In other words, the independent Gaussian random variables are derived from the MODGD spectrum, instead of their direct estimation from the Discrete Fourier Transform (Parthasarathi et al., 2011), to improve the MMSE algorithm by exploiting the information contained in the phase spectrums.

The proposed modification is motivated by two considerations.

- In general, the speech signal is a mixed phase signal, because a speaker's vocal tract is a minimum phase system (Akande & Murphy, 2005), and for minimum phase systems, information can be extracted from the phase or magnitude spectrum. Thus, in terms of analysis, the group delay of a mixed phase signal is the sum of the group delay of its minimum phase components (Hegde et al., 2004). In this study, the MODGD spectrum is thus processed by computing the mean of the posteriori density given in Lu and Loizou (2011), to exploit the properties of MODGDs (high resolution formants) on the MMSE method;
- Furthermore, Parthasarathi et al. (2011) indicated that, the group delay spectrum retains most of the formants information even at low SNR_s of environmental noise. The MODGD spectrum is less affected by noise than the magnitude spectrum.

The contribution of this work is threefold; first, the MMSE estimators based on the MODGD are better adapted to the noisy speech segments of the tests in many applications of speaker recognition systems. Then, the exploitation of the information contained in the phase as well as in the amplitude spectrum can be noted for the proposed MMSE-MODGD. Finally, extensive testing and experimental validation of the proposed MMSE were carried out.

2 Forensic Automatic Speaker Recognition (FASR)

In the FASR systems, the use of scientific tools is necessary to meet the needs of a court for a crime or civil litigation (Roux et al., 2012). The main fields used in forensic science are: biology, chemistry, and medicine (Forest et al., 1983). Despite the predominance of the latter, other disciplines used such as: physics, computer science, geology, and psychology (Forest et al., 1983). For example, traditional biometric parameters, such as DNA and fingerprints, are often used in many forensic cases. The nature of the evidence, whether found at the crime scene or collected during investigations dictates the scientific methods or disciplines needed to study it. In the context of the FASR, experts are interested in methods of identifying a recorded voice. This is based on the fact that each person can be identified from a sample of his/her voice. In addition, a suspect can leave recordings of his/her voice on the phone, voicemail, an answering machine or a hidden recorder, which can then be used as evidence. Three databases are generally required to establish a FASR system: Potential population database (P), suspected speaker Reference database (R) and suspected speaker Control database (C). They allow calculating and evaluating the evidence from the questioned recording (trace) (Drygajlo et al., 2003; Kenai et al., 2019).

There is also another methodology adopted in FASR systems, which requires a statistical model capable of computing a likelihood value, when feature vectors are compared against such a model. This method uses only two databases: the suspected speaker Reference database and the relevant Population (Drygajlo, 2012). These two databases can be used to create two statistical models: (1) statistical model of the suspected speaker and (2) statistical model of the relevant population. The Universal Background Model (UBM) (Kenai et al., 2019), trained with the relevant population database, can also be used as model of the statistical model of the relevant population (Drygajlo, 2012). The multivariate evidence represented by the ensemble of feature vectors extracted from the questioned recording is compared to model of the statistical model of the suspected speaker and statistical model of the relevant population to calculate the likelihood ratio. The first comparison gives the similarity likelihood score (numerator of LR) and the second one gives the typicality likelihood score (denominator of LR) (Drygajlo et al., 2003; Drygajlo, 2012). Figure 1 shows the principle of this methodological approach.

However, in real forensic scenarios, the speech signal left by the suspects (trace) is often corrupted by the environmental noise, which degrades the performances of the FASR system (Alexander et al., 2004). To this end, this paper discusses the MMSE-MODGD estimator used in speech enhancement (Gerkmann & Hendriks, 2012) to improve the FASR system under snoisy environments (Figs. 2, 3, 4).

3 Minimum mean square error estimator of the noisy short-time power spectrum

The spectral subtraction method (Berouti et al., 1979) based on MMSE (Gerkmann & Hendriks, 2012) and minimum noise statistics (MS) (Martin, 2001) was used to enhance the speech signal damaged by the additive noise. The amplitude of the noisy signal was multiplied with a certain gain factor. Spectral subtraction introduced by Boll (1979), is the oldest method to remove the noise. It operates in the frequency domain, and its principle is to subtract a noise estimate from the observed signal. Noise is assumed to be additive, stationary or slightly

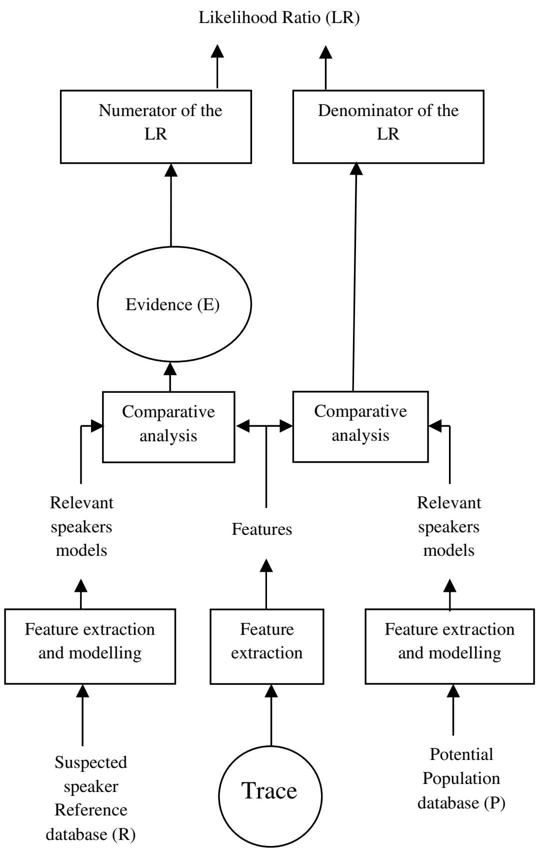

Fig. 1 The principle of the FASR methodological approach

varying, which allows to estimate it during silence periods. The noisy signal $y(n)$ can be written as (Lu & Loizou, 2011):

$$y(n) = x(n) + d(n) \tag{1}$$

where $x(n)$ and $d(n)$ represent the clean speech and noise signals, respectively.

Taking the short-time Fourier transform of $y(n)$, we obtain:

$$Y\left(w_k\right) = X\left(w_k\right) + D\left(w_k\right) \tag{2}$$

Equation (2) can be expressed in a polar form as follows:

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{3}$$

where, $\left\{Y_k, X_k, D_k\right\}$ denotes the magnitudes and $\left\{\theta_y(k), \theta_x(k), \theta_d(k)\right\}$ denotes the phases at frequency bin k of the noisy speech, clean speech and noise, respectively.

The MMSE estimator of the short-time power spectrum (MMSE) is given by (Wolfe & Godsill, 2003) as follows:

$$\begin{aligned}\hat{X}_k^2 &= E\left\{X_k^2/Y\left(w_k\right)\right\} \\ &= \int_0^\infty X_k^2 f_{X_k}(X_k/Y(w_k))dX_k \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1+\xi_k}\right)Y_k^2\end{aligned} \tag{4}$$

and,

$$\xi_k \equiv \frac{\sigma_x^2(k)}{\sigma_d^2(k)}, \gamma_k \equiv \frac{Y_k^2}{\sigma_d^2(k)} \tag{5}$$

$$\sigma_x^2(k) \equiv E\left\{X_k^2\right\}, \sigma_d^2(k) \equiv E\left\{D_k^2\right\} \tag{6}$$

where, ξ_k and γ_k denote the a priori and a posteriori SNRs, respectively.

The derivations of the above MMSE estimator were based on the following Rician posterior density $f_{X_k}\left(X_k/Y\left(w_k\right)\right)$:

$$f_{X_k}\left(X_k/Y\left(w_k\right)\right) = \frac{X_k}{\sigma_k^2}\exp\left(-\frac{X_k^2 + s_k^2}{2\sigma_k^2}\right)I_0\left(\frac{X_k s_k}{\sigma_k^2}\right) \tag{7}$$

where,

$$\frac{1}{\lambda'(k)} \equiv \frac{1}{\sigma_x^2(k)} + \frac{1}{\sigma_d^2(k)} \tag{8}$$

$$\upsilon_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k \tag{9}$$

$$\sigma_k^2 \equiv \frac{\lambda'(k)}{2}, s_k^2 \equiv \upsilon_k \lambda'(k) \tag{10}$$

$I_0(.)$ is the first kind modified Bessel function of zeroth order.

However, the analysis of the suppression curves revealed that the MMSE spectral power suppression rule of Eq. (4) provides less suppression in regions of low a priori SNR (Wolfe & Godsill, 2003). Lu and Loizou 2011) proposed the improved MMSE estimator of the short-time power-spectrum, to remedy the problem of less suppression in regions of low a priori SNR.

The power spectrum of the noise-corrupt signal is assumed to be the sum of the power spectra of the clean speech and noise, written as follows:

$$P_y(w) = P_x(w) + P_d(w) \tag{11}$$

In addition, an assumption is used in the derivation of these estimators based on Eq. (11) by approximating the power spectrum using the magnitude squared spectrum, which is the sample estimate of the ensemble average. Therefore, Eq. (11) can be written as follows:

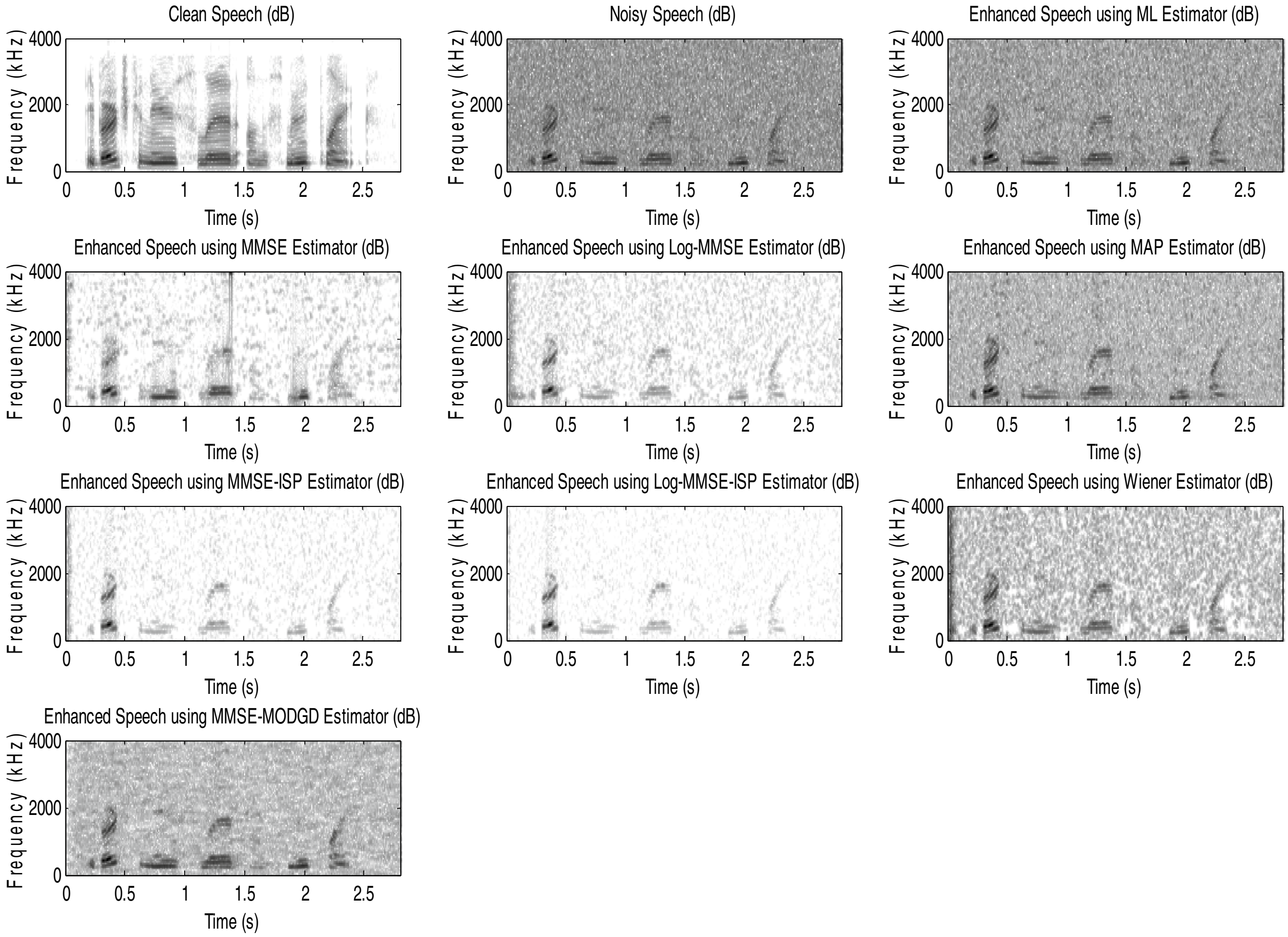

Fig. 2 Spectrograms of clean speech, noisy speech corrupted with white noise at 0 dB input SNR and speech enhancement methods

$$Y_k^2 \approx X_k^2 + D_k^2 \tag{12}$$

Moreover, assuming that the real and imaginary parts of the Discrete Fourier Transform (DFT) coefficients are modeled as independent Gaussian random variables with equal variance (Ephraim & Malah, 1984), the probability density of X_k^2 is exponential and can be written as follows:

$$f_{X_k^2}\left(X_k^2\right) = \frac{1}{\sigma_x^2(k)} e^{-\frac{X_k^2}{\sigma_x^2(k)}} \tag{13}$$

Similarly, the density of D_k^2 is given by Eq. (14):

$$f_{D_k^2}\left(D_k^2\right) = \frac{1}{\sigma_d^2(k)} e^{-\frac{D_k^2}{\sigma_d^2(k)}} \tag{14}$$

where, $\sigma_x^2(k)$ and $\sigma_d^2(k)$ are given by Eq. (6).

The posterior probability density of the clean speech magnitude-squared spectrum is obtained using the Bayes' rule as follows:

$$f_{X_k^2}\left(X_k^2/Y_k^2\right) = \frac{f_{Y_k^2}\left(Y_k^2/X_k^2\right) f_{X_k^2}\left(X_k^2\right)}{f_{Y_k^2}\left(Y_k^2\right)} = \left\{ \begin{array}{ll} \psi_k e^{-\frac{X_k^2}{\lambda(k)}}, & if\ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{Y_k^2}, & if\ \sigma_x^2(k) = \sigma_d^2(k) \end{array} \right\} \tag{15}$$

$\lambda(k)$ is defined as:

$$\frac{1}{\lambda(k)} \equiv \frac{1}{\sigma_x^2(k)} - \frac{1}{\sigma_d^2(k)}, \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \tag{16}$$

and

$$\psi_k \equiv \frac{1}{\lambda(k)\left\{1 - exp\left[-\frac{Y_k^2}{\lambda(k)}\right]\right\}}. \tag{17}$$

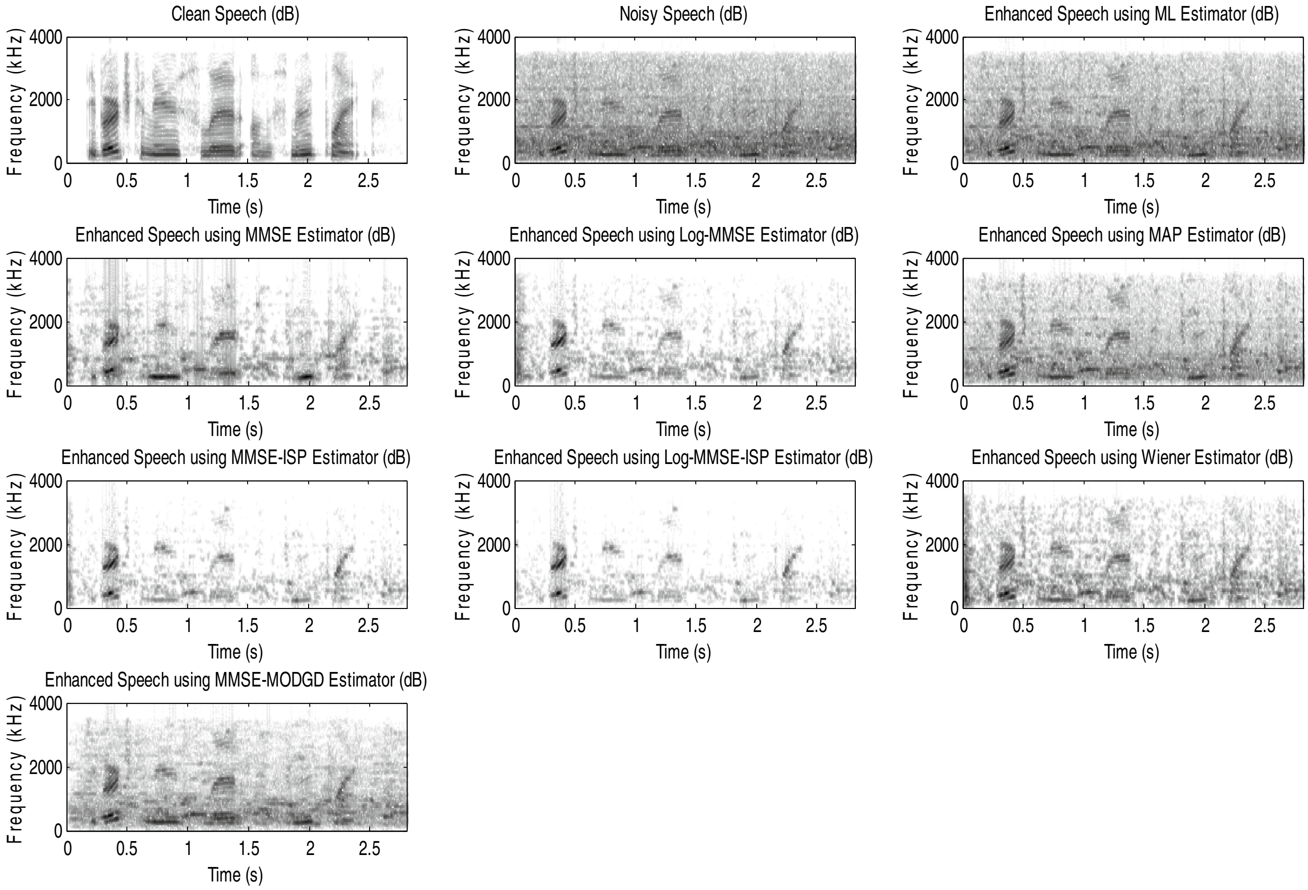

Fig. 3 Spectrograms of clean speech, noisy speech corrupted with factory noise at 0 dB input SNR and speech enhancement methods

Using Eqs. (12)–(15), the MMSE estimator is obtained by computing the mean of the posteriori density given in Eq. (15) as follows:

$$\begin{aligned}\hat{X}_k^2 &= E\{X_k^2/Y_k^2\} \\ &= \int_0^{Y_k^2} X_k^2 f_{X_k^2}\left(X_k^2/Y_k^2\right) dX_k^2 \\ &= \begin{cases} \left(\frac{1}{\upsilon_k} - \frac{1}{e^{\upsilon_k}-1}\right) Y_k^2, & if\ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} Y_k^2, & if\ \sigma_x^2(k) = \sigma_d^2(k) \end{cases}\end{aligned} \tag{18}$$

where, υ_k is defined as:

$$\upsilon_k \equiv \frac{1-\xi_k}{\xi_k}\gamma_k \tag{19}$$

4 The proposed modified group delay functions for the MMSE estimator of the noisy short-time power spectrum

A speech signal can be represented completely in the spectral domain only if the amplitude and phase information is specified. However, the information extracted from the phase spectrum is more complex than the information extracted from the amplitude spectrum, as the phase spectrum is generally discontinuous (orwrapped) between $[-\pi, \pi]$ (Murthy & Yegnanarayana, 2011). A multi-valued function is used to make it into a continuous function; this is called the unwrapped phase (unwrapping) (Parthasarathi et al., 2011). The processing of its derivative (i.e., the phase derivative), the "group delay function" (Parthasarathi et al., 2011), is mainly used to extract the information contained in the phase spectrum.

Let $x(n)$ a speech signal, its Fourier transform is given by Eq. (3).

The group delay function $\tau(\omega)$ of a signal $x(n)$ is defined as the negative derivative of the phase spectrum $\theta(\omega)$ as follow:

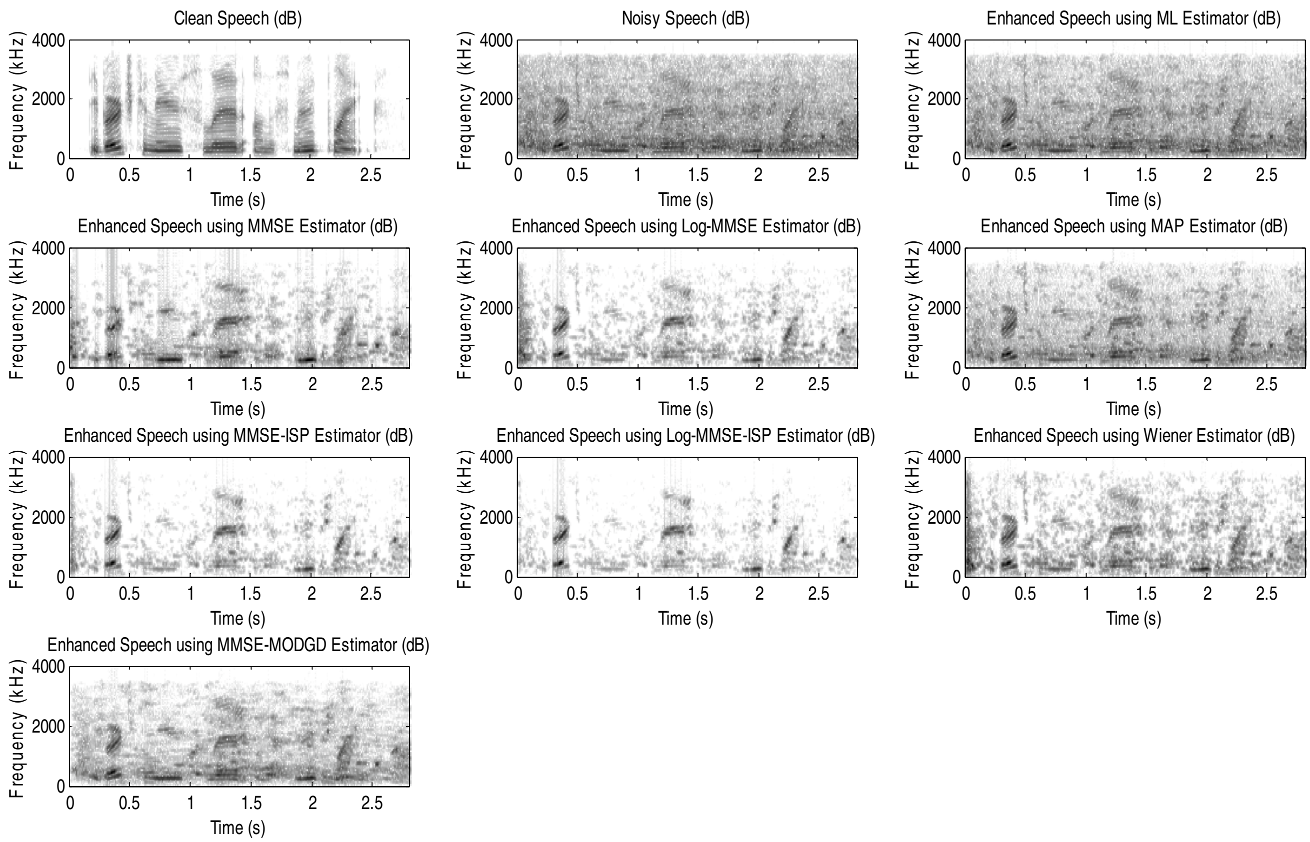

Fig. 4 Spectrograms of clean speech, noisy speech corrupted with babble noise at 0 dB input SNR and speech enhancement methods

$$\tau_X(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{20}$$

The group delay function can also be estimated from the speech signal using Eq. (21) (Asbai & Amrouche, 2017):

$$\tau_X(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_1(\omega)\hat{X}_1(\omega)}{|X(\omega)|^2} \tag{21}$$

where, R and I denote the real part and imaginary part respectively, $x(n) \leftrightarrow X(\omega)$ and $\widehat{x}(n) \leftrightarrow \widehat{X}(\omega)$ are Fourier Transform pairs, and $\widehat{x}(n) = nx(n)$.

The group delay function requires that the speech signal must be a minimum phase or that the poles of the transfer function be within the unit circle (Asbai & Amrouche, 2017).

By smoothing the amplitude $X(\omega)$ (Asbai & Amrouche, 2017) spectrum in Eq. (21), we define a MODified Group Delay function (MODGD) which given as follows:

$$\tau_X(\omega) = \left(\frac{\tau_s(\omega)}{\left|\tau_s(\omega)\right|}\right)\left(\left|\tau_s(\omega)\right|^{\alpha}\right) \tag{22}$$

where,

$$\tau_X(\omega) = \frac{X_R(\omega)\hat{X}_R(\omega) + X_1(\omega)\hat{X}_1(\omega)}{|S(\omega)|^{2Y}} \tag{23}$$

and $|S(\omega)|$ is a smoothed version of $|X(\omega)|$; the parameters α and γ are introduced to control the dynamic range. The length of the cepstral smoothing window is controlled by the parameter lifter_{ω}.

Therefore, based on Eqs. (22) and (23), Eq. (4) can be written as follows:

$$\begin{aligned} \hat{X}_k^2 &= E\left\{\tau_{Xk}^2/\tau_Y\left(w_k\right)\right\} \\ &= \int_0^{\infty} \tau_{Xk}^2 f_{\tau_{Xk}}\left(\tau_{Xk}/\tau_{Yk}\left(w_k\right)\right) d\tau_{Xk} \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1+\xi_k}\right)\tau_{Yk}^2 \end{aligned} \tag{24}$$

where,

$$\xi_k \equiv \frac{\sigma_x^2(k)}{\sigma_d^2(k)}, \gamma_k \equiv \frac{\tau_{Yk}^2}{\sigma_d^2(k)} \tag{25}$$

$$\sigma_x^2(k) \equiv E\left\{\tau_{Xk}^2\right\}, \sigma_d^2(k) \equiv E\left\{\tau_{Dk}^2\right\} \tag{26}$$

Finally, the Rician posterior density $f_{X_k}\left(X_k/Y\left(w_k\right)\right)$ becomes:

$$f_{\tau_{Xk}}\left(\tau_{Xk}/\tau_{YK}\left(w_k\right)\right) = \frac{\tau_{Xk}}{\sigma_k^2} \exp\left(-\frac{\tau_{Xk}^2 + s_k^2}{2\sigma_k^2}\right) I_0\left(\frac{\tau_{Xk} s_k}{\sigma_k^2}\right) \tag{27}$$

Moreover, Eqs. (12), (13) and (14) can be written as follows:

$$\tau_{Yk}^2 \approx \tau_{Xk}^2 + \tau_{Dk}^2 \tag{28}$$

$$f_{\tau_{Xk}^2}\left(\tau_{Xk}^2\right) = \frac{1}{\sigma_x^2(k)} e^{-\frac{\tau_{Xk}^2}{\sigma_x^2(k)}} \tag{29}$$

$$f_{\tau_{Dk}^2}\left(\tau_{Dk}^2\right) = \frac{1}{\sigma_d^2(k)} e^{-\frac{\tau_{Dk}^2}{\sigma_d^2(k)}} \tag{30}$$

where, $\sigma_x^2(k)$ and $\sigma_d^2(k)$ are given by Eq. (26).

The posterior probability density of the clean speech magnitude-squared spectrum become as follows:

$$\begin{aligned} f_{\tau_{Xk}^2}\left(\tau_{Xk}^2/\tau_{Yk}^2\right) &= \frac{f_{\tau_{Yk}^2}\left(\tau_{Yk}^2/\tau_{Xk}^2\right) f_{\tau_{Xk}^2}\left(\tau_{Xk}^2\right)}{f_{\tau_{Yk}^2}\left(\tau_{Yk}^2\right)} \\ &= \begin{cases} \psi_k e^{-\frac{\tau_{Xk}^2}{\lambda(k)}}, & if\ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{\tau_{Yk}^2}, & if\ \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \end{aligned} \tag{31}$$

where, $\lambda(k)$ is given by Eqs. (16) and (26).and

$$\psi_k \equiv \frac{1}{\lambda(k)\left\{1 - exp\left[-\frac{\tau_{Yk}^2}{\lambda(k)}\right]\right\}} \tag{32}$$

Finally, the modified MMSE estimator is given by:

$$\begin{aligned} \hat{X}_k^2 &= E\left\{\tau_{Xk}^2/\tau_{Yk}^2\right\} \\ &= \int_0^{\tau_{Yk}^2} \tau_{Xk}^2 f_{\tau_{Xk}^2}\left(\tau_{Xk}^2/\tau_{Yk}^2\right) d\tau_{Xk}^2 \\ &= \begin{cases} \left(\frac{1}{\upsilon_k} - \frac{1}{e^{\upsilon_k}-1}\right)\tau_{Yk}^2, & if\ \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2}\tau_{Yk}^2, & if\ \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \end{aligned} \tag{33}$$

υ_k is given by Eq. (19), where, γ_k is given by Eq. (25).

5 Experimental protocol for speech enhancement

Extensive objective quality tests were carried out to evaluate the performance of the proposed MMSE-MODGD estimation method using ten (10) sentences extracted from the NOIZEUS database (Hu & Loizou). In this database, the noise signals are generated by adding the noise from the AURORA and NOISEX-92 databases to the clean signals, to an overall SNR of 0 dB, 5 dB and 10 dB. The frame size chosen is 20 ms with a 50% overlap. A sampling frequency of 8 kHz and a Hamming window were used. The methods used for comparison with the proposed MMSE-MODGD are the maximum-likelihood (ML) estimator, the MMSE estimator, the log MMSE estimator, the maximum a posteriori (MAP) estimator, incorporating speech presence probability in MMSE (MMSE-ISP) estimator, incorporating speech presence probability in log MMSE (log MMSE-ISP) estimator and Wiener estimator (Loizou, 2007). The objective assessment was carried out as proposed in Hu and Loizou (2008). The tests carried out to evaluate the proposed method include measures related to the perception of the speech signal on a five-point (1–5) scale of signal distortion (SIG), background noise on a five-point (1–5) scale (BAK) and overall quality (OVRL) based on the Mean Opinion Score (MOS) ranging from 1 to 5. The other measures used are segmental SNR (SegSNR), weighted-slope spectral (WSS), perceptual evaluation of speech quality (PESQ) and log-likelihood ratio (LLR) (Hu & Loizou, 2008).

5.1 Results and discussion

Based on a comparative study using spectrograms, it can be noticed that the proposed MMSE-MODGD method gives good results compared to ML, MMSE, log MMSE, MAP, MMSE-ISP, log MMSE-ISP and Wiener. This good performance achieved by the proposed approach is confirmed by the objective evaluation.

Tables 1, 2 and 3 show the results of the evaluations using the objective measures: SIG, BAK, OVRL, PESQ, SegSNR, WSS and LLR using 10 sentences extracted from the NOIZEUS database. The proposed MMSE-MODGD method is compared with ML, MMSE, log MMSE, MAP, MMSE-ISP, log MMSE-ISP and Wiener, in the context of degradation by a white, factory and babble noises, respectively. The LLR and WSS scores indicate speech loss and should therefore be minimal. The results presented in the tables clearly show that the SIG, BAK and OVRL scores, which reflect the level of perception of the speech signal and the overall quality, are generally higher

Table 1 Objective evaluations of the MMSE-MODGD technique compared with ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP and Wiener and corrupted with white noise

Objective measures	Input SNR dB	White noise							
		ML	MMSE	Log-MMSE	MAP	MMSE-ISP	Log-MMSE-ISP	Wiener	MMSE-MODGD
SIG [1 to5]	0	1.31	1.29	1.09	1.19	1.19	0.91	0.93	**1.36**
	5	1.83	1.90	1.61	1.68	1.68	1.27	1.30	**1.95**
	10	2.43	**2.46**	2.12	2.19	2.20	1.74	1.84	2.38
BAK [1 to 5]	0	1.60	1.59	1.55	1.62	1.64	1.57	1.58	**1.71**
	5	1.94	1.99	1.84	1.85	1.86	1.70	1.72	**2.03**
	10	2.33	2.36	2.17	2.16	2.17	1.98	2.01	**2.39**
OVRL [1 to 5]	0	**1.30**	1.29	1.16	1.24	1.25	1.06	1.08	**1.30**
	5	1.72	1.75	1.59	1.63	1.62	1.29	1.33	**1.78**
	10	2.22	2.29	2.09	2.06	2.05	1.72	1.80	**2.32**
PESQ	0	1.54	1.70	1.57	1.68	1.62	1.63	1.68	**1.72**
	5	1.81	2.08	1.82	1.94	1.76	1.71	1.87	**2.10**
	10	2.16	**2.40**	2.14	2.23	2.08	1.98	2.21	**2.40**
SegSNR	0	– 4.18	– 1.95	– 1.95	– 3.03	– 1.46	**– 1.13**	**– 1.13**	– 1.32
	5	– 1.79	0.35	– 1.26	– 1.61	– 0.73	– 0.34	– 0.34	**1.23**
	10	0.78	2.27	– 0.43	0.70	– 0.01	0.34	0.34	**4.03**
WSS	0	72.54	104.33	108.63	72.30	109.70	105.64	108.12	**71.18**
	5	63.80	94.14	104.25	64.21	102.32	98.81	100.74	**59.00**
	10	54.20	80.55	93.20	55.23	91.86	89.76	85.98	**53.12**
LLR	0	1.99	1.83	2.10	1.90	2.12	2.12	2.13	**1.81**
	5	1.72	1.55	1.86	1.60	1.85	1.89	1.87	**1.45**
	10	1.42	1.31	1.71	1.29	1.69	1.72	1.62	**1.24**

Average values were obtained using 10 sentences extracted from the NOIZEUS database. The best performance is shown in bold

for the MMSE-MODGD method than for the other methods. The results also show that these assessments confirm that speech improvement based on the MMSE-MODGD method produces a higher segmental SNR, higher PESQ and lower WSS than other methods.

6 Experimental protocol for FASR setup

Generally, there are two constraints in the FASR scenarios. The first is the non-collaboration of the suspects and the second one is the limited number of suspects known by the target person (person who suffers from the actions of others). Due to these constraints, the number of suspects used to develop such systems (FASR) is really limited.

In this work, all the experiments were performed on the NIST 2000 corpus, which consists of the spontaneous telephone speech sampled at 8 kHz. For feature extraction, a 23 MFCC vector is found from pre-emphasized speech every 10 ms using a 20 ms Hamming window.

Twenty speakers were chosen as suspects from this corpus; the suspected speaker Reference database (R) was recorded with 1 recording of 2 min duration which was chosen for each suspect, and 75% of the duration of this recording was intended for modeling and 25% for tests (traces).

The test segment is divided into 4 sections, to have 4 traces for each suspect. The Potential database (P) used was a subset of 420 speakers from the same corpus cited below. The GMM-UBM consisted of 256 mixture components trained via Expectation Maximization (EM) algorithm using 10 iterations (Reynolds & Rose, 1995).

Twenty suspects models were created through the GMM-UBM using maximum a posteriori (MAP) adaptation with factor relevance r = 16, 256 mixtures and an adaptation data amount of 14 h is used (Reynolds et al., 2000).

According to the Fig. 1, which explains the FASR methodological approach adopted in our work, we need 3 databases:

1. Potential-database (UBM database): contains 420 speakers (420*2 min = 14 h);
2. Trace-database (T): contains 20 speakers, each speaker has 4 traces of (0.25*2 min)/4 = 7.5 s. So, the total of true trials (H_0) is 20*4 = 80 and the total of false trials (H_1) is 4*20*20 – 80 (true trials) = 1520;
3. Reference-database: contains 20 speakers, each speaker has 0.75*2 min = 1.5 min.

Table 2 Objective evaluations of the MMSE-MODGD technique compared with ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP and Wiener and corrupted with factory noise.

Objective measures	Input SNR dB	Factory noise							
		ML	MMSE	Log-MMSE	MAP	MMSE-ISP	Log-MMSE-ISP	Wiener	MMSE-MODGD
SIG [1 to 5]	0	1.41	1.33	1.22	1.25	1.23	1.01	1.13	**1.58**
	5	1.95	1.93	1.60	1.72	1.80	1.34	1.39	**2.05**
	10	2.49	2.55	2.18	2.29	**2.64**	1.89	2.04	2.47
BAK [1 to 5]	0	1.65	1.67	1.62	1.77	1.44	1.30	1.72	**1.82**
	5	2.04	2.11	1.92	1.95	1.96	1.74	1.54	**2.13**
	10	2.45	2.47	2.37	2.46	2.16	2.08	2.21	**2.56**
OVRL [1 to 5]	0	1.82	1.49	1.15	1.22	1.28	1.17	1.22	**1.87**
	5	1.83	1.50	1.19	1.33	1.31	1.34	1.33	**1.92**
	10	2.12	2.20	2.07	2.00	2.01	1.55	1.88	**2.34**
PESQ	0	1.04	1.07	1.76	1.83	1.72	1.74	1.98	**2.02**
	5	1.81	2.01	1.80	1.93	1.77	1.81	2.07	**2.29**
	10	2.36	2.45	2.04	2.32	2.17	1.99	2.47	**2.54**
SegSNR	0	− 3.13	− 1.73	− 1.82	− 3.13	− 1.70	− 1.10	− **1.03**	− 1.77
	5	− 0.99	1.05	− 1.06	− 1.33	− 0.22	0.04	− 0.19	**1.28**
	10	1.02	2.50	0.88	1.09	0.30	1.50	0.95	**3.73**
WSS	0	72.03	98.33	102.63	67.30	103.10	100.60	102.02	**66.11**
	5	53.70	84.12	94.52	54.22	98.12	88.80	100.01	**50.99**
	10	45.20	78.52	90.21	52.22	81.87	83.73	83.93	**44.14**
LLR	0	1.88	1.72	2.05	1.87	2.02	2.09	2.17	**1.49**
	5	1.39	1.67	**1.12**	1.57	1.66	1.98	1.77	1.33
	10	1.30	**1.01**	1.05	1.19	1.59	1.53	1.44	1.11

Average values were obtained using 10 sentences extracted from the NOIZEUS database. The best performance is shown in bold

Performance metrics provided a single numerical value that described the performance in terms of accuracy, discriminating power and calibration of the LR method (Probabilities of Misleading Evidence, $PMEH_0$ and $PMEH_1$), Equal Proportion Probability (EPP) (Drygajlo et al., 2016; Haraksim & Drygajlo, 2016; Kenai et al., 2019).The values used for the MODGD functions are the length of cepstral lifter window $lifter_\omega = 8$ and $\alpha = 0.4$, $\gamma = 0.9$.

7 Classical Forensic Automatic Speaker Recognition results

This Section evaluates the results obtained in clean and noisy environments.

7.1 FASR performance under clean conditions

An evaluation of FASR based on GMM-UBM performance in terms of EPP, $PMEH_0$ and $PMEH_1$ was performed in a clean environment.

According to Table 4, the results are very satisfying, in terms of EPP, $PMEH_0$ and $PMEH_1$. Therefore, EPP = 1, 25%, the LR exceeds 1 in 96% of cases when H_0 is true and in only 0.4% of cases when H_1 is true.

7.2 FASR performance in noisy conditions

Different noises were arbitrarily chosen in this study (babble, factory and white) that were added to the corpus of the questioned recording (traces) to produce noisy feature vectors. Table 5 presents the performances of FASR, at SNR = 0 dB and SNR = 5 dB.

Table 5 summarizes the performances of the FASR under noisy environment, in terms of EPP, $PMEH_0$ and $PMEH_1$. It can be noticed that the performance metrics decrease with decreasing SNR, and increase with increasing SNR, and the performance of noisy speech corrupted with babble noise is less degraded compared to the other noises. This can be explained by the fact that, the babble noise is an overlap of several sounds that comes from two or more speakers (Djeghiour et al., 2018). Its features are like those of the voice. It covers only the low frequency spectrum. Therefore, only the information in low frequency regions is affected by this noise. Whereas, the factory and white noises are characterized by a high intensity. They cover the low and high frequency spectrum and they affect all the existing information

Table 3 Objective evaluations of the MMSE-MODGD technique compared with ML, MMSE, Log-MMSE, MAP, MMSE-ISP, Log-MMSE-ISP and Wiener and corrupted with babble noise.

Objective measures	Input SNR dB	Babble noise							
		ML	MMSE	Log-MMSE	MAP	MMSE-ISP	Log-MMSE-ISP	Wiener	MMSE-MODGD
SIG [1 to5]	0	2.33	1.88	1.92	2.29	2.22	1.84	1.95	**2.67**
	5	2.74	2.73	2.27	2.59	2.51	2.04	2.25	**3.09**
	10	**3.52**	2.93	2.77	3.03	3.06	2.63	2.81	3.28
BAK [1 to 5]	0	1.77	1.48	1.55	1.73	1.71	1.58	1.62	**1.84**
	5	2.07	1.85	1.84	1.99	1.99	1.82	1.89	**2.19**
	10	**2.42**	2.25	2.16	2.28	2.29	2.09	2.14	**2.42**
OVRL [1 to 5]	0	**2.17**	1.52	1.60	1.90	1.84	1.54	1.64	1.93
	5	2.30	1.99	1.95	2.18	2.13	1.79	1.95	**2.54**
	10	2.76	2.48	2.39	2.57	2.59	2.27	2.40	**2.89**
PESQ	0	1.85	1.55	1.79	**1.87**	1.69	1.64	1.83	**1.87**
	5	2.12	1.95	1.98	2.13	1.93	1.85	2.12	**2.15**
	10	2.34	2.28	2.30	2.38	2.28	2.16	2.38	**2.48**
SegSNR	0	− 4.19	− 3.09	− 2.01	− 3.46	− 1.67	**− 1.42**	**− 1.42**	− 2.59
	5	− 1.72	− 1.10	− 1.49	− 1.30	− 1.02	− 0.57	− 0.57	**− 0.23**
	10	0.93	0.67	− 1.00	0.82	− 0.39	0.04	0.04	**2.81**
WSS	0	69.77	100.07	107.02	65.20	108.35	105.06	104.81	**58.80**
	5	**50.10**	92.07	92.55	58.87	95.87	95.17	87.61	64.38
	10	51.12	73.35	84.21	49.12	83.88	83.96	73.73	**41.42**
LLR	0	0.98	1.21	1.20	0.99	1.25	1.26	1.14	**0.96**
	5	0.80	1.33	1.23	0.85	1.28	1.29	1.07	**0.78**
	10	0.69	0.85	1.06	0.63	0.98	1.03	0.79	**0.60**

Average values were obtained using 10 sentences extracted from the NOIZEUS database. The best performance is shown in bold

Table 4 Evaluation results obtained in clean environment

EPP (%)	H_0 true (%)		H_1 true (%)	
	LR < 1	LR > 1	LR < 1	LR > 1
1.25	4	96	99.60	0.40

in the speech signal. The performance is worse when using these two types of noises (factory and white) than those obtained under babble.

8 Enhanced Forensic Automatic Speaker Recognition results

In this Section, the performance of this system was calculated using the MMSE magnitude enhancement processing and our approach proposed MMSE-MODGD enhancement processing.

Table 5 Evaluation results obtained under noisy environments

Type of noise	SNR (dB)	EPP (%)	H_0 true (%)		H_1 true (%)	
			LR < 1	LR > 1	LR < 1	LR > 1
Babble	0	20	85	15	100	0
	5	13.75	71	29	100	0
Factory	0	37.50	91	09	100	0
	5	20	86	14	100	0
White	0	43.75	95	05	100	0
	5	33.75	85	15	100	0

Table 6 Evaluation results obtained with MMSE enhancement processing

Type of noise	SNR (dB)	EPP (%)	H_0 true (%)		H_1 true (%)	
			LR < 1	LR > 1	LR < 1	LR > 1
Babble	0	15.39	85	15	100	0
	5	10	76.25	23.75	100	0
Factory	0	17.82	90	10	100	0
	5	11.25	77.50	22.50	100	0
White	0	20	91.25	8.75	100	0
	5	8.81	80	20	100	0

8.1 FASR performance using MMSE-magnitude enhancement processing

Table 6 indicates the results obtained when using the MMSE-magnitude enhancement processing (only the information contained in the magnitude), at SNR = 0 dB and SNR = 5 dB.

The results presented in Table 6, when applying the MMSE speech enhancement algorithm on noisy tests (traces) speech, indicate an improvement of the performances represented by the decreasing of the EPP with the evolution of LR.

This improvement is explained by the fact that, the MMSE based magnitude spectrum estimator discards all the broadband noise by eliminating most of the wide peaks that constitute the undesirable variances of the spectrum ordinates (Loizou, 2007).

Moreover, the MMSE based magnitude spectrum estimator provides the posterior Probability Density Function (PDF) of the clean signal given the noisy signal. This PDF is an optimal estimator for a large class of difference distortion measures between clean and noisy signal. This distortion measure assigns zero distortion for estimates in the immediate neighborhood of the clean signal, and uniform distortion for the ones outside this neighborhood (Loizou, 2007; Lu & Loizou, 2011). Therefore, the separation between noise and speech components is better.

8.2 FASR performance using the proposed improved MMSE-MODGD enhancement processing

Table 7 summarizes the results obtained when using the proposed algorithm (improved MMSE-MODGD enhancement processing), taking into account the information contained in the magnitude and phase, at SNR = 0 dB and SNR = 5 dB.

Based on the results in Table 7, it can be observed that when comparing these results with those obtained in Sect. 8.1, a significant improvement of FASR performance metrics in terms of EPP and Probabilities of Misleading Evidence ($PMEH_0$ and $PMEH_1$) is observed, for the three kinds of noises (babble, factory and white). Therefore, in terms of EPP, the improvements represent 1.84% reduction for babble noise and 1.25% reduction for other noises. These results are encouraging given that 1% improvement is significant for high security systems such as FASR systems, as the innocence or indictment of individuals is at stake.

This improvement given by the addition of the MMSE-MODGD estimator to the FASR system is explained by the fact that, the subtraction of the noise from the noisy speech signal, when using MMSE-magnitude spectrum cannot eliminate the deep valleys surrounding the narrow peaks, which remain in the noise spectrum. Therefore, the excursion of noise peaks remains large. However, MMSE-MODGD discards these deep valleys by well preserving the peaks and valleys (depth reduction) of the clean magnitude spectrum in the presence of additive noise (properties of the group delay function of a minimum-phase signal).

Table 7 Evaluation results obtained with the proposed improved MMSE-MODGD enhancement processing

Type of noise	SNR (dB)	EPP (%)	H_0 true (%)		H_1 true (%)	
			LR < 1	LR > 1	LR < 1	LR > 1
Babble	0	13.55	88.75	11.25	100	0
	5	8.75	72.50	27.50	100	0
Factory	0	17.50	90	10	100	0
	5	10	79	21	100	0
White	0	18.75	92	8	100	0
	5	8.75	79	21	100	0

Moreover, in (Parthasarathi et al., 2011), the authors indicated that the MODGD spectrum is inversely proportional to the noise power at frequencies corresponding to high noise regions, and directly proportional to the signal power. This indicates that, the MODGD spectrum tends to follow the magnitude spectrum of the signal, rather than that of the noise.

Thus, on the basis of experiments, it was found that noise distorts the shape of the MODGD spectrum less than the FFT spectrum, changes its slopes and reduces the dynamic range of the MODGD spectrum less than the FFT. Most of the time, the frequency locations of the peaks of the higher formants are preserved to some extent in the MODGD spectrum compared to the FFT spectrum in the presence of noise. Therefore, our proposal for MMSE-MODGD retains more information contained in the noisy speech signal than conventional MMSE (Gerkmann & Hendriks, 2012), to avoid any degradation in speech intelligibility and FASR performance.

9 Conclusion

In this work, speech enhancement estimators of noisy speech signal were studied under the assumption that the spectrum of the noisy speech signal can be represented in complex plane as sum of clean signal spectrum and noise spectrum. In addition to the traditional estimator, which is based on the MMSE principles, the improved estimator was proposed by incorporating modified group delay spectrums. Furthermore, compared to the FASR performance using the classical MMSE spectral power estimators, the FASR using the proposed MMSE-MODGD resulted in significantly better speech enhancement quality.

The results of the experiments show that MODGD spectrum has the potential to reduce noise components in the noisy speech signal, since the MODGD spectra tends to follow the magnitude spectrum of speech and opposes the noise spectrum. Therefore, it can be concluded that the important information retained in the enhanced speech using the MODGD spectrum can complement that given by FFT spectrum and give more reliability and robustness to the FASR system under noisy environments.

In future work, we intend to apply a state of the art technique during the parametrization or training phase, which should be an interesting approach to refine the speaker models to obtain a better performance for the proposed forensic system. Subsequently, the latter will be applied to another database specific to the forensic field to compare the two systems.

References

Akande, O. O., & Murphy, P. J. (2005). Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication, 46*(1), 15–36.

Alexander, A., Botti, F., Dessimoz, D., & Drygajlo, A. N. D. R. Z. E. J. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International, 146,* S95–S99.

Asbai, N., & Amrouche, A. (2017). Boosting scores fusion approach using Front-End Diversity and adaboost Algorithm, for speaker verification. *Computers & Electrical Engineering, 62,* 648–662.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79*. (Vol. 4, pp. 208–211). IEEE.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*(2), 113–120.

De Forest, P. R., Gaensslen, R. D., & Lee, H. C. (1983). *Forensic science: an introduction to criminalistics*. New York: McGraw-Hill Humanities/Social Sciences/Languages.

Deshpande, M. S., & Holambe, R. S. (2011). Robust speaker identification in babble noise. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*. (pp. 635–640). ACM

Dixit, S., & Mulge, D. M. Y. (2014). Review on speech enhancement techniques. *International Journal of Computer Science and Mobile Computing, 3*(8), 285–290.

Djeghiour, S., Asbai, N., Kenai, O., & Guerti, M. (2018). *Forensic Automatic Speaker Recognition under Noisy Environments. IC3E'2018*.University of Bouira (pp. 1–5).

Drygajlo, A. (2012). Automatic speaker recognition for forensic case assessment and interpretation. In *Forensic Speaker Recognition* (pp. 21–39). New York: Springer.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2016). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. VerlagfürPolizeiwissenschaft.

Drygajlo, A., Meuwly, D., & Alexander, A. (2003). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. In *Eighth European Conference on Speech Communication and Technology*.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121.

Gerkmann, T., & Hendriks, R. C. (2012). Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(4), 1383–1393.

Gustafsson, H., Lindgren, U., Claesson, I., & Nordholm, S. (2004). U.S. Patent No. 6,717,991. Washington, DC: U.S. Patent and Trademark Office.

Haraksim, R., & Drygajlo, A. (2016). *Measuring performance in forensic automatic speaker recognition: VQ, GMM-UBM, i-vectors*. Biosig 2016.

Hegde, R. M., Murthy, H. A., & Rao, G. R. (2004). Application of the modified group delay function to speaker identification and discrimination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04)*. (Vol. 1, pp. I-517). IEEE.

Hu, Y., & Loizou, P. C. *Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms*. Retrieved from http://www.utdallas.edu/~loizou/speech/noizeus/.

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement, IEEE Trans. *Audio, Speech Language Process., 16*(1), 229–238.

Kenai, O., Djeghiour, S., Asbai, N., & Guerti, M. (2019). Forensic gender speaker recognition under clean and noisy environments. *Procedia Computer Science, 151,* 897–902.

Kwatra, N., Milani, A. A., & Alderson, J. (2017). U.S. Patent No. 9,824,677. Washington, DC: U.S. Patent and Trademark Office.

Loizou, P. C. (2007). *Speech enhancement theory and practice* (1st ed.). Boca raton: CRC Press.

Lu, Y., & Loizou, P. C. (2011). Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1123–1137.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing, 9*(5), 504–512.

Murthy, H. A., & Yegnanarayana, B. (2011). Group delay functions and its applications in speech technology. *Sadhana, 36*(5), 745–782.

Parthasarathi, S. H. K., Padmanabhan, R., & Murthy, H. A. (2011). Robustness of group delay representations for noisy speech signals. *International Journal of Speech Technology, 14*(4), 361.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(1–3), 19–41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.

Roux, C., Crispino, F., & Ribaux, O. (2012). From forensics to forensic science. *Current Issues Criminal Justice, 24,* 7.

Scheffer, N., Ferrer, L., Lawson, A., Lei, Y., & McLaren, M. (2013). Recent developments in voice biometrics: Robustness and high accuracy. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, (pp. 447–452). IEEE.

Verschuur, C., Lutman, M., & Wahat, N. H. A. (2006). Evaluation of a non linear spectral subtraction noise suppression scheme in cochlear implant users. *Cochlear Implants International, 7*(4), 193–196.

Wolfe, P. J., & Godsill, S. J. (2003). Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP Journal on Applied Signal Processing, 2003,* 1043–1051.

Zhang, Y., & Abdulla, W. H. (2007). Robust speaker identification in noisy environment using cross diagonal GTF-ICA feature. In *2007 6th International Conference on Information, Communications & Signal Processing*, (pp. 1–4). IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.