Modern Speech Enhancement Techniques in Text-Independent Speaker Verification

César A. Medina, José A. Apolinário Jr., and Abraham Alcaim
IME and CETUC/PUC-Rio

***Abstract*— The noise robustness of speaker verification systems is crucial for real applications although only few articles have tackled this problem. In this paper, we study the performance of several modern speech enhancement solutions including wavelet–based speech denoising. We use these algorithms as a preprocessing stage in a text–independent speaker verification system. The results are presented after exhaustive simulations.**

I. Introduction

The effect of additive noise in a speaker verification system is a critical problem for real applications [1]. One possible approach to improving the Equal Error Rate (EER) of the verification system is the use of a preprocessing stage with speech enhancement capability. Due to its simplicity, spectral subtraction has been widely used to reduce the effect of additive noise in several areas of speech processing. Recently, wavelet transforms were proposed for denoising speech data [2]. In this paper, we study the performance of modern spectral subtraction and wavelet denoising techniques in a GMM text–independent speaker verification system.

The paper is organized as follows: Section II reviews the most common speech enhancement techniques, including the wavelet–based denoising approach. In Section III, some basic speaker verification concepts are presented. The simulations results are shown in Section IV, and, finally, Section V summarizes some conclusions.

II. Speech Enhancement Techniques

In the technical literature, there are several approaches to the speech enhancement problem, some examples are short–time spectral modification [3]–[7], adaptive filtering [8], and wavelet thresholding [2]. In this section, we review the fundamental concepts of short–time spectral modification and wavelet thresholding.

Due to the non-stationarity of the speech signal, its processing is carried out on a frame-by-frame basis and its reconstruction with an overlap–add procedure. Proper synthesis is also a key to obtain good performance in any noise reduction method.

C. A. Medina and J. A. Apolinário Jr. are with the Departamento de Engenharia Elétrica, Instituto Militar de Engenharia, Praça General Tibúrcio 80, Rio de Janeiro, RJ, 22.290-270 (e-mail: csmedina@epq.ime.eb.br and apolin@ieee.org).

A. Alcaim is with CETUC/PUC-Rio, Rua Marquês de São Vicente, 225, Rio de Janeiro, RJ, 22.453-900 (e-mail: alcaim@cetuc.puc-rio.br).

A. Short–Time Spectral Modification

This type of algorithm attempts to estimate the short–time spectral magnitude (STSM) of the noise embedded in a noisy speech signal and multiply the STSM of the noisy speech by a gain function that depends of the noise estimation. The phase of the noisy speech, on the other hand, is not processed, based on the assumption that phase distortion is not perceived by the human ear [5].

These algorithms constitute the traditional approach for removing background noise in single channel systems. They present a trade-off among the amount of noise reduction, the speech distortion, and the level of residual musical noise.

1) **Classical Power Subtraction***:* Consider a speech signal $x(n)$ corrupted by uncorrelated additive stationary background noise $d(n)$. The noisy speech can be expressed as follows.

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

The enhanced speech STSM $|\hat{X}(w)|$ is obtained by subtracting, from the noisy speech short-time magnitude $|Y(w)|$, the noise spectral magnitude estimate $|\hat{D}(w)|$ computed during speech pauses. For the particular case of power subtraction, this is expressed as follows:

$$|\hat{X}(w)|^2 = \begin{cases} |Y(w)|^2 - |\hat{D}(w)|^2, & \text{if } |Y(w)|^2 > |\hat{D}(w)|^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $|\hat{D}(w)|^2$ represents the noise power spectrum estimate. The phase of the noisy speech is not modified.

Once the subtraction is computed in the spectral domain with (2), the enhanced speech signal is obtained with the following relation.

$$\hat{x}(n) = IFFT[|\hat{X}(w)| \cdot \exp(j \arg Y(w))] \tag{3}$$

2) **Spectral Subtraction Based on Masking Properties of the Human Auditory System***:* This algorithm was proposed by [5] and uses the masking properties—where strong sounds make weaker sounds inaudible—calculated from auditory models. In order to attain better performance, the spectral subtraction algorithm became a Generalized Parametric Spectral Subtraction Algorithm as follows.

$$|\hat{X}(w)| = G(w)|Y(w)| \tag{4}$$

$$G(w) = \begin{cases} \left(1 - \alpha \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{if } \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1} < \frac{1}{\alpha+\beta} \\ \left(\beta \left[\frac{|\hat{D}(w)|}{|Y(w)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{otherwise} \end{cases} \tag{5}$$

This algorithm allows a variation of the trade-off between noise reduction and speech distortion with the variations of the free parameters of (5):

- Over-subtraction factor α ($\alpha > 1$): the short-time spectral is attenuated more than necessary. This leads to a reduction of residual noise peaks but also to an increased audible distortion.
- Spectral flooring β ($0 \leq \beta \ll 1$): addition of background noise in order to mask the residual noise. This leads to a reduction of residual noise but an increased level of background noise remains in the enhanced speech.
- Exponent $\gamma = \gamma_1 = 1/\gamma_2$: determines the sharpness of the transition from the $G(w) = 1$ to the $G(w) = 0$. The choice of this parameter is not as critical as that of α and β.

The choice of the parameters is then based on the concept of noise masking or masking properties of the human hearing system. The auditory spectral subtraction scheme (ASS) is composed of the following stages:

- Spectral Decomposition;
- Speech/noise detection and estimation of noise during speech pauses;
- Calculation of the noise masking threshold, $T(w)$;
- Adaptation in time and frequency of the subtraction parameters, α and β, based on the masking threshold $T(w)$;
- Calculation of the enhanced signal via (5);
- Inverse transform.

Masking is present because the auditory system is incapable of distinguishing two signals close in time or frequency domain. This is manifested by an elevation of the minimum threshold of audibility due to the masker signal. Here, it is only considered the frequency domain masking or simultaneous masking. This phenomenon is modeled via a noise masking threshold which can be found in [5], [9], [10].

Following, the steps to calculate the masking threshold is briefly described. Initially, the power spectrum of the signal (estimated from the magnitude squared of the DFT), $P(w)$, is used to compute the energy present in each critical band of a Bark scale—see the frequency bands in Tab. I—as in

$$B_i = \sum_{w=bl_i}^{bh_i} P(w) \tag{6}$$

where the summation limits are the critical band boundaries. Also note that the range of the index i, in the above equation, is sample–rate dependent. $P(w)$ must be computed from the unknown clean signal, then, we use an estimate of the signal obtained form a speech enhancement algorithm such as power spectral subtraction.

In the next step, the spreading function, given in (7), is convolved with the discrete Bark spectrum, as in (8), to account for the spread of masking.

$$SF(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2} \tag{7}$$

$$C_i = B_i * SF_i \tag{8}$$

We then compute $T_i = 10^{\log_{10}(C_i) - (O(i)/10)}$, where $O(k)$ is the relative threshold offset which depends on the noise–like (higher critical bands) or tone–like (lower critical bands) nature of C_i. In [5], we see that $O(k)$ can be easily obtained from the values of Fig. 1.

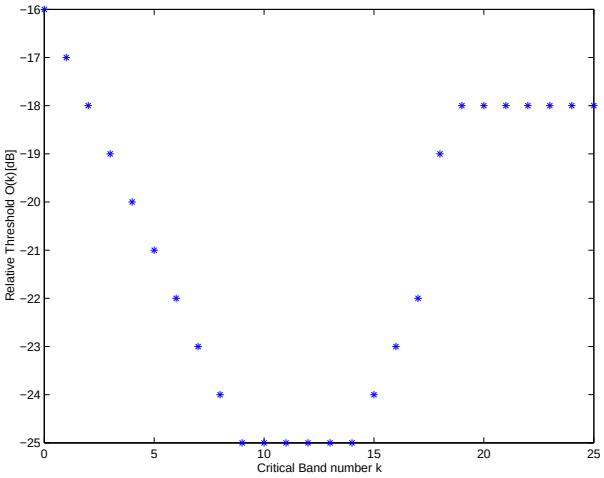

Fig. 1. Relative Threshold Offset.

Finally, each T_i is renormalized and is checked against the absolute threshold of hearing and replaced by $\max(T_f, T_q(f))$, $T_q(f)$ being the absolute threshold, given in [9].

$$T_q(f) = 3.64f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 10^{-3}f^4 \tag{9}$$

and f is the frequency in KHz.

The adaptation parameters, α and β, are based on the masking threshold, $T(w)$, and on the following assumption: if the masking threshold is high, residual noise will be naturally masked and inaudible. Hence, there is no need to reduce it in order to keep distortion as low as possible.

The adaptation parameters is then performed with the following relations:

$$\alpha_m(w) = F_\alpha[\alpha_{min}, \alpha_{max}, T(w)] \tag{10}$$

$$\beta_m(w) = F_\beta[\beta_{min}, \beta_{max}, T(w)] \tag{11}$$

where $F_\alpha = \alpha_{max}$ if $T(w) = T(w)_{min}$, and $F_\alpha = \alpha_{min}$ if $T(w) = T(w)_{max}$, $T(w)_{max}$ and $T(w)_{min}$ are the maximum and minimum value of $T(w)$ updated frame by frame. The minimum and maximum values of α and β are experimentally calculated. The values of F_α and F_β between these extreme cases are interpolated based on the value of $T(w)$.

B. Denoising by Wavelets

The denoising with wavelets is a technique based on the thresholding of the "detail" coefficients of the wavelet transform of a given signal [11]. This technique can be summarized as follows.

Using wavelet transforms, (1) can be expressed as

$$Y = X + D \tag{12}$$

where $Y = Wy$, $X = Wx$, and $D = Wd$, W being the wavelet transform matrix ($WW^T = I$). The transformed signal estimation, $\hat{X}$, is based on the thresholding of the detail coefficients through a thresholding function, such as:

TABLE I

CRITICAL BAND FILTER BANK (CO=CENTER FREQ., BW=BANDWIDTH)

Band No.	Co(Hz)	BW (Hz)	Band No.	Co(Hz)	BW (Hz)	Band No.	Co(Hz)	BW (Hz)
1	50	-100	9	1000	$920-1080$	17	3400	$3150-3700$
2	150	$100-200$	10	1175	$1080-1270$	18	4000	$3700-4400$
3	250	$200-300$	11	1370	$1270-1480$	19	4800	$4400-5300$
4	350	$300-400$	12	1600	$1480-1720$	20	5800	$5300-6400$
5	450	$400-510$	13	1850	$1720-2000$	21	7000	$6400-7700$
6	570	$510-630$	14	2150	$2000-2320$	22	8500	$7700-9500$
7	700	$630-770$	15	2500	$2320-2700$	23	10500	$9500-12000$
8	840	$770-920$	16	2900	$2700-3150$	24	13500	$12000-15500$
						25	19500	$15500-$

- Soft–Thresholding

$$\eta_S(Y,t) = \begin{cases} sgn(Y)(|Y|-t), & |Y| \geq t \\ 0, & otherwise \end{cases} \quad (13)$$

- Hard–Thresholding

$$\eta_H(Y,t) = \begin{cases} Y, & |Y| \geq t \\ 0, & otherwise \end{cases} \quad (14)$$

Finally, the signal estimation, $\hat{x}(n)$, is obtained from $\hat{x}(n) = W^{-1}\hat{X}$.

There are several techniques to compute the threshold, t, in (14) and (13). The most classical are:

- VisuShrink [12]: The threshold is chosen as $t = \hat{\sigma}\sqrt{2\log N}$, where N is the number of coefficients at the finest level and $\hat{\sigma} = m/.6745$, m being the median absolute deviation. This method can be level dependent if we update t and m for each level or can be level independent if we update t and m based only in the coefficients of the finest level.
- SureShrink [11], [12]: Given the noisy detail coefficients expressed by $Z_k = \beta_k + \sigma_k\xi_k$, $k = 1, \cdots, N$, where $\sigma_k > 0$ are unknown parameters and ξ_k are independent Gaussian random variables (mean=0, variance =1), and a risk function defined by:

$$\mathcal{R} = \sum_{k=1}^{N} E[(\hat{\beta}_k - \beta_k)^2] \quad (15)$$

where $\hat{\beta}_k = Z_k + H_t(Z_k)$ is an estimator and $H_t[\cdot]$ is a weakly differentiable function, the Stein's Unbiased Risk Estimator (SURE) is an unbiased estimate of the risk function in (15), and is given by:

$$\widehat{\mathcal{R}} = \sum_{k=1}^{N} R(\sigma_k, Z_k, t) \quad (16)$$

where

$$R(\sigma,x,t) = \sigma^2 + 2\sigma^2\frac{\partial}{\partial x}H_t(x) + H_t^2(x). \quad (17)$$

This estimate depends on the thresholding function: for Soft–Thresholding,

$$H_t(x) = -xI\{|x| < t\} - tI\{|x| \geq t\}\text{sign}(x) \quad (18)$$

where $I\{\cdot\}$ is the indicator function defined by:

$$I\{x\} = \begin{cases} 1, & x \text{ true;} \\ 0, & x \text{ false} \end{cases} \quad (19)$$

from (18), we rewrite (17) as:

$$\begin{aligned} R(\sigma,x,t) &= (x^2-\sigma^2)I\{|x|<t\} + (\sigma^2+t^2)I\{|x|\geq t\} \\ &= [x^2-\sigma^2] + (2\sigma^2 - x^2 + t^2)I\{|x|\geq t\} \end{aligned} \quad (20)$$

The expression in square brackets in (20) does not depend on t. Thus, it is equivalent to

$$\hat{t} = \min_{t\geq 0}\sum_{k=1}^{N}(2\sigma_k^2 + t^2 - Z_k^2)I\{|Z_k|\geq t\}. \quad (21)$$

The main advantage of these wavelet denoising techniques is the absence of residual musical noise.

III. SPEAKER VERIFICATION

In this paper we use a system of speaker verification based on Gaussian Mixture Models (GMM) [13]. A mixture of Gaussian probabilities is a weighted sum of M Gaussian densities and is given by $p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x})$, where $\boldsymbol{x}$ is a $D \times 1$ random vector, $b_i(\boldsymbol{x})_{i=1,\ldots,M}$ are the densities components, and $p_{ii=1,\ldots,M}$ are the mixture weights.

Each component density is a D variate Gaussian function with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{K}_i$. The model parameters $\lambda = \{p_i, \boldsymbol{\mu}_i, \boldsymbol{K}_i\}_{i=1,\ldots,M}$ are estimated by an EM algorithm such as the one used in [14]. For a set of training data, the model parameters are determined in order to maximize the likelihood of the GMM.

For a sequence of T independent training vectors $\boldsymbol{X} = \{\boldsymbol{x}_s, \ldots, \boldsymbol{x}_T\}$, the likelihood of the GMM for modeling a true speaker (λ model) is calculated through $\log p(\boldsymbol{X}|\lambda)$. The scale factor $\frac{1}{T}$ is used in order to normalize the likelihood according to the duration of the utterance (number of feature vectors).

The speaker verification problem requires a binary decision, accepting or rejecting a pretense speaker. The system computes the normalized logarithmic likelihood for two models: one from the pretense speaker and another one trying to minimize the variations not related to the speaker (*background* model) The background model provides a more stable decision threshold []. If the system output value (difference between two likelihoods) is higher than a given threshold, θ, the speaker is accepted; otherwise, it is rejected. The background is built with a hypothetical set of false speakers and modeled via GMM (universal background model). The thresholds is calculated on the basis of experimental results.

IV. SIMULATIONS

This section presents the performance evaluation of the reviewed speech enhancement algorithms in terms of SNR gain as well as in the task of text–independent speaker verification.

In our experiments, the parametrization of the speech was carried out by $mel-cepstrum$ coefficients, under the conditions listed in Tab. II.

TABLE II
FEATURES EXTRACTION CONDITIONS USED IN THE SIMULATIONS

Parameter	Value
Pre-emphasis	$1-0.95z^{-1}$
Window length	$32ms$
Window shift	$16ms$
Window type	Hamming
MFCC order	$15\ mel-cepstrum$
Sampling frequency	$8KHz$

The background noise signals used in the simulations are the artificially generated Gaussian white noise and three distinct noises obtained from the *NOISEX*-92 database [15]: factory noise, speech like noise, and aircraft cockpit noise. Phonetically balanced speech sentences were extracted from the *IME*-2001 database. This database is composed of 66 different male speakers and the utterances were recorded with $8KHz$ sampling frequency, electret microphones, and a low noise environment.

Tab. III shows the SNR gain over 100 speech signals of 25 seconds each, taken from different speakers of the IME-2001 database and corrupted in different noise conditions. The SNR gain (in dB) is computed with [5]:

$$G_{SNR} = \frac{1}{L}\sum_{m=0}^{L-1} 10\cdot \log_{10}\frac{\frac{1}{N}\sum_{n=0}^{N-1} w^2(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}\left[s(n+Nm)-\hat{s}(n+Nm)\right]^2} \quad (22)$$

where L is the number of processed frames in each signal, N is the number of samples in each frame and $w(n)$, $s(n)$ and $\hat{s}(n)$ are the added noisy, the clean signal, and the enhanced signal, respectively.

The algorithms used were: classical power subtraction (PS), spectral subtraction based on properties of the human auditory system (ASS), VisuShrink wavelet denoising (VS), and SureShrink wavelet denoising (SUS).

In [5], the performance of the ASS algorithm is better than other spectral subtraction techniques; here, however, we have chosen the free min–max parameters of (5) such that better performance was obtained not in terms of SNR gain, but Equal Error Rate (EER) in the Speaker Verification task.

We build the speaker verification database by taking 120 seconds of speech/silence signal to train the GMM model, and 25 seconds of speech/silence signal to test the GMM model. The training signals were corrupted by Gaussian white noise (SNR=5dB). Both, training and test signals, were preprocessed by the same speech enhancement algorithm. The simulation results of the speaker verification system are shown in Tab. IV.

TABLE III
MEAN SNR GAIN FOR DIFFERENT NOISE TYPES

Input SNR	PS	ASS	SUS	VS
White Gaussian Noise				
-5	4.73	**5.96**	0.67	2.14
0	4.94	**5.14**	-0.11	0.05
5	**4.52**	4.05	-2.16	-3.60
10	**3.60**	2.64	-5.54	-7.99
Factory Noise				
-5	3.78	**4.16**	0.17	-0.61
0	3.64	**3.88**	-0.57	-1.96
5	**3.14**	2.72	-2.49	-4.59
10	**2.51**	0.84	-5.70	-8.38
Aircraft Cockpit Noise				
-5	3.54	**4.25**	0.51	-0.09
0	3.38	**3.84**	-0.24	-1.39
5	**2.97**	2.65	-2.29	-4.29
10	**2.50**	1.26	-5.60	-8.23
Speech Like Noise				
-5	5.36	**5.77**	1.40	1.35
0	**4.42**	4.29	0.385	-0.67
5	**3.49**	2.51	-1.906	-3.95
10	**2.74**	0.97	-5.42	-8.10

TABLE IV
EER (%) FOR DIFFERENT NOISE TYPES (SNR=5 dB IN TRAINING SIGNAL).

Input SNR	PS	ASS	SUS	VS
White Gaussian Noise				
-5	36.27	35.94	34.61	**33.78**
0	24.46	17.80	**16.31**	16.47
5	15.97	6.49	**6.32**	**6.32**
10	13.81	8.32	8.15	**7.82**
Factory Noise				
-5	**42.76**	48.42	48.92	48.59
0	**35.44**	44.09	46.59	48.42
5	**23.96**	36.27	46.42	47.75
10	**19.47**	35.44	42.26	46.42
Aircraft Cockpit Noise				
-5	**42.10**	48.42	49.58	48.42
0	**34.11**	47.42	46.76	48.75
5	**22.46**	40.27	43.76	45.59
10	**16.14**	40.60	44.43	44.59
Speech Like Noise				
-5	**36.77**	47.09	50.08	48.75
0	**29.78**	44.93	50.25	48.25
5	**21.80**	45.09	48.09	45.92
10	**16.64**	43.76	45.09	45.76

V. CONCLUSIONS

A database exclusively composed of male speakers was corrupted by additive noisy of different types from the NOISEX-92 noisy database and several algorithms of speech enhancement, used as preprocessing stages, were applied to these corrupted signals in order to improve the performance of a text–independent speaker verification system. The results show that, for low SNR, the speech enhancement approach using masking properties of the human auditory system (ASS) provided better SNR gain when compared to (classical) power subtraction and wavelet–based denoising. Furthermore, in the speaker verification application, the PS algorithm, in general, performs better than the others. Only for the case of Gaussian white noise, wavelet based denoising performs better than spectral subtraction techniques. Nevertheless, this remarkable

difference in the SNR gain between ASS or PS and wavelet based denoising was not reflected in the $EER(\%)$ results of the speaker verification task; this could be due to the difference of residual noise present in each algorithm. In the PS and ASS, the so-called musical noise remains but with a low background noise. VS and SUS, on the other hand, do not present residual musical noise but the background noise level is considerably high. The use of different noises, other than white Gaussian, to be added to the training signals with different amplitudes (different SNR's) is currently under investigation.

REFERENCES

[1] N. B. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted viterbi algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, Mar. 2002.

[2] D. L. Donoho, "De–noising by soft–thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP–27, Apr. 1979.

[4] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft–decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, Apr. 1980.

[5] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[6] S. L. Gay and J. Benesty, *Acoustic Signal Processing for Telecomunications*, Kluwer international series in engineering and computer science, 2001.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error short–time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[8] M. Gabrea and C. Tadj, "Speech enhancement for speaker identification," in *International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, Sept. 2001.

[9] T. Painter and A. Spanias, "Perceptual coding of digital audio," in *Proceedings of the IEEE*. Apr. 2000, vol. 88, pp. 452–513, IEEE.

[10] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[11] W. Härdle, G. Kerkyachrian, D. Picard, and A. Tsybakov, "Wavelets approximation and statistical applications," Web version of the book: Wavelets, Approximation and Statistical Applications, Lecture Notes in Statistics (Springer-Verlag), Vol. 129. Springer-Verlag, New York., Sept. 1997.

[12] D. L. Donoho and I. M. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," in *Proceedings of the 16th Annual Conference of the IEEE Engineering in Medicine and Biology Society*, Maryland, USA, Nov. 1994, pp. 24a–25a.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, vol. 10, pp. 19–41. 2000.

[14] D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, 1992.

[15] A. P. Varga, M. Steeneken, H. J. M.and Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., Speech Research Unit, Defense Research Agency, Malvern, U.K., 1992.