NON-STATIONARY NOISE POWER SPECTRAL DENSITY ESTIMATION BASED ON REGIONAL STATISTICS

Xiaofei Li[1], *Laurent Girin*[1,2], *Sharon Gannot*[3] *and Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes
[2]GIPSA-Lab & Univ. Grenoble Alpes
[3]Faculty of Engineering, Bar-Ilan University

ABSTRACT

Estimating the noise power spectral density (PSD) is essential for single channel speech enhancement algorithms. In this paper, we propose a noise PSD estimation approach based on regional statistics. The proposed regional statistics consist of four features representing the statistics of the past and present periodograms in a short-time period. We show that these features are efficient in characterizing the statistical difference between noise PSD and noisy speech PSD. We therefore propose to use these features for estimating the speech presence probability (SPP). The noise PSD is recursively estimated by averaging past spectral power values with a time-varying smoothing parameter controlled by the SPP. The proposed method exhibits good tracking capability for non-stationary noise, even for abruptly increasing noise level.

Index Terms— noise PSD, speech presence probability, regional statistics.

1. INTRODUCTION

Noise power spectral density (PSD) estimation is an essential prerequisite for single channel speech enhancement algorithms [1, 2, 3]. For non-stationary noise, the PSD is generally estimated locally in the time-frequency domain. Local minimum of the smoothed noisy signal power spectrogram is often employed, such as the minimum statistics algorithm [4, 5], the minima controlled recursive averaging (MCRA) [6] and improved MCRA (IMCRA) [7] algorithms. If the speech signal is continuously present in the noisy speech mixture signal, these minimum-based methods are prone to overestimation of the minimum, especially if the search window is too short. Conversely, when the noise power is rising, the minimum detection may provide underestimated noise PSD, and the tracking delay will be large if the search window is too long. Other techniques, such as subspace-DFT [8] and minimum mean-squared error (MMSE) based estimators [9, 10], do not depend on minimum tracking, but rather use a weighting function optimal in the MMSE sense for estimating the noise PSD. When the noise power abruptly rises, the so-called *safety-net* [9] is adopted to shorten the adaptation time of the noise PSD tracking, i.e., if the minimum value of the periodograms in a search window is larger than the current noise PSD estimation, the latter will be replaced with the corresponding minimum value. However, the adaptation time is still too long due to the large search window.

Besides, IMCRA [7] uses a time-varying smoothing parameter, adjusted by the SPP, to average the past spectral power values. The SPP is estimated based on the *a posteriori* and *a priori* signal-to-noise ratios. Here, we adopt such a time-varying smoothing approach for estimating the noise PSD. However, we propose to estimate the SPP by using four statistical features, namely *Normalized Variance*, *Normalized Differential Variance*, *Normalized Average Variance* and *Median Crossing Rate*. These features are recursively computed, and represent the statistics of the periodograms in the region of approximately past 0.2 s. They will be therefore referred to as *regional statistics*. In this short-time period, the noise-only PSD is assumed to be an uncorrelated stationary process, whereas the speech PSD is considered to be non-stationary to some extent, and correlated between adjacent frames. The regional statistics of periodograms can efficiently characterize the statistical difference between noise PSD and noisy speech PSD, and hence enable to infer a reliable SPP estimator. Since the estimation of SPP refers only to the signal PSD in the past 0.2 s, the proposed noise PSD estimator has a short response time and is therefore suitable for tracking the non-stationary noise. Although the validity of the regional statistics is based on the assumption that the noise PSD is stationary in a short-time period, experiments show that the proposed noise PSD estimator works well even for abruptly changing noise power, e.g., with changing rate of 10 dB/s.

2. NOISE PSD ESTIMATION

Let us consider an additive speech + noise single-channel mixture signal. In time-frequency domain, the signal is written as

$$X(k,l) = S(k,l) + N(k,l), \tag{1}$$

This research has received funding from the EU-FP7 STREP project EARS (#609465).

where $X(k,l)$, $S(k,l)$ and $N(k,l)$ are the short-time Fourier transform (STFT) coefficients of the noisy speech, clean speech and noise signal, respectively, k is the frequency bin and l is the frame index. $S(k,l)$ and $N(k,l)$ are assumed to be independent random variables. It is shown in [11] that the successive noise spectral magnitudes are approximately uncorrelated if the frame overlap is not larger than 50%, while the successive speech spectral magnitudes are correlated along frames. For stationary signals, the probability density function (pdf) of periodogram bins $|X(k,l)|^2$ obeys the exponential distribution [4]. Let $\lambda_s(k,l) = E\{|S(k,l)|^2\}$ and $\lambda_n(k,l) = E\{|N(k,l)|^2\}$ denote the PSDs (or variances) of the speech and the noise signals, respectively. If speech is absent, the mean and variance of $|X(k,l)|^2$ are $\lambda_n(k,l)$ and $\lambda_n^2(k,l)$, respectively.

The noise PSD is estimated in this paper by using SPP-based recursive averaging [7]:

$$\hat{\lambda}_n(k,l) = \tilde{\alpha}_n(k,l)\hat{\lambda}_n(k,l-1) + (1-\tilde{\alpha}_n(k,l))|X(k,l)|^2$$

where $\tilde{\alpha}_n(k,l) = \alpha_n + (1-\alpha_n)p(k,l)$ is a time-varying smoothing parameter, adjusted by the SPP $p(k,l)$. The parameter α_n is empirically set to 0.8 in this paper. The smoothing parameter $\tilde{\alpha}_n(k,l)$ therefore increases from 0.8 to 1 together with the increase of the SPP from 0 to 1.

In what follows, we propose an alternative SPP estimation method based on regional statistics.

2.1. Regional statistics

First-order recursively smoothed periodogram is defined as:

$$P(k,l) = \alpha_x P(k,l-1) + (1-\alpha_x)|X(k,l)|^2 \quad (2)$$

where the smoothing parameter $\alpha_x = 0.85$ is equivalent to a smoothing window with the length of 0.2s (the signal sampling rate is 16kHz, the window length and the overlap between successive STFT frames are 512 and 256, respectively). This gives a good tradeoff between noise smoothing and non-stationary speech signal tracking [4]. In a time period of about 0.2s, the noise PSD is assumed to be an uncorrelated stationary process, whereas the noisy speech PSD is non-stationary and correlated. Four regional statistical features are proposed to distinguish the noise and noisy speech PSD.

Figure 1(a)-(d) shows the probability density of the regional statistics for noise-only and noisy speech. Stationary white Gaussian noise (WGN) is added to clean speech signal with SNR of 10dB. Signal duration is set to 300s. The frames with *a posteriori* SNR $|X(k,l)|^2/\lambda_n(k,l) > 9.2$ are considered as noisy speech frames. For the exponential distribution, 9.2 corresponds to the significance level of 0.01. The histograms of regional statistics for noise-only and noisy speech frames are calculated to represent the probability density.

The recursive *Normalized Variance*,

$$\vartheta_{nv}(k,l) = \alpha_x\vartheta_{nv}(k,l-1) + (1-\alpha_x)\times (|X(k,l)|^2 - P(k,l))^2/P^2(k,l) \quad (3)$$

is an estimation of the variance of the past periodograms normalized by $P^2(k,l)$. The smoothed periodogram $P(k,l)$ is considered as the local mean value of the past periodograms. If speech is absent, as mentioned above, the stationary exponential process has the variance $\lambda_n^2(k,l)$, and $P^2(k,l)$ is an estimation of $\lambda_n^2(k,l)$. Thence the *Normalized Variance* is a value around 1. When speech is present, specifically when the speech PSD changes abruptly, the variance of periodograms will be prominently larger than 1. This phenomenon can be observed in Fig. 1(a): a large number of noisy speech frames have larger *Normalized Variance* than noise frames. Compared with the real noise PSD $\lambda_n(k,l)$, the locally estimated mean value $P(k,l)$ is closer to the present periodogram. This causes *Normalized Variance* values of noise to be less than 1, as can be verified from Fig. 1(a).

The recursive *Normalized Differential Variance*

$$\vartheta_{ndv}(k,l) = \alpha_x\vartheta_{ndv}(k,l-1) + (1-\alpha_x)\times (|X(k,l)|^2 - |X(k,l-1)|^2)^2/P^2(k,l) \quad (4)$$

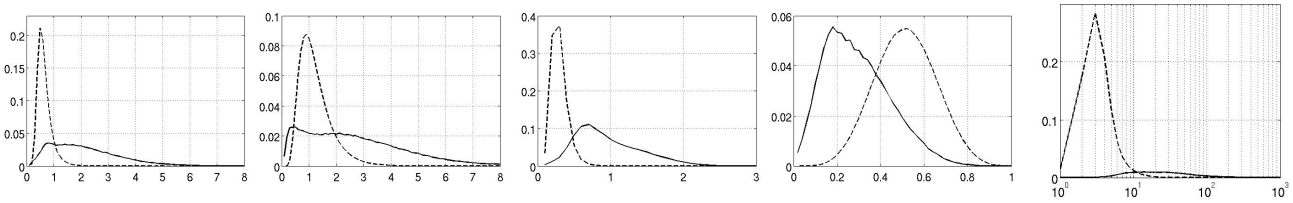
is an estimation of the normalized variance of the differential periodogram. When speech is absent, the noise periodograms are supposed to be i.i.d. random variables, thence the differential variable $|X(k,l)|^2 - |X(k,l-1)|^2$ is zero-mean and its variance is $2\lambda_n^2(k,l)$. Therefore, $\vartheta_{ndv}(k,l)$ is a value around 2. When speech is present, if the periodogram sequence is smooth due to the correlation between two adjacent noisy speech frames, $\vartheta_{ndv}(k,l)$ will tend to be smaller than 2, especially for harmonic components of voiced phonemes. Fig. 1(b) clearly shows that some noisy speech frames have smaller *Normalized Differential Variance* than noise frames, but that this is not always the case, probably due to speech onsets and offsets.

The *Normalized Average Variance* represents the normalized variance of the average periodogram:

$$\vartheta_{nav}(k,l) = \alpha_x\vartheta_{ndv}(k,l-1) + (1-\alpha_x)\times ((|X(k,l)|^2 + |X(k,l-1)|^2)/2 - P(k,l))^2/P^2(k,l). \quad (5)$$

When speech is absent, the average variable $(|X(k,l)|^2 + |X(k,l-1)|^2)/2$ has the mean value of $\lambda_n(k,l)$ and the variance of $0.5\lambda_n^2(k,l)$. Therefore, $\vartheta_{nav}(k,l)$ is a value around 0.5. For noisy speech, if two adjacent periodograms are correlated, the average periodogram $(|X(k,l)|^2+|X(k,l-1)|^2)/2$ has a value close to $|X(k,l)|^2$ or $|X(k,l-1)|^2$. Therefore, the variance of the average periodogram will be comparable with the one of original periodogram. This is different from the noise-only case. Comparing the probability density curve of the *Normalized Variance* to the one of *Normalized Average Variance* in Fig. 1(c) shows that the latter is a more distinctive feature between noise-only and noisy speech frames.

The *Median Crossing Rate* represents the rate at which the periodogram changes from positive to negative (and vice versa) with respect to the median value of the noise-only periodograms. The median is approximately set to $0.69P(k,l)$



(a) ϑ_{nv} (b) ϑ_{ndv} (c) ϑ_{nav} (d) ϑ_{mcr} (e) $\bar{d}$

Fig. 1: The probability density of regional statistics for noise-only (dashed line) and noisy speech (solid line). The vertical axis and horizontal axis are the probability density and the value of regional statistics, respectively.

according to the exponential distribution. It is computed as

$$\begin{aligned}\vartheta_{mcr}(k,l) &= \alpha_x \vartheta_{mcr}(k,l-1) + (1-\alpha_x) \\ &\times I\{(|X(k,l)|^2 - 0.69P(k,l)) \\ &\times (|X(k,l-1)|^2 - 0.69P(k,l-1)) < 0\} \quad (6)\end{aligned}$$

where the indicator function $I\{\cdot\}$ is 1 if its argument is true and 0 otherwise. In a short time period, compared with noise-only signal, the *Median Crossing Rate* is generally smaller for noisy speech periodograms. This can be due to the high correlation between successive speech frames as opposed to the uncorrelated noise frames. Even during speech onsets or offsets, the number of expected crossings in the frame is still very small. This phenomenon can be observed in Fig. 1(d).

Compared with stationary WGN, the behavior of non-stationary noise, such as babble noise, resembles the behavior of speech signal. Therefore, the probability density curves of the regional statistics will have more overlap between non-stationary noise and speech frames. This makes it more difficult to distinguish between noise frames and noisy speech frames.

2.2. Speech presence probability

In order to estimate the SPP using the four regional statistics features, we first concatenate them in a vector: $\boldsymbol{\vartheta}(k,l) = [\vartheta_{nv}(k,l), \vartheta_{ndv}(k,l), \vartheta_{nav}(k,l), \vartheta_{mcr}(k,l)]^T$, where ^{T} denotes vector transpose. Let $\bar{\boldsymbol{\vartheta}}$ and $\boldsymbol{\Sigma}$ denote the expectation vector and the covariance matrix of $\boldsymbol{\vartheta}(k,l)$ for noise-only signal, respectively. Since the regional statistics of noisy speech is expected to be significantly different from the ones of WNG (as opposed to non-stationary noise, e.g. babble), we have inferred $\bar{\boldsymbol{\vartheta}}$ and $\boldsymbol{\Sigma}$ from WNG signal. The normalized distance between an instantaneous $\boldsymbol{\vartheta}(k,l)$ and the expectation vector can then be computed as

$$d(k,l) = (\boldsymbol{\vartheta}(k,l) - \bar{\boldsymbol{\vartheta}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\vartheta}(k,l) - \bar{\boldsymbol{\vartheta}}). \quad (7)$$

Then the normalized distance is smoothed using the adjacent frequency bins and past frames, i.e. taking the average value of the normalized distances in the frame range of $[l-3, l]$ and frequency range of $[k-1, k+1]$. The smoothed distance is denoted as $\bar{d}(l,k)$. Fig. 1(e) shows the probability density of the smoothed distance. A small number of noisy speech frames with the smoothed distance in the range of 4-20 are ambiguous with respect to noise/speech frames classification.

The SPP can be computed using the smoothed distance:

$$p(k,l) = \begin{cases} 0, & \text{if } \bar{d}(l,k) \le \delta_1 \\ & \text{or } P(k,l) \le \hat{\lambda}_n(k,l-1) \\ \frac{\bar{d}(l,k)-\delta_1}{\delta_2-\delta_1}, & \text{if } \delta_1 < \bar{d}(l,k) < \delta_2 \\ & \text{and } |X(k,l)|^2/\hat{\lambda}_n(k,l-1) < 9.2 \\ 1, & \text{otherwise.} \end{cases}$$

where δ_1 and δ_2 are set to 4 and 8, respectively. This is motivated as follows: (1) If $\bar{d}(l,k) \le \delta_1$ (see Fig. 1(e)) or $P(k,l) \le \hat{\lambda}_n(k,l-1)$, there is a high confidence that speech is absent, thence $p(k,l)$ is set to 0. (2) If $\delta_1 < \bar{d}(l,k) < \delta_2$ (see Fig. 1(e)) and the *a posteriori* SNR is less than 9.2, it is uncertain whether speech is present or not. Thence $p(k,l)$ is set to be a value between 0 and 1. (3) For other cases, there is a high confidence that speech is present, thence $p(k,l)$ is set to 1.

3. EXPERIMENTS

In order to test the performance of the proposed method, two state-of-the-art noise PSD estimation methods are compared, i.e. the IMCRA method [7] and the MMSE-based method [9]. The clean speech originates from the TIMIT database [12] with a duration of 300s. Four types of noise are tested: computer generated stationary WGN, non-stationary WGN, factory noise and babble noise from the NOISEX92 database [13]. The non-stationary WGN is generated from the stationary WGN with fluctuating variance with amplitude of 20dB. Two change rates of the noise power are designed, i.e. 10dB/s and 2dB/s, respectively. Fig. 2 shows a cycle of non-stationary WGN, in which the noise is stationary between the increasing period and the decreasing period with a duration of 5s. The speech signal is contaminated by the various types of noise with SNRs 0, 5, 10 and 15dB, respectively.

To initialize our algorithm we have assumed that speech is absent at the first frame and that $P(k,1) = |X(k,1)|^2$, $\hat{\lambda}_n(k,1) = |X(k,1)|^2$, $p(k,1) = 0$ and $\boldsymbol{\vartheta}(k,1) = \bar{\boldsymbol{\vartheta}}$.

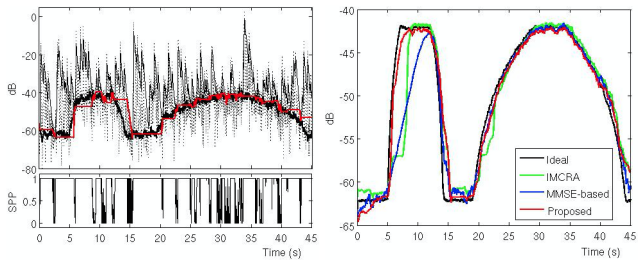


Fig. 2: An instance of noise PSD estimation for non-stationary WGN with 10dB input SNR. **Left:** Periodogram (dotted), smoothed periodogram (fine solid), smoothed ideal noise PSD (heavy solid) and the noise PSD estimation by the proposed method (red) for a single frequency bin with the center frequency 844 Hz. The SPP is shown in the lower graph. **Right:** The averaged noise PSD across frequency.

The symmetric segmental logarithmic error (LogErr, in dB) [8], [9] is taken as the criterion for evaluating the noise PSD tracking performance. The frequency range of up to 8 kHz is taken into consideration for performance evaluation. The ideal noise PSD is obtained using the smoothed noise periodogram with smoothing parameter $\alpha_x = 0.85$.

In order to evaluate the performance of speech enhancement, the noise PSD estimation is used in a speech enhancement algorithm [1]. For all three PSD estimation methods, the speech coefficients are estimated by the well-known MMSE amplitude estimator [1], and the *a priori* SNR is estimated by the decision-directed approach [1] with the smoothing parameter 0.98. After noise reduction, the output segmental SNR (SNR_{seg}, in dB) [8] is taken as the performance criterion.

Significance of each regional statistic. The function of regional statistics is differentiating noise components and noisy speech components. To assess the usefulness of each feature of the regional statistics, we respectively remove each feature, and only the three remaining features are used for noise PSD estimation. The average LogErr (for all four types of noise and four SNRs) for the four feature selection options are 2.191, 2.190, 2.203 and 2.185 dB, respectively. The average LogErr when all four features are used is 2.171 dB. This demonstrates that each feature contributes the discrimination between noise-only and noisy speech components.

Noise PSD estimation results. Fig. 2 shows an instance of noise PSD estimation. From Fig. 2(a), it can be seen that the proposed method tracks the noise PSD changes quite well, even for the case when the noise power abruptly rises. Fig. 2(b) depicts the averaged noise PSD. When the noise power rises slowly (from 19s to 29s, with increasing rate of 2dB/s), the MMSE and the proposed method can reliably track the increasing noise power, and achieve similar performance measures. The IMCRA method exhibits a slight tracking delay. When the noise power rises rapidly (from 5s to 7s, with increasing rate of 10dB/s), the IMCRA method has a long response time due to the search window, and the MMSE method also takes a long period of time to keep up with the ideal noise power by its *safety-net* procedure. The proposed method adjusts the smoothing parameter using the accurate estimated SPP, and obtains a better tracking performance.

noise source	input SNR(dB)	LogErr (dB) IMCRA	MMSE	Prop.	SNR_{seg} (dB) IMCRA	MMSE	Prop.
stationary WGN	15	1.16	1.21	1.18	18.03	17.41	17.00
	10	1.03	1.12	0.97	15.16	14.54	14.48
	5	0.94	1.05	0.80	12.46	11.85	11.63
	0	0.88	1.00	0.67	9.82	9.25	9.27
non-stationary WGN	15	2.86	2.46	2.33	16.14	15.77	15.66
	10	2.65	2.23	1.74	12.31	12.33	12.91
	5	2.52	2.02	1.31	7.86	8.70	10.43
	0	2.43	1.85	1.02	2.73	4.66	8.09
factory noise	15	3.48	2.99	3.29	16.19	16.13	15.71
	10	3.56	2.95	3.01	12.43	12.75	12.45
	5	3.69	2.93	2.82	8.32	9.28	9.16
	0	3.83	2.93	2.67	3.76	5.56	5.70
babble noise	15	3.53	2.95	3.79	10.89	15.55	15.32
	10	3.33	2.85	3.34	5.85	11.73	11.66
	5	3.38	2.85	3.06	0.60	7.76	7.82
	0	3.53	2.88	2.84	-4.82	3.21	3.62

Table 1: Performance in terms of LogErr and SNR_{seg} for various noise sources and input SNRs (dB).

Table 1 summarizes the results of the three noise PSD estimation methods. For stationary WGN, the proposed method obtains the smallest error, since most of noise/speech frames are correctly classified. However, it achieves a similar output SNR compared with the other methods. For nonstationary WGN, the proposed method achieves the smallest error and the largest output SNR. The reason is that the proposed method has a better tracking capability for the rapidly increasing noise, as shown in Fig. 2. The expectation vector $\bar{\boldsymbol{\vartheta}}$ and the covariance matrix $\boldsymbol{\Sigma}$ are inferred from stationary WGN. However, the distinction between regional statistics feature vector of nonstationary noise and speech becomes smaller, thence the performance of the proposed method for nonstationary noise is worse than the performance for stationary WGN. For nonstationary factory noise and babble noise, the MMSE-based and the proposed method outperform the IMCRA method considerably, and they achieve comparable performance. Roughly speaking, the proposed method performs better when the input SNR is low (less than 10dB).

4. CONCLUSION

In this paper, we have proposed a noise PSD estimation method, in which the speech presence probability is estimated by the regional statistics. The statistics of the past periodograms in a short-time period are valid to differentiate the noise-only and noisy speech components. This SPP estimator refers only to the information of the past (0.2 s approximately), therefore it can rapidly respond to the change of noise level. Experiments demonstrate that the regional statistics efficiently track the non-stationary noise.

5. REFERENCES

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

[2] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[3] Jan S Erkelens, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[4] Rainer Martin, "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, 2001.

[5] Rainer Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, no. 6, pp. 1215–1229, 2006.

[6] Israel Cohen and Baruch Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[7] Israel Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[8] Richard C Hendriks, Jesper Jensen, and Richard Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 541–553, 2008.

[9] Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "MMSE based noise psd tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4266–4269.

[10] Timo Gerkmann and Richard C Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[11] Israel Cohen and Sharon Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, pp. 873–902. Springer, 2008.

[12] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.

[13] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.