A SUPERVISED APPROACH TO GLOBAL SIGNAL-TO-NOISE RATIO ESTIMATION FOR WHISPERED AND PATHOLOGICAL VOICES

Amir Hossein Poorjam [1] *, Max A. Little* [2,3]*, Jesper Rindom Jensen* [1] *and Mads Græsbøll Christensen* [1]

[1] Audio Analysis Lab, CREATE, Aalborg University, Aalborg, DK
[2] Engineering and Applied Science, Aston University, Birmingham, UK
[3] Media Lab, MIT, Cambridge, Massachusetts, USA
[1] {ahp, jrj, mgc}@create.aau.dk , [2] max.little@aston.ac.uk

ABSTRACT

The presence of background noise in signals adversely affects the performance of many speech-based algorithms. Accurate estimation of signal-to-noise-ratio (SNR), as a measure of noise level in a signal, can help in compensating for noise effects. Most existing SNR estimation methods have been developed for normal speech and might not provide accurate estimation for special speech types such as whispered or disordered voices, particularly, when they are corrupted by non-stationary noises. In this paper, we first investigate the impact of stationary and non-stationary noise on the behavior of mel-frequency cepstral coefficients (MFCCs) extracted from normal, whispered and pathological voices. We demonstrate that, regardless of the speech type, the mean and the covariance of MFCCs are predictably modified by additive noise and the amount of change is related to the noise level. Then, we propose a new supervised method for SNR estimation which is based on a regression model trained on MFCCs of the noisy signals. Experimental results show that the proposed approach provides accurate estimation and consistent performance for various speech types under different noise conditions.

Index Terms— Global SNR estimation, pathological voice, whispered speech, MFCC, support vector regression

1. INTRODUCTION

The performance of many speech-based systems is degraded by acoustical background noise in signals. Information about the noise level can help in compensating for its effects. The signal-to-noise ratio (SNR), which measures the level of noise in a speech signal, is defined as the ratio of signal power to noise power, typically in decibels (dB). In practice, the speech SNR should be estimated since we only have access to the noisy signals. This estimation is, however, challenging since speech, which is a highly non-stationary signal, is typically corrupted by a variety of unknown noise.

The speech SNR, in a broad sense, can be classified into two categories, namely segmental and global SNR. Techniques calculating the segmental SNR estimate the noise level at short frames of approximately 30 ms in which the signal is assumed to be stationary. This type of SNR has attracted more attention in decades due to its direct application in speech enhancement, noise estimation and noise suppression [1, 2]. Global SNR estimation algorithms, on the other hand, consider the entire signal and provide information about the effect of noise on the whole recording. In this study we focus on the global SNR estimation which is beneficial in many SNR-specific applications such as environmental sniffing [3], speech recognition [4] and speaker verification [5] in which the noise level at the entire signal affects the model selection process.

This work was funded by Independent Research Fund Denmark: DFF 4184-00056

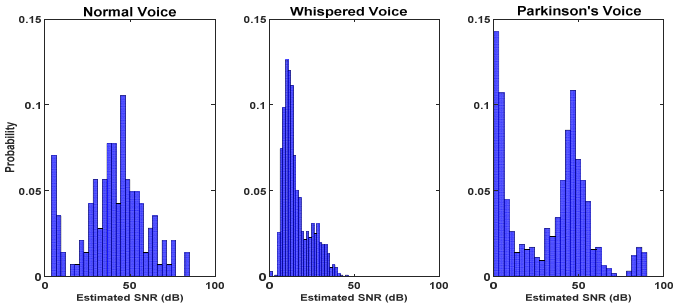

Fig. 1: The normalized histograms of estimated SNR values for three different clean databases using the NIST algorithm [7].

Most existing global SNR estimation algorithms are based on measuring the energy content of speech and non-speech regions in a signal. These regions can be identified by a variety of techniques such as speech activity detection [6], sequential Gaussian mixture estimation [7], Gaussian mixture modeling in the log-power domain [8], ideal binary masks [9–12], and long-term acoustic features [13]. Kim and Stern have proposed an approach, called waveform amplitude distribution analysis (WADA), to estimate the SNR from the value of the shaping parameter of a Gamma distribution fitted to a noisy speech [14]. However, these methods have difficulties dealing with some speech types such as sustained vowels (in which there is no regular pauses), whispered speech (from which it is difficult to accurately identify speech and non-speech regions), and pathological voice (in which the distortion due to vocal disorder is considered as noise even if it is recorded in a noise-free environment). Fig.1 shows the normalized histograms of estimated SNR values for three clean databases, namely normal speech, whispered speech and Parkinson's voice (described in Section 4.1) using the NIST algorithm [7]. Since the signals of these databases have been recorded in noise-free environments, we expect high SNR values for the recordings. However, we can observe from the plots that the NIST SNR estimation algorithm has failed to correctly estimate the global SNR for a large amount of whispered and disordered recordings. Therefore, a more robust algorithm is required to deal with various speech types.

In this study, we consider mel-frequency cepstral coefficients (MFCCs), which are known to be very sensitive to a small change in signal characteristics due to noise and other variabilities [15], and investigate the effect of stationary and non-stationary noise at different levels on the behavior of MFCCs extracted from normal speech, whispered speech and disordered voice recordings. We show

through the experimental analysis that the mean and the covariance of MFCCs are predictably modified by additive noise and the amount of change is related to the level of noise in a speech signal. Motivated by the experimental observations, a new supervised approach to estimate the global SNR is then developed. In this method, instead of identifying speech and non-speech regions in a signal, MFCCs extracted from noisy recordings are utilized to train a regression model. The SNR value for an unseen recording is then estimated using the trained model. This is useful, for example, in SNR-specific applications in which the train-test SNR conditions should be matched.

2. IMPACT OF ADDITIVE NOISE ON MFCCS

MFCCs are based on the source-filter theory of speech production [16]. To compute MFCCs, we take the discrete Fourier transform (DFT) of the speech frames. Then, the power spectrum is computed and passed through a set of triangular filter banks, linearly spaced on the mel-frequency scale. The log-energy output of the filter bank, which is sensitive to small changes in signal characteristics due to noise or articulatory movements, is then passed through the discrete cosine transform (DCT). The MFCCs are the amplitudes of the DCT coefficients. The effects of additive noise on MFCCs are complex since MFCC calculations involve several nonlinear functions. In [17], the effect of additive white Gaussian noise on the MFCC parameters is studied. It has been shown that the variance of the error in MFCC computation due to adding white Gaussian noise is related to the variance of the noise.

In this section, we investigate the effect of stationary and non-stationary noises on normal, whispered and disordered speech recordings. Specifically, for normal voices, we take 100 clean speech recordings of the LibriSpeech database [18]. For whispered speech, we take 36 clean recordings from the CHAINS database [19]. For pathological voice, we take clean recordings of the sustained vowel /a/ produced by 100 PD patients. We then corrupt them by white Gaussian and babble noises under different SNR conditions ranging from -20 dB to 60 dB in 1 dB steps. Using a Hamming window, recordings are segmented into frames of 30 ms. For each frame of a signal, 12 MFCCs together with the log energy are calculated along with *delta* and *double-delta* coefficients. They are concatenated to produce a 39-dimensional vector. We then evaluate the shift in the sample mean and covariances of the MFCCs computed from the noisy signals. Specifically, the mean shift can be defined as:

$$\xi(i) = \frac{1}{M}\sum_{m=1}^{M} \parallel \boldsymbol{\mu}_m^{\mathrm{n}_i} - \boldsymbol{\mu}_n^{\mathrm{c}} \parallel_2, \qquad (1)$$

where M is the number of speakers, $\|\cdot\|_2$ represents the 2-norm, and $\boldsymbol{\mu}_m^{\mathrm{c}}$ and $\boldsymbol{\mu}_n^{\mathrm{n}_i}$ are the means of the MFCCs computed respectively from the clean and noisy signals from the m^{th} speaker subject to the i^{th} noise level. The larger the value of ξ, the farther the MFCC vector is moved with respect to the clean one. The change in the covariance matrix of the MFCC under the i^{th} noise level is measured as:

$$\delta(i) = \frac{1}{M}\sum_{m=1}^{M} \frac{\parallel \boldsymbol{\Sigma}_m^{\mathrm{n}_i} - \boldsymbol{\Sigma}_m^{\mathrm{c}} \parallel_F}{\parallel \boldsymbol{\Sigma}_m^{\mathrm{c}} \parallel_F}, \qquad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm which, for an arbitrary matrix $\boldsymbol{A}$ with elements a_{pq}, is defined as $\|\boldsymbol{A}\|_F = \sqrt{\sum_{p=1}^{P}\sum_{q=1}^{Q}|a_{pq}|^2}$ which maps a matrix to a single real number, and $\boldsymbol{\Sigma}_n^{\mathrm{c}}$ and $\boldsymbol{\Sigma}_n^{\mathrm{n}_i}$ are respectively the covariance matrices of the MFCCs extracted from the clean and the noisy utterances of the m^{th} speaker. $\delta = 0$ represents no change in covariance. A value of $\delta < 1$ indicates a reduction in covariance with respect to the covariance of the clean MFCC. That is, the MFCCs become more compact in the feature space.

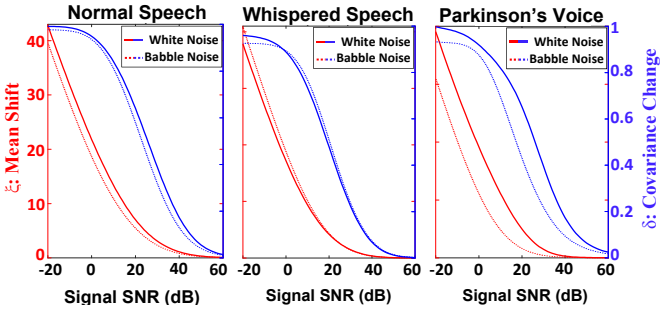

Fig. 2: Impact of stationary and non-stationary noises at different levels on the mean and the covariance matrix of MFCCs of the normal, whispered and pathological voices. The left vertical axes represent the amount of mean shift defined in (1). The right vertical axes represent the relative change in the covariance matrix defined in (2).

Fig.2 shows the impact of the stationary and non-stationary noises at different levels on the mean and the covariance matrix of MFCCs. The left vertical axes represent the amount of mean shift as defined in (1) and the right vertical axes represent the relative change in the covariance matrix as defined in (2). The plots suggest that variable noise levels shift the mean of MFCCs to different, but predictable, regions in the feature space. It can be noticed that the amount of shift monotonically increases as the noise level increases and it has a linear relation with the noise level for the SNRs almost below 30 dB, except for the Parkinson's voices corrupted by the babble noise where it is below 15 dB. Notice that, for the tested conditions, the covariance of the noisy MFCCs is always smaller than that of the clean one. A linear relation between the noise level and the covariance change can be observed when the noise level is almost between 0 dB and 40 dB.

3. THE PROPOSED SNR ESTIMATION APPROACH

The experimental analysis above illustrates similar trends for the behavior of MFCCs of various speech types under different noise conditions. Motivated by this observation, we develop a new method for global speech SNR estimation based on the MFCCs. In this approach, instead of identifying speech and non-speech regions in a signal, we create a regression model for each speech type based on the MFCCs extracted from noisy signals under different SNR conditions. The SNR value for an unseen recording is then estimated using the trained regression model. This section briefly describes the main components of the proposed global SNR estimation approach.

3.1. Feature Extraction

Each recording in the database is segmented into frames of 30 ms (with 10 ms overlap) using a Hamming window. Setting the DFT size equal to 512 and using 27 mel filters, 12 MFCCs are extracted for each frame and appended with the frame energy and concatenated with *delta* and *double-delta* coefficients, resulting in a 39-dimensional feature vector. Finally, to have a fixed-length vector per recording, the feature vectors of each utterance are averaged.

3.2. Support Vector Regression (SVR)

In this study, the support vector regression (SVR) is used as a function approximation to estimate the speech SNR. Developed as

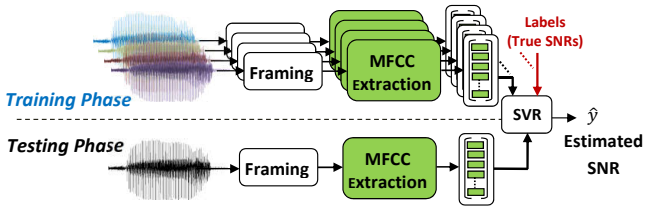

Fig. 3: Block diagram of the proposed method for global speech SNR estimation in training and testing phases.

the regression version of the support vector machines, SVRs perform the regression task by finding the optimal regression hyperplane in which most of training observations lie within an ϵ-margin around this hyperplane. In the ϵ-SVR [20], we are given a set of training data, $S^{\text{tr}} = \{\boldsymbol{x}_j, y_j\}_{j=1}^{J}$, where $\boldsymbol{x}_j$ denotes a feature vector of the j^{th} sample and y_j is the corresponding target value. The goal is to determine a function $g(\boldsymbol{x}_j)$, so that for all the training data, the outputs have at most ϵ deviation from the actual outputs. In the SVR framework, we consider the following relation for g:

$$g(\boldsymbol{x}_j) = \boldsymbol{\nu}^T f(\boldsymbol{x}_j) + c\,, \quad j = 1, \cdots, J, \tag{3}$$

where $f(\boldsymbol{x}_j)$ is a mapping function in the feature space, $\boldsymbol{\nu}$ is a row vector of the same dimension as $f(\boldsymbol{x}_j)$, c is a real-valued constant, and ^{T} represents the transpose operator. The estimation of $\boldsymbol{\nu}$ and c is formulated as the following optimization problem [20]:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\boldsymbol{\nu}\|^2 + \gamma \sum_{j=1}^{J} (\kappa_j + \kappa_j^*) \\ \text{subject to} \quad & \begin{cases} y_j - \boldsymbol{\nu}^T f(\boldsymbol{x}_j) - c & \leq \epsilon + \kappa_j, \\ \boldsymbol{\nu}^T f(\boldsymbol{x}_j) + c - y_j & \leq \epsilon + \kappa_j^*, \\ \kappa_j, \kappa_j^* & \geq 0 \end{cases} \end{aligned} \tag{4}$$

where κ_j and κ_j^* are slack variables, $\epsilon > 0$ controls the ϵ-insensitive zone used for fitting the training data and $\gamma > 0$ determines the trade-off between the flatness of g and the amount up to which deviations larger than ϵ are tolerated. Since the optimization problem is convex, a unique optimal solution can be found [21].

In this study, the LIBSVM toolbox [22] is used to implement SVR models. We select the kernel type and tune the hyper-parameters of the model, namely the ϵ and γ using the 5-fold cross-validation (CV).

3.3. Training and Testing

The block diagram of the proposed approach in training and testing phases is shown in Fig.3. During the training phase, the recordings in the training subset are converted into fixed-length feature vectors using the approach described in Section 3.1. The obtained vectors along with their corresponding SNR values are then used to train a regression model. In the testing phase, the feature extraction approach, applied in the training phase, is used to extract a feature vector from an unseen test recording from which the trained regression model estimates the SNR of the signal.

4. EXPERIMENTAL SETUP

4.1. Database

The proposed SNR estimation system has been developed and validated using three different databases, namely the LibriSpeech database [18] for normal speech, the CHAINS database [19] for whispered speech, and the Parkinson's voice database for pathological voice. The LibriSpeech database consists of 1000 hours of read English speech based on LibriVox's audio books. The recordings are sampled at 16 kHz. From this database, we have chosen 426 recordings of 10 s average duration range from 2 s to 28 s uttered by 142 speakers of both genders. The CHAINS database contains speech recordings of 36 speakers reading a text in a whisper in two different environments, namely a sound-proof booth and a quiet office, and sampled at 44.1 kHz. From this database, we have chosen 8 recordings per each speaker, selected equally from both environments, to form a data set of 288 whispered speech samples of 20 s average duration range from 2 s to 76 s. To evaluate the proposed system on the pathological voice signals, we used the Parkinson's voice database since the wast majority of people with Parkinson's disease (PD) exhibit some form of vocal disorder [23]. This database is generated through collaboration between Sage Bionetworks, PatientsLikeMe and Dr. Max Little as part of the Patient Voice Analysis study[1], and contains telephone recordings of the sustained vowels /a/ uttered by 750 patients of both genders, sampled at 8 kHz and range from 3 s to 30 s long with 16 s average duration. The last two databases are challenging since most existing speech SNR estimation algorithms have been developed based on the normal and healthy voices.

The voice recordings in each database are divided into non-overlapping training and test subsets consisting of 80% and 20% of the speakers, respectively. To create databases for SNR estimation, we corrupted all clean recordings by adding stationary and non-stationary noises at different SNRs, ranging from -5 dB to 30 dB in 1 dB steps, and appended them to the databases. Specifically, for stationary noise we used white Gaussian and car engine noises, and for non-stationary one we used babble, street and keyboard noises. Therefore, the training subsets of the extended normal speech, whispered speech and Parkinson's voice databases for each noise type respectively contain 12180, 5816 and 18576 recordings. The test subsets of the extended databases for each noise type contain 3045, 1454 and 4644 recordings, respectively.

4.2. Performance Metric

In order to evaluate the effectiveness of the proposed method, we use the mean-absolute-error (MAE) of the estimated SNRs which is defined as:

$$E_{\text{MA}} = \frac{1}{L} \sum_{l=1}^{L} |\hat{y}_l - y_l| \tag{5}$$

where $\hat{y}_l$ is the l^{th} estimated SNR, y_l is the l^{th} actual SNR and L is the total number of test samples.

5. RESULTS

In this study, we compare our proposed approach with the NIST SNR measurement method [7] and the WADA method [14]. We used 5-fold CV to evaluate the performance of different methods in terms of the MAE, E_{MA}, of the estimated SNR (in dB). First, we assume that the noise type is known. In this case, we train different regression models for each noise and speech type and use the corresponding noise-dependent model to estimate the SNR. In the next step, assuming that the noise type is unknown, we train a noise-independent regression model using the recordings corrupted by all five mentioned noise types. To this aim, we have randomly selected 20% of the

[1] Obtained through Synapse ID [syn2321745]

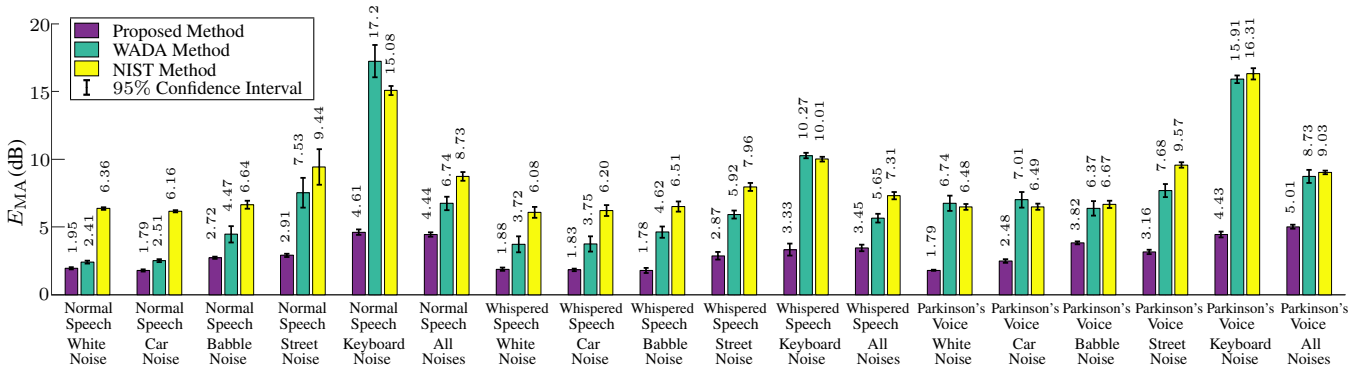

Fig. 4: Comparison of the MAE, E_{MA}, (in dB) of the proposed method and the baseline systems for speech SNR estimation using three different speech types under various noise conditions, along with 95% confidence intervals.

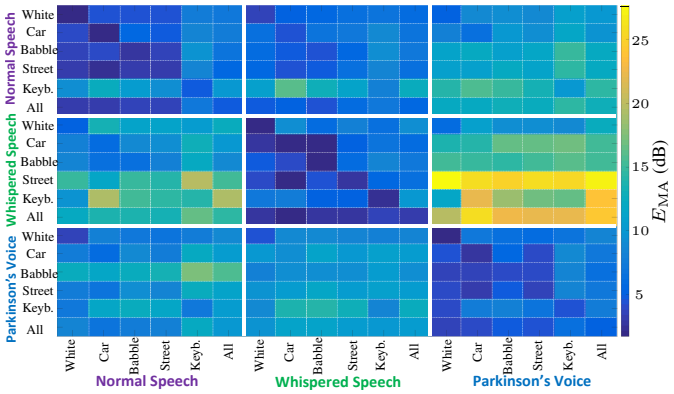

Fig. 5: Performance of every trained regression model (represented in rows) on every other noise and speech conditions (represented in columns) in terms of E_{MA} (in dB).

noisy recordings from each noise class to have the same number of samples as the noise-dependent experiments. In both cases, we use SVR models with a linear kernel.

The results of the baseline systems and the proposed approach using normal, whispered and Parkinson's voices under different stationary and non-stationary noise conditions over all CV repetitions along with 95% confidence intervals are presented in Fig.4. The results show that the proposed approach outperforms baseline methods for all the cases we tested and it provides consistent and accurate estimation for different speech types under various noise conditions. It can be noticed that WADA gives a comparable performance to the proposed method in the case where normal speech recordings are corrupted by stationary noises since the Gaussian and gamma distribution assumptions respectively for background noise and speech signals are satisfied. However, both baseline methods have failed to provide accurate estimation when they are applied to the whispered and disordered voices. It is probably due to the fact that these types of speech do not satisfy the underlying assumptions in the algorithms. Moreover, the poor SNR estimation for Parkinson's voices using the baseline methods might also be due to the fact that they consider distortion in pathological voices, due to vocal disorders, as noise even if they are recorded in a noise-free environment. Notice that as the non-stationarity of the noise increases, the performance of all tested methods degrades. Furthermore, when signals are corrupted by the keyboard noise, which has impulsive characteristics, the baseline systems fail to estimate the SNR, while the proposed method still exhibits a good performance. The results suggest that if the noise type is unknown, the proposed method can still provide a satisfactory accuracy if the noise-independent regression model is trained with a variety of noise types.

To investigate the generalization of the proposed method in diverse conditions, we compare, in Fig.5, the performance of every trained regression model (represented in each row) on every other noise and speech condition (denoted in each column). It can be noticed that noise-dependent regression models in matched train-test conditions provide the best performance. Moreover, there are some noise-dependent models in each speech class that can be utilized to estimate SNR for other noise conditions. For example, the models trained with street noise, car noise, and babble noise respectively in normal speech, whispered speech and Parkinson's voice classes, can provide good estimations for other noise types within those classes, except the keyboard noise. Notice that the keyboard models fail to accurately estimate SNR for other noise conditions in both within and between speech classes. We observe that the white Gaussian models give satisfactory results when they are applied to the recordings of other speech types corrupted by white Gaussian noise. We notice that the noise-independent models can provide a good performance in all noise conditions within each speech class. It means that they can be utilized for both known and unknown noise conditions, resulting in avoiding the need for noise-type classification prior to SNR estimation.

6. CONCLUSION

In this study, we investigated the impact of stationary and non-stationary noises on the behavior of MFCCs of normal, whispered and pathological voices. It has been demonstrated experimentally that, regardless of the speech type, introducing additive noise to the recordings results in predictable modification in mean and covariance matrix of the MFCCs and the amount of change is related to the level of noise. Motivated by this observation, we proposed a new supervised method to estimate the global speech SNR which uses MFCCs of the noisy signals to train a regression model for each speech type. The proposed approach avoids the need for identification of speech and non-speech regions in signals facilitating dealing with special speech signals such as sustained vowels, whispered speech and pathological voices. Experimental results show the consistent performance of the proposed approach in accurately estimating SNR for various speech types under different known and unknown noise conditions.

7. REFERENCES

[1] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, 2006.

[2] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, 2003.

[3] M. Akbacak and J. H. L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 2, pp. 465–477, 2007.

[4] J. A. Morales-Cordovilla, N. Ma, V. Sanchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with Missing Data," in *ICASSP*, 2011, pp. 4808–4811.

[5] M.-w. Mak, X. Pang, and J.-t. Chien, "Mixture of PLDA for Noise Robust I-Vector," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 130–142, 2016.

[6] M. Vondrasek and P. Pollak, "Methods for Speech SNR estimation: Evaluation Tool and Analysis of VAD Dependency," *Radioengineering*, vol. 14, no. 1, pp. 6–11, 2005.

[7] "The NIST speech signal-to-noise ratio measurement." [Online]. Available: https://www.nist.gov/information-technology-laboratory/iad/mig/nist-speech-signal-noise-ratio-measurements

[8] T. H. Dat, K. Takeda, and F. Itakura, "On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," *Speech Commun.*, vol. 48, no. 11, pp. 1515–1527, 2006.

[9] D. Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis," in *Speech Sep. by Humans Mach.*, 2005, pp. 181–197.

[10] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1306–1319, 2008.

[11] A. Narayanan and D. Wang, "A CASA-Based System for Long-Term SNR Estimation," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 9, pp. 2518–2527, 2012.

[12] K. Hu and D. Wang, "Unvoiced Speech Segregation From Nonspeech Interference via CASA and Spectral Subtraction," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 6, pp. 1600–1609, 2011.

[13] P. Papadopoulos, A. Tsiartas, and S. Narayanan, "Long-Term SNR Estimation of Speech Signals in Known and Unknown Channel Conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2495–2506, 2016.

[14] C. Kim and R. M. Stern, "Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," in *INTERSPEECH*, 2008, pp. 2598–2601.

[15] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 289–293.

[16] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, 2nd ed. New York: IEEE Press, 2000.

[17] M. L. Narayana and S. K. Kopparapu, "Effect of Noise-in-speech on MFCC Parameters," in *WSEAS*, 2009, pp. 39–43.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[19] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers," *SPECOM*, pp. 431–435, 2006.

[20] V. N. Vapnik, *The nature of statistical learning theory*, 2nd ed. Springer, 2000.

[21] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, aug 2004.

[22] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," National Taiwan University, Tech. Rep., 2016.

[23] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease." *Behav. Neurol.*, vol. 11, no. 3, pp. 131–137, 1998.