Selflearning Codebook Speech Enhancement

Florian Heese, Christoph Matthias Nelke, Markus Niermann, and Peter Vary

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, 52056 Aachen, Germany
E-Mail: {heese|nelke|niermann|vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

A novel speech enhancement system is presented which exploits a codebook for noise estimation. In contrast to state-of-the-art noise estimators which usually rely on the assumption that the noise signal is only slightly time-varying, codebook approaches allow also non-stationary environments. The basic concept of the proposed codebook noise estimation is a superposition of a scaled speech and noise codebook entry. In order to be independent of *a priori* noise knowledge, the new estimator is able to learn new noise types online. Training vectors for codebook updates are identified using a speech activity detector (VAD) and a codebook mismatch measure. The VAD is realized as part of the codebook matching. A Wiener filter or any state-of-the-art weighting rule can be applied subsequently for speech enhancement. Experiments confirmed that the new system is able to learn new noise types and provides improved performance compared to state-of-the-art algorithms.

1 Introduction

Since communication is mobile it takes place at many different locations. As a result speech intelligibility and quality may significantly be degraded by the presence of background noise such as traffic, wind, engine, babble and construction site noise.

There are established techniques for enhancing degraded speech. One of the popular methods represents the noisy signal in the short-time Fourier domain and applies individual adaptive gains to each frequency bin based on a noise power spectral density (PSD) estimation, e.g., [1–3]. Other approaches such as beamforming exploit multi-channel techniques, where multiple microphones are placed at different positions to exploit the spatial information [4–6]. Their common aim is to suppress noise while preserving the speech as purely as possible.

Single channel systems usually rely on the assumption that background noise is stationary or only slightly time-varying [2, 7, 8] which is often not fulfilled. The class of codebook based enhancement systems [9–11] faces this constraint by *a priori* knowledge about speech and noise. Spectral speech and noise estimates are obtained by a linear combination or a weighted sum of entries from pre-trained codebooks. However, the performance is limited depending on the quality of the codebook matching which is mainly degraded either by missing *a priori* knowledge (especially with respect to noise) or deviations due to the signal transmission path, i.e., the (changing) acoustic and electrical (recording equipment, microphone) path. In [11] the adaption of codebooks focuses on compensating the influence of the transmission path while in [12] fixed delta codebooks between the actual and a conventional noise estimate (e.g., [2, 7, 8]) are employed to reduce the effect of missing *a priori* noise knowledge.

In this contribution a codebook speech enhancement system is developed which adapts new noise types online and therefore relies only on speech *a priori* knowledge. The remainder of this paper is organized as follows. A brief overview of the proposed speech enhancement system is given in Sec. 2. In Sec. 3 the concept of the codebook based noise PSD estimation is presented. Experimental results are presented in Sec. 4 and conclusions are drawn in Sec. 5.

2 System Overview

In Fig. 1 a simplified block diagram of the proposed noise reduction system is given. It is assumed that the noisy input signal $x(k)$ consists of a clean speech signal $s(k)$ degraded by an additive noise component $n(k)$ according to:

$$x(k) = s(k) + n(k). \tag{1}$$

The samples $x(k)$ are obtained by analog-digital conversion with a sampling frequency of $f_s = 16\,\text{kHz}$.

The noise suppression relies on a codebook based noise PSD estimation which is performed in the frequency domain. Hence, $x(k)$ is segmented into overlapping frames of length L_F, followed by windowing (square root Hann-window) and zero-padding. Subsequently each frame is transformed by applying the Fast Fourier Transform (FFT) of length M_F. The spectral coefficients of the input signal $x(k)$ at frequency bin μ and frame λ are given by:

$$X(\lambda,\mu) = S(\lambda,\mu) + N(\lambda,\mu), \tag{2}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ correspond to the spectral coefficients of the speech and noise signal.

The minimization of the distance between the noisy input frame $X(\lambda,\mu)$ and an estimate $\hat{X}(\lambda,\mu)$ which is a superposition of scaled speech and noise codebook entries according to

$$\hat{X} = \sigma_s \tilde{S}_l e^{i\varphi_s} + \sigma_n \tilde{N}_m e^{i\varphi_n}, \tag{3}$$

delivers the estimate $|\hat{N}(\lambda,\mu)|^2 = \sigma_n^2 \tilde{N}_m^2$ of the current noise PSD, where l,m denote the codebook indices and σ_s, σ_n the gain factors of speech and noise, respectively. While the speech codebook is pre-trained in advance, the noise codebook is adapted to new noise types online. Essential for a noise codebook update is a mismatch measure and the absence of speech. This requires a voice activity detection (VAD) in addition, which is provided by means of the speech codebook. The codebooks are created by training of vector quantizers (refer to Sec. 3.1).

Using the noise PSD estimate $|\hat{N}(\lambda,\mu)|^2$ two SNR parameters, namely the *a posteriori* SNR $\gamma(\lambda,\mu)$ and the *a priori* SNR $\xi(\lambda,\mu)$ defined as:

$$\gamma(\lambda,\mu) = \frac{|X(\lambda,\mu)|^2}{|\hat{N}(\lambda,\mu)|^2}, \xi(\lambda,\mu) = \frac{\mathscr{E}\{|S(\lambda,\mu)|^2\}}{|\hat{N}(\lambda,\mu)|^2}, \tag{4}$$

ISBN 978-3-8007-3640-9

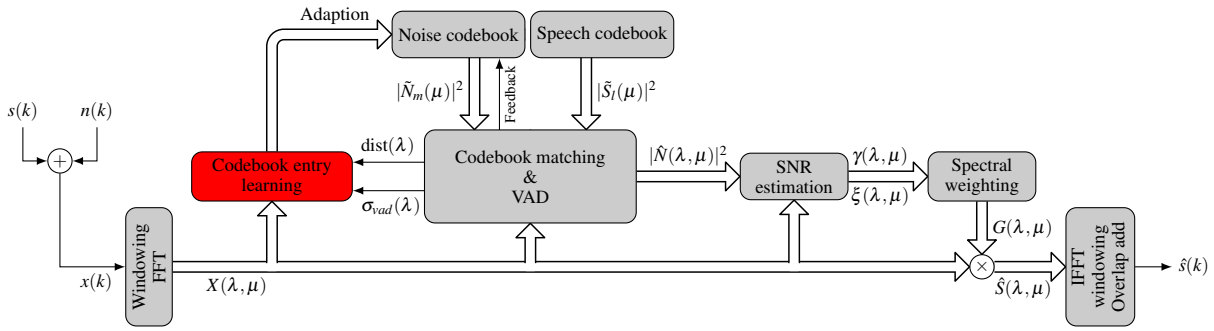

Figure 1: Proposed speech enhancement system using a codebook based noise PSD estimator

can be derived. For the *a priori* SNR estimation the decision-directed estimation approach [1] is applied. Finally, by multiplying the noisy spectrum $X(\lambda,\mu)$ with spectral gains $G(\lambda,\mu)$:

$$\hat{S}(\lambda,\mu) = G(\lambda,\mu)\cdot X(\lambda,\mu), \tag{5}$$

the noise suppression is achieved. As spectral gains the well-known Wiener filter is utilized which is dependent on the SNR estimates. The enhanced time domain signal $\hat{s}(k)$ is obtained by applying an Inverse Fast Fourier Transform (IFFT), windowing (square root hann window) and overlap-add.

3 Noise Estimation

Given an additively disturbed speech signal the magnitude square of the resulting noisy signal can be expressed as:

$$\begin{aligned} |X(\mu)|^2 &= \left|\left(|S|e^{i\varphi_s} + |N|e^{i\varphi_n}\right)\right|^2 \\ &= |S|^2 + |N|^2 + 2\,|S|\,|N|\cos\left(\varphi_s - \varphi_n\right). \end{aligned} \tag{6}$$

The phase difference $\varphi = \varphi_s - \varphi_n$ is unknown. According to measurements with plain speech and noise it is considered to be an equally distributed random variable on $[0, 2\pi]$. Because of $E\{\cos(\varphi)\} = 0$ the cross-term is omitted in the following due to averaging in the SNR estimation. Experiments have confirmed that the resulting error with respect to the noise reduction is negligible.

The noise estimation is performed by estimating the PSD of current noisy frame $|X(\lambda,\mu)|^2$ by an optimized superposition of a speech and noise codebook entry:

$$|\hat{X}(\mu)|^2 = \sigma_s^2|\tilde{S}_l(\mu)|^2 + \sigma_n^2|\tilde{N}_m(\mu)|^2 \tag{7}$$

where l, m denote the codebook entry indices and $\sigma_s \geq 0$, $\sigma_n \geq 0$ the corresponding scaling factors. The codebook entries $\tilde{S}_l(\mu)$, $\tilde{N}_m(\mu)$, with $l \in \{1,\ldots,L\}$, $m \in \{1,\ldots,M\}$ are normalized to one with respect to their energy and consist of spectral magnitudes. Thus, the gain factors σ_n^2 and σ_s^2 represent the speech and short-term noise energy. In order to reduce the speaker dependence only spectral envelopes are stored as speech codebook entries. An estimate of the noise PSD for the current frame thus is given by $\sigma_n^2\left|\tilde{N}_l(\mu)\right|^2$. With the assumption

$$\sum_{\mu}\left|\hat{X}_{l,m,\sigma_s,\sigma_n}(\mu)\right|^2 \approx \sum_{\mu}|X(\mu)|^2 =: \sigma_x^2, \tag{8}$$

the speech gain σ_s can be substituted and Eqn. 7 simplifies to:

$$\left|\hat{X}_{l,m,\sigma_n}(\mu)\right|^2 = \left(\sigma_x^2 - \sigma_n^2\right)\left|\tilde{S}_l(\mu)\right|^2 + \sigma_n^2\left|\tilde{N}_m(\mu)\right|^2, \tag{9}$$

which reduces the number of parameters to be optimized and the computational expense. Techniques known from gain shape quantization to determine the codebook entry and gain independently cannot be applied since the shape of $\left|\hat{X}_{l,m,\sigma_n}(\mu)\right|^2$ strongly depends on the gain factors. Hence, all permutations of the parameters l, m, and σ_n must be taken into account, which can be realized by a quantization of σ_n according to:

$$\sigma_n = \frac{i}{p-1}\sigma_x \quad i = 0,\ldots,p-1. \tag{10}$$

Finally, the optimal parameters $l_{opt}, m_{opt}, \sigma_{n,opt}$ can be found by minimizing:

$$\arg\min_{l,m,\sigma_n} \operatorname{dist}\left(|X(\mu)|^2, \left|\hat{X}_{l,m,\sigma_n}(\mu)\right|^2\right). \tag{11}$$

We use as in [11] the Itakura-Saito-distance measure:

$$\operatorname{dist}\left(P(\mu),\hat{P}(\mu)\right) = \sum_{\mu=0}^{M-1}\left[\frac{P(\mu)}{\hat{P}(\mu)} - \log\frac{P(\mu)}{\hat{P}(\mu)} - 1\right]. \tag{12}$$

3.1 Codebook Training

A training sequence segmented in overlapping and windowed frames is transformed into the frequency domain according to Sec. 2. After taking the magnitude square operation, all resulting PSD frames below a certain energy threshold are discarded. This removes silent parts of the training data which may be over-represented in the later vector quantization and prevents for frames with upscaled recording noise after the subsequent energy normalization.

Applying this procedure, a large amount of vectors (frames) exists which are used for the training of a vector quantizer (VQ). The result of the VQ training is used as codebook. In this work the LBG algorithm [13] is employed together with the Itakuro Saito distance (Eqn. 12) as distance measure.

While this procedure is useful for the noise codebook creation, an extension is necessary for the speech codebook training to reduce the speaker dependence. Therefor, the spectral envelope is calculated using a cepstrum approach as in [11] before the quantization process, resulting in a speaker independent codebook.

3.2 Voice Activity Detection

The performance of many noise robust VAD systems rely on the quality of an underlying noise PSD estimation. Since the VAD is required by the noise PSD estimation system a VAD is developed which is independent of a noise

ISBN 978-3-8007-3640-9

estimation. The new VAD is based on the codebook matching with the speech gain, named σ_{vad} in this context, as speech presence indicator. In speech pauses the gain is assumed to be very small since no suitable codebook entry can be found whereas in frames with speech activity the gain will be significantly larger.

In contrast to Sec. 3 the distance minimization is carried out using only the speech codebook:

$$\arg\min_{l,\sigma_s} \text{dist}\left(|X(\mu)|^2, \left|\sigma_{vad}\tilde{S}_l(\mu)\right|^2\right), \tag{13}$$

with $\sigma_s = \frac{i}{p-1}\sigma_x$, $i = 0,...,p-1$. For simplicity we use the relative energy difference

$$\text{dist}\left(P(\mu), \hat{P}(\mu)\right) = \frac{1}{\sum_{\mu=0}^{M-1} P(\mu)} \sum_{\mu=0}^{N-1} \left|P(\mu) - \hat{P}(\mu)\right|, \tag{14}$$

as distance measure. Afterwards the speech gain σ_{vad} is smoothed recursively by:

$$\bar{\sigma}^2_{vad}(\lambda) = \alpha\bar{\sigma}^2_{vad}(\lambda-1) + (1-\alpha)\sigma^2_{vad}(\lambda). \tag{15}$$

The smoothing parameter $0 < \alpha < 1$ determines the smoothing intensity and is chosen different for falling or rising values:

$$\alpha = \begin{cases} \alpha_+ & \sigma^2_{vad}(\lambda) \geq \bar{\sigma}^2_{vad}(\lambda-1) \\ \alpha_- & \sigma^2_{vad}(\lambda) < \bar{\sigma}^2_{vad}(\lambda-1). \end{cases} \tag{16}$$

The parameters are set to $\alpha_+ = 0.6$ and $\alpha_- = 0.95$ in the following. This ensures a fast rising of the speech indicator at sudden speech activity with a slow decay resulting in a slight overestimation of speech presence which is desired in the case of noise codebook adaption to assure no adaption while speech presence.

3.3 Noise Codebook Adaptation

Since the noise environment is not known *a priori* an online adaption of the noise codebook is required. Therefore, training sequences with the unknown noise types are essential. They can be found if speech is absent and the mismatch defined as $\text{dist}(|X(\mu)|^2, |\hat{X}(\mu)|^2)$ exceeds a threshold. In addition the following conditions must match:

- A frame is assumed of voice activity if $\bar{\sigma}_{vad} > 2.7$,
- Training frames must not contain speech. Therefor, a hangover frame distance to the last VAD frame of L_s is introduced,
- A frame is classified as new noisy type if the mismatch $\text{dist}(|X(\mu)|^2, |\hat{X}(\mu)|^2) > 1$
- The distance measure evaluation of the last L_T frames must have detected an unknown noise sound, i.e., T percent of the last L_T frames exceed the distance threshold T_D
- A safety margin between two adaptions of frame length L_A has to be kept.

Given at least L_T frames in the past which satisfy these conditions the same vector training as in Sec. 3 is utilized to obtain M_Δ new codebook entries which are then combined with the noise codebook. If the maximum defined noise codebook size M_{max} is exceeded, the less used entries of the last L_R frames are discarded.

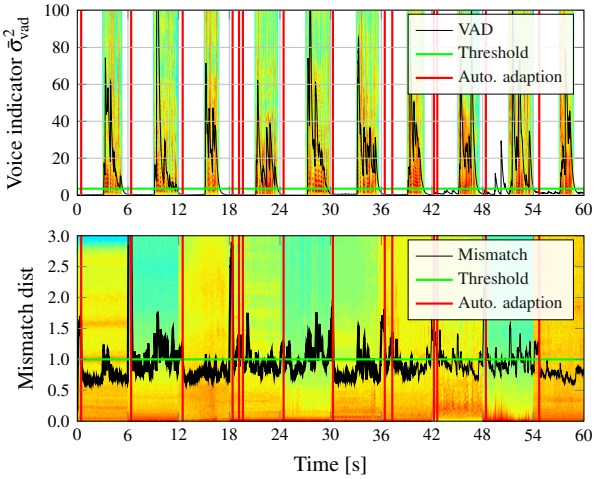

Figure 2: Example of online noise codebook adaption learning ten unknown noise types (SNR 0 dB). The upper plot shows the VAD performance while the lower plot presents the codebook mismatch. Red vertical lines indicate codebook update times

4 Results

A prove of concept which illustrates the performance of the online codebook adaption is given in Fig. 2. A noisy input signal is generated consisting of ten different six seconds long stationary and non-stationary noise types mixed with five male and female english speakers taken from the TIMIT corpus [14] at an SNR of 0 dB. Since the noise codebook is initialized with a white noise codebook entry it has to be adapted every six seconds. Tab. 1 shows the parameters for the simulation and the codebook algorithm settings are summarized in Tab. 2, with a maximum noise codebook size of $M = 28$. The upper plot depicts the clean speech spectogram of the input signal to emphasize the performance of the VAD in terms of $\bar{\sigma}_{vad}$ (black line). Apart from wind noise (arround 49 s) the presented new VAD algorithm provides reliable decisions. Vertical red lines indicate a codebook adaption which uses each time the past 40 frames as training sequence, while green lines indicate the thresholds for VAD frame indicator and codebook mismatch, respectively. In the lower plot of Fig. 2 the noise-only spectogram of the input signal can be seen together with the codebook mismatch indicator (black line). It is obvious that each noise change is detected and the noise codebook is adapted accordingly. In seven out of ten cases a single adaptation is sufficient while repeatedly adapting is necessary in the remaining cases, which reflects a fast changing characteristic of the noise. It is also apparent that adaptation takes exclusively place in speech pauses while always a certain safety distance to speech activity frames is maintained which preserves the noise codebook from speech.

Parameter	**Settings**
Sampling frequency	16 kHz
Frame length L_F	320 ($\hat{=}$ 20 ms)
FFT length M_F	512 (including zero-padding)
Frame overlap	50% ($\sqrt{\text{Hann}}$ – window)
SNR estimation	Decision-directed approach [1]

Table 1: System settings

ISBN 978-3-8007-3640-9

Parameter	Settings
Training frames L_T	40
Distance threshold T_D	1
VQ output size M_Δ	4 codebook entries
Hangover VAD margin L_S	60 frames
Adaption margin L_A	40 frames
Hit rate T	95%
Speech codebook size L	128 entries
Histogram window L_R	500 frames

Table 2: Codebook algorithm parameters

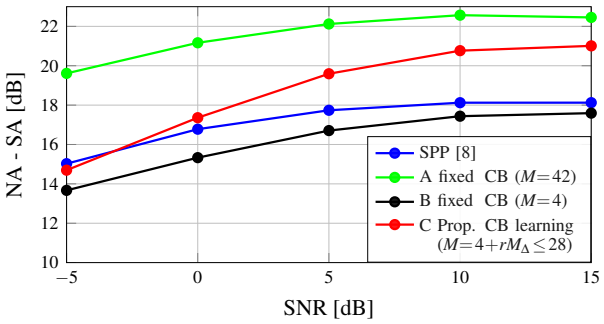

Figure 3: Difference between noise attenuation and speech attenuation plotted over input SNR

In addition a benchmark is carried out to compare the proposed noise PSD estimator in three different configurations two fixed noise codebooks and the proposed adaptive with the state of the art SPP noise tracker [8]. For all configurations the speech codebook was trained with 400 sentences taken from the TIMIT corpus [14]. The performance is measured in terms of segmental noise attenuation (NA) minus speech attenuation (SA) [15] using the noise reduction system depicted in Fig. 1. Higher values indicate a better performance. Therefor the noise estimate for the SNR estimation stage is adapted for the different algorithms. The benchmark consists of all permutations of the following parameters: the input SNR varies from -5 to 15 dB in 5 dB steps and four male and female english speakers (disjunct with the training data) are mixed with ten different stationary and non-stationary noise types. Each permutation is performed independently and begins with 3 s of speech pause to allow for codebook adaption. Configuration A exhibits a pre-trained large ($M=42$) noise codebook consisting of four entries for each noise type (6 s training sequence) together with white and pink noise entries while in configuration B the codebook from A is condensed to $M=4$ entries as representative for a small fixed codebook. In C the proposed online learning is applied with the noise codebook from B as initialization and a codebook size of $M=4+rM_\Delta \leq 28$, with r the number of online updates. The parameters of the SPP algorithm are chosen as suggested in [8]. The configurations of the simulation and the codebook matching algorithm remain (see Tab. 1, 2).

In Fig. 3 the result of the benchmark is depicted. As expected, configuration A with the best *a priori* knowledge defines the upper bound over the complete SNR range while for B the performance is only comparable to the SPP approach. The system with online adaption (C) is superior compared to system B and the SPP. This clearly demonstrates the advantage of the proposed algorithm. Since a reliable VAD (needed for adaption) can be expected for SNRs greater than 0 dB the performance gain grows with increasing SNR.

5 Conclusions

A codebook based speech enhancement system was presented which is capable of learning new noise types online under the constraint of a given maximum codebook size. By minimizing a distance between the noisy input and a scaled superposition of clean speech and noise codebook entries, a noise PSD estimate can be found. As mismatch the Itakuro Saito distance between the codebook choice and the noisy input turned out to be favorable. Adapting the noise codebook online requires for a mismatch measure and a reliable VAD to determine training sequence. Given a training sequence new noise codebook entries can be calculated by vector quantizer training. The VAD can be computed easily by a slight modification of the codebook matching algorithm and thus is inherent integrated. It reduces the codebook matching to only the speech codebook and employs the determined smoothed speech energy as speech activity indicator providing a stable VAD. Instrumental measurements confirmed a consistent improvement compared to the state-of-the-art SPP algorithm and a fixed noise codebook system with averaged *a priori* knowledge.

References

[1] 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. 32, no. 6, p. 1109–1121, 1984.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, p. 504–512, 2001.

[3] T. Esch, *Model-based speech enhancement exploiting temporal and spectral dependencies*. No. 32 in Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN), Verlag Mainz in Aachen, 2012.

[4] F. Heese, T. Esch, and P. Vary, "Dual channel reduction of rapidly varying harmonic and random noise using a spot microphone," in *ESSV, Aachen, Germany*, 2011.

[5] S. Haykin and K. R. Liu, *Handbook on array processing and sensor networks*, vol. 63. John Wiley & Sons, 2010.

[6] F. Heese, M. Schäfer, P. Vary, E. Hadad, S. Golan, and S. Gannot, "Comparison of supervised and semi-supervised beamformers using real audio recordings," in *2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, pp. 1 –5, Nov. 2012.

[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, p. 4266–4269, 2010.

[8] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 145–148, 2011.

[9] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 441 –452, Feb. 2007.

[11] T. Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv*, 2010.

[12] T. Rosenkranz and H. Puder, "Improving robustness of codebook-based noise estimation approaches with delta codebooks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, p. 1177–1188, 2012.

[13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, p. 84–95, 1980.

[14] J. S. Garofolo and L. D. Consortium, *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1993.

[15] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ, 1988.