AUDIO ANALYSIS LAB

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], Mads Græsbøll Christensen [1]

[1] Audio Analysis Lab, CREATE, Aalborg University, 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

AALBORG UNIVERSITY DENMARK

Introduction

- Most existing global SNR estimation algorithms are based on measuring the energy contents of speech and non-speech regions in a signal.
- These methods have difficulties dealing with some speech types:
  - **Sustained Vowels:** there is no regular pauses.
  - **Whispered Speech:** difficult to identify speech and non-speech regions.
  - **Pathological Voice:** the distortion due to vocal disorder is considered as noise even if it is recorded in a noise-free environment.

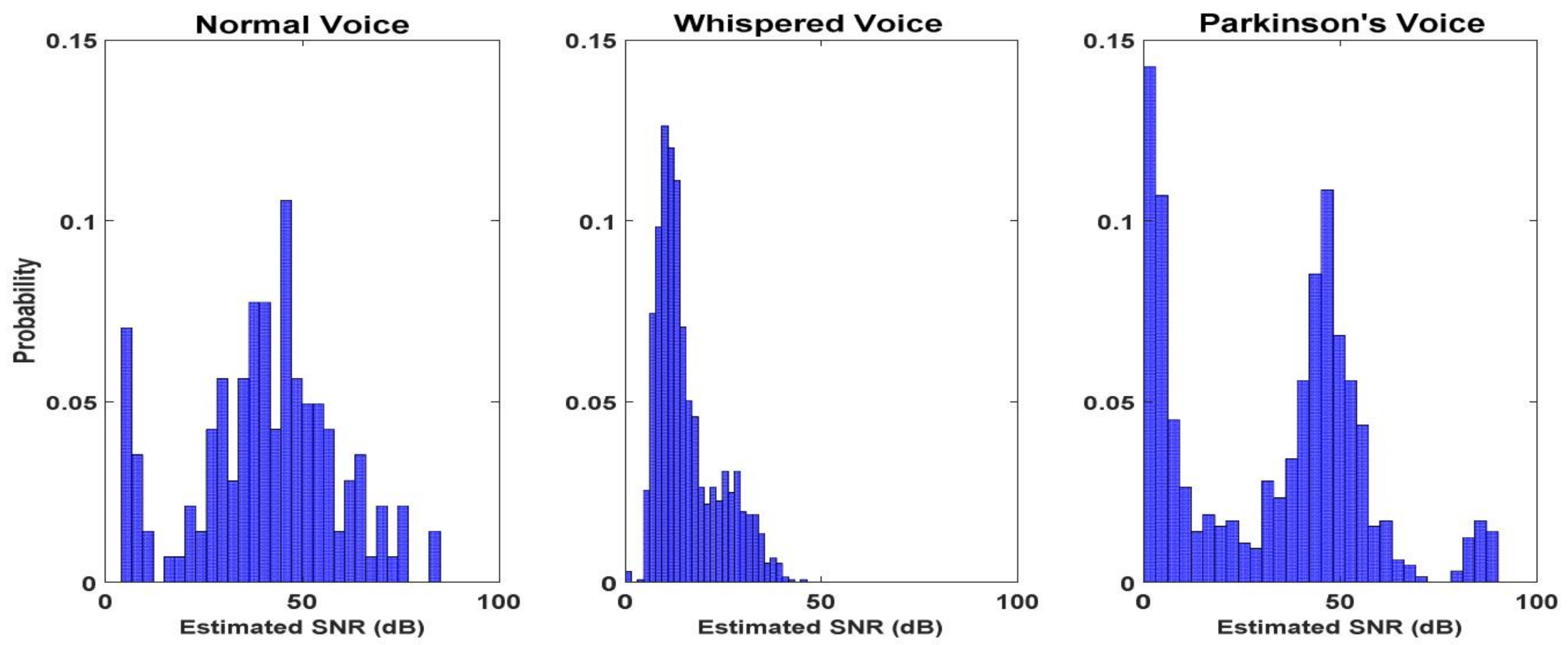

Fig. 1: The normalized histograms of estimated SNR values for three different clean databases using the NIST algorithm [1].

- High SNR values are expected since these databases are collected in very low-noise environments.

Impact of Additive Noise on MFCCs

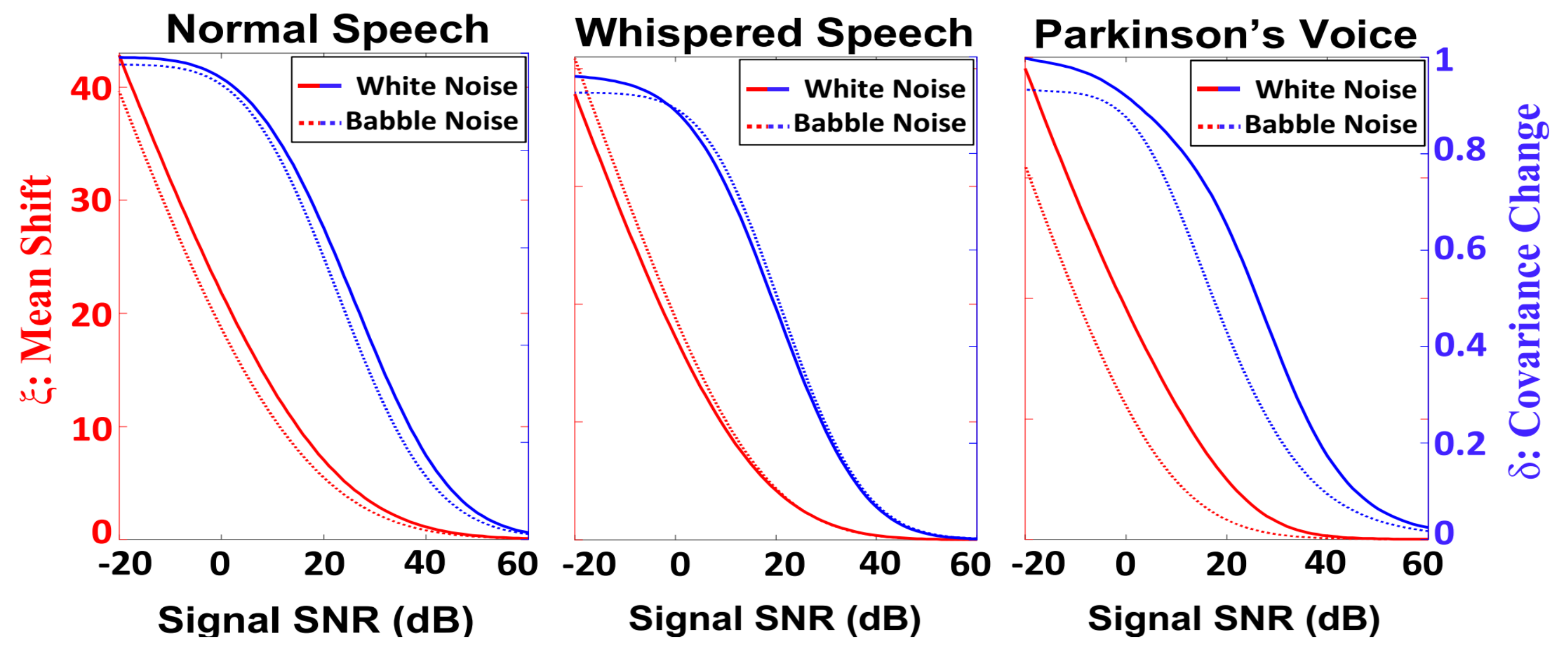

Fig. 2: Impact of noise at different SNR levels on the mean and the covariance matrix of MFCCs of the normal, whispered and pathological voices.

$$\zeta(i) = \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{\mu}_m^{n_i}-\boldsymbol{\mu}_m^{c}\right\|_2 \quad , \qquad \delta(i)=\frac{1}{M}\sum_{m=1}^{M}\frac{\left\|\boldsymbol{\Sigma}_m^{n_i}-\boldsymbol{\Sigma}_m^{c}\right\|_F}{\left\|\boldsymbol{\Sigma}_m^{c}\right\|_F} \quad .$$

where M is the number of speakers, $\boldsymbol{\mu}_m^c$ and $\boldsymbol{\mu}_m^{n_i}$ are the means of the MFCCs computed from the clean and noisy signals from the m^{th} speaker subject to the i^{th} noise level, $\boldsymbol{\Sigma}_m^c$ and $\boldsymbol{\Sigma}_m^{n_i}$ are the covariance matrices of the MFCCs extracted from the clean and the noisy utterances of the m^{th} speaker.

SNR Estimation Method

- **Principle:** Instead of identifying speech and non-speech regions in a signal, the global SNR of a signal is directly estimated using a regression model trained by MFCCs of noisy signals at different SNRs.
- **Features:** A 39-D feature vector per recording (12 MFCCs + frame energy + Δ + ΔΔ, averaged over frames).
- **Regression model:** Support Vector Regression (SVR) with a linear kernel.

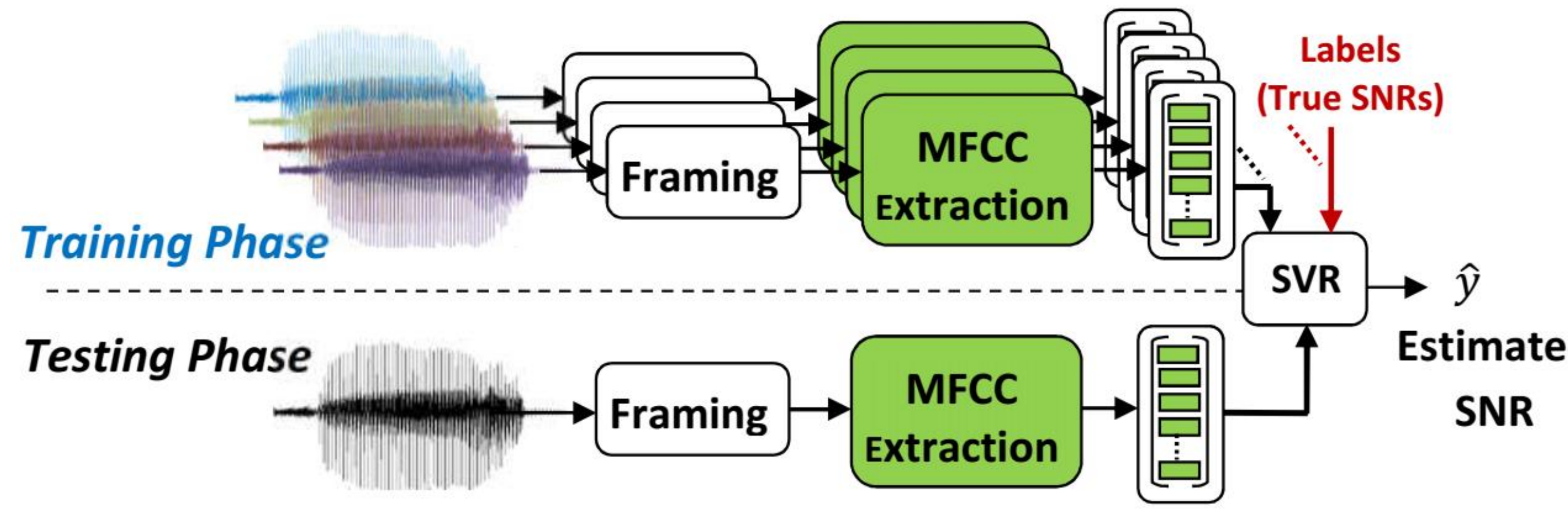

Fig. 3: Block diagram of the proposed global speech SNR estimation method.

Experimental Setup

- **Databases:**
  - **Normal speech:** 426 recordings of 10 s average duration uttered by 142 speakers of both genders selected from LibriSpeech database.
  - **Whispered speech:** 288 whispered speech samples of 20 s average duration selected from CHAIN database.
  - **Pathological voice:** Telephone recordings of the sustained vowels /a/ by 750 Parkinson's patients of both genders, with 16 s average duration.
- **Noise types (SNR range from -5 dB to 30 dB in 1 dB steps):**
  - **Stationary:**
    - White Gaussian noise
    - Car engine noise
  - **Non-stationary:**
    - Babble noise
    - Street noise
    - Keyboard noise
- **Performance Metric:** Mean-absolute-error of the estimated SNRs

Results

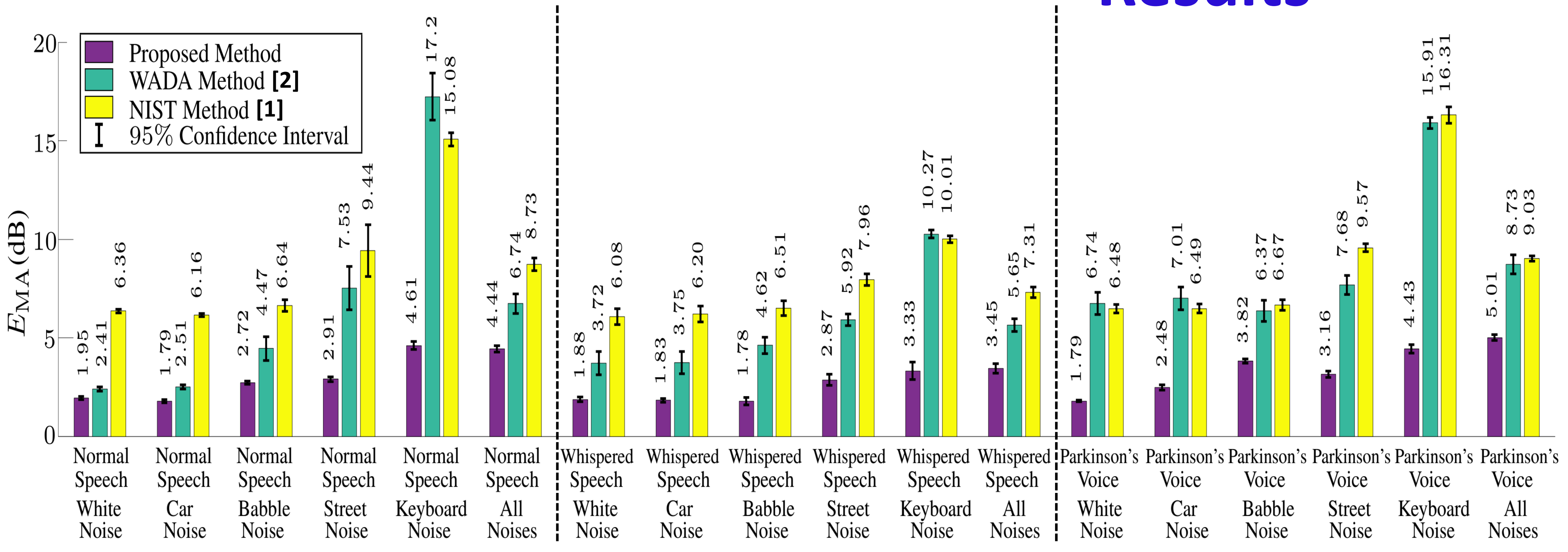

Fig. 4: Comparison of the MAE, EMA, (in dB) of the proposed method and the baseline systems for speech SNR estimation using 3 different speech types under various noise conditions, along with 95% confidence intervals.

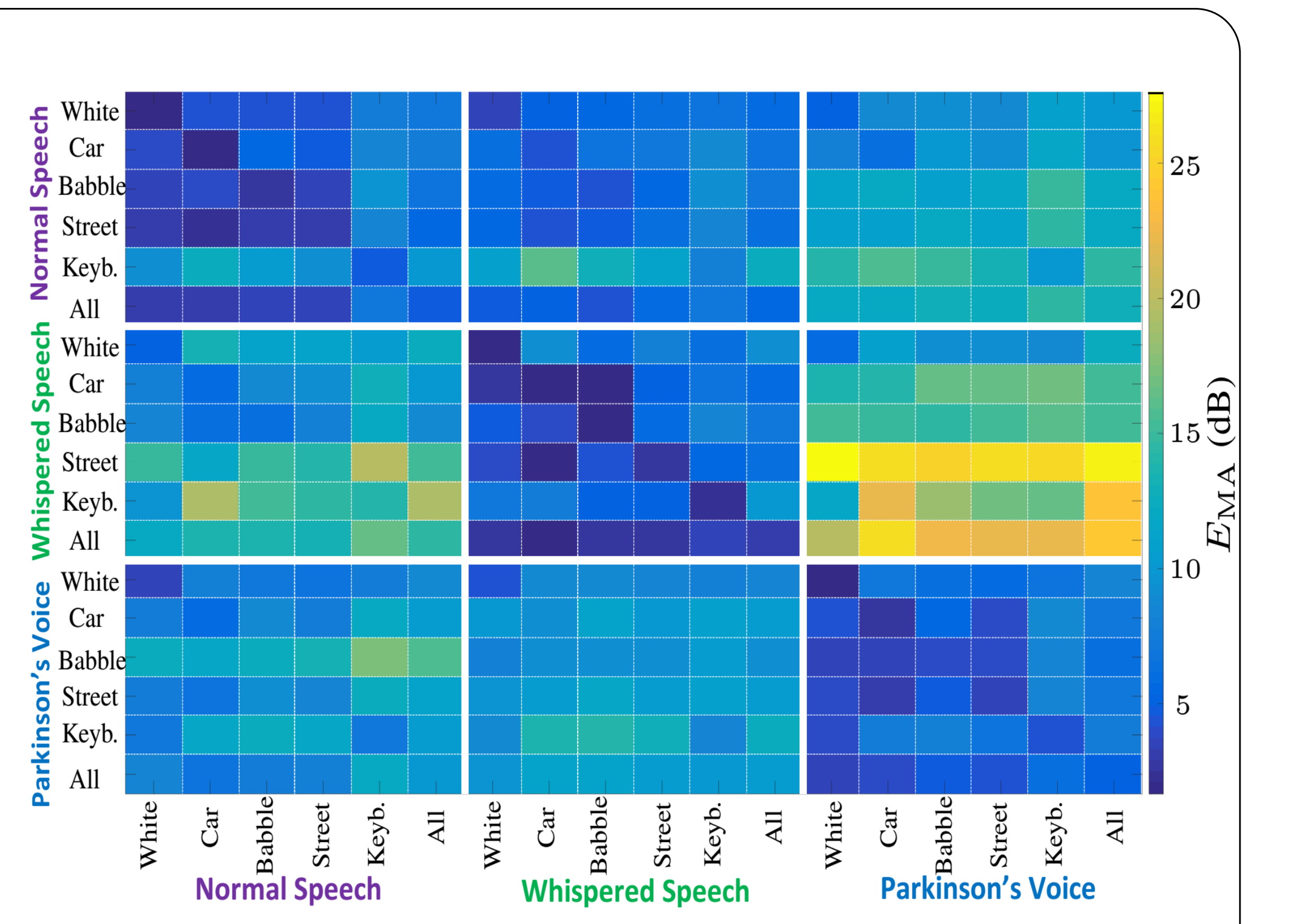

Fig. 5: Performance of trained regression models (rows) on every other noise and speech conditions (columns) in terms of EMA.

Conclusion

- The presence of additive noise in speech signals results in predictable modification in mean and covariance matrix of the MFCCs and the amount of change is related to the level of noise, regardless of the speech type.
- We proposed a supervised approach to estimate the global speech SNR that uses MFCCs to train a regression model for each speech type.
- The proposed method avoids the need for identification of speech and non-speech regions in signals facilitating dealing with special speech signals.

References

[1] "The NIST speech signal-to-noise ratio measurement." [Online]. Available: https://www.nist.gov/informationtechnology-laboratory/iad/mig/nist-speech-signal-noise-ratiomeasurements.

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