AALBORG UNIVERSITY
STUDENT REPORT

Department of Architecture, Design and Media Technology

Sound and Music Computing, 10th Semester

Title:

Wind noise reduction in speech signals using non-negative matrix factorisation

Project Period:

02/02/18 - 29/05/18

Semester Theme:

Master´s Thesis

Supervisors:

Mads Græsbøll Christensen &

Jesper Kjær Nielsen

Projectgroup no.

SMC181032

Members:

Benjamin Elif Larsen

Abstract:

The purpose of this project was to implement a non-negative matrix factorisation (NMF) method and evaluate its ability to extract speech from a mixed signal with non-stationary noise, the noise being wind. It was compared to the state-of-the-art (Non-negative sparse coding, NNSC), the non-processed signals, and two noise reduction methods for stationary noise. The methods were tested and evaluated over different conditions. The conditions being the number of wind and speech components, the signal-to-noise ratio (SNR), and two different β-divergence values.

Two different dictionaries were trained, a speech and a wind dictionary.

The measurements used for the evaluate were the PESQ and the STOI. The SNR_{out} was measured for the NMF and the state-of-the-art.

The results indicate that the NMF failed at extracting the speech and wind from the mixed signals as it overall scored lower than the non-processed signals and the two stationary noise reduction methods, while for the most of the time it did similar to the NNSC method.
The NNSC had been found to give good results, which could indicate that the number of signals used for the training of the speech and wind dictionaries was not high enough to allow the NMF and NNSC to be able to extraction untrained speech and wind signals. At the same time, it was noticed that a lot of distortion was present in the audio signals, which could indicate that the dictionaries extracted parts of the wrong source.

Abstract

The purpose of this project was to implement a non-negative matrix factorisation (NMF) method and evaluate its ability to extract speech from a mixed signal with non-stationary noise, the noise being wind. It was compared to the state-of-the-art (Non-negative sparse coding, NNSC), the non-processed signals, and two noise reduction methods for stationary noise (Spectral subtraction and minimum mean square error estimate of short-time log-spectral amplitude, MMSE STSA). The methods were tested and evaluated over different conditions. The conditions being the number of wind and speech components, the signal-to-noise ratio (SNR), and two different β-divergence values.

Two different dictionaries were trained, a speech and a wind dictionary.

The measurements used for the evaluate were the PESQ and the STOI. The SNR_{out} was measured for the NMF and the state-of-the-art.

The results indicate that the NMF failed at extracting the speech and wind from the mixed signals as it overall scored lower than the non-processed signals and the two stationary noise reduction methods, while for most of the time it did similar to the NNSC method. The NNSC had been found to give good results, which could indicate that the number of signals used for the training of the speech and wind dictionaries was not high enough to allow the NMF and NNSC to be able to extraction untrained speech and wind signals. At the same time, it was noticed that a lot of distortion was present in the audio signals, which could indicate that the dictionaries extracted parts of the wrong source.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Audio noise reduction have been possible for a while and been quite successful, at least when the noise is stationary, however, it have been proven harder to remove non-stationary noise from audio signals [1, 2]. Non-negative matrix factorisation have been found to be able to reduce non-stationary babble noise (multiple humans speaking simultaneously) from speech signals [2]. A non-stationary noise that is common in most audio sampling recorded outside is the sound of wind, a sound that can differ a lot depending on the whether the sound is a mild breeze or the sounds of a hurricane, hence the signal-to-noise ratio (SNR) between the wind and e.g. speech is far from constant. At the same time wind can be close to stationary by being stagnant and more or less constant in time and frequency, while at other times the wind can change a lot over time in frequency, this can be seen in figure 1.1. This problem with wind can be noticed if speaking with a person who are outside using telecommunication.

The problem to solve, the removal of wind in a mixed signal with speech, can be viewed in figure 1.2. As it can be seen the wind spectrogram has some similar frequencies to the speech spectrogram (a spectrogram being the visual representation of frequencies over time).

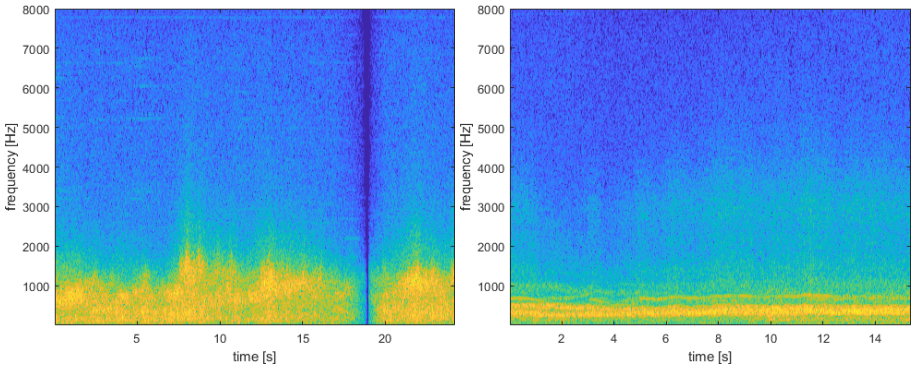


Figure 1.1: Spectrograms of two different wind signals

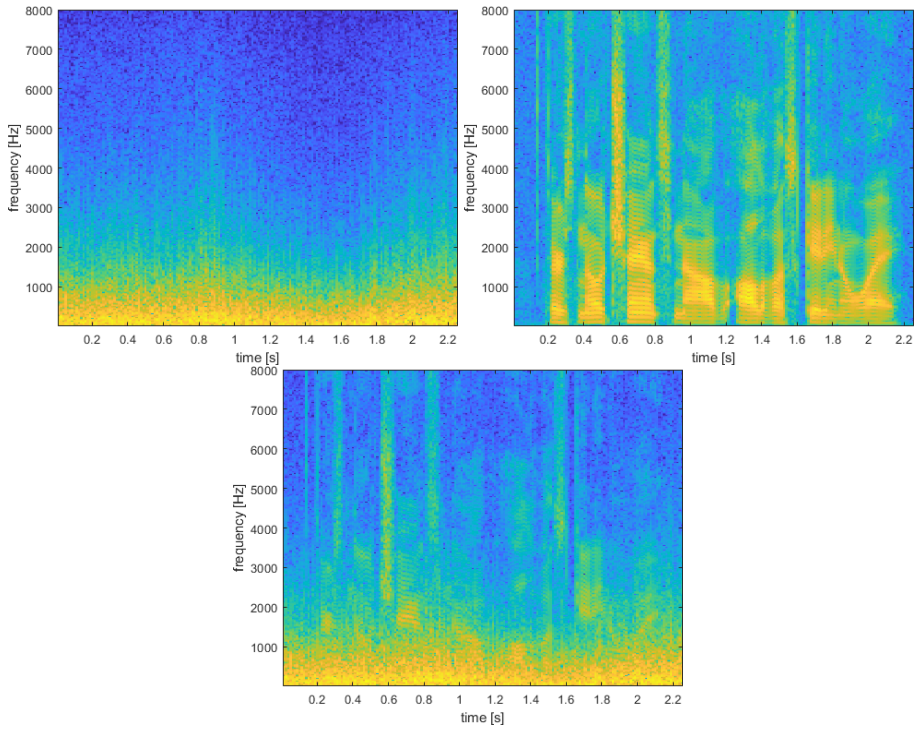


Figure 1.2: Top left: The spectrogram of a wind signal. Top right: The spectrogram of a speech signal. Bottom: The spectrogram of the mixed signal

Chapter 2

Problem Statement

Removing stationary noises from speech signals have been solvable with great results for a long time, however, removing non-stationary noises from speech signals have been proved to be a harder problem as most of the methods used for stationary noises do not work well, when the noises are non-stationary [1, 2]. However, recently non-negative matrix factorisation (NMF) methods have been developed that can be used in situations with babble noise, which is a non-stationary noise and thus hard to remove [2].

Thus the hypothesis for the study is: How well can the implemented NMF algorithm separate mixed signals, the mixed signals consisting of speech and wind, by evaluating the NMF by comparing it to other methods and itself over different conditions, while figuring out why it generates the specific outputs, what affects the ability to extract speech and wind from a mixed signal, and why it may or may not work well compared to the other methods. The NMF will be evaluate by measuring the quality and intelligibility of the output speech signals and the SNR before and after the processing.

Hence, the purpose of this project is to apply the NMF method on signals with speech and non-stationary noise, where the noise is classified as wind noises, by training two dictionaries, one for noise and one for speech, and then apply them to mixed signals consisting of speech and noise.

The conditions given in the list below are the different conditions the NMF will be evaluated over.

- Performance over different amount of wind components
- Performance over different amount of speech components
- Performance over different signal-to-noise ratios (SNR)
- Performance between Kullback-Leibler divergence and Itakura-Saito divergence

Note that in the report, noise and interference is considered the same thing.

Chapter 3

Theory

The first part of the theory chapter focus on signal model, then the non-negative matrix factorisation. The second part of the chapter focus on the state-of-the-art for separation of non-stationary noise and lastly methods for stationary noise reduction.

3.1 Signal Model

The signals used for this study is in the time domain given as

$$y(n) = s(n) + v(n) \tag{3.1}$$

where y being the mixed signal, s being the speech signal, v being the noise/interference, where in this study this would be the wind, and n is the discrete time index [3].

In the frequency domain, which is part of the domain that is used by the non-negative matrix factorisation [4], a signal is described using the following equation

$$Y(j\omega) = S(j\omega) + V(j\omega) \tag{3.2}$$

where j is the imaginary unit, ω is the normalised frequency index, and $Y(j\omega)$, $S(j\omega)$, and $V(j\omega)$ are the discrete-time Fourier transform of their respectively time domain version [3]. It should be noticed that in equation 3.2 it is assumed that the signals are stationary [3], which they are not in this study. In this study they will be represented in the time-frequency domain using the short-time Fourier transform given in 3.3.

$$Y(n,k) = \sum_{m=0}^{L-1} s(m+nR)w(m)e^{-j\frac{2\pi}{N}km} + \sum_{m=0}^{L-1} v(m+nR)w(m)e^{-j\frac{2\pi}{N}km} \tag{3.3}$$

where $w(n)$ is the window function of length L, L being the amount of discrete time samples used to calculate $\cdot(n,k)$, N is the amount of discrete frequency bins used and is usually the same as or bigger than L, R is the hop size that control the amount of samples the window is moved with and is normally smaller or the same as L. The STFT is normally not calculated for each discrete time index n, rather it only calculated for $n = 0, R, 2R, \ldots$ and this case the n is called the frame. This is because the frequency content is not likely to change much from sample to sample. The angular frequency of each frequency bin k is given as $2\pi k/N, k = 0, 1, 2, \ldots, N-1$, thus each frequency bin k, also written as ω_k, covers multiple frequencies. It should be noticed that it is common to only calculate frequency bins up to N/2 (in the case of an uneven N, only up to (N+1)/2) as it is the half of the sampling rate and the frequency bins $N/2+1$ to $N-1$ above this are complex conjugate symmetry with frequency bin 1 to N/2 -1 mirrored given the Nyquist frequency (For an uneven N, (N+1)/2 to N-1 are the complex conjugate symmetry mirrored of frequency bin 1 to (N+1)/2-1). In the case of an even N the last calculated frequency bin $k = N/2$ contains the energy at the Nyquist frequency and consist of real values. The first frequency bin, $k = 0$, consist of only real values. Thus for an even N the unique frequency bins are from 0 to N/2, while for uneven N they are from 0 to (N+1)/2-1.

In this project each n, k (frame, frequency bin) index is called the energy component to differ it from the usage of word component in the non-negative matrix factorisation.

3.2 Non-negative Matrix Factorisation

Non-negative matrix factorisation (NMF) is a method to identify and extract the components a non-negative data matrix consist of in what is called compositional analysis [4]. The data matrix can e.g. be the Short-time Fourier Transform (STFT), of a signal y, y being given in eq.3.1, transformed into a magnitude $|\mathbf{Y}|$ or power $|\mathbf{Y}|^2$ spectrogram. It should be noticed that spectrogram $\mathbf{Y}$ is transposed for the NMF algorithm [4, 5], that is $\mathbf{Y}(k, n)$ instead of $\mathbf{Y}(n, k)$.

The components are made up of additive combinations that results in no subtraction or diminishment of any of the components, thus a component can be extracted without any effect on the other components of the data matrix [4]. This is refereed to as compositional data [4]. The models that makes use of this represent the data as non-negative linear combinations of parts, the non-negative data is to ensure that the combination of data do not result in subtraction or diminishment and the models are referred as compositional models [4]. Thus the non-negative matrix factorisation belongs to this class of models [4]. The components can be referred to as atoms or atomic units, where each component is a column vector that represents the spectral vector in the spectrogram that have been decomposed into a linear combination of these non-negative components [4].

Given a matrix of data $\mathbf{V}$, e.g. magnitude or power spectrogram, of di-

mensions $F \times N$, with non-negative entries, the problem of NMF is to find a factorisation where

$$\mathbf{V} \approx \mathbf{WH} \stackrel{\text{def}}{=} \mathbf{V}_{min} \tag{3.4}$$

where $\mathbf{W}$ and $\mathbf{H}$ are the basic matrix of dimensions $F \times K$ and the activation matrix of dimensions $K \times N$, respectively, and consist solely of nonnegative entries [4]. For audio, the dimensions will be F frequencies or frequency bins, N segments of the STFT, and K components [4]. The V_{min} is the non-negative data matrix of same size of $\mathbf{V}$ and being the product of $\mathbf{W}$ and $\mathbf{H}$, the reason for $\mathbf{V}_{min}$ is that the $\mathbf{W}$ and $\mathbf{H}$ rarely will describe $\mathbf{V}$ fully, some residual is normally left if $\mathbf{V}_{min}$ is subtracted from $\mathbf{V}$ [4]. In this report $\mathbf{V}_{min}$ will instead being given as $\hat{\mathbf{V}}$ to simplify the equations.

The NMF can be conducted in two different ways, supervised and non-supervised [4, 5]. In the case of the supervised version, the dictionary $\mathbf{W}$ is known before hand by training on other data matrices and only the activation $\mathbf{H}$ is need to be established [4]. The training is either done by running each component of $\mathbf{W}$ on a single, different, sound, e.g. a note, and then collect them all into the a single dictionary if possible, if not possible, the training can be done by running multiple components over a data matrix consisting of only specific kind of data, e.g. a single or multiple speaker(s) or wind sound(s) [4]. For the non-supervised version neither $\mathbf{W}$ nor $\mathbf{H}$ is known and both are needed to be established out from a single data matrix $\mathbf{V}$ [4]. If the amount of components are not known beforehand, it is possible to estimate the amount needed to properly estimate $\mathbf{V}$ [4]. This is done using methods like enforcing sparsity on activations and remove any component that display constant low activation [4] or by applying more complicated methods like Bayesian formulations or Markov chain Monto Carlo [4, 6].

The number of components in the dictionary should preferably equal that of the number of latent compositional units in the signal, however, the number of units might not been known [4]. A mathematical restriction exists for the number of components in the dictionary, that if $K \geq F$ and if no other restriction are present trivial solutions for V can be found, that is the dictionary can model any data matrix fully and the decomposition is not unique [4]. Some of the restrictions that exist are the sparsity on activations as mentioned earlier, group sparsity to promote sparsity of a group of components, and temporal continuity, which promotes on the activations smooth temporal variation [4].

For the calculation of $\mathbf{W}$ and $\mathbf{H}$, the multiplication update rule given in eq. 3.5 and 3.6 can be used [5]. Both are initialised to random positive values [5].

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{((\mathbf{WH})^{\beta-2}\mathbf{V})\mathbf{H}^T}{(\mathbf{WH})^{\beta-1}\mathbf{H}^T} \tag{3.5}$$

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T((\mathbf{WH})^{\beta-2}\mathbf{V})}{\mathbf{W}^T(\mathbf{WH})^{\beta-1}} \tag{3.6}$$

where $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the matrix $\mathbf{A}.\mathbf{B}^{\cdot-1}$, the dots indicate element-wise multiplication, and the superscript ^{T} denotes the transpose of a vector or a matrix

[5]. In the case of supervised NMF only $\mathbf{H}$ should be updated [4]. If both $\mathbf{W}$ and $\mathbf{H}$ are to be established, both are required to be normalised after each iteration [4]. Normally the number of iterations are either set to a fixed amount or stopped when the error between $\mathbf{V}$ and $\hat{\mathbf{V}}$ are under a threshold, however, a combination of both can be used. It should be noticed that supervised NMF normally converge to the global minimum, or close to it, while non-supervised NMF might converge to a local minimum [5]. To get around the converging to a local minimum, a Monto Carlo simulation might be added, where a number of chain (the amount of times the Monto Carlo simulation is run) are added, where each chain runs the NMF with new randomised initialised values for $\mathbf{W}$ and $\mathbf{H}$. Then $\mathbf{W}$ and $\mathbf{H}$ from the chain that minimises the error the most is selected. The possibility of all the chains converging on local minimums still exist.

It is possible to have a mixture of a supervised and non-supervised NMF, where part of the dictionary $\mathbf{W}$ is known beforehand, but with added not-trained components that has to be estimated. In this case only the not-trained components in $\mathbf{W}$ and $\mathbf{H}$ should be updated and normalised [7].

Regarding the β value to use in eq. 3.5 and eq. 3.6, it depends on the specific β-divergence cost function used, where the most common three cost functions are given in eq. 3.7.

$$d_\beta(x|y) \overset{\text{def}}{=} \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}), & \beta \in \mathbb{R}\backslash\{0,1\} \\ x \log x/y + (y-x), & \beta = 1 \\ \frac{x}{y} - \log \frac{x}{y} - 1, & \beta = 0 \end{cases} \tag{3.7}$$

For $\beta = 1$ the β-divergence is the Kullback-Leibler divergence (KL), if $\beta = 0$ it is the Itakura-Saito divergence (IS) [4, 5]. For any other β it is the general β-divergence with the special case of $\beta = 2$ being the Euclidean distance (EUC) [5].

The purpose of the β-divergences are about solving the following equation

$$D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{f,t} d(\mathbf{V}_{f,t}, \hat{\mathbf{V}}_{f,t}) \tag{3.8}$$

where $v_{f,t}$ and $\hat{v}_{f,t}$ are the (f,t)th element of $\mathbf{V}$ and $\hat{\mathbf{V}}$, respectively, and $d()$ is the divergence between the two scalars [4].

The cost function is then used to solve eq. 3.9.

$$\begin{gathered} \mathbf{W}^*, \mathbf{H}^* = \arg\min_{\mathbf{W},\mathbf{H}} \mathrm{D}(\mathbf{V}\|\mathbf{W}\mathbf{H}) \\ \mathbf{W} \geq 0 \; \mathbf{H} \geq 0 \end{gathered} \tag{3.9}$$

where $\mathbf{W}^*$ and $\mathbf{H}^*$ are the $\mathbf{W}$ and $\mathbf{H}$ that minimises the error between $\mathbf{V}$ and $\mathbf{WH}$ from all iterations.

A thing to note regarding the β-divergences are that the IS divergence is scale invariant, that is the same relative weight is given to small and large coefficients of $\mathbf{V}$ in the cost function, while the KL and EUC divergence are not

scale invariant as more emphasize is given on the largest coefficients, hence less precision in the estimation of low coefficient energy components [5]. This is important given that in some cases the small coefficients might still be important, e.g. being part of the frequencies of a single note, while in other cases the small coefficients might be sensor noise and thus not important [4].

3.2.1 Phase Reconstruction

The phase reconstruction is an important part for source extraction as the NMF does not keep the phase of the signal, which is needed to convert the time-frequency signal back into the time domain [4, 8]. For solving this problem, the most simple method is to use the Wiener phase reconstruction filter [4, 8] given by eq. 3.10

$$\mathbb{E}[\mathbf{V}^k_{mn} * |\mathbf{V}_{mn}*] = \frac{\hat{\mathbf{V}}^k_{mn}}{\hat{\mathbf{V}}_{mn}}\mathbf{V}*_{mn} = \frac{\mathbf{W}_{mk}\mathbf{H}_{kn}}{\sum_k \mathbf{W}_{mk}\mathbf{H}_{kn}}\mathbf{V}*_{mn} \tag{3.10}$$

where k indicated the kth source, $\hat{\mathbf{V}}$ being the estimated data matrix and $\mathbf{V}*$ being the complex data matrix with the phase information, e.g. a STFT, and m and n being the matrix indexes. Equation 3.10 is the element-wise multiplication and division and can also be written like equation 3.11.

$$\mathbb{E}[\mathbf{V}^k * |\mathbf{V}*] = \mathbf{V} * . \frac{\mathbf{W}_k\mathbf{H}_k}{\mathbf{WH}} \tag{3.11}$$

This method makes use of $\mathbf{V}*$ as a mask to determinate how much phase belong to each source by using a scalar value between zero and one, that is if only the kth source is present than eq. 3.10 becomes one multiplied $\mathbf{V}*$, if the kth source is not present then it becomes zero multiplied with $\mathbf{V}*$, if the kth source is present with non-k sources then the scalar becomes a decimal number between zero and one multiplied with $\mathbf{V}*$ [8]. This means that the source signals $\mathbb{E}[\mathbf{V}^k_{mn} * |\mathbf{V}_{mn}*]$ can be resynthesized in the time-domain using the overlap add synthesis (OLA) method [8]. However, the complex spectrogram of the separated source signals are unlikely to be equal to $\mathbb{E}[\mathbf{V}^k_{mn} * |\mathbf{V}_{mn}*]$ as the spectrogram is an inconsistent spectrogram, which does not match an actual time-domain signal [8].

3.3 State of the Art

A compositional model that have been used for wind reduction in mixed speech signals is the non-negative sparse coding (NNSC), which is similar to NMF in how it works, but with an added sparsity parameter in the activation matrix calculations, that controls the sparsity in the activation matrix [7]. This specific method has been successful used to separate wind and speech from a mixed signal and to increase the output SNR compared to the input SNR [7]. This study by Schmidt et al. [7] is considered the state-of-the-art as the two

methods (NNSC and NMF) are quite similar and both studies focus on the extracting speech from a signal mixed with wind. It should be pointed out that the multiplication update rule equations for the basic matrix and the activation matrix are different from those given in eq. 3.5 and 3.6. The cost function used for the NNSC was the squared error [7].

In the study by Schmidt et al. only one of the dictionaries was pre-calculated [7]. This dictionary was the wind dictionary, trained on a single wind sound file of half of a minute length, while the speech dictionary was calculated out from the mixed signal by allocating an empty speech dictionary with 64 components for the extraction of the non-wind part of the mixed signals [7]. They had selected the amount of components for pre-allocation from testing with different amounts of components for the speech part of the mixed signals [7].

For the experiment they made use of a 100 sentence from the GRID database mixed with wind interference under six different SNR values, ranging from zero to six dB [7]. The amount of iterations was 500 or when the square error was less than 10^{-4} [7]. A Hanning window of 35 ms and 75 % overlap was used, while the sampling rate was 16 kHz [7]. It should be noticed that the study by Schmidt et al. did not make use of the power spectrogram or the magnitude spectrogram, rather they lifted their STFT using a value of 0.6 or $|\mathbf{Y}|^{0.6}$ [7]. For comparison they used spectral subtraction and Qualcomm-ICSI-OGI, a method based on adaptive Wiener filtering and used for automatic speech recognition [7]. The measurements used in the study was the SNR_{out} and the word recognition rate [7]. They found that their NMF algorithm performed well in the terms of SNR as it increased the SNR_{out} and mainly did better than the comparison methods [7]. For word recognition rate, it did better than the spectral subtraction method over all SNR conditions, however, the Qualcomm-ICSI-OCG did better than their proposed method [7]. When the SNR value was 3 dB their method performed the same as the signals with no noise reduction and above 3 dB their method did worse than the signals with no noise reduction [7].

The differences between their study and this study was that Schmidt et al. only measured the SNR difference and the word recognition rate using an automatic speech recognition method, where this study measured the SNR, the quality, and the intelligibility. More different SNR values was tested, most importantly negative SNR values. Both dictionaries were trained using separated wind and speech signals, rather than having only a single trained dictionary.

3.4 Noise Reduction Methods for Comparison

Two other methods, beside the NNSC, for noise reduction were used in this study for comparison. These were the spectral subtraction and the minimum mean square error estimate of log-spectral amplitude, which will be explained in their own subsection.

3.4.1 Minimum Mean Square Error Estimate of Short-Time Log-Spectral Amplitude

Minimum mean square error estimate (MMSE) of short-time log-spectral amplitude (STSA) is a method used for stationary noise reduction [9, 10]. This method works by using a statistical model for the probability distribution of the speech and the noise Fourier energy component [9, 10]. The assumption of the model is that the Fourier energy components for each process can be modeled as a statistically independent Gaussian random values, where the mean of each component is also assumed to be zero [9, 10]. Because of the non-stationarity, it means that the variance of each Fourier energy component is time-varied [9, 10]. The estimated noise is removed using two probability density functions (PDF), one PDF uses the assumed statistical model parameters of the estimated speech log-spectral amplitude and the estimated noise log-spectral amplitude, while the other uses the assumed statistical model parameters of the estimated speech log-spectral amplitude, the estimated noise log-spectral amplitude, and the mixed signal log-spectral amplitude [9, 10]. It is important to notice that MMSE STSA assumes only that the speech part a of a mixed signal to be non-stationary, the noise is assumed to be stationary [9, 10], thus this method can suffer when the noise is non-stationary. When the speech is not visible present in the mixed signal the model consider the possibility that a speech component might appear with insignificant coefficient compared to the coefficient of the noise in a specific energy component as the author behind the MMSE STSA found this model to be more appropriated for speech signals when the low speech signal energy components are considered as if they were to be not present at all [9, 10]. The MMSE STSA makes use of a MMSE to estimate the complex exponential of the phase as the MMSE STSA is unable to estimate both parts simultaneously in an optimal way [9, 10]. The log-spectral amplitude was found to produce better results than the original STSA [10].

3.4.2 Spectral Subtraction

Spectral subtraction is a method that works on the STSA and the estimator is derived from an optimal variance estimator, where the estimator is optimised for maximum likelihood [9]. The spectral subtraction method is developed for improving speech signals mixed with broadband white noise [11] and the method works by subtracting the estimated noise power spectrum from the mixed power spectrum [11]. Typically the subtraction method suffer from musical noise, however, by subtracting an overestimated noise power spectrum and preventing the components from reaching under a threshold, it is possible to eliminate this problem [9, 11]. The method has the assumption that the power spectrum of the mixed signal is equal to the sum of the power spectrum of the speech and the power spectrum of noise alone [11]. The noise is estimated from periods of a STFT where no speech is present [11]. The noise is averaged and then smoothed in frequencies to create an averaged noise spec-

trum for subtraction [11]. The original phase of the mixed signal is retained for resynthesis [11]. Spectral subtraction can adapt by itself to any SNR as long time the method can reasonable estimate the noise power spectrum [11]. Lastly, it have been found that both quality and intelligibility are improved using this method [11].

Chapter 4

Implementation

This chapter will explain processing done to the audio signals and the implementation of the noise reduction methods. The code was implemented in Matlab R2018a and with a 4.0 GHz CPU.

4.1 Audio Signals

The speech audio signals came from the TIMIT corpus[12], while the wind signals were either gathered from different locations in Denmark/Greenland, from Nonspeech Sounds Corpus [13] or from soundsnap.com. A total of 99 speech signals was used and 28 wind signals were used for this study.

The sampling rate of the audio signals was 16000 Hz. Any signals above or under this was resampled to 16000 Hz. The resampling used an anti-aliasing finite impulse response filter and compensated for delay introduced using the filter. The decision behind using 16000 Hz sampling rate was that one of the measurements in this study, the PESQ, required a sampling rate of 16000 Hz, while most of the signals used had a sampling rate of 16000 Hz, hence no need to resample most of the signals. The Short-Time Objective Intelligibility measurement required a different sampling rate, however, the code used to implement this measurement would resample the signals automatically.

Given that the noise reduction methods were tested under different levels of SNR, the SNR was changed using the two following equations, eq. 4.1 and eq. 4.2.

$$\sqrt{\sigma_i^2} = \frac{\sigma_s^2}{10^{\left(\frac{SNR_{in}dB}{10}\right)}} \tag{4.1}$$

$$x_{i,after} = \frac{x_{i,before}\sigma_{i,after}}{\sigma_{i,before}} \tag{4.2}$$

where $\text{SNR}_{in}dB$ is the wanted signal-to-noise ratio, σ_s^2 is the variance of the speech signal, σ_i^2 is the variance of the noise/interference signal, and x_i is the noise/interference signal.

4.2 Noise Reduction Methods

To compare the ability of the NMF to remove the interferences, three other methods was implemented. Two of these were the spectral subtraction and minimum mean square error estimate of log-spectral amplitude, these two methods are developed to remove stationary noise and wind sounds can contain both stationary and non-stationary frequencies as given in figure 1.1. The third one was the non-negative sparse coding.

4.2.1 Noise Reduction Method Implementation

The implemented non-negative matrix factorisation was based upon the math given in [5], the math is given in section 3.2.

Spectral subtraction and minimum mean square error estimate of log-spectral amplitude was implemented using the Matlab toolbox Voicebox using their default options.

The NNSC method was implemented using the equations given in [7]. These equations were

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T\mathbf{W}\mathbf{H}+\lambda} \tag{4.3}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{\mathbf{V}\mathbf{H}^T+\mathbf{W}.(1(\mathbf{W}\mathbf{H}\mathbf{H}^T.\mathbf{W}))}{\mathbf{W}\mathbf{H}\mathbf{H}^T+\mathbf{W}.(1(\mathbf{V}\mathbf{H}^T.\mathbf{W}))} \tag{4.4}$$

λ is the sparsity parameter, lower values allows for more sparsity. 1 is square matrix of ones.

After [7] a Monto Carlo simulation could be used with the NNSC, however, it was decided not to use a Monto Carlo simulation as it was not used for the NMF either. The chosen λ values in this study were for noise 0.1 and for speech 0.3 under the training. For testing the selected λ value was zero. The reason for going with different values than those in [7] came from trail-and-errors to find values that modeled the spectrograms well and that the signals for training were longer than those in [7]. The reason for setting the λ value to zero under the testing was that Schmidt et al. had set the only trained dictionary to zero when conducting their testing [7]. Lastly, it was decided to use the power spectrogram to ensure both NNSC and NMF used the same spectrograms.

Chapter 5

Measurements and Experiment

This chapter will cover the measurements used in this study and the experiment set-up and conditions.

5.1 Measurements

To measure the impact of the NMF it was decided to measure the quality and intelligibility of the output signal by using the The Perceptual Evaluation of Speech Quality (PESQ) and the Short-Time Objective Intelligibility (STOI), both which are being described in the two subsections. The SNR was also implemented to measure if the dB ratio between the speech signal and the wind signal was changed.

5.1.1 The Perceptual Evaluation of Speech Quality

The Perceptual Evaluation of Speech Quality is a ITU-T (International Telecommunication Union) recommendation, know also as P.862, and is a model for assessment of speech quality [14]. This method works by comparing a reference signal and a degraded version of the reference signal to each other [14]. Both of the signals are aligned in intensity to a standard listening level, then filtered using a Fourier-transform witch uses a model of a standard telephone handset [14]. Then both signals are aligned to each other in the time domain, before being processed through a psychoacoustic model [14]. The psychoacoustic model involves equalising for linear filtering, for gain variation, and more, thus the score is not affected by differences in intensity between the two signals [14]. Two distortion parameters are extracted from the difference between the signals, these being calculated in time and frequency, before being mapped to a prediction of the subjective mean opinion score (MOS) [14]. The MOS score

is then calculated into the MOS objective listening quality (MOS-LQO) which ranges from 1 (bad) to 5 (excellent) [14].

It should be mentioned that the PESQ code used in this study was of an older version and gave results from 0.5 (bad) to 4.5 (excellent), however, there was no difference between the minimum and maximum meaning of the values between the old version and the new version.

The PESQ require the sampling rate to be either 8000 Hz or 16000 Hz. Lastly, the PESQ is a prediction and an actual person might rate differently.

5.1.2 Short-Time Objective Intelligibility

The Short-Time Objective Intelligibility Measure (STOI) is useful for methods where mixed signals, consisting of speech and noise, is processed using a time-frequency weighting [15, 16]. The STOI analyses the signals using 15 one-third octave bands in the time-frequency domain by measuring the signal-to-distortion rate on each one-third octave band, then a intermediate intelligibility measure is conducted on each one-third octave band, and finally an the objective intelligibility measurement is calculated out from the measurements of all one-third octave bands and frames [15, 16]. The STOI gives a score between 0 and 1, where a higher score indicates better intelligibility of the processed signal [15, 16]. The STOI expects the sampling rate to be 10000 Hz and thus other sampling rates have to be resampled. The clean and the processed signal are assumed to be time-aligned with each other [15, 16].

The measurement that STOI gives is a prediction of the intelligibility of the signal post-processed compared to the signal pre-processed [15, 16]. However, it should be noticed that the STOI is a prediction and thus scores by a person could differ. The code to implement this method was given in [15].

5.1.3 Signal-to-Noise Ratio

The signal-to-noise ratio (SNR) is a quite important measurement when it comes to noise reduction [3]. The SNR is the relevant ratio of the intensity of wanted signal over the intensity of the noise [3]. The less noise compared to the wanted signal, the higher SNR value. The relevant improvement in SNR is given by SNR_{in} and SNR_{out} [3] and is defined as

$$\Delta\text{SNR} = \frac{\text{SNR}_{out}}{\text{SNR}_{in}} \tag{5.1}$$

The input SNR is the ratio of the wanted signal over the unwanted noise and is defined as

$$\text{SNR}_{in} = \frac{\sigma_s^2}{\sigma_v^2} \tag{5.2}$$

where σ_s^2 and σ_v^2 are the variances of the signal $\mathbf{s}$ and $\mathbf{v}$, $\mathbf{s}$ being the wanted signal and $\mathbf{v}$ being the noise signal.

The output SNR is the ratio of the wanted signal to the noise signal and is given by

$$\text{SNR}_{out} = \frac{\sigma_{ys}^2}{\sigma_{yv}^2} \tag{5.3}$$

where σ_{ys}^2 and σ_{ys}^2 are the processed wanted signal **s** and processed noise signal **v**.

Eq. 5.2 and eq. 5.3 can be combined into one equation [3], equation being

$$\Delta\text{SNR} = \frac{\sigma_{ys}^2}{\sigma_{yv}^2}\frac{\sigma_v^2}{\sigma_s^2} \tag{5.4}$$

The noise in this project for the SNR_{in} was the wind noise added over the wanted signal and for SNR_{out} the noise was the audio that was extracted using the noise components of **W** and **H**, hence the noise for SNR_{out} could consist of actual speech.

The SNR can be converted to decibels (dB) [3] using

$$\text{SNR}_{db} = 10\log_{10}(\text{SNR}) \tag{5.5}$$

Decibels is a logarithmic unit used to express the level of one value relative to another value, in SNR being the SNR_{out} and SNR_{in} [3].

Given that the experiment of the study would test the NMF under different conditions of SNR_{in}, the SNR_{in} was not calculated, but rather changed to fit the specific SNR using eq. 4.1 and eq. 4.2. However, the SNR_{out} was calculated to see if the dB ratio between the extracted speech and extracted wind sounds changed, that is if the SNR_{out} was different from the SNR_{in}. In a "perfect" separation of speech and wind without any distortion the SNR_{out} should be the same as the SNR_{in} (However, there would be a chance that some speech/wind is modeled by the other dictionary with a mix that allows the variance of the two non-mixed signals to give the same SNR). If the SNR is higher it indicate that, at least part, of the speech is extracted using the wind dictionary and if the SNR was lower, that part of the wind is extracted by the speech dictionary, or that something has happen to the signals like distortion or reduction in intensity.

5.2 Experiment

Multiple conditions were used in the experiment, these being the amount of speech and wind components, the SNR_{in}, and the β-divergence. The speech components for testing ranged from 30 to 150 with an incrementation of 20 (a total of 7 different amounts), while for the wind the amount of components ranged from 10 to 90 with an incrementation of 20 (a total of 5 different amounts). The components were extracted from 90 speech signals (45 females and 45 males) that had been combined into one long signal (length of around 270 seconds) and 20 wind signals that had been combined similar to the speech

SNR (dB)	-15 - 5, incrementation of 5
Wind components	10 - 90, incrementation of 20
Speech components	30 - 150, incrementation of 20
β-divergences	Kullback-Leibler and Itakura-Saito
Window	Hanning
Segment length	25 ms (401 samples)
Overlap	50 %
fs	16 kHz
Spectrogram	Power
Iterations Testing	1500
Iterations Training	1500

Table 5.1: Experiment Conditions

signals (length of around 205 seconds). The reason for the selected component ranges came from looking through the NMF spectrograms to get an idea when there where to few or to many components to get any useful extraction. The incrementation was chosen as a change of 10 components did not have much of an impact of the NMF spectrograms to justify a lower incrementation. The β-divergence used was the Kullback-Leibler and Itakura-Sauto divergence to evaluate which β-divergence could separate the mixed signals best.

For testing nine speech and wind signals were used (five males and four females speakers). Each speech signal was mixed with a single wind signal, where the shortest length of either the speech or the wind signal determined the overall mixed signal length.

Each unknown mixed signal, consisting of data that had not been trained on, was changed to have different levels of SNR, ranging from -15 dB speech to wind to 5 dB speech to wind with an increase of 5 dB each time. This was to evaluate the ability of the NMF to extract the speech from the mixed signal under different SNRs.

For the STFT a Hanning window was chosen with 50 % overlap + one sample, to ensure it obeyed the constant overlap-add. The segment length was 25 ms (401 samples, uneven because of the Hanning window), the number of discrete-Fourier-transform (DFT) bins was the same as the segment length. The spectrogram used was the power spectrogram.

To get an overview of the experiment conditions, see table 5.1. A total of 350 different combinations of conditions were present in the experiment.

Chapter 6

Results

This chapter goes through the PESQ results first, then the STOI results, and the SNR results. Lastly, an informal listening test, with only the author, was conducted as the results of the three measurements contradicted each other regarding PESQ and STOI compared the SNR_{out} and is given at the end of the result chapter.

The results given in section 6.1, 6.2, and 6.3 are the mean values of the nine mixed signals of the testing data. Each plot show the standard deviations of the mean values. It should be noticed that the SNR results only exist for the NMF and NNSC in this study and not the two comparison methods or the non-processed signals.

Three different kinds of plots will be presented in each section: The first kind of plots has the speech components fixed and show the effects of the different wind components with a plot for different SNR value. In the second kind of plots the wind component is fixed and the effects of the different speech components are showed for different SNR values. The last kind of plots have the wind and speech components fixed and shows the effects of the different SNR values. Lastly only some of the result plots are showed in the result chapter, however, all plots are given in addendum A, B, and C.

The names in the legends of the plots are as followed: NP = non-processed, IS = Itakura-Saito, KL = Kullback-Leiber, NNSC = non-negative sparse coding, SS = spectral subtraction, and STSA = minimum mean square error estimate of short-time log-spectral amplitude.

6.1 PESQ

As the number of speech components and SNR values were fixed some noticeable trends appeared, see figure 6.1, as it can be observed that higher numbers of wind components would in most SNR cases lower the PESQ scores. Lower numbers of speech components and higher numbers of wind components increased the difference between the two β-divergences. The NNSC did similar

to Kullback-Leibler in most conditions. This pattern existed for all the different amounts of SNR values and amount of speech components. In almost all conditions the NMF and the NNSC did worse than the non-processed signals and the two other methods. However, for low dB values Kullback-Leibler and NNSC could, depending on the number of wind components, do better than the non-processed signal, see figure 6.2, and for - 15 dB even do better or similar to spectral subtraction and MMSE STSA.

When the wind components were fixed, figure 6.3, it can be observed that more speech components could improve the PESQ score, mainly for Itakura-Saito while Kullback-Leibler did not really improve with higher amount of speech components. For low amount of wind components less amount of speech components were needed for the two different β-divergence to be similar in score. The NNSC did overall better than the two β-divergences and similar to when the amount of speech components was fixed it could do better than the non-processed signals, when the SNR was -15 and -10 dB. It should be noticed that generally the PESQ values would start lower when the amount of wind components increased.

When the SNR was fixed as seen in figure 6.4 the PESQ scores increased close to linearly for spectral subtraction and MMSE STSA, and partly linearly for the non-processed signals, when the SNR becomes higher. This was expected as the PESQ measurement compare the mixed signal to the pure speech signal and with higher SNR values the less wind was present compared to speech. The results for the two non-negative measurements did not not have this linear increment, at least to the same amount as the other measurements and the non-processed signals. They did improve in most conditions.

Hence, the PESQ results indicated that the two non-negative methods caused problems with the quality of the outputs, something that spectral subtraction and MMSE STSA did not do, rather they improved the quality. Under some few specific conditions Kullback-Leibler and NNSC could generate PESQ scores that were better than the non-processed PESQ values. Lastly, it should be noticed that the PESQ results of the NMF and the NNSC had a bigger standard deviation than the other methods and the non-processed signals. No pattern could be found, related to the amount of speech and wind components or SNR values that could explain why some conditions generated bigger standard deviations than other conditions. A possible reason for the lower PESQ scores will be mentioned in section 6.4.

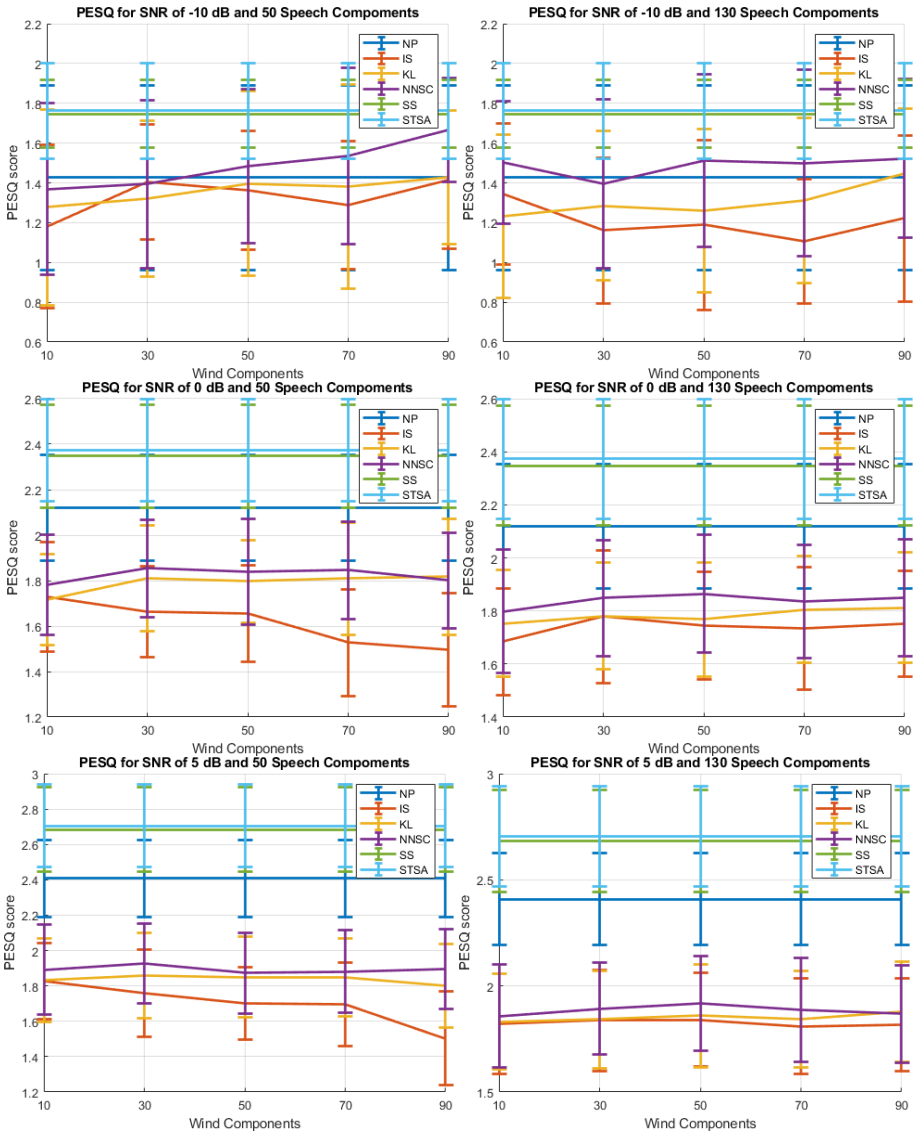


Figure 6.1: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

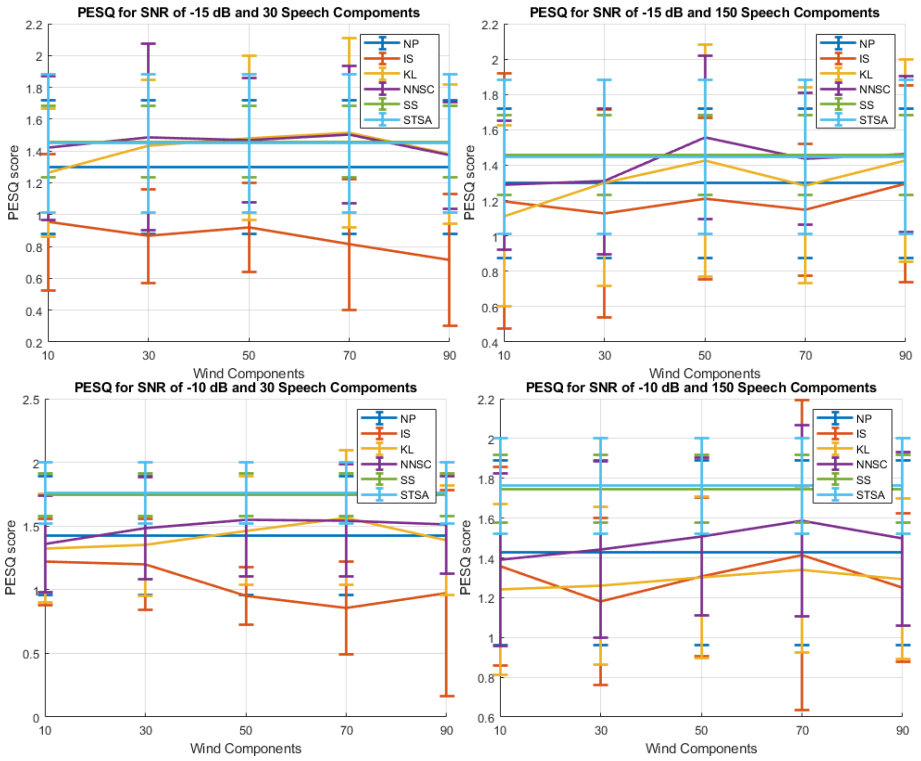


Figure 6.2: Two of the non-negative methods surpassing the non-processed signal in PESQ score. Top: -15 dB. Bottom: -10 dB

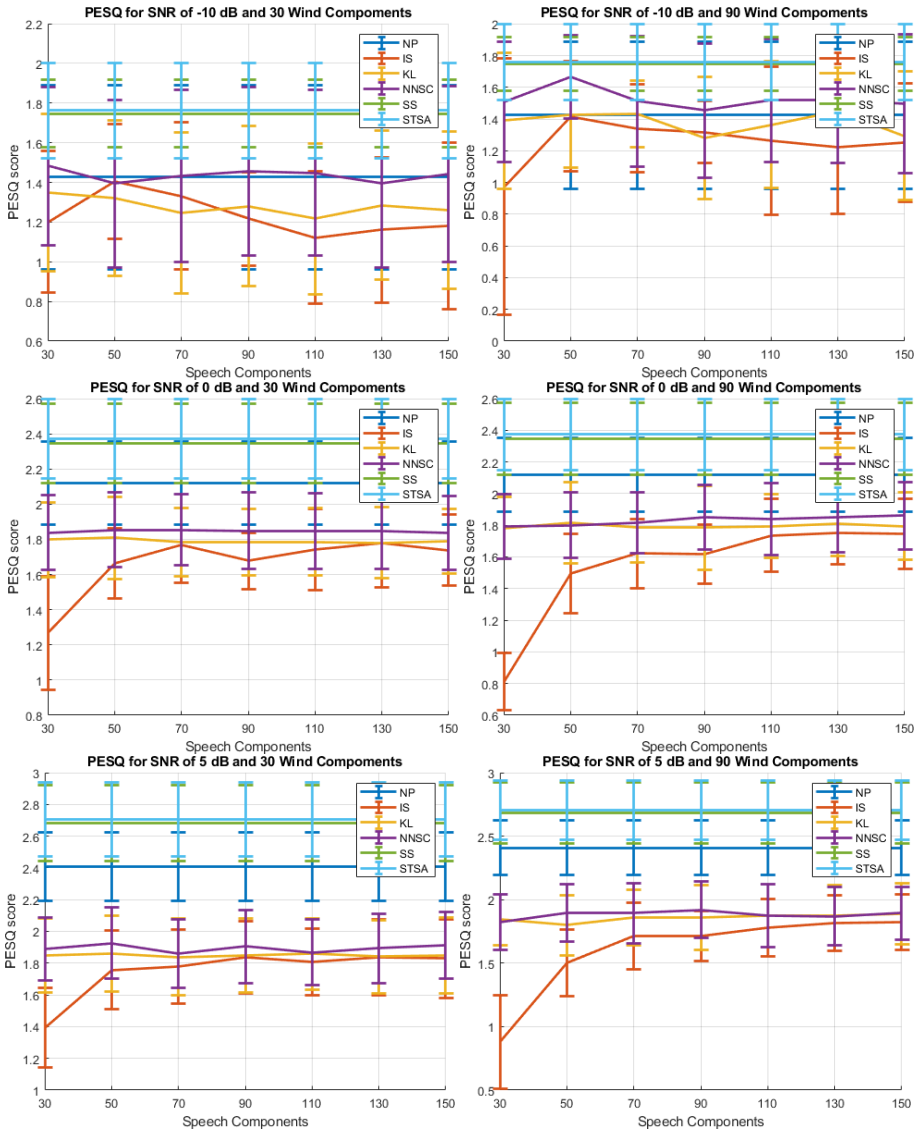


Figure 6.3: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

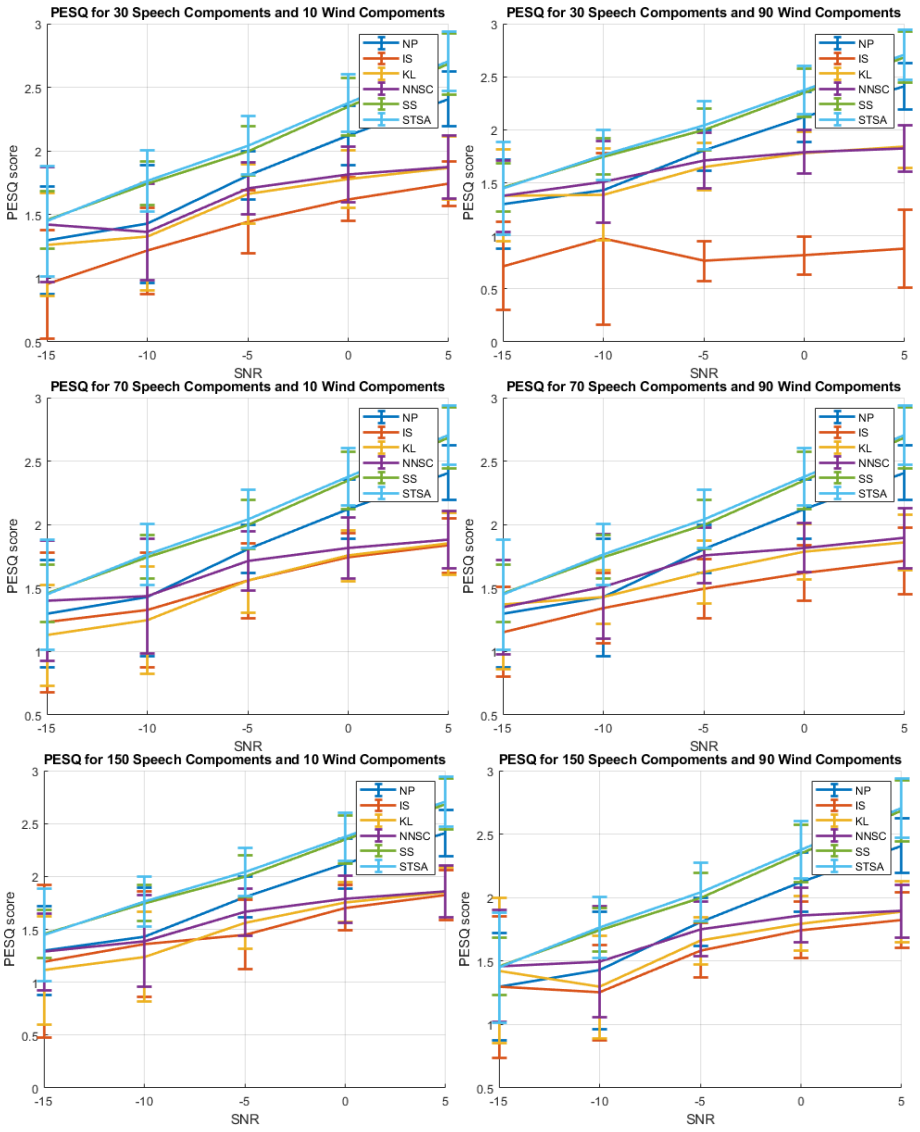


Figure 6.4: Top plots: 30 Speech components. Middle plots: 70 Speech components. Bottom plots: 150 Speech components

6.2 STOI

The STOI scores shared similarities with the PESQ scores in that the three non-negative methods did overall worse than the non-processed signals and the two stationary noise reduction methods. As it can be observed in 6.5 and 6.6 the non-processed signals and the two stationary noise reduction methods

had very similar scores. It can also be observed in both figures that Kullback-Leibler and NNSC had very similar results most of the times, while Itakura-Saito did mainly slightly worse, but with high amount of speech components the three non-negative methods ended up being fairly similar regarding the STOI scores.

It can be observed that the amount of wind components did not have much of an impact on the scores for both Kullback-Leibler and NNSC, however, when the SNR increased the Itakura-Saito started to suffer when the number of wind components was increased, but only for lower amounts of speech components as it can be seen in figure 6.5. When the number of wind components was fixed the effect of the number of speech components was more noticeable for the Itakura-Saito divergence where lower numbers of speech components gave much worse results as seen in figure 6.6.

From figure 6.7, the number of speech and wind components was fixed and the STOI values were plotted over the different SNR values, it can be observed that all of the different methods had fairly linear incrementation,which was expected as the base non-processed signals had better intelligibility, as there was less wind compared to speech.As it be noticed the non-processed signals, the spectral subtract, and MMSE STSA had almost the same STOI scores.The two non-negative measurements also had very similar scores most of the times, in some conditions the Itakura-Saito divergence was quite below Kullback-Leibler and NNSC when there were few speech components. It should be mentioned that effect of the SNR values decreased after the 0 dB mark as the values below this point increase in score quicker than after the 0 dB mark. Most of the result plots show NNSC and the two β-divergence approaching a similar point at the 5 dB mark, however, this is as mentioned not for all plots.

A possible reason for the lower STOI scores will be mentioned in section 6.4.

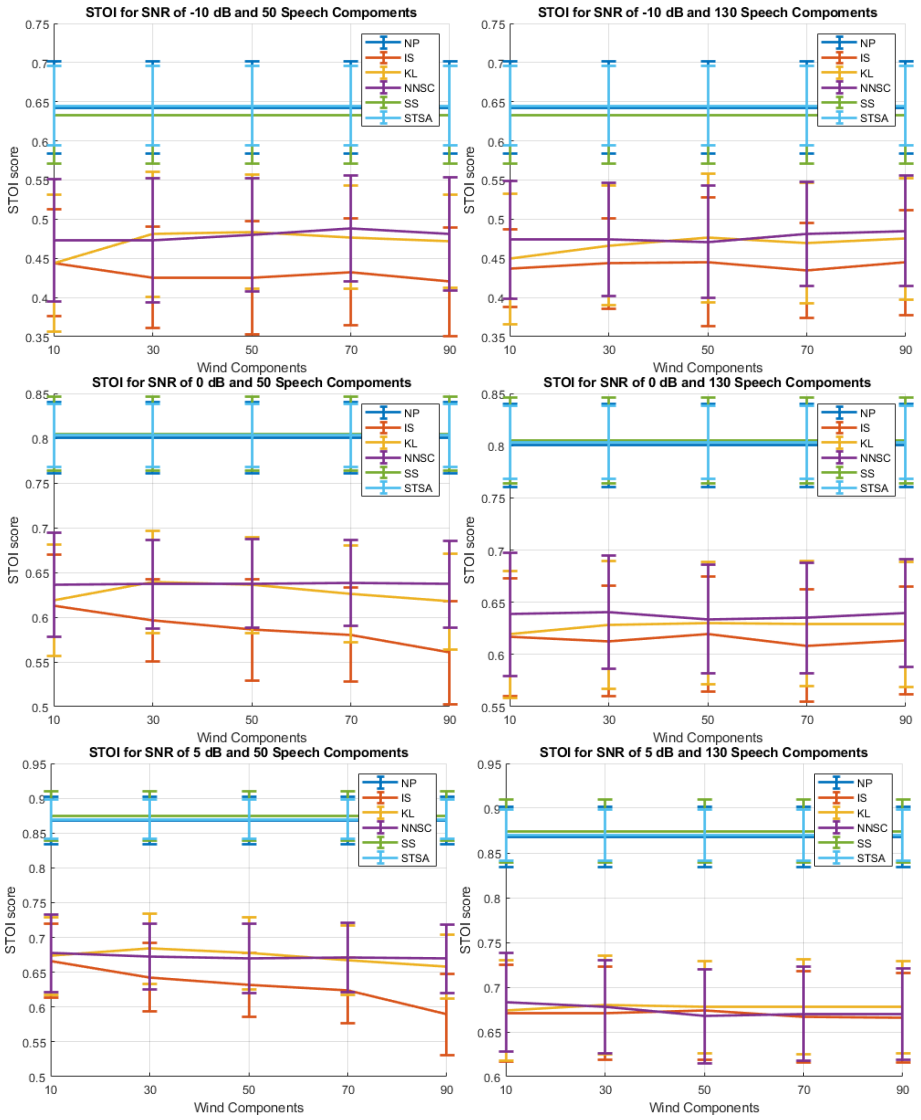


Figure 6.5: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

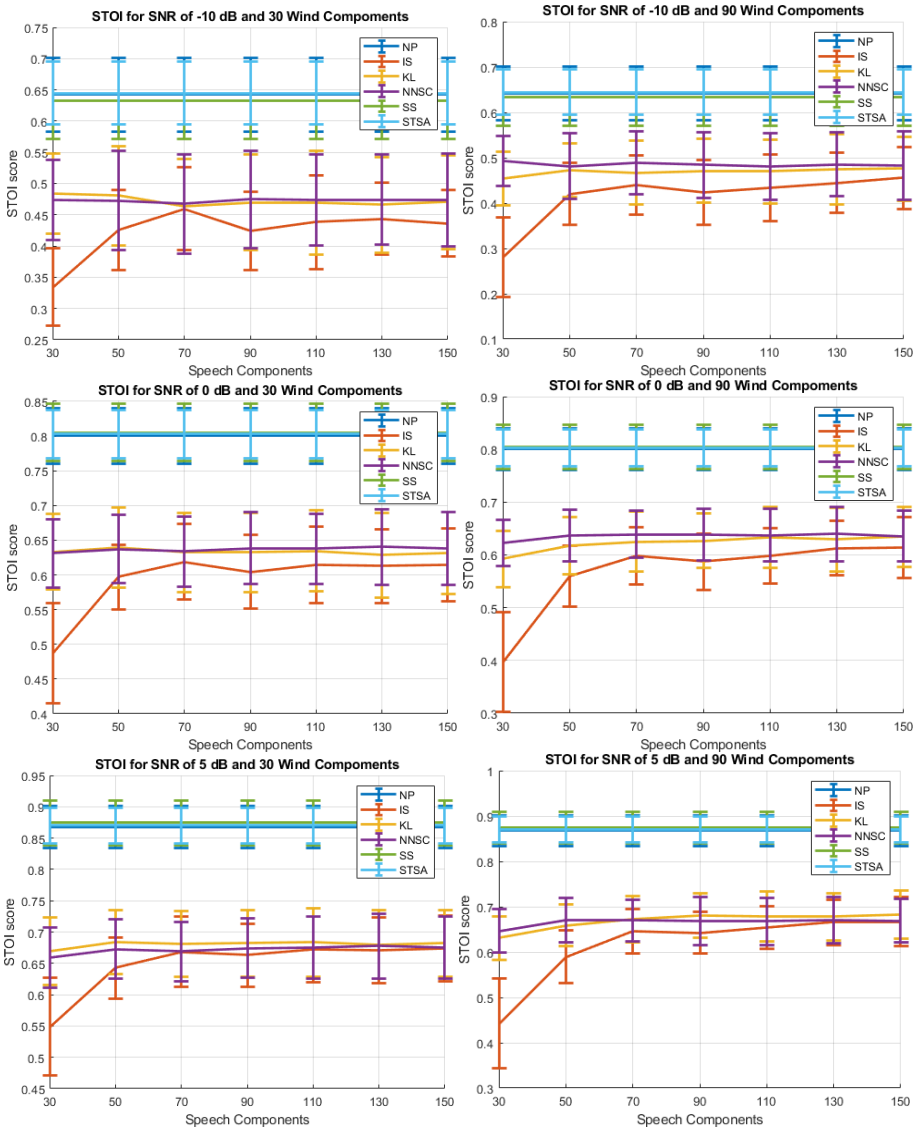


Figure 6.6: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

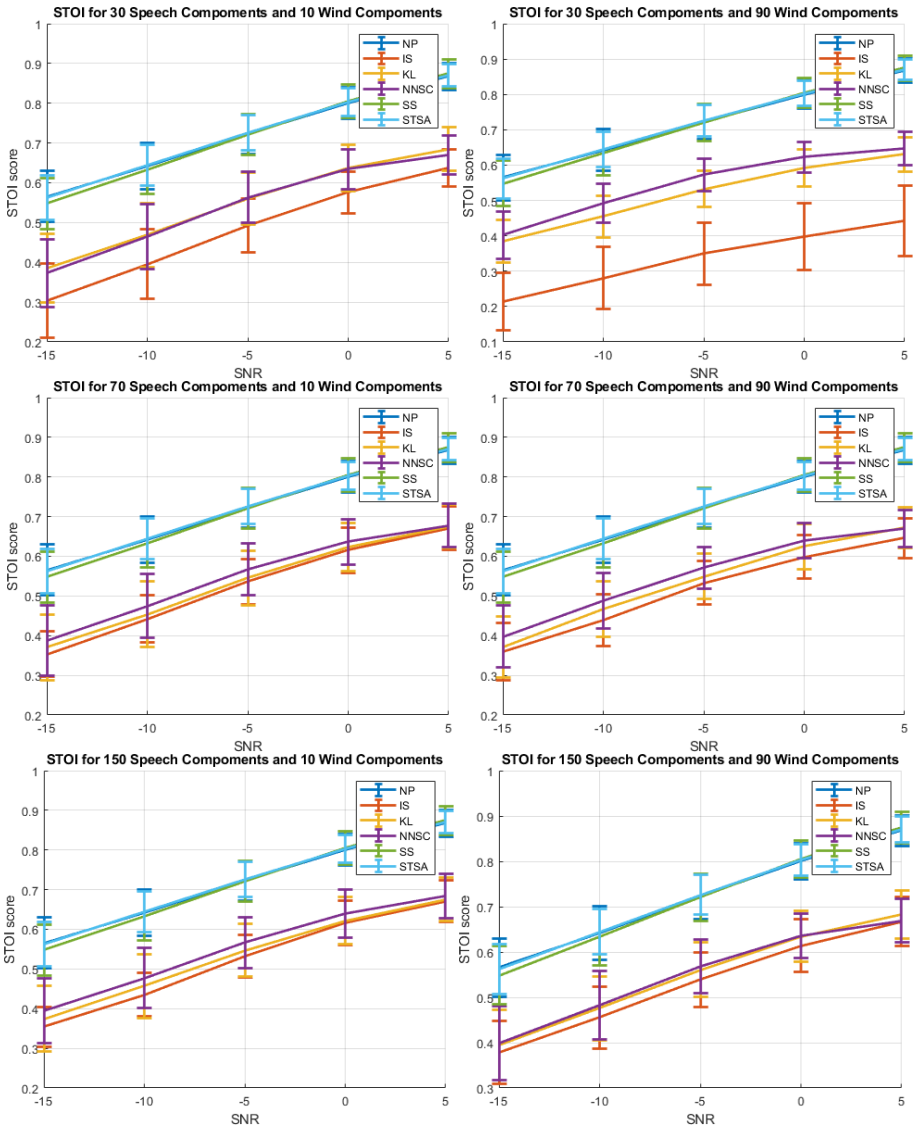


Figure 6.7: Top plots: 30 Speech components. Middle plots: 70 Speech components. Bottom plots: 150 Speech components

6.3 Signal-to-Noise Ratio Out

As mentioned in the beginning of this chapter this measurements only has scores for NNSC and NMF. This is because the implementations of spectral subtraction and MMSE STSA only gave the the estimated signal $\hat{y}(k)$, not the estimated $\hat{s}(k)$ and estimated $\hat{v}(k)$.

When the number of speech components are fixed as seen in figure 6.8 the SNR_{out} decreased as the number of wind components are increased. When the number of speech components was below the 70 - 90 components the Itakura-Saito divergence, mainly, did worse than when it was above the 70 - 90 components. Kullback-Leibler and NNSC had similar SNR_{out} results.

Looking at the results when the number of wind components was fixed, figure 6.9, the SNR_{out} mainly increased when the number of speech components was increased.

When both the speech and wind components were fixed, as seen in figure 6.10, the SNR_{out} increased when the SNR_{in} increased, which is expected as there are less intensity of the wind signals compared to the intensity of the speech signals. When the amount of wind components increased, the closer the two β-divergences and the NNSC were to each other.

It can be viewed in figure 6.9, figure 6.9, and figure 6.10that the SNR_{out} mainly differ from the SNR_{in}.

A possible reason for the SNR_{out} scores will be mentioned in section 6.4.

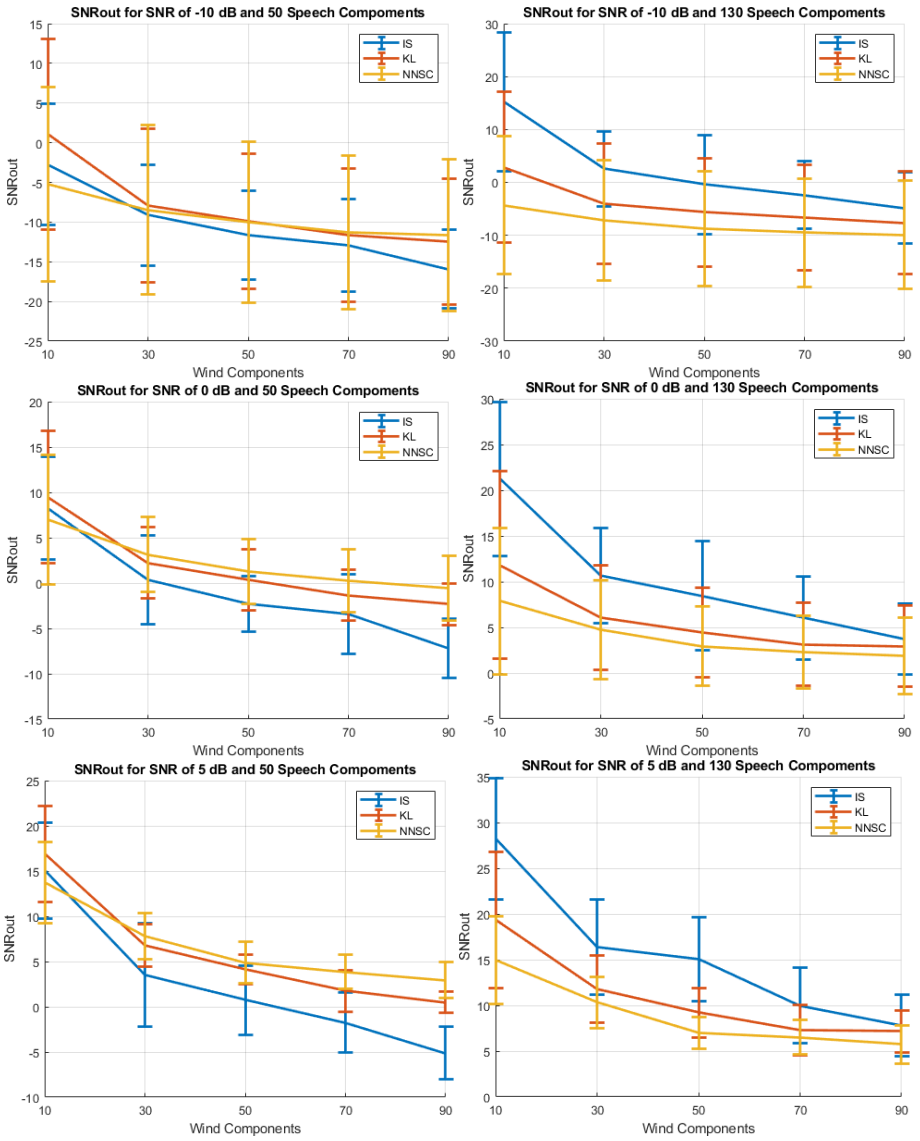


Figure 6.8: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

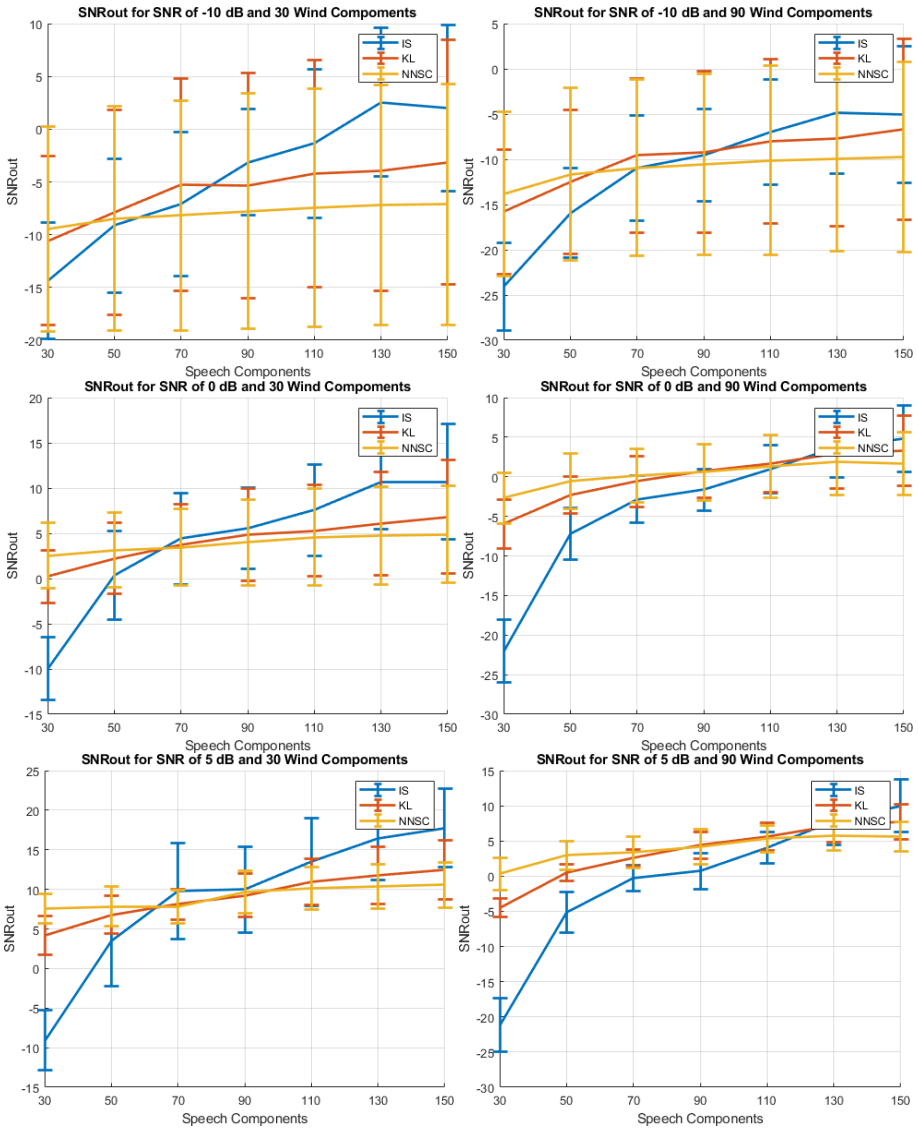


Figure 6.9: Top plots: SNR of -10 dB. Middle plots: SNR of 0 dB. Bottom plots: SNR of 5 dB

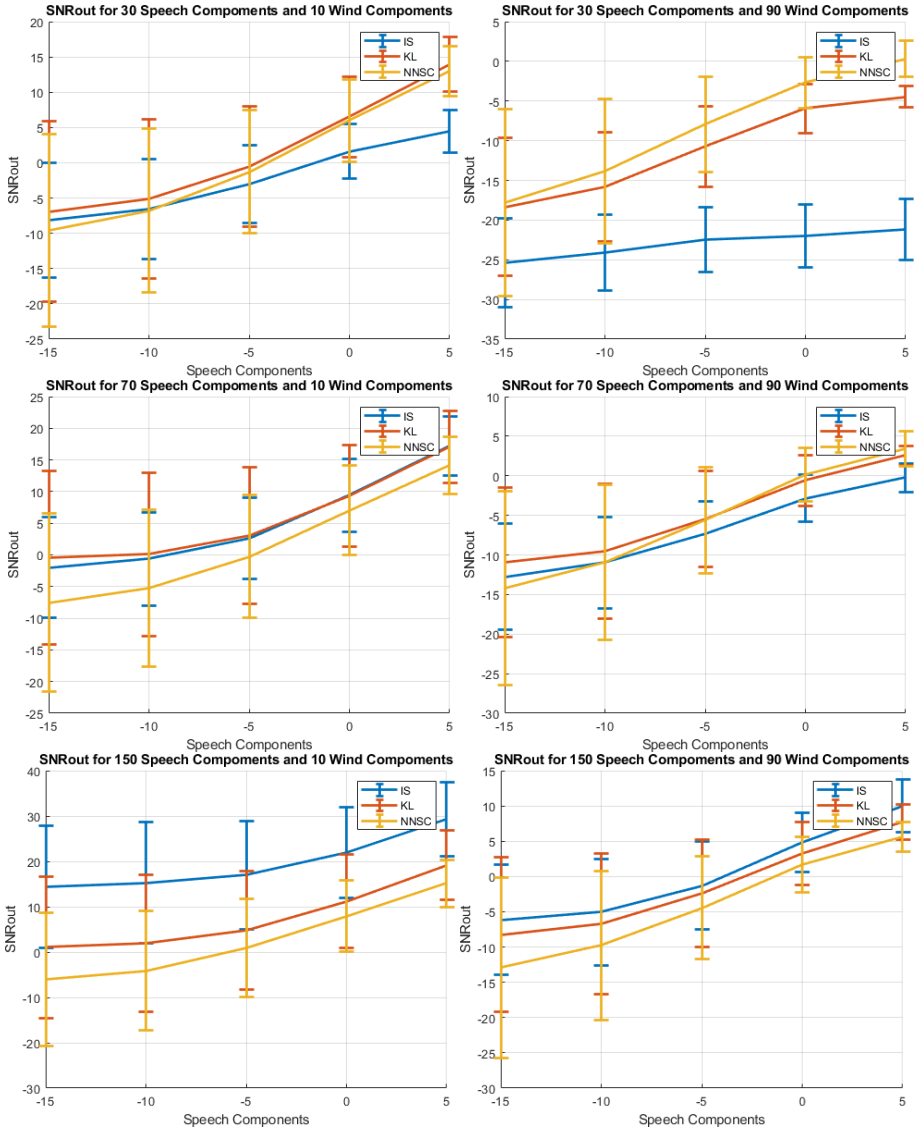


Figure 6.10: Top plots: 30 Speech components. Middle plots: 70 Speech components. Bottom plots: 150 Speech components

6.4 Informal Listening Test

Because of the SNR_{out} indicated that the NNSC and NMF improved the outputs compared to the non-processed signals, while the PESQ and the STOI indicated that the NNSC and NMF did worse than the non-processed signals it was decided to listen to the output signals of all four methods.

Both spectral subtraction and MMSE STSA removed part of the wind sounds, but most of it stayed. This was expected as both methods were developed for stationary noises [9, 10, 11] and thus would suffer when trying to remove non-stationary noises as they could not remove the non-stationary noise, however, these methods are designed to not affect the speech part of the mixed signals [9, 10, 11]. The speech sounded quite similar to the non-processed mixed signals, which would explain their better PESQ and STOI scores.

Regarding the two non-negative methods, the NMF and NNSC, a lot of distortion could be heard in all of the processed outputs and most of the distortion sounded like it followed the pattern of the non-processed wind in the mixed signals. At the same time the speech was also distorted, but not as much as the wind. The distortion in the NNSC outputs sounded like it had a higher frequency than the outputs of the NMF. Hence this could explain the worse PESQ and STOI scores than the non-processed signals. The signal outputs created from the extracted spectrograms using the speech dictionary contained a lot of the distorted wind sounds, while the spectrograms extracted using the wind dictionary had some mix of wind and speech. The amount of speech to wind sounded like it depended on the amount of speech/wind components, the more components present in one dictionary, the more of the other source that dictionary sounded like it extracted. This could explain the overall improvement in SNR_{out} as the speech signals would contain a lot of wind and speech, while the wind signals would contain some wind and speech, in most cases only a little amount of speech. That could be why increasing the number of wind components lowered the SNR_{out} and increasing the number of speech components increased the SNR_{out} as the SNR_{out} was calculated as the intensity of speech over the intensity of the noise.

Chapter 7

Discussion

The discussion chapter starts with discussing the NMF and the different conditions it was tested under, the different SNR, the β-divergence, and the combination of components. Then the noise reduction methods, the data used in the NMF, the measurements used for this study, listening, novelty of the study, the limitations of the study, and lastly possible improvements.

7.1 Non-Negative Matrix Factorisation

The basic NMF was implemented for this study as given in chapter 3.2. Other versions of the NMF that exist is updates of $\mathbf{W}$ and $\mathbf{H}$ only for a specific β-divergence, which are computational optimised, but lacks the diversity, but if the specific β-divergence is known to solve the problem well, the specialised updates are better [4]. Some versions of the NMF are designed around specific restrictions, e.g. the sparsity restriction, as normally they require changes to the multiplication update rule to ensure the cost function decreases [4, 17]. Other versions of the NMF makes use of either a Bayesian modified version [6] or has a statistical model of the speech and the noise to apply co-occurrence statistics on the basic matrix $\mathbf{W}$ to encourages the output signals to have statistics that are similar to the statistics of the priors [18]. However, these were not implemented as it was decided to test the ability of the basic NMF method to solve the problem of separating speech from a signal mixed with wind and to figure out why it either worked or did not work.

7.1.1 Signal-to-Noise Ratio

In this study multiple kinds of SNR values were used, ranging from -15 dB to 5 dB with an incrementation of 5 dB. The reason for the selected incrementation was to get a more overall idea about the impact of the SNR had on the NMF, rather than an idea about effect of each possible specific SNR value between

-15 dB and 5 dB had on the separation of wind and speech from the mixed signal.

7.1.2 Dictionary Components

Only some few specific number of components were used for the dictionary creation, for wind a minimum of 10 and maximum of 90 components and for speech a minimum of 30 and a maximum of 150 components, both with an incrementation of 20 components. The reason for this decision was that from looking at the results only using some few signals and with an incrementation of 10 components, the more combinations did not yield a result that really differed from having the incrementation of 20 components, thus it was decided to use fewer conditions when running with the full testing set.

The amount of dictionary components could have been increased, however, because of the STFT segment length the total frequency bins below the Nyquist limit was 201 bins, which limited the total amount of components in each dictionary and incrementation could lead to components that described very little of the data and could not be generalised. If the number of components reached the number of frequency bins, no unique solution could be found anymore and everything could be modeled correctly, however, the ability to separate the different parts of a signal, e.g. speech and wind, would be lost [4]. At the same time it was found that increasing the number of components also increased the change of modeling the wrong source and even a fairly low amount of components could model the wrong source as for most signals the wind dictionary could model, to some extent, understandable speech with only 30 components.

7.1.3 Kullback-Leibler and Itakura-Saito

Two different kinds of β-divergences, the Kullback-Leibler and Itakura-Saito, were used. However, the math used for the cost function and the multiplication update rule allowed for any β value and not just 1 and 0. The decision behind only using those two specific β values was that these were found to be the most common used for audio processing [1, 2, 4, 5, 18].

When looking through the spectrograms for different amount of components, it was noticed that Kullback-Leibler produced NMF spectrograms $\hat{\mathbf{V}}$ that had correctly placed coefficients in the energy components that in spectrogram $\mathbf{V}$ contained high coefficients, with fewer components than what Itakura-Saito did, as already around the 20-40 components most of the high coefficients of the energy components were similar to spectrogram $\mathbf{V}$. However, Kullback-Leibler was more imprecise as a lot of small coefficients in different energy components were modeled, which did not exist in $\mathbf{V}$. On this point Itakura-Saito did better, it did not capture the energy components with high coefficients as well as Kullback-Leibler, but it was better at modeling the energy components with low coefficients [5].

An interesting trend was noticed for Kullback-Leibler when the amount of components was increased. For each incrementation more components were

used to model higher and higher frequencies correctly with little change for lower frequencies with a clear line between where it did well modeling and where it did not. This could indicate that the NMF made use of the increased amount of components by separating the different columns into different components rather than using the increased amount of components to model an entire column of frequencies with a single components. This would affect the ability of separation as the NMF could then use the non-generalised components to modeled specific parts of each column in the mixed signal spectrogram, i.e. it could model parts of overlapping wind and speech in a column with the same component instead of using one generalised wind component for the wind part and one generalised speech component for the speech part of the mixed column. This could explain the distortion present in the audio output of the NMF as it would have place part of the speech and the wind columns into the wrong dictionary and not the entire column of either speech or wind. Hence, this was a negative reason for just adding more components for training without informing the NMF about how to use the increased number of components, which could have been solvable with restriction of component usage.

Itakura-Saito suffered from distortion too, however, this could not be explained in the same way as the distortion in Kullback-Leibler might have been able to be explained. The reason for this was from looking through the spectrograms for different amount of components, no notable pattern for whether it used the increase amount of components for modeling specific frequencies could be observed. Rather it had focused on modeling the pattern of the original data matrix, which was expected [5]. There did exist a possibility that it used the increased amount of components to separate what used to be modeled using a single basic vector into multiple basic vectors, thus being able to model other kind of sources than what it had been trained on, however, higher numbers of components made $\hat{\mathbf{V}}$ less blurry regarding the energy components with high coefficients.

A simple case is given in figure 7.1. In this case only 30 components have been used for training of the wind dictionary. As it can be observed Kullback-Leibler has focused on modeling the spectrogram energy components with high coefficients correctly, while the energy components with low coefficients had been modeled imprecise. Itakura-Saito, however, has focused just as much on modeling the low coefficients energy components as the high coefficients energy components as it can be observed that the high coefficients energy components are more blurry than in the case of Kullback-Leibler, but the low coefficients energy components are modeled much better and follow the shape of the original spectrogram. This was expected as Itakura-Saito is scale invariant, while Kullback-Leibler is not [5].

7.2 Noise Reduction Methods

In this study, four different methods of noise reduction was used. The non-negative matrix factorisation using two different β-divergence values, the non-

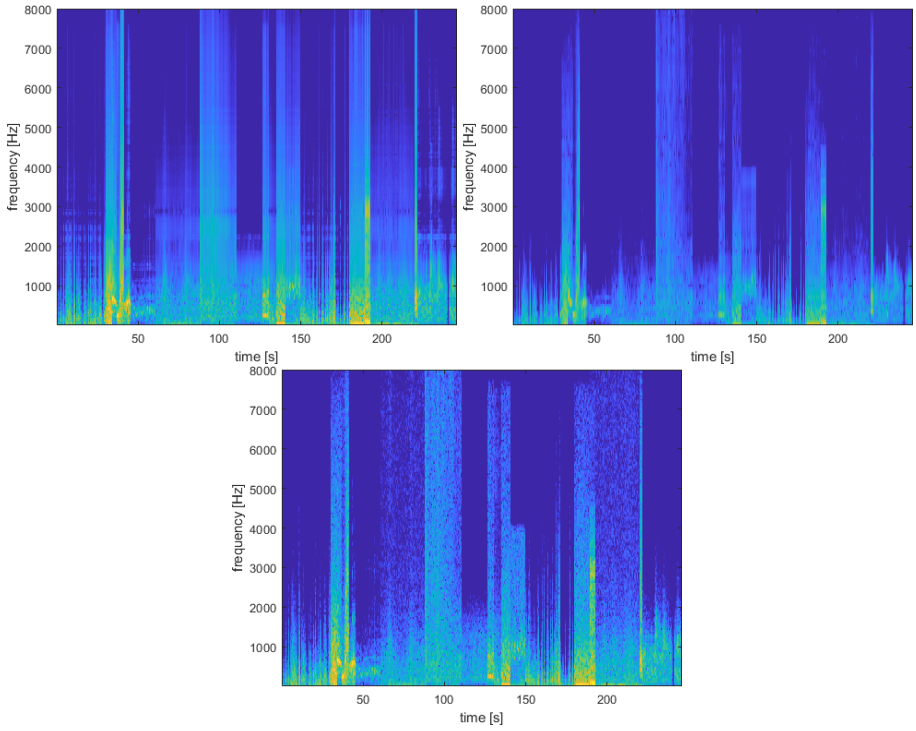


Figure 7.1: Top left: Kullback-Leibler spectrogram with 30 components. Top right: Itakura-Saito spectrogram with 30 components. Bottom: Original spectrogram

negative sparse coding, MMSE STSA, and spectral subtraction. Out from the results and from listening to the signals it was noticed that MMSE STSA and spectral subtraction did only partly or not all remove the wind sounds, at least between 0 and -15 dB SNR. It was hard to state of the the NMF did a better job at removed the wind as the NMF distorted the speech and wind under its processing, the same applied to the NNSC, but the NNSC did an overall better job than the NMF with less distortion from what could be heard. This could indicate that NMF and NNSC were not useful for extracting speech from a mixed signal, however, the NNSC had been found to do a good job in another study [7], thus it could indicate a problem with the dictionaries or that the signals in this study were more of a challenge.

As mention the MMSE STSA and spectral subtraction methods failed in most cases of removing the wind sounds, but this was expected as these two methods were developed to reduce stationary noises and not non-stationary winds that were present in this study. For some of the mixed signals processed by MMSE STSA an echo effect had appeared on the speech part of the signal. In the cases where the wind sounds were close to be stationary MMSE STSA

and spectral subtraction succeeded to remove almost all of the wind sounds. After [11] the spectral subtraction method, which is implemented by the Matlab Voicebox toolbox using [11], the worst SNR under testing was -5 dB which is much better than the worst dB in this study which was -15 dB.

7.3 Training and Testing Data

As the results of the two non-negative methods were worse than the other noise reduction methods and the non-processed signals the signals were listen to. From listening to the signals distortion was noticed. Given the distortion in the test signals after having been processed by the NMF and the NNSC, it could indicate that the speech dictionary was not able to fully identify unknown speech, while it also contained wind sounds, thus it could allow trivial solutions to some extent. The same goes for the wind dictionary as the wind dictionary signals contained, at least, some speech, while also being distorted.

Other methods, e.g. sparsity on the activation matrix, might have been needed to deal with the distortion by ensuring less wind was modeled by the speech dictionary and the other way around. The sparsity on the activation matrix was implemented in the form of the NNSC method and it still suffered distortion. Hence, it was expected, even though the λ value used for controlling the sparsity might had needed tweaking, that the number of wind signals used in the training of the wind dictionary was to low for it to be able to capture the different wind signals used in the testing set. Thus most of the distortion might have been caused by a lack of wind signals for training of the wind dictionary.

7.4 Measurements

The measurements used in this study was the Perceptual Evaluation of Speech Quality (PESQ), the Short-Time Objective Intelligibility Measure (STOI) and the Signal-to-Noise Ratio$_{out}$ (SNR$_{out}$), all being objective measurements and not subjective measurements. This could be a problem as a listener might rate the outputs of the three implemented noise reduction methods than the objective measurements did. Of course the SNR$_{out}$ would not have been done in a subjective way, however, both PESQ and STOI could have been replaced with a listening test like the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA). The reason for not conducting a MUSHRA was that any results where not expected to be different from the objective measurements, because of the distortion in the output of the NMF algorithm and thus no new knowledge would have been gained. However, the results for the MMSE STSA and Spectral Subtraction might have been different in some conditions as these two methods were close to each other in score, but their outputs sounded slightly different.

An informal listening test was conducted with only the author. The reason

for using a informal listening test and not to conduct a proper listening test was the figure out why the three measurements gave the results they did as the PESQ and STOI indicated the NMF did a worse job than doing nothing and the SNR_{out} was improved in most cases.

It was found that some of the signals did better than what a person might would have rated them, at least in the term of quality, so the results of the PESQ measurements was not fully trusted, e.g. a signal was rated to a quality of 3.3514 even though the author, if the author had to rate it, would have rated it much lower, while a the same time the signal processed with only 20 more components got a PESQ score of 2.4228, where the author thought it had better quality. The non-processed PESQ value was 1.2283. It is not expected that a user group would have rated the two NMF processed signals higher than the non-processed signal (it is possible they might have), so the results were not expected to be fully correct and thus not fully trustworthy. The STOI also had some "weird" results. It was noticed that these results only seemed to happen for one of the signals and not the rest, thus the PESQ and STOI code might have had problems with conducting their calculations on that given signal under some specific test conditions values as it was mainly noticeable with very low SNR (-15 dB) and few speech (30-50) and wind (30-50) components. However, this is the reason for using multiple signals for testing as it would limit the effect of a single or a few signals would have on the overall results given in the form of the mean scores, while also having the standard deviation to help give an idea about the range of scores for each condition. It should mentioned that the lower SNR values, mainly at -10 and -15 dB, had higher standard deviations at specific conditions for some of the measurements, mainly the SNR_{out} than the other SNR values, which again could indicate that the measurements suffered under these conditions depending on the mixed signals that were being processed.

Lastly, the PESQ code used to implement the PESQ measurement was an older version from before the wideband version of PESQ was implemented. From its paper nothing could be found that would indicate it would generate a different score compared to the newer versions of the PESQ other than the old version used in this study rated a signal on a scale going from 0.5 to 4.5 rather than from 1 to 5. Thus it was necessary to remember, when reading the scores, that a 0.5 had to be added to help comparing them to the newer versions.

7.5 Results

The results for the NMF outputs indicated that the NMF algorithm had problems with separating the speech and the wind sources from each other. The NMF and the NNSC always scored lower than the non-processed signals and the two other noise reduction methods, expect in some very few cases regarding the PESQ measurement, where they were above the non-processed mean. Overall the Itakura-Saito divergence seemed to do worse than the Kullback-Leibler divergence and the other noise reduction methods, while the NNSC

mostly did the best of the two non-negative methods, but most of the time it had similar scores to that of the Kullback-Leibler divergence. Kullback-Leibler might have done better as in speech signals most of the medium and high coefficients only exist in some "few" energy components and are more important for understanding speech than the low coefficients that exist in most of the energy components. Part of the reason that the results of the two non-negative methods was so worse than the other methods and the non-processed signal could have been the dictionaries had not been good enough, thus better dictionaries could have changed the results of the experiment.

In the listening test it was noticed that the NMF and NNSC produced distortion in their outputs, which would explain the lower PESQ and STOI scores, while also changing the SNR_{out} compared to the SNR_{in} as the intensity of the extracted speech and wind signals would have contained parts of the wrong source. The few signals that scored better in the PESQ and the STOI, compared to the non-processed signals and the two stationary noise reduction methods, nothing could be heard that could explain these better scores as they suffered from the same problems as the other signals in any other conditions, thus it was expected that the PESQ and the STOI measurements had suffered and not worked properly in these specific conditions.

A reason for why the spectral subtraction and MMSE STSA did better was that they were developed to removed non-stationary noise, which means that they should have a minimum effect on the speech part of the mixed signals as speech is non-stationary. Of course this meant that they did not work well with the non-stationary wind noises, however, any part of the wind noises that had been stationary enough would still have been reduced, thus improved the quality and intelligibility of the mixed signals. From the SNR_{out} scores it could be seen that the two non-negative methods had problems completely separating the speech and wind into their own signals and/or that some effect happened to the extracted signals. From the informal listening test of the output signals it could be heard that the signals had distortions, which sounded like they followed a pattern similar to the wind in the non-processed signals and the speech sounded distorted. Thus the non-negative methods had not separated the two sources, the speech and the wind, into two audio files rather the speech components had extracted most of the speech and the wind, while the wind components had only extracted a little of the speech and the wind, which would explain the improved SNR_{out} and the lower PESQ and STOI scores.

As mentioned in 7.4 a single mixed signal had results, for some specific conditions, that differed from the other signals scored. These scores would have a minor affect on overall mean score for the measurements PESQ and STOI as each mean score was the mean of nine signals, however, it still meant that the results could be screwed to be either more positive or negative if more signals have had the same weird scores.

A potential reason for the lower scores was because of the distortion caused by the two methods. A reason for the distortion will be mentioned in section 7.6.

7.6 Spectrograms

All outputs signals (speech components only and wind components only) of the NNSC and the NMF method suffered from distortion. Thus it was decided to look at the spectrograms for why this distortion might have happened. When looking at the spectrograms consisting of only speech components, figure 7.2, it can be observed that the wind is being modeled by the speech components. This problem also exist for the wind components. In figure 7.2 it can be viewed that the NNSC did a better job at modeling most of the lower frequencies of the wind compared to the two β-divergences, both which modeled most of the lower frequencies using the speech components. This would explain the distortion in the outputs of the NNSC and the NMF. It can also be observed that the Kullback-Leibler and the NNSC did not model higher frequencies well compared to Itakura-Saito, but they did better at model the high coefficient energy components in the spectrogram, which was expected given whether they were scale invariant or not. At the same time Itakura-Saito modeled the frequencies that lied below that of speech, something that was only modeled a little by Kullback-Leibler and not at all by NNSC. Another figure of spectrograms, figure 7.3, the speech components of another mixed signal shows the wind is more clearly modeled by the wrong components.

The last spectrograms that will be shown are in figure 7.4. These spectrograms differ from the others as these are the spectrograms modeled by the wind components and they belong to the same mixed signal and conditions as those in figure 7.2 do. Here the differences between both the two β-divergences and the NNSC are more clear. Both β-divergences and the NNSC have clearly modeled most of the lower frequencies of the wind signal, however, this is the only point they have behaved similarly. The NMF method has modeled a lot more of the speech than the NNSC method has, while the Itakura-Saito divergence modeled what kind of seems like random frequencies at random frames, while the Kullback-Leibler divergence modeled frequencies around the 3000 and below more precise, but the frequencies above this point are blurry and imprecise similarly to the spectrogram of speech components. An interesting thing to notice about the NNSC wind spectrogram is the almost similar coefficients in most of the energy components, mainly, under the 3100 frequency mark. This square of similar coefficients did not exist in the original mixed signal, only in the output of the NNSC and no wind signal for the training of the NNSC had a similar pattern. The pattern of similar coefficients around and below the 3100 frequency mark was noticed for all other signals, however, this pattern of which frequency the box stopped at did change slightly depending on the amount of wind components.

7.7 Listening

From listening to the audio signals it was noticed that no matter the amount of wind components that the wind components could model the speech to some

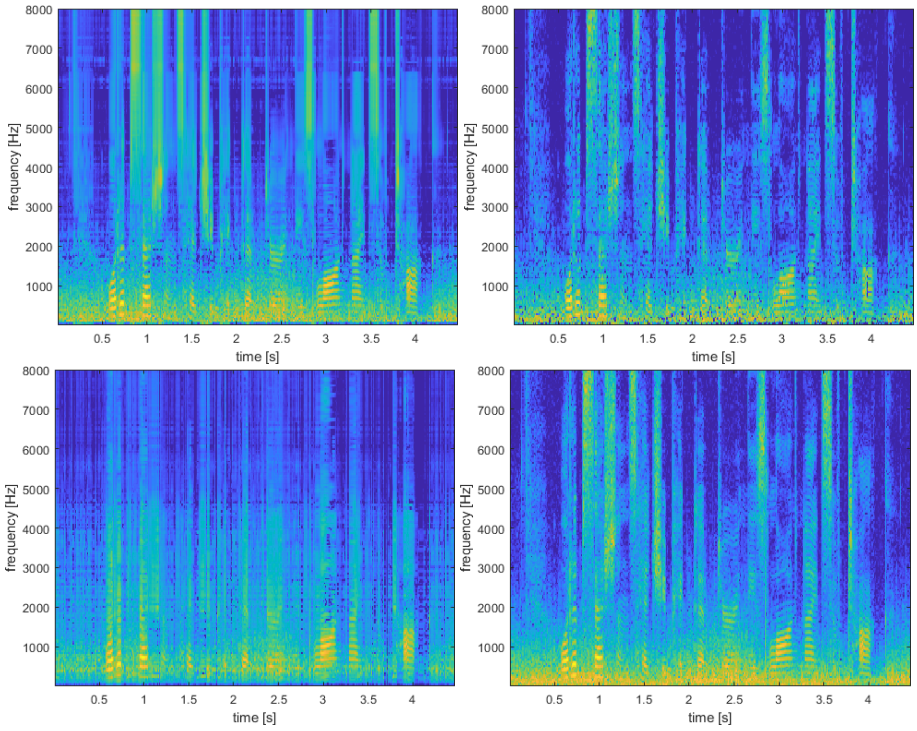


Figure 7.2: Top left plots: Kullback-Leibler. Top right: Itakura-Saito. Bottom Left: NNSC. Bottom Right: Original spectrogram. 70 wind components and 110 speech components

extent, higher amount of components did better at extracting speech. Thus the NMF require a fine tuning of the amount of wind components, this being a trade between being generalised to model as much wind as possible and the amount of speech it can model. When it came to the speech the problem was that not even high number of speech components, e.g. 150, could properly model some of the higher frequencies and higher amounts of components were more likely to model the wind. This also existed for the wind components. Even low amount of speech components could model wind, however, the wind was highly distorted, while higher amount of speech components had slightly less distortion in most cases.

It was noticed that most of the distortion in a single signal followed the pattern of the wind in the specific mixed signal, thus it was determined that the distortion was, most likely, caused by the wind being partly extracted using the speech components and not fully extracted by the wind dictionary.

A difference between the distortion caused by the NMF and NNSC method was noticed. The difference was that distortion in the NNSC signals had a higher frequency than the NMF signals.

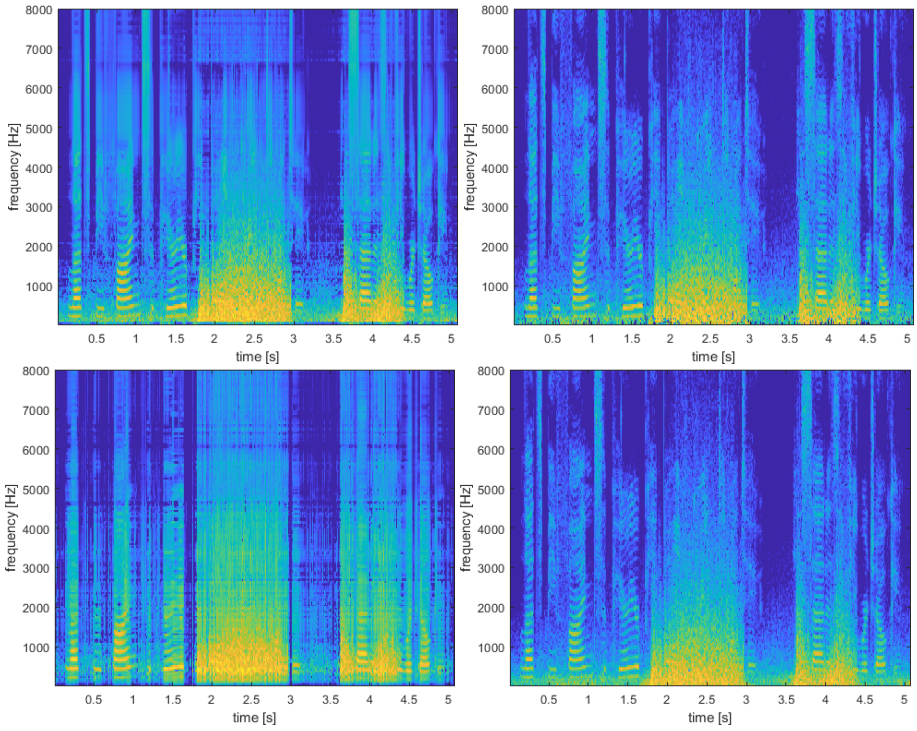


Figure 7.3: Top left plots: Kullback-Leibler. Top right: Itakura-Saito. Bottom Left: NNSC. Bottom Right: Original spectrogram. 30 wind components and 150 speech components

7.8 Novelty of the Study

The main novelty of this study was the focus of removing non-stationary interfering sounds in the form of wind by using a basic NMF algorithm. Other novelties were the focus of performance on the measurements PESQ and STOI over different conditions, these being the amount of wind and speech components, different SNR, and Kullback-Leibler and Itakura-Saito divergence, while evaluating the NMF algorithm for its limitations and why it produced the results it did. Another study had used a similar method, the NMSC, however, it was unsure how generalised their dictionaries had been regarding the wind sounds used [7] and they did not measure the change to quality and intelligibility of the processed signals.

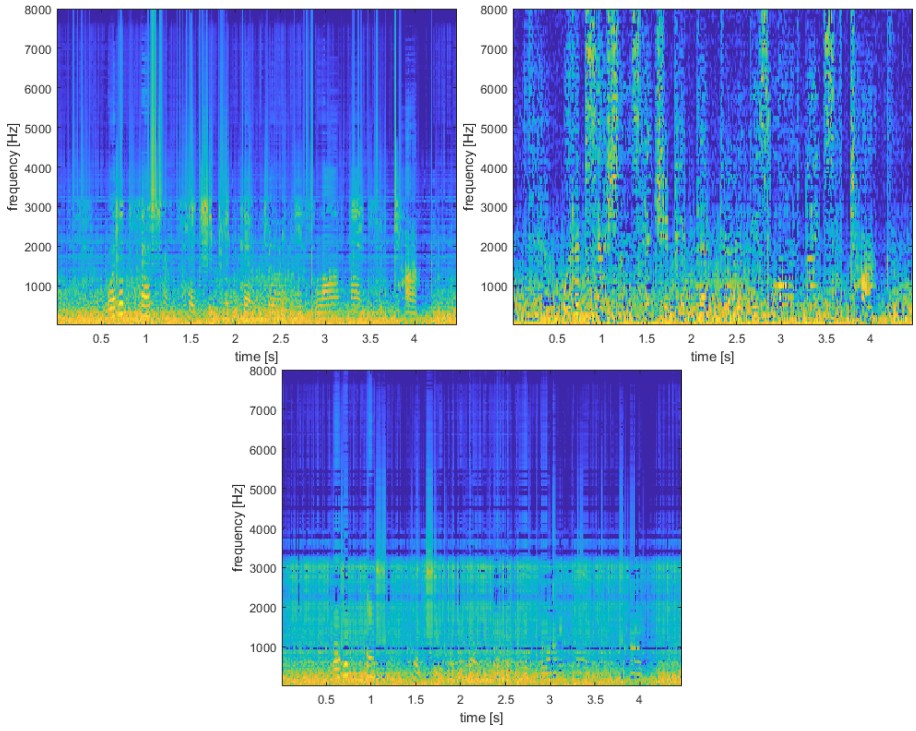


Figure 7.4: Top left plots: Kullback-Leibler. Top right: Itakura-Saito. Bottom: NNSC. 70 wind components and 110 speech components

7.9 Limitations of the Study

The study had some limitations. Regarding the representation of wind sounds, nothing was done to ensure that the chosen wind sounds represented all possible wind sounds. This was mainly because of the difficulty of collecting usable audio samples as, naturally, more than wind is present in most cases of audio samples of wind, e.g. leaves, cars, birds etc. More audio samples of the wind signals could be have been a useful thing as 99 speech signals were used, but only 29 wind signals were used and from the spectrograms of these wind signals they seemed to differ more from each other than the speech signals did.

The measurements in this study was only objective measurements, no subjective measurements, e.g. a listening test like the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) could have been conducted with a user group as it would have given more data and allowed to evaluate the trustworthiness of the objective measurements.

Only some few specific SNR values were used for the testing of the NMF method, the values ranged from -15 - 5 dB with an increment of 5 dB. More precise results could have been acquired by increasing the number of values.

7.10 Possible Improvements

The study had areas that could improved in further studies.

A possible improvement could have been an increased sampling rate as this study used 16000 Hz and the human ear can pick up frequencies to 22050 Hz, thus needing a sampling rate of 44100 Hz.

Another improvement could be a bigger training set for wind as only 20 wind signals was used for training, while the speech training set consisted of 90 signals. This could have helped on lowering the amount of distortion as the processed wind did not sound like the non-processed wind sounds, speech signals sounded similar to the non-processed speech signals, hence more wind signals for training could be an improvement.

The implementation of restrictions could have allowed for improvement in the output of the NMF algorithm, e.g. the sparsity restriction on activation could have helped on the modeling of wind in the speech dictionary and the other way around. However, it was decided not to implement this to test the ability of the basic NMF could solve the problem given in the problem statement, while the NNSC had sparsity on the activation matrix.

Under the training nothing was done to ensure that the SNR of all wind/speech signals had the same SNR, thus each of the signals in the two long training audio signals could swing in SNR when compared to the other parts, this could, maybe, affect the estimation of the basic matrix $\mathbf{V}$ depending on the chosen β-divergence.

Chapter 8

Conclusion

The purpose of this project was to implement and evaluate the basic non-negative matrix factorisation (NMF) algorithm's ability to separate speech and wind from a mixed signal, while at the same time comparing it to the state-of-the-art, the non-negative sparse coding (NNSC), and two stationary noise reduction methods, the spectral subtract and the minimum mean square error estimate of short-time log-spectral amplitude (MMSE STSA).

Two perceptual measurements were implemented, these being the Perceptual Evaluation of Speech Quality (PESQ) and the Short-Time Objective Intelligibility (STOI). The SNR_{out} was also implemented to see what affect the separation had on the SNR. Multiple conditions were used for the testing of the NMF's ability for separation. The conditions were different SNR values, different numbers of wind and speech components, and two different β-divergences, these being the Kullback-Leibler and Itakura-Saito.

Out from the results of the PESQ, STOI, and SNR_{out} it was found that increased amounts of wind components would affect them negatively in most cases by lowering the scores. By increasing the number of speech components the PESQ, STOI, and SNR_{out} improved in most of the conditions.

However, the NMF and NNSC did not do as well in either PESQ or STOI compared to the non-processed signals and the two stationary methods. For the PESQ measurement the MMSE STSA and spectral subtraction did better than the non-processing signals, while for STOI MMSE STSA, spectral subtraction, and the non-processing signals did almost similar. For the STOI measurement MMSE STSA, spectral subtraction and the non-processed signals did very similar, while the NMF and NNSC did much worse. The SNR_{out} was rarely the same as the SNR_{in}.

To figure out why the NMF (and the NNSC) did worse the signal outputs were listen to, in an informal listening test, and from them it was noticed the extracted speech and wind signals contain parts of the wrongs sources, e.g. wind in the signal extracted using the speech dictionary, and the present of a fairly amount of distortion. From the different mixed signals used in the testing it was determined that the distortion was caused by the wind source in

each mixed signals being extracted partly by the speech components and not fully by the wind components. Hence, this could explain the lower PESQ and STOI scores and the difference in the SNR_{out} compared to the SNR_{in}. Hence, higher numbers of either wind or speech components increased how much of the wrong source was extracted.

Thus, it can be conclude that with the data used to train the two dictionaries, the speech and the wind dictionary, that the NMF failed in doing better than the non-processed signals and the two stationary methods after the measurements. The NNSC did slightly better than the than the NMF using either β-divergences and worse than the non-processed signals and the two stationary methods, thus it was considered that the training data might not have been good enough to extract not-trained data. However, with the data that was used for study the final conclusion can be that the NMF did not succeed in extracting speech and wind from the mixed signals and did worse than the stationary noise reduction methods, hence the NMF was considered to have failed.

Bibliography

[1] K. Kwon, J. W. Shin, S. Sonowat, I. Choi, and N. S. Kim, "Speech enhancement combining statistical models and NMF with update of speech and noise bases," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5 2014, pp. 7053–7057. [Online]. Available: http://ieeexplore.ieee.org/document/6854968/

[2] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 10 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6544586/

[3] I. Cohen, Y. Huang, J. Chen, and J. Benesty, *Noise Reduction in Speech Processing*, ser. Springer Topics in Signal Processing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 2. [Online]. Available: http://link.springer.com/10.1007/978-3-642-00296-0

[4] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional Models for Audio Processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 3 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7038275/

[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 3 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18785855http://www.mitpressjournals.org/doi/10.1162/neco.2008.04-08-771

[6] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian Non-negative Matrix Factorization." Springer, Berlin, Heidelberg, 2009, pp. 540–547. [Online]. Available: http://link.springer.com/10.1007/978-3-642-00599-2_68

[7] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind Noise Reduction using Non-Negative Sparse Coding," in *2007 IEEE Workshop on Machine*

Learning for Signal Processing. IEEE, 8 2007, pp. 431–436. [Online]. Available: http://ieeexplore.ieee.org/document/4414345/

[8] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "BEYOND NMF: TIME-DOMAIN AUDIO SOURCE SEPARATION WITHOUT PHASE RECONSTRUCTION," 2013. [Online]. Available: http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/32_Paper.pdf

[9] 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, pp. 1109–1121, 12 1984. [Online]. Available: http://ieeexplore.ieee.org/document/1164453/

[10] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 4 1985. [Online]. Available: http://ieeexplore.ieee.org/document/1164550/

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. Institute of Electrical and Electronics Engineers, pp. 208–211. [Online]. Available: http://ieeexplore.ieee.org/document/1170788/

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa Timit Acoustic-Phonetic Continuous Speech Corpus CD-ROM {TIMIT}," *NIST Interagency/Internal Report (NISTIR) - 4930*, 1993. [Online]. Available: https://www.nist.gov/publications/darpa-timit-acoustic-phonetic-continuous-speech-corpus-cd-rom-timit

[13] "A corpus of nonspeech sounds." [Online]. Available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html

[14] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2. IEEE, pp. 749–752. [Online]. Available: http://ieeexplore.ieee.org/document/941023/

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217. [Online]. Available: http://ieeexplore.ieee.org/document/5495701/

[16] ——, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and*

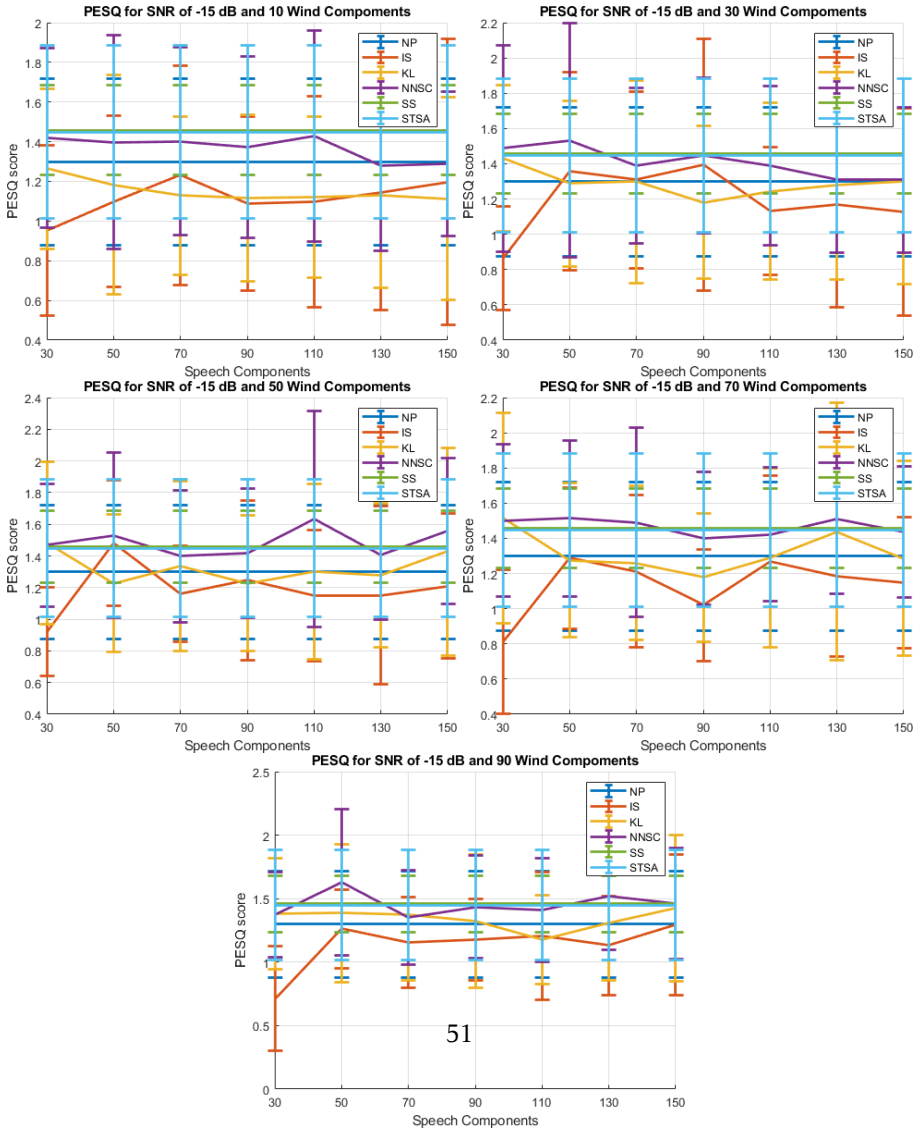
Language Processing, vol. 19, no. 7, pp. 2125–2136, 9 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5713237/

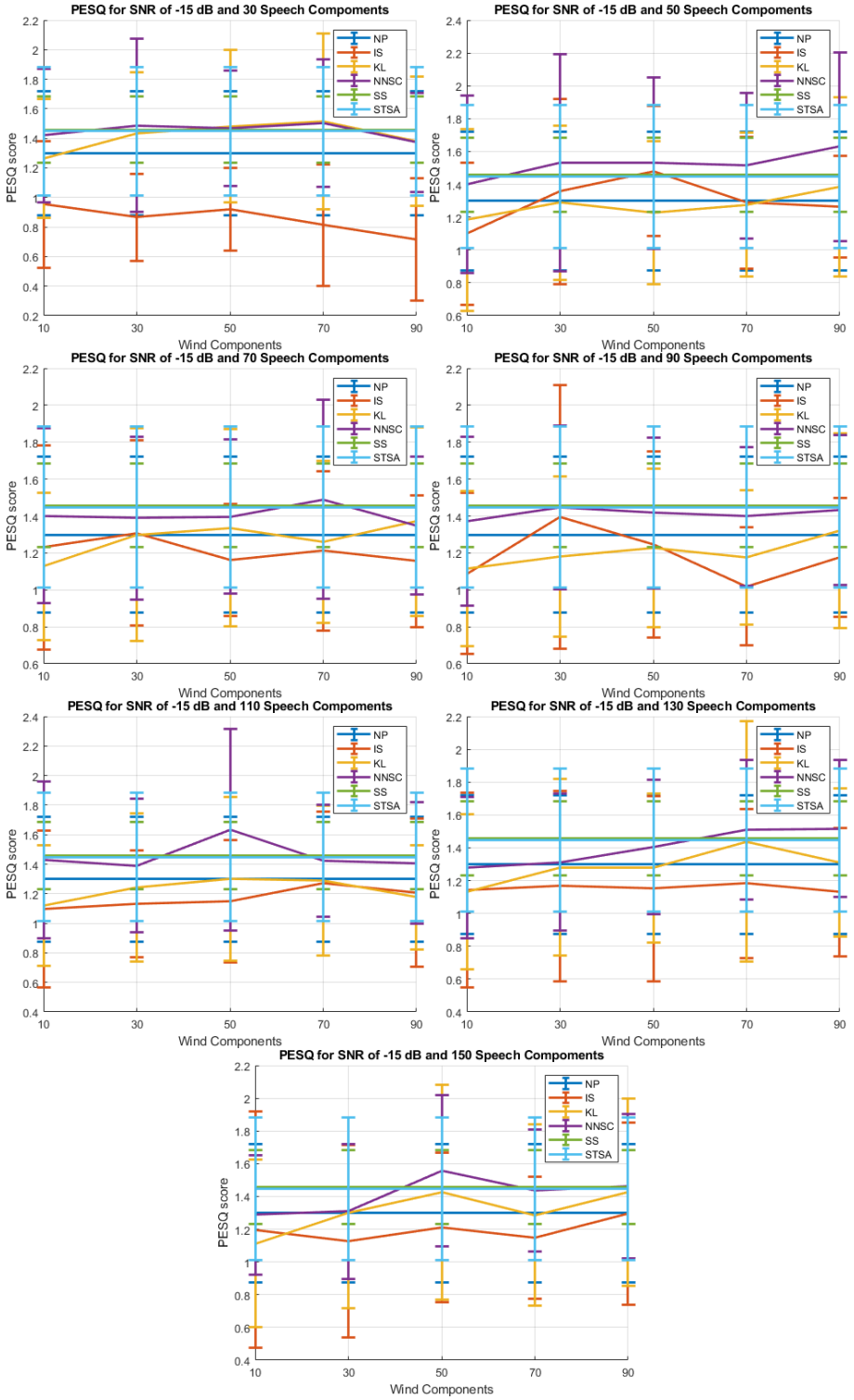
[17] L. Roux, J. . Weninger, and F. J. . Hershey, "Sparse NMF – half-baked or well done?" 2015. [Online]. Available: http://www.merl.com/publications/docs/TR2015-023.pdf

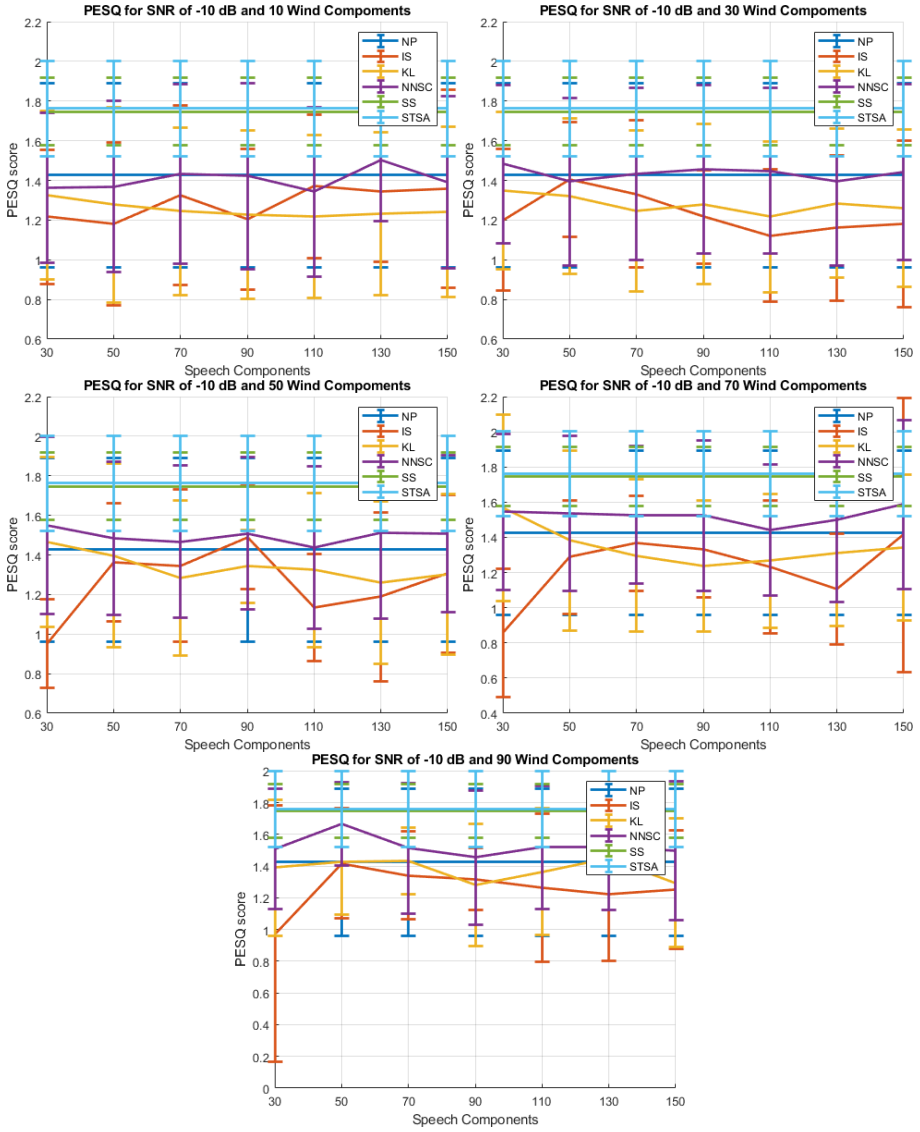
[18] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3 2008, pp. 4029–4032. [Online]. Available: http://ieeexplore.ieee.org/document/4518538/

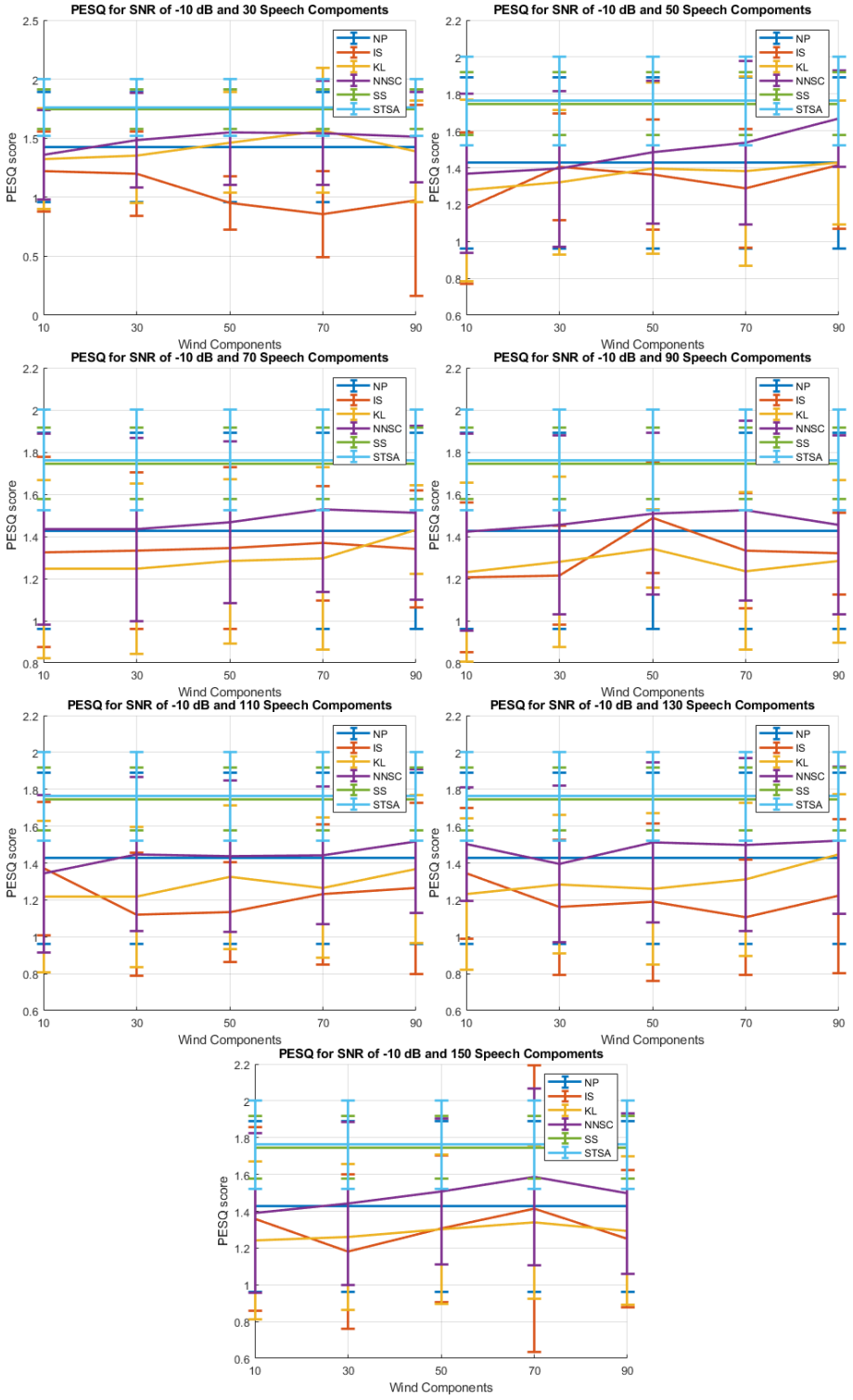
Appendix A

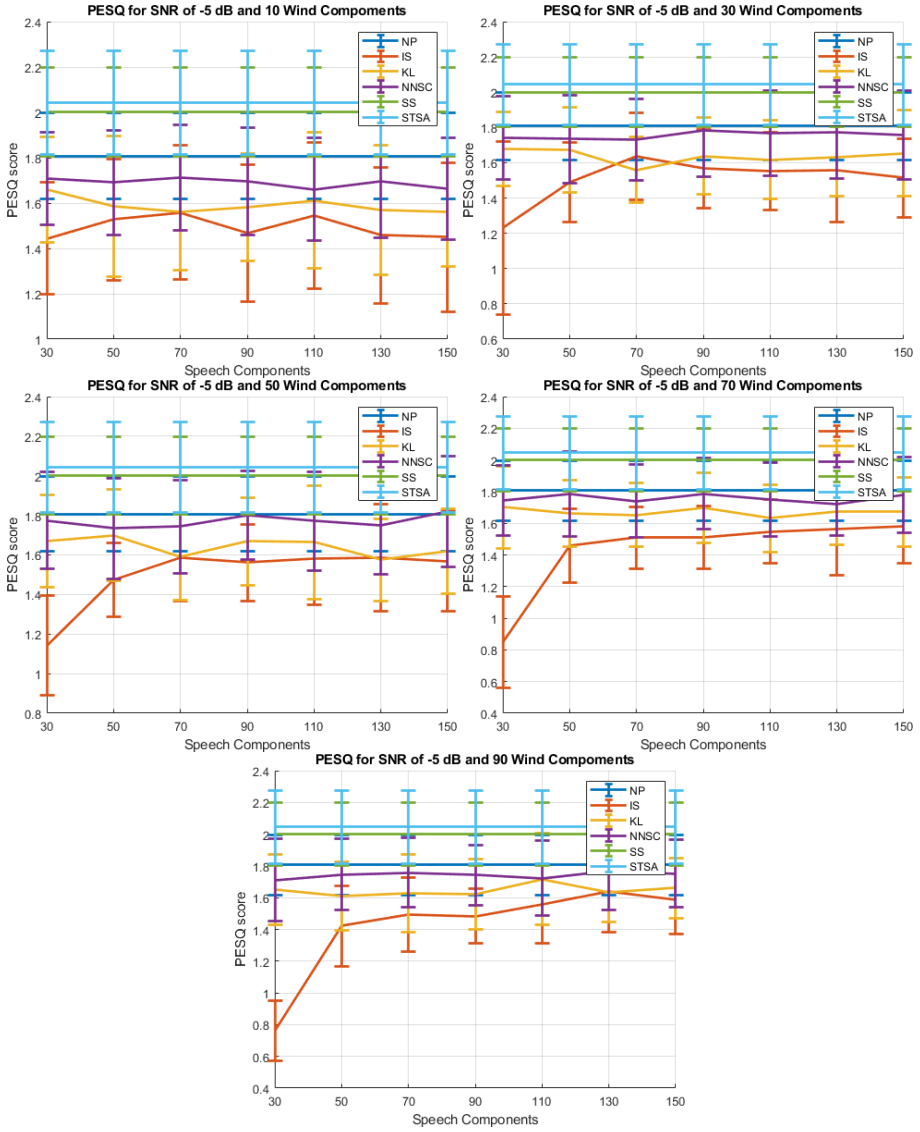
PESQ Plots

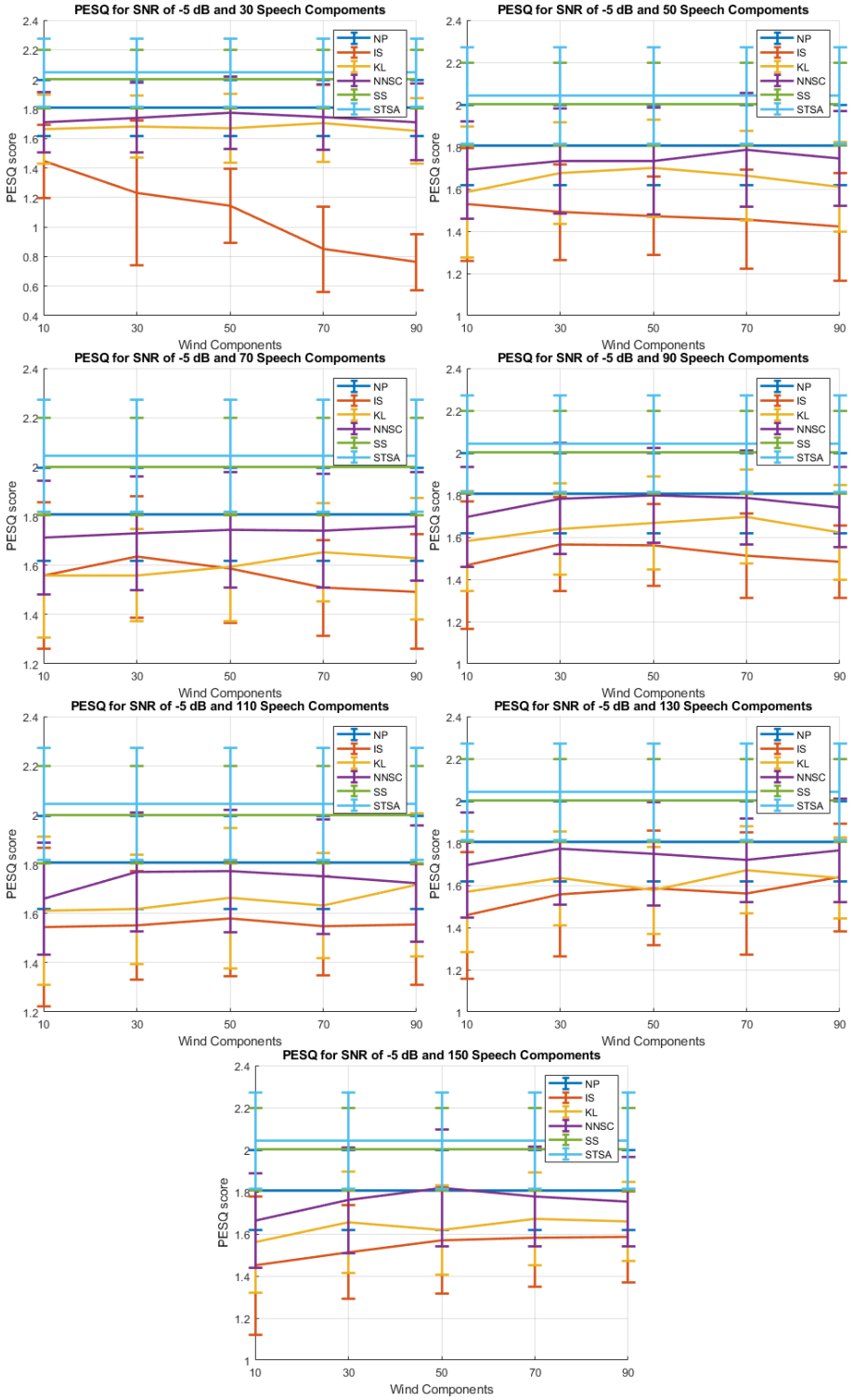


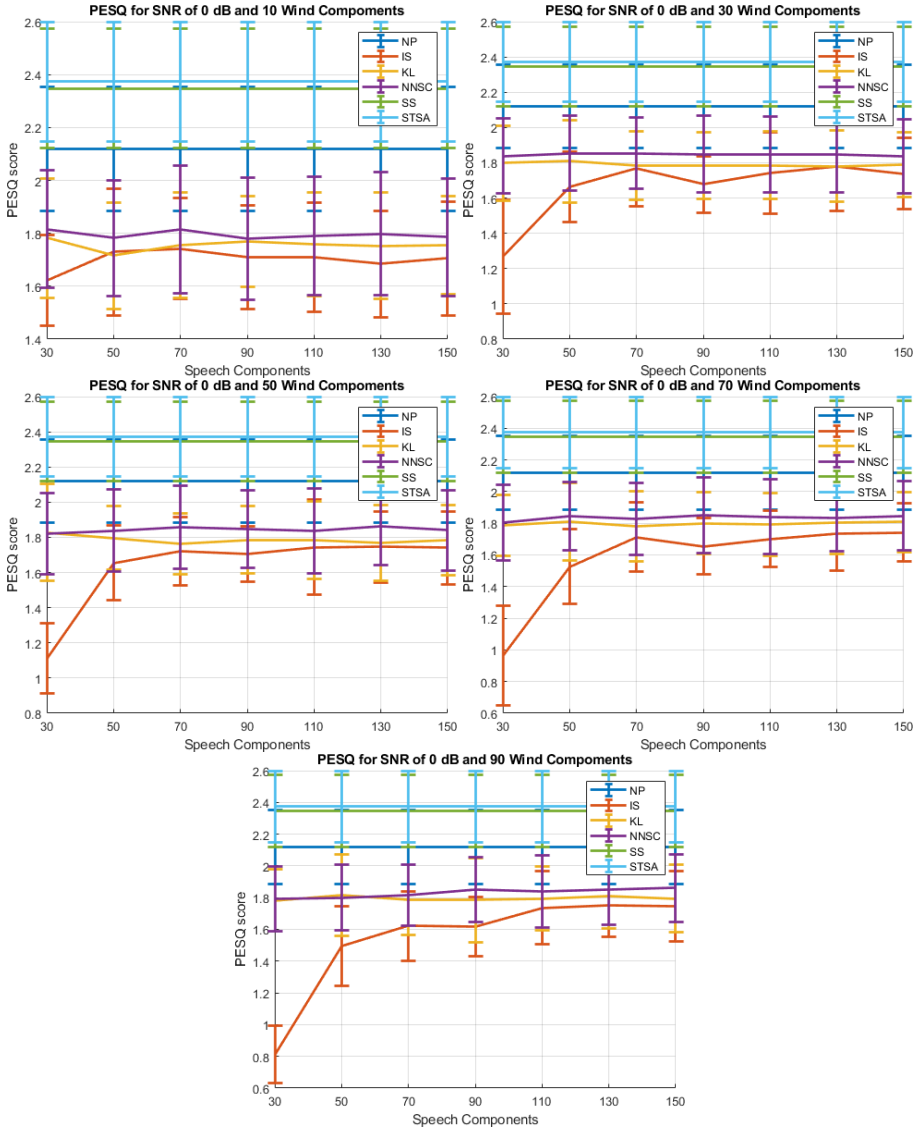

PESQ for SNR of -15 dB and 30 Speech Compoments
PESQ for SNR of -15 dB and 50 Speech Compoments
PESQ for SNR of -15 dB and 70 Speech Compoments
PESQ for SNR of -15 dB and 90 Speech Compoments
PESQ for SNR of -15 dB and 110 Speech Compoments
PESQ for SNR of -15 dB and 130 Speech Compoments
PESQ for SNR of -15 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Wind Components

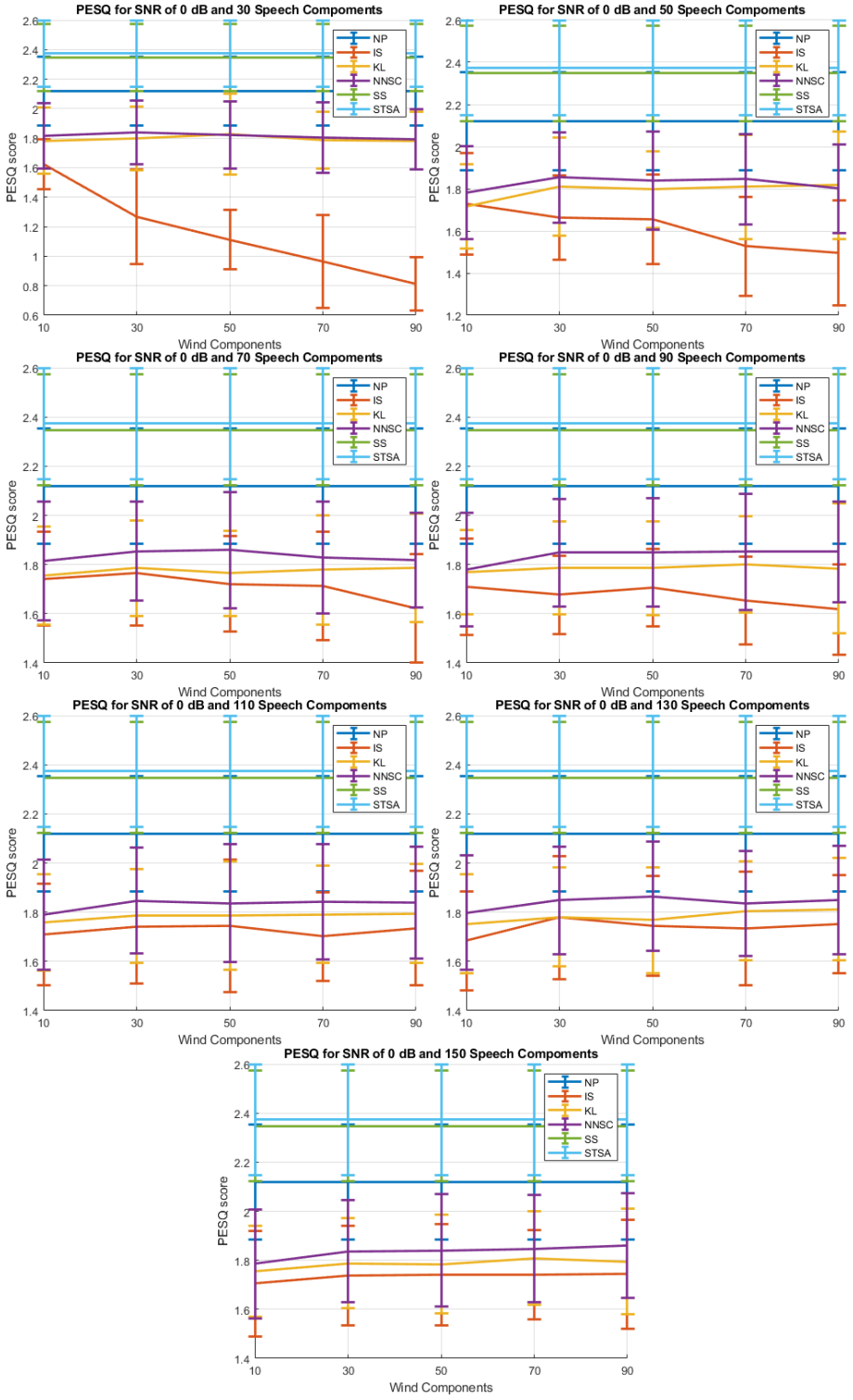

PESQ for SNR of -10 dB and 10 Wind Compoments
PESQ for SNR of -10 dB and 30 Wind Compoments
PESQ for SNR of -10 dB and 50 Wind Compoments
PESQ for SNR of -10 dB and 70 Wind Compoments
PESQ for SNR of -10 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Speech Components

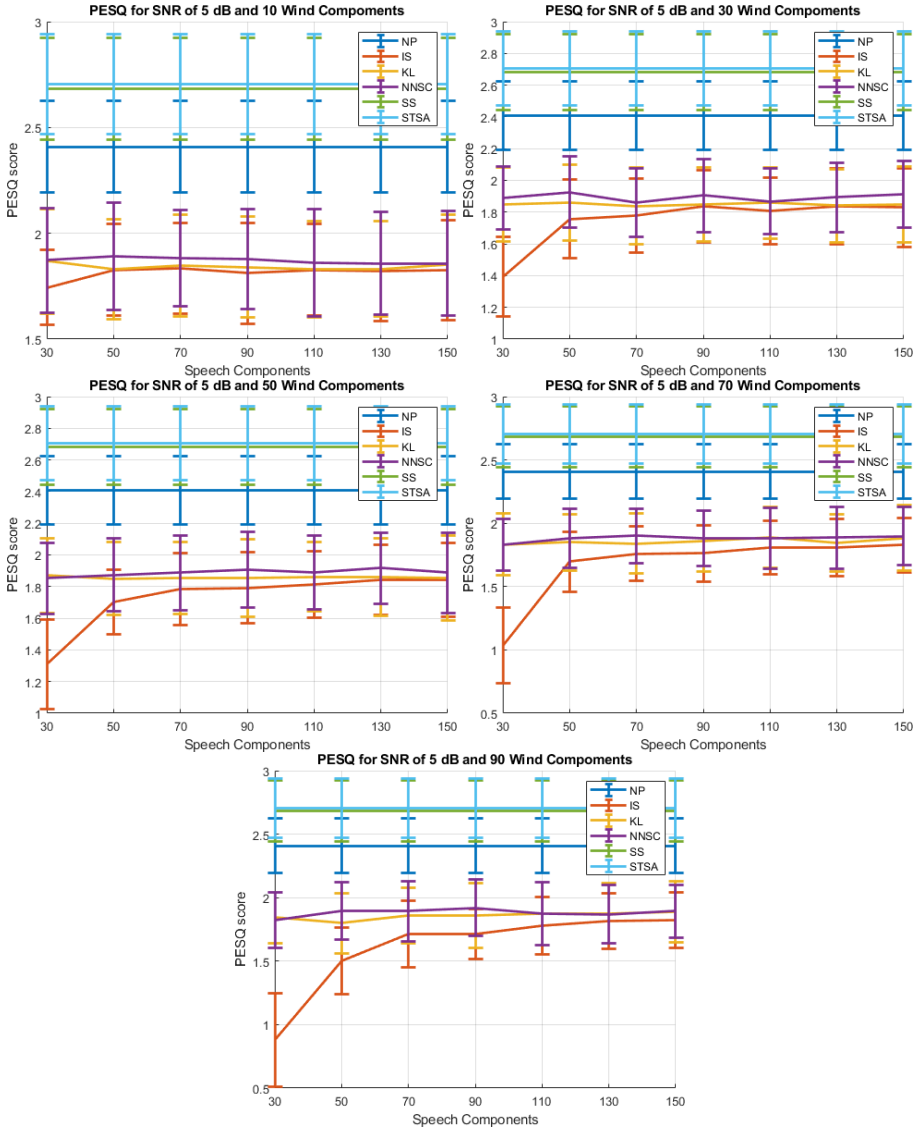

PESQ for SNR of -10 dB and 30 Speech Compoments
PESQ for SNR of -10 dB and 50 Speech Compoments
PESQ for SNR of -10 dB and 70 Speech Compoments
PESQ for SNR of -10 dB and 90 Speech Compoments
PESQ for SNR of -10 dB and 110 Speech Compoments
PESQ for SNR of -10 dB and 130 Speech Compoments
PESQ for SNR of -10 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Wind Components

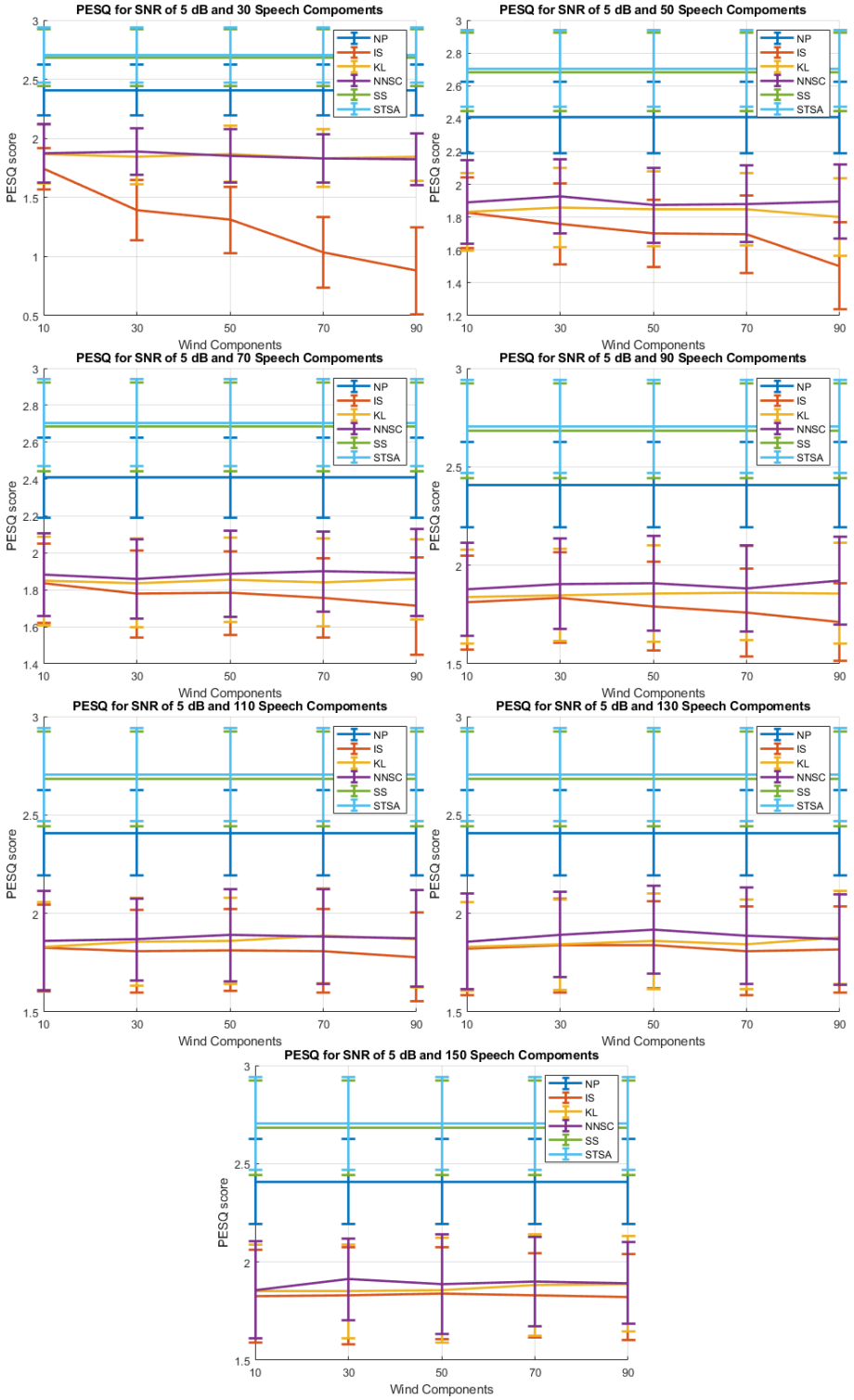

PESQ for SNR of -5 dB and 10 Wind Compoments
PESQ for SNR of -5 dB and 30 Wind Compoments
PESQ for SNR of -5 dB and 50 Wind Compoments
PESQ for SNR of -5 dB and 70 Wind Compoments
PESQ for SNR of -5 dB and 90 Wind Compoments
PESQ score
Speech Components
NP
IS
KL
NNSC
SS
STSA

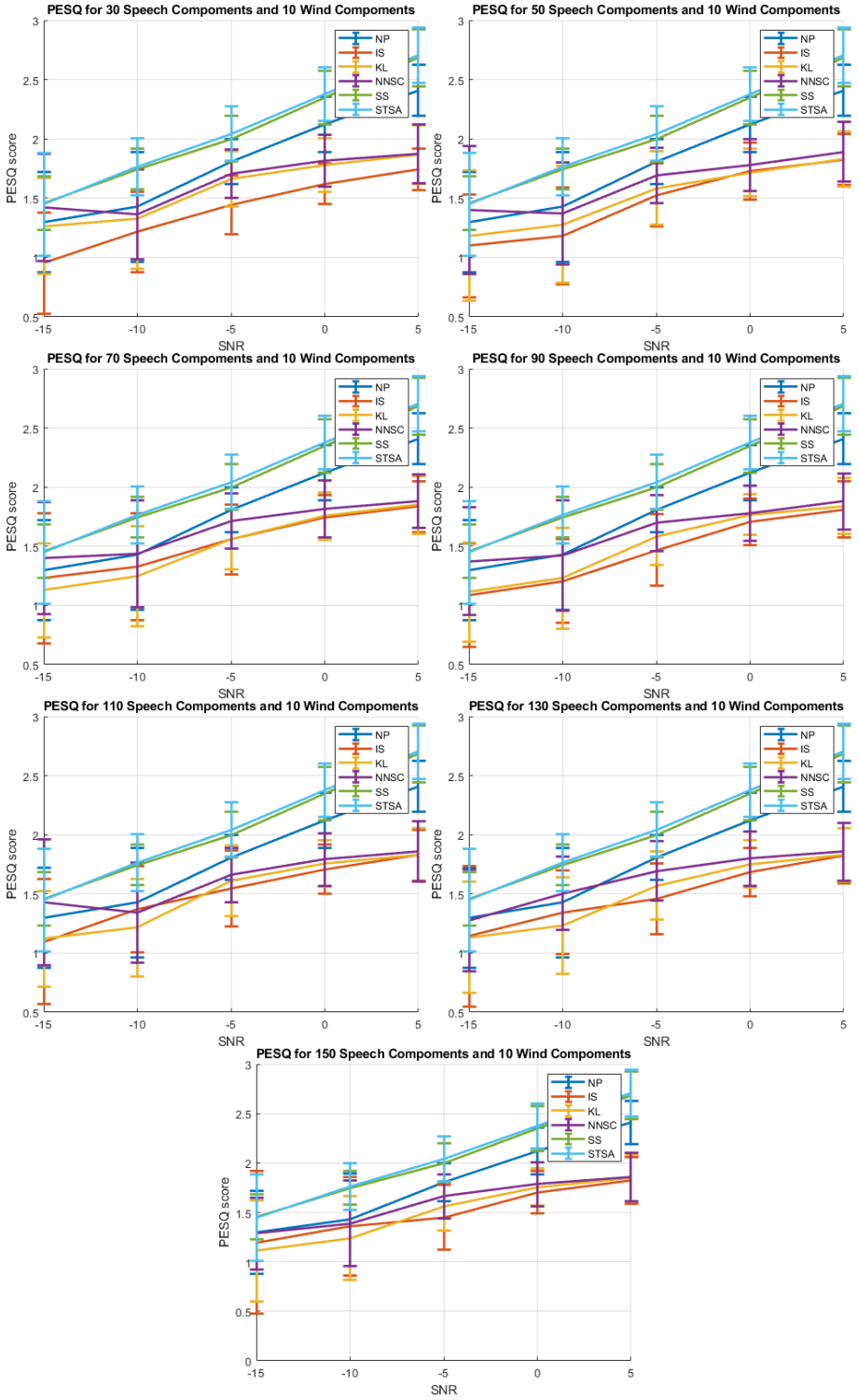

PESQ for SNR of -5 dB and 30 Speech Compoments
PESQ for SNR of -5 dB and 50 Speech Compoments
PESQ for SNR of -5 dB and 70 Speech Compoments
PESQ for SNR of -5 dB and 90 Speech Compoments
PESQ for SNR of -5 dB and 110 Speech Compoments
PESQ for SNR of -5 dB and 130 Speech Compoments
PESQ for SNR of -5 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Wind Components

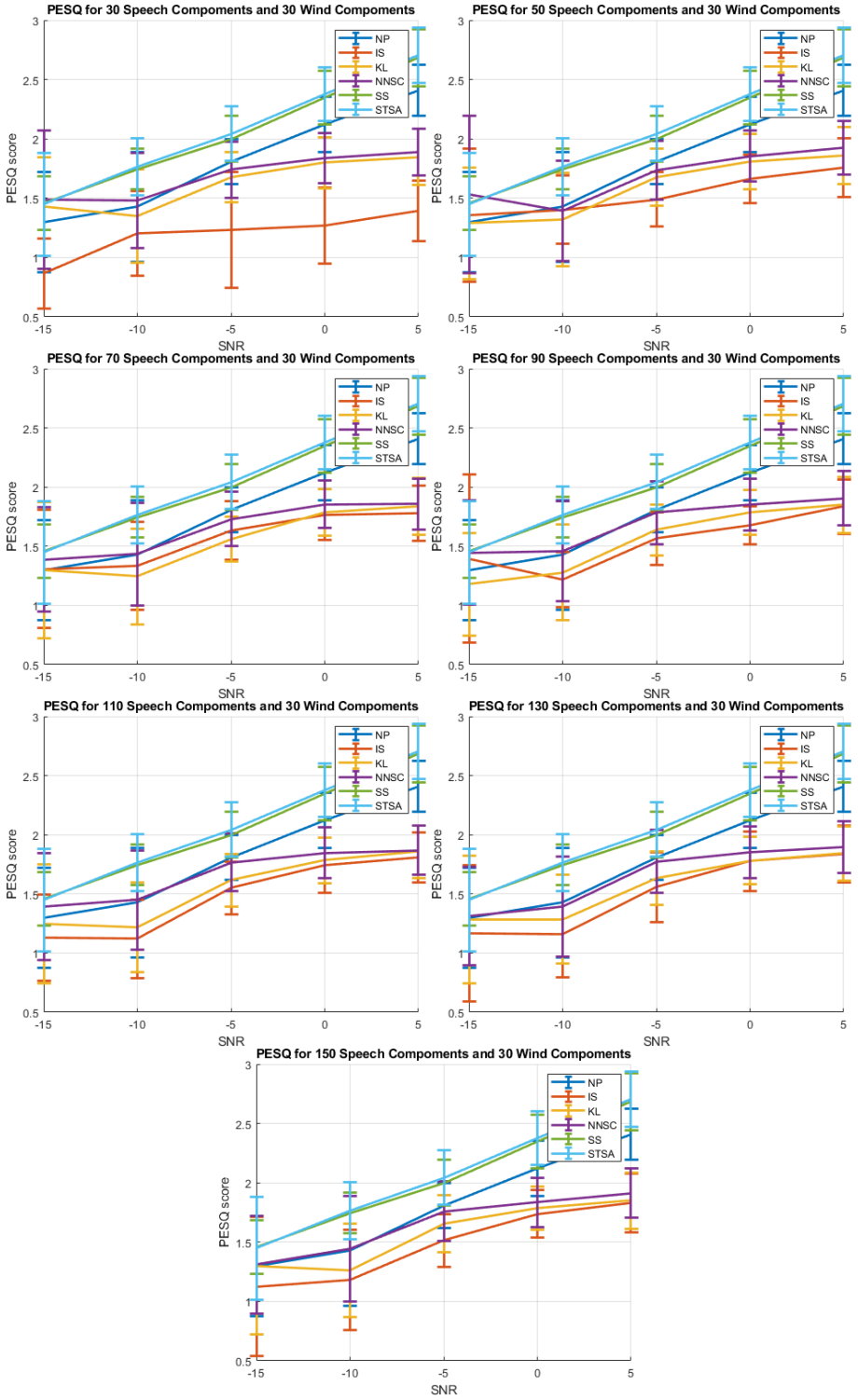

PESQ for SNR of 0 dB and 10 Wind Compoments
PESQ for SNR of 0 dB and 30 Wind Compoments
PESQ for SNR of 0 dB and 50 Wind Compoments
PESQ for SNR of 0 dB and 70 Wind Compoments
PESQ for SNR of 0 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Speech Components

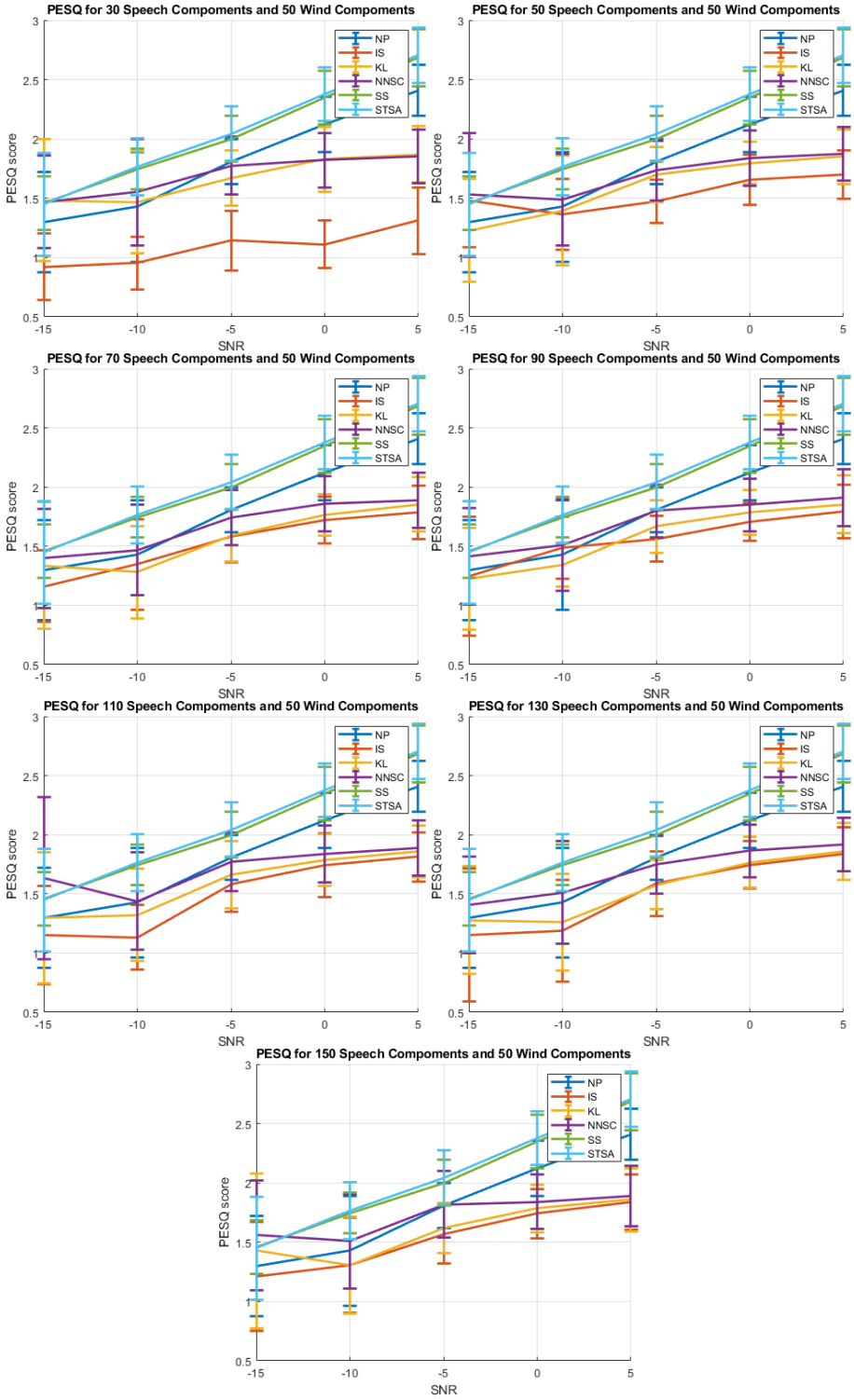

PESQ for SNR of 0 dB and 30 Speech Compoments
PESQ for SNR of 0 dB and 50 Speech Compoments
PESQ for SNR of 0 dB and 70 Speech Compoments
PESQ for SNR of 0 dB and 90 Speech Compoments
PESQ for SNR of 0 dB and 110 Speech Compoments
PESQ for SNR of 0 dB and 130 Speech Compoments
PESQ for SNR of 0 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Wind Components

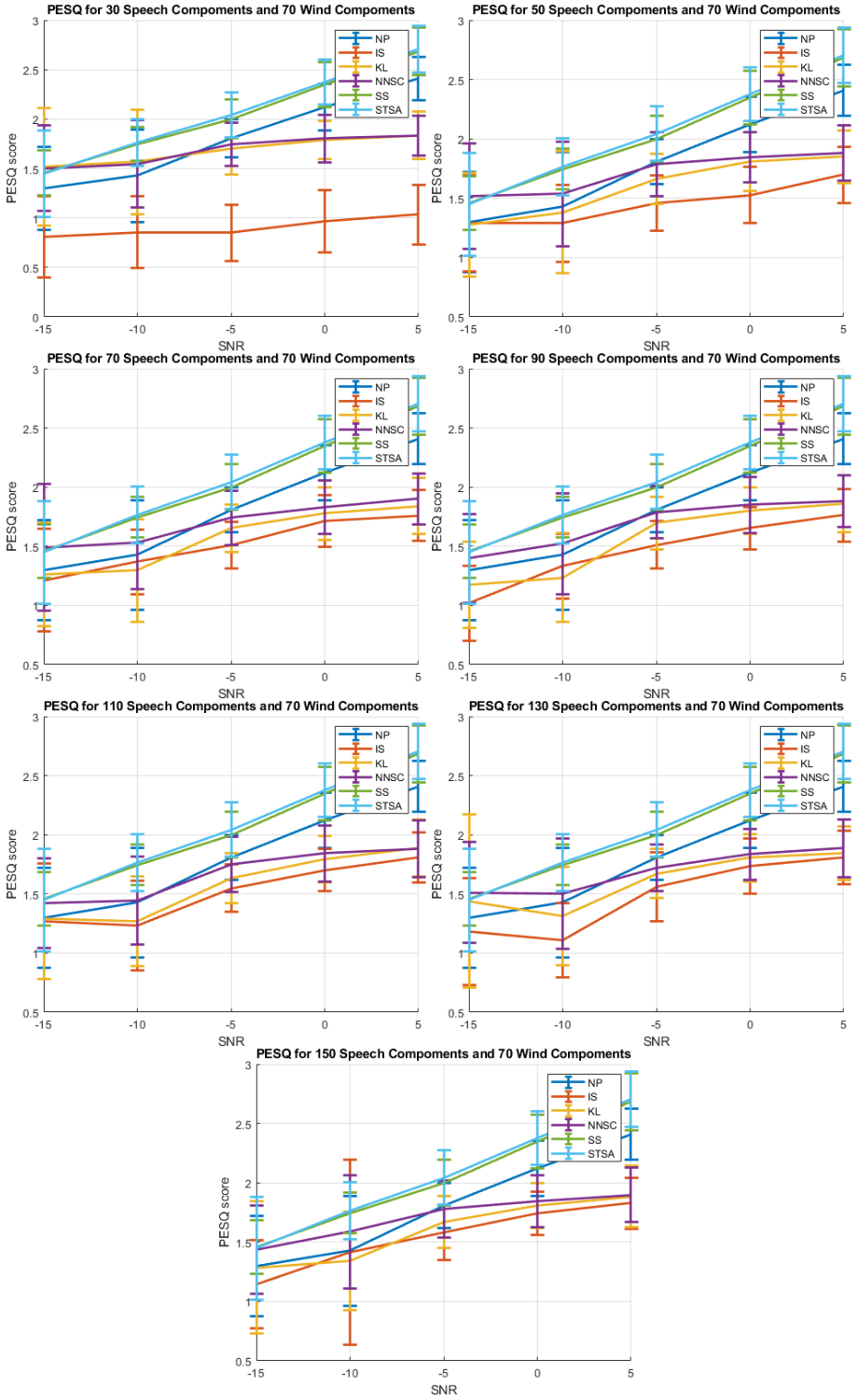

PESQ for SNR of 5 dB and 10 Wind Compoments
PESQ for SNR of 5 dB and 30 Wind Compoments
PESQ for SNR of 5 dB and 50 Wind Compoments
PESQ for SNR of 5 dB and 70 Wind Compoments
PESQ for SNR of 5 dB and 90 Wind Compoments
PESQ score
Speech Components
NP
IS
KL
NNSC
SS
STSA

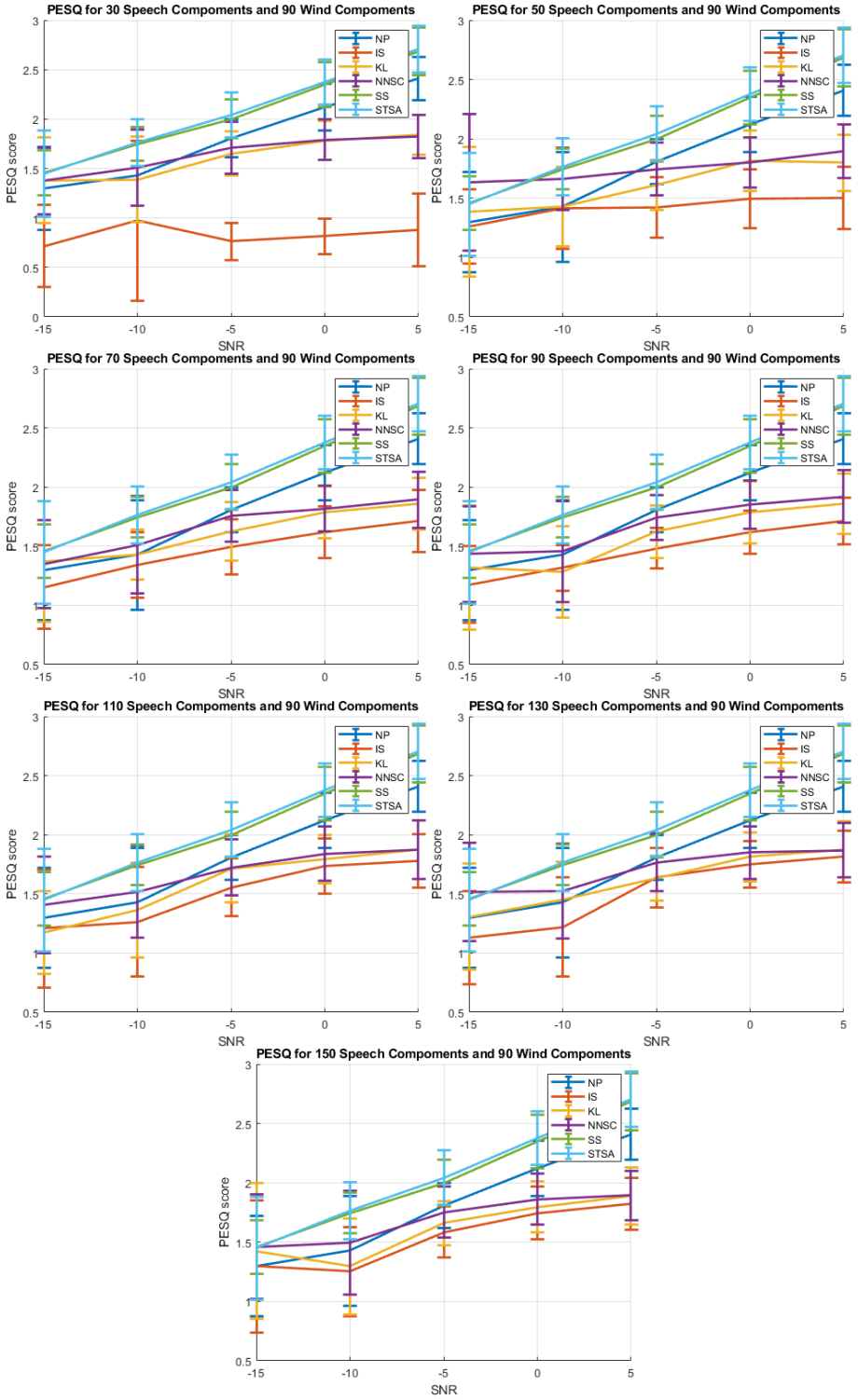

PESQ for SNR of 5 dB and 30 Speech Compoments
PESQ for SNR of 5 dB and 50 Speech Compoments
PESQ for SNR of 5 dB and 70 Speech Compoments
PESQ for SNR of 5 dB and 90 Speech Compoments
PESQ for SNR of 5 dB and 110 Speech Compoments
PESQ for SNR of 5 dB and 130 Speech Compoments
PESQ for SNR of 5 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
Wind Components


PESQ for 30 Speech Compoments and 10 Wind Compoments
PESQ for 50 Speech Compoments and 10 Wind Compoments
PESQ for 70 Speech Compoments and 10 Wind Compoments
PESQ for 90 Speech Compoments and 10 Wind Compoments
PESQ for 110 Speech Compoments and 10 Wind Compoments
PESQ for 130 Speech Compoments and 10 Wind Compoments
PESQ for 150 Speech Compoments and 10 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
SNR


PESQ for 30 Speech Compoments and 30 Wind Compoments
PESQ for 50 Speech Compoments and 30 Wind Compoments
PESQ for 70 Speech Compoments and 30 Wind Compoments
PESQ for 90 Speech Compoments and 30 Wind Compoments
PESQ for 110 Speech Compoments and 30 Wind Compoments
PESQ for 130 Speech Compoments and 30 Wind Compoments
PESQ for 150 Speech Compoments and 30 Wind Compoments
PESQ score
SNR
NP
IS
KL
NNSC
SS
STSA

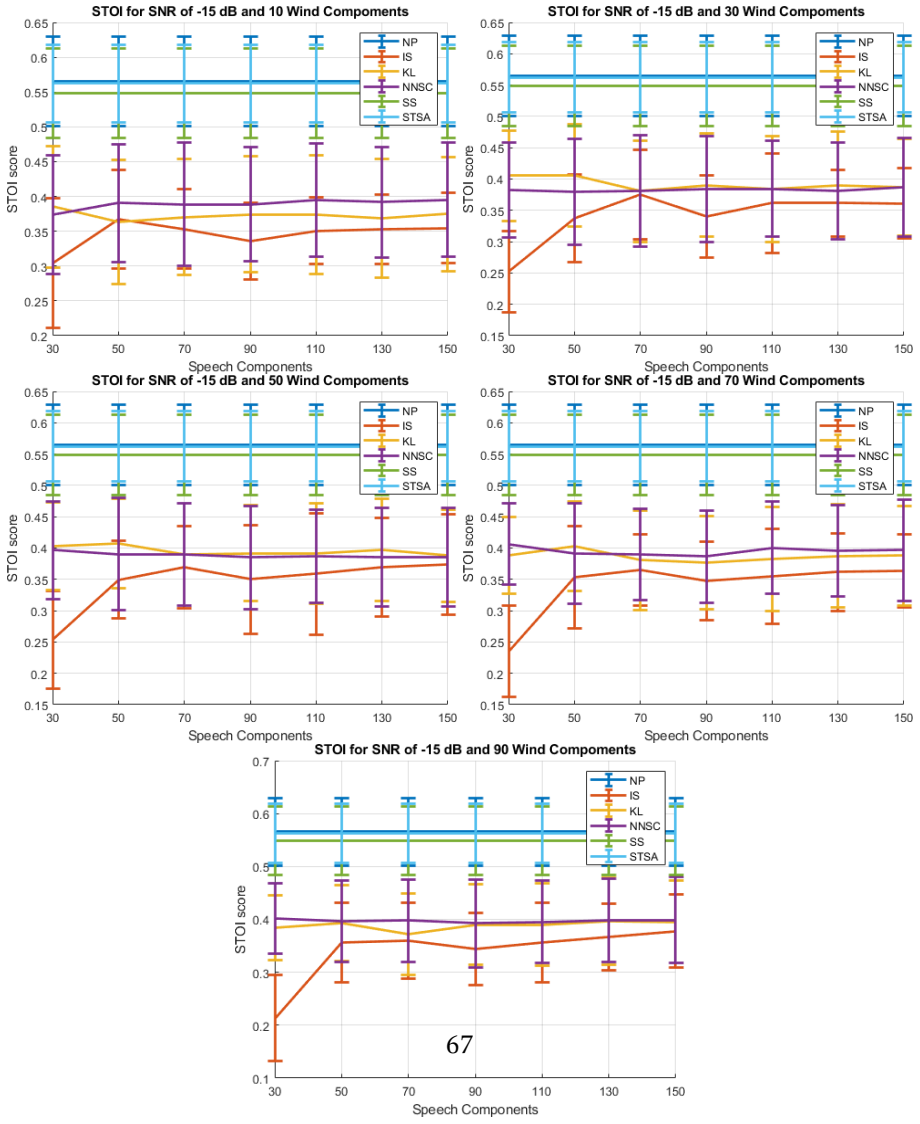

PESQ for 30 Speech Compoments and 50 Wind Compoments
PESQ for 50 Speech Compoments and 50 Wind Compoments
PESQ for 70 Speech Compoments and 50 Wind Compoments
PESQ for 90 Speech Compoments and 50 Wind Compoments
PESQ for 110 Speech Compoments and 50 Wind Compoments
PESQ for 130 Speech Compoments and 50 Wind Compoments
PESQ for 150 Speech Compoments and 50 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
SNR
0.5
1
1.5
2
2.5
3
-15
-10
-5
0
5

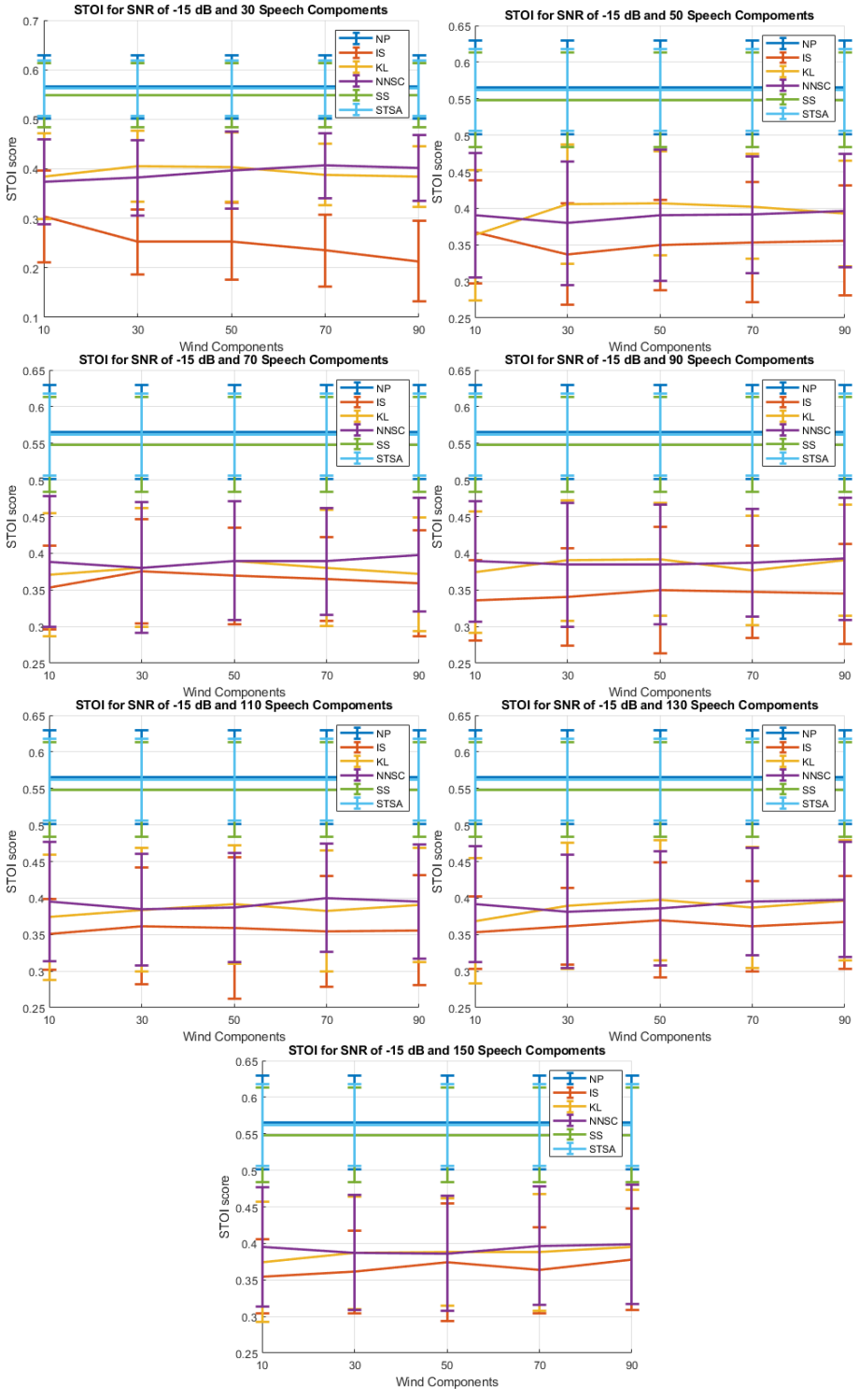

PESQ for 30 Speech Compoments and 70 Wind Compoments
PESQ for 50 Speech Compoments and 70 Wind Compoments
PESQ for 70 Speech Compoments and 70 Wind Compoments
PESQ for 90 Speech Compoments and 70 Wind Compoments
PESQ for 110 Speech Compoments and 70 Wind Compoments
PESQ for 130 Speech Compoments and 70 Wind Compoments
PESQ for 150 Speech Compoments and 70 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
SNR

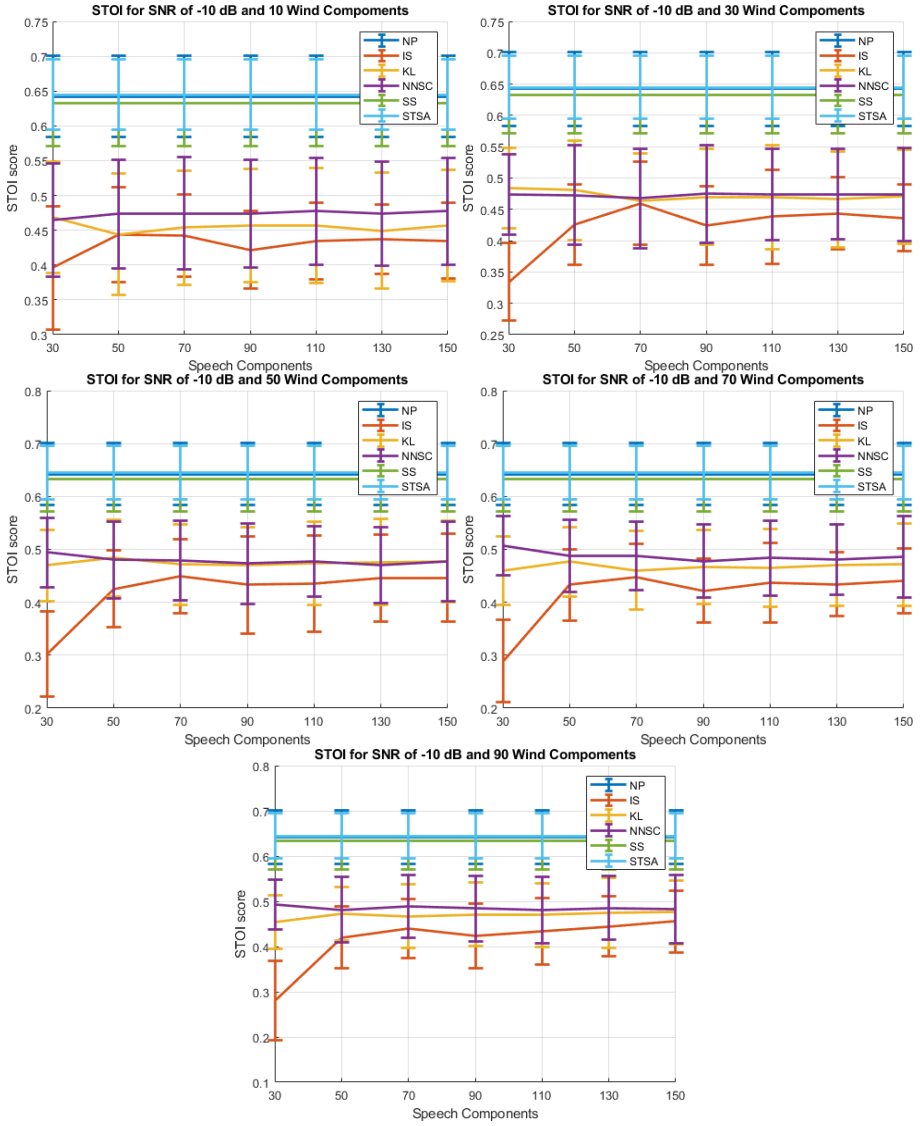

PESQ for 30 Speech Compoments and 90 Wind Compoments
PESQ for 50 Speech Compoments and 90 Wind Compoments
PESQ for 70 Speech Compoments and 90 Wind Compoments
PESQ for 90 Speech Compoments and 90 Wind Compoments
PESQ for 110 Speech Compoments and 90 Wind Compoments
PESQ for 130 Speech Compoments and 90 Wind Compoments
PESQ for 150 Speech Compoments and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
PESQ score
SNR

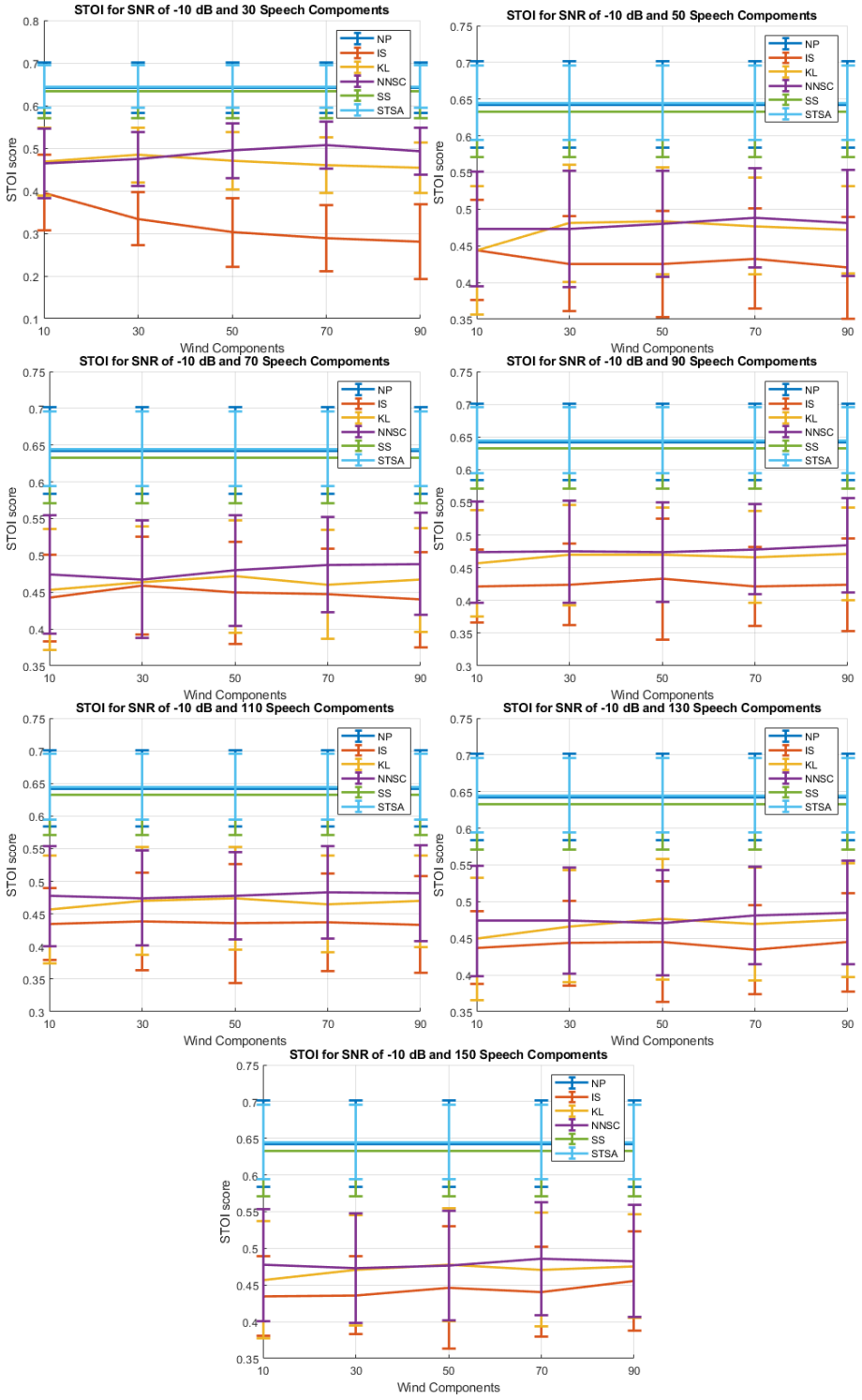
Appendix B

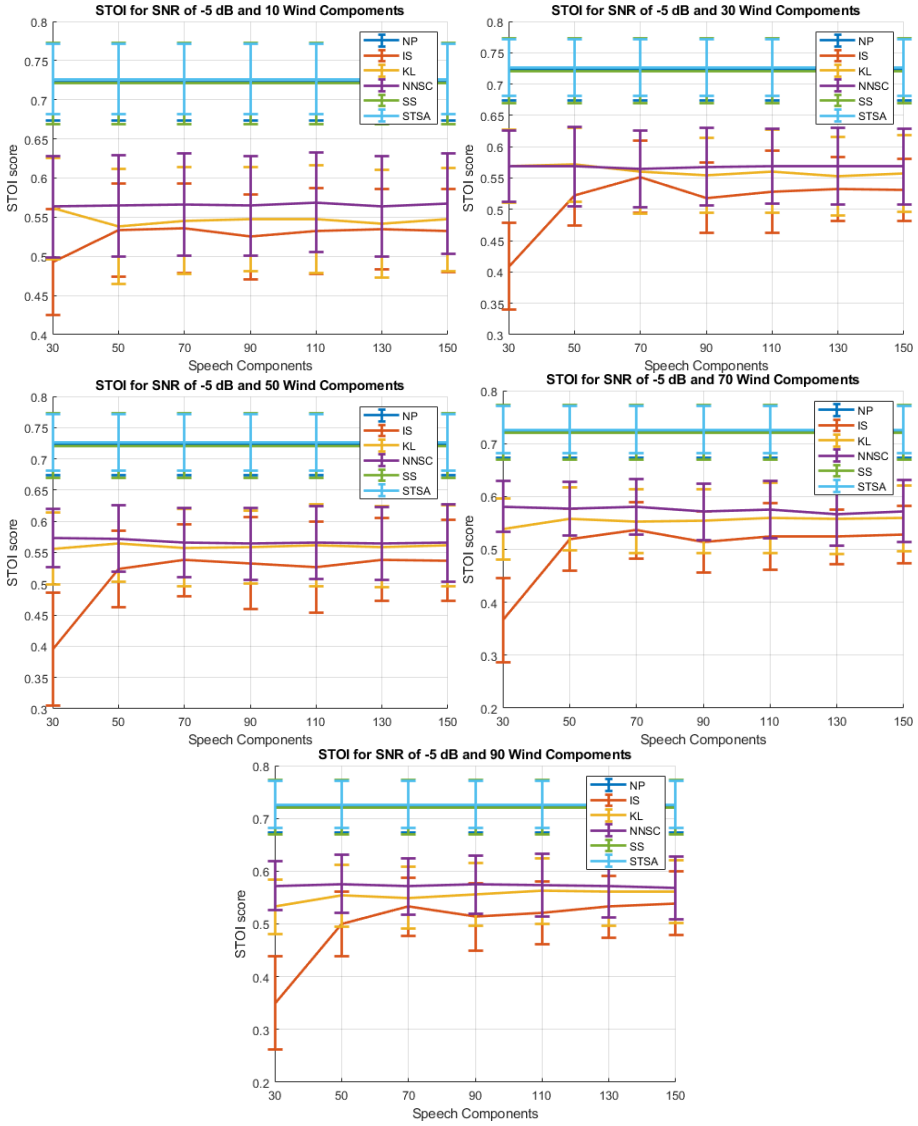
STOI Plots

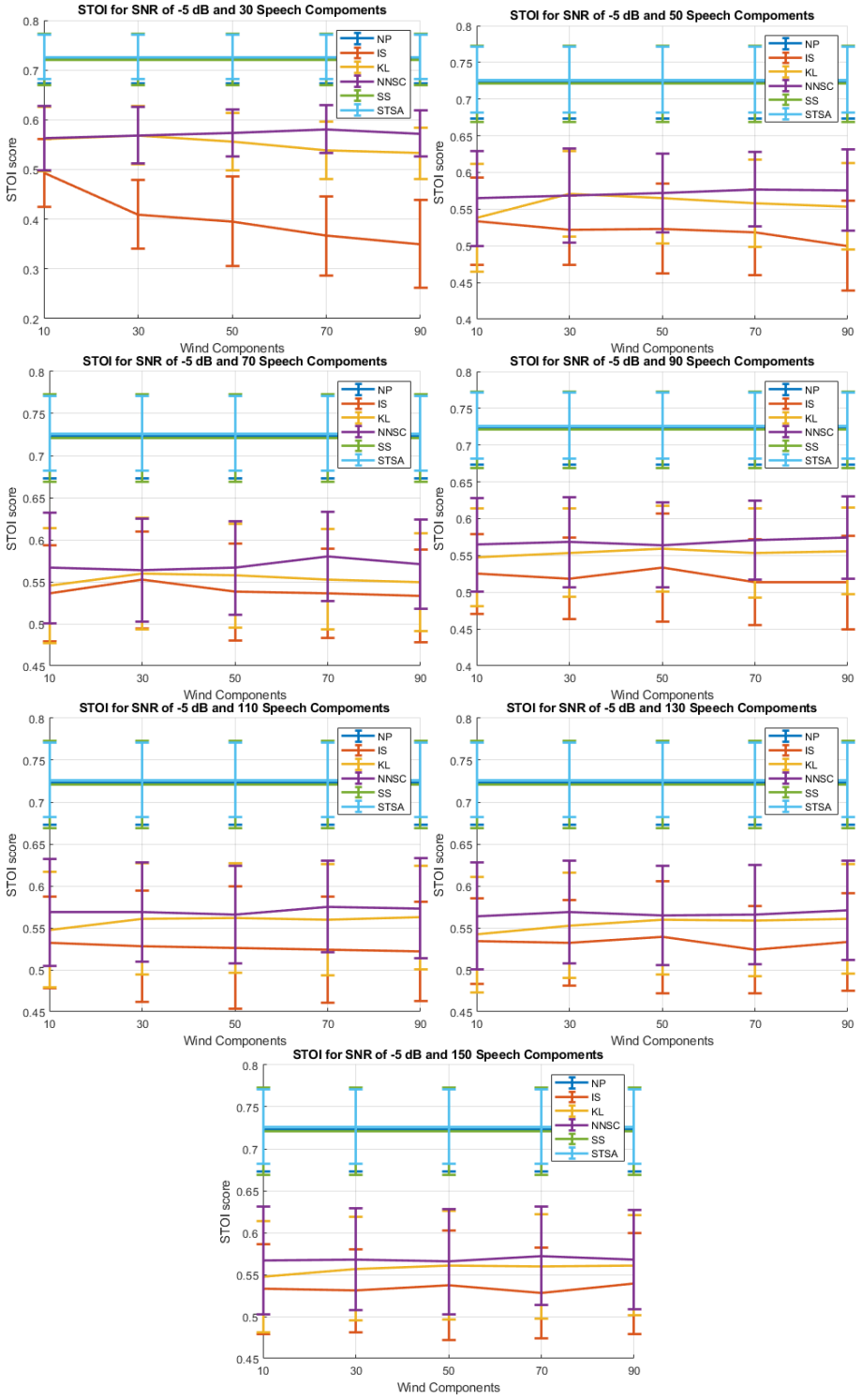


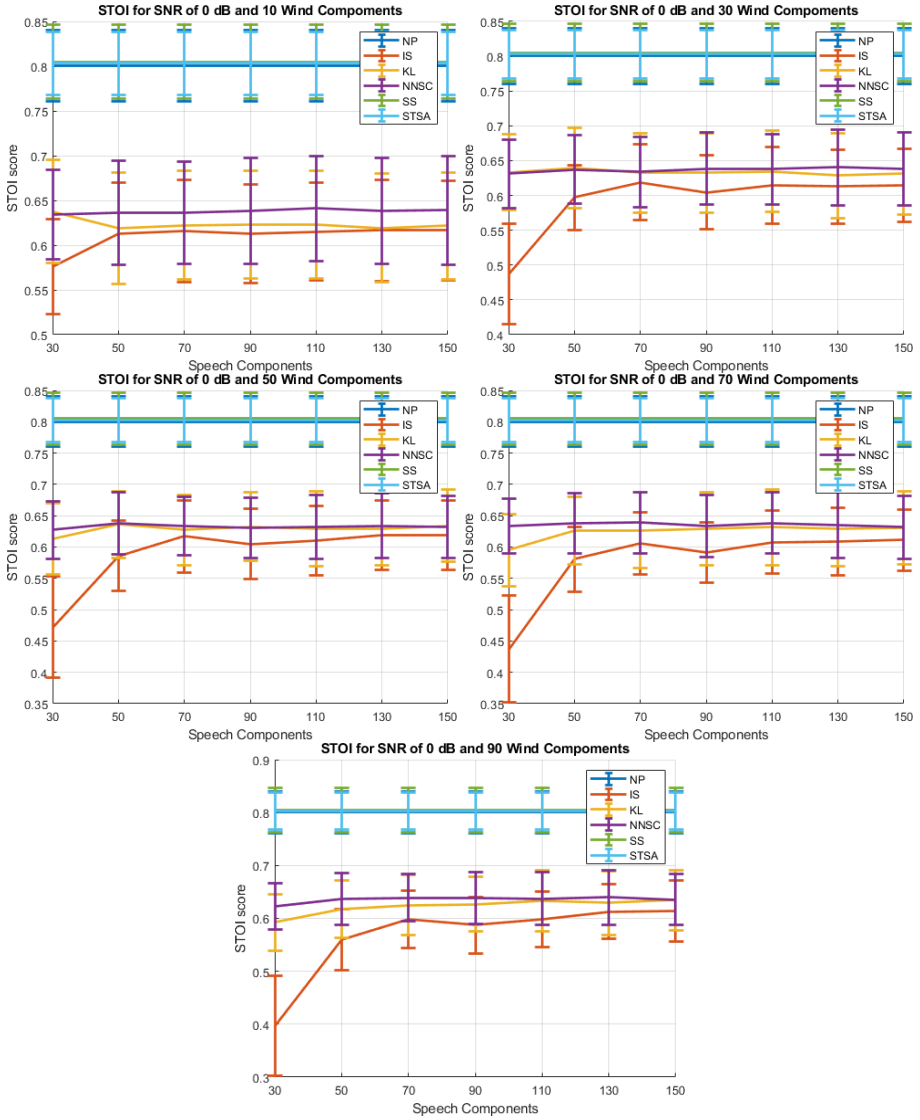

STOI for SNR of -15 dB and 30 Speech Compoments
STOI for SNR of -15 dB and 50 Speech Compoments
STOI for SNR of -15 dB and 70 Speech Compoments
STOI for SNR of -15 dB and 90 Speech Compoments
STOI for SNR of -15 dB and 110 Speech Compoments
STOI for SNR of -15 dB and 130 Speech Compoments
STOI for SNR of -15 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Wind Components

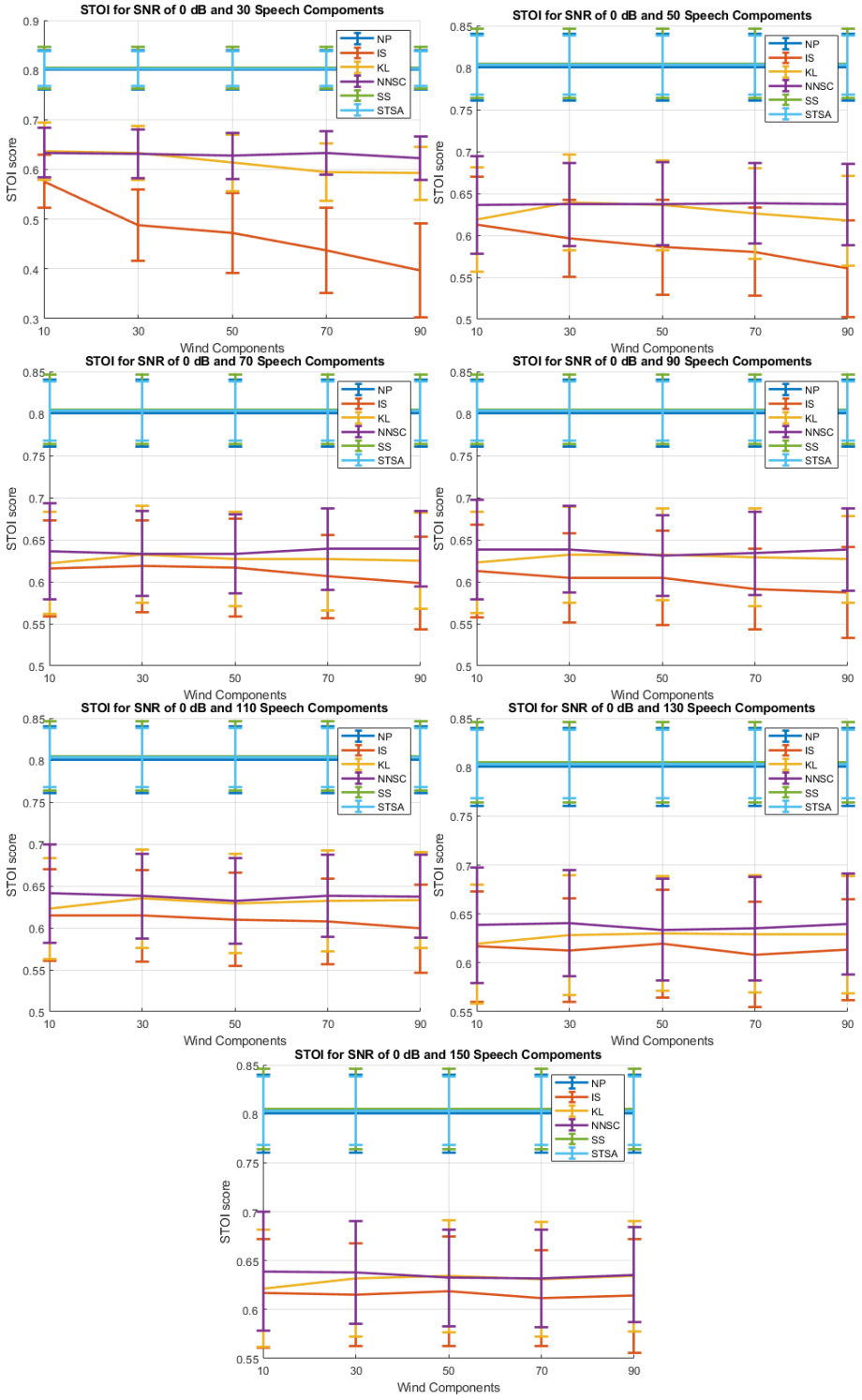

STOI for SNR of -10 dB and 10 Wind Compoments
STOI for SNR of -10 dB and 30 Wind Compoments
STOI for SNR of -10 dB and 50 Wind Compoments
STOI for SNR of -10 dB and 70 Wind Compoments
STOI for SNR of -10 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Speech Components

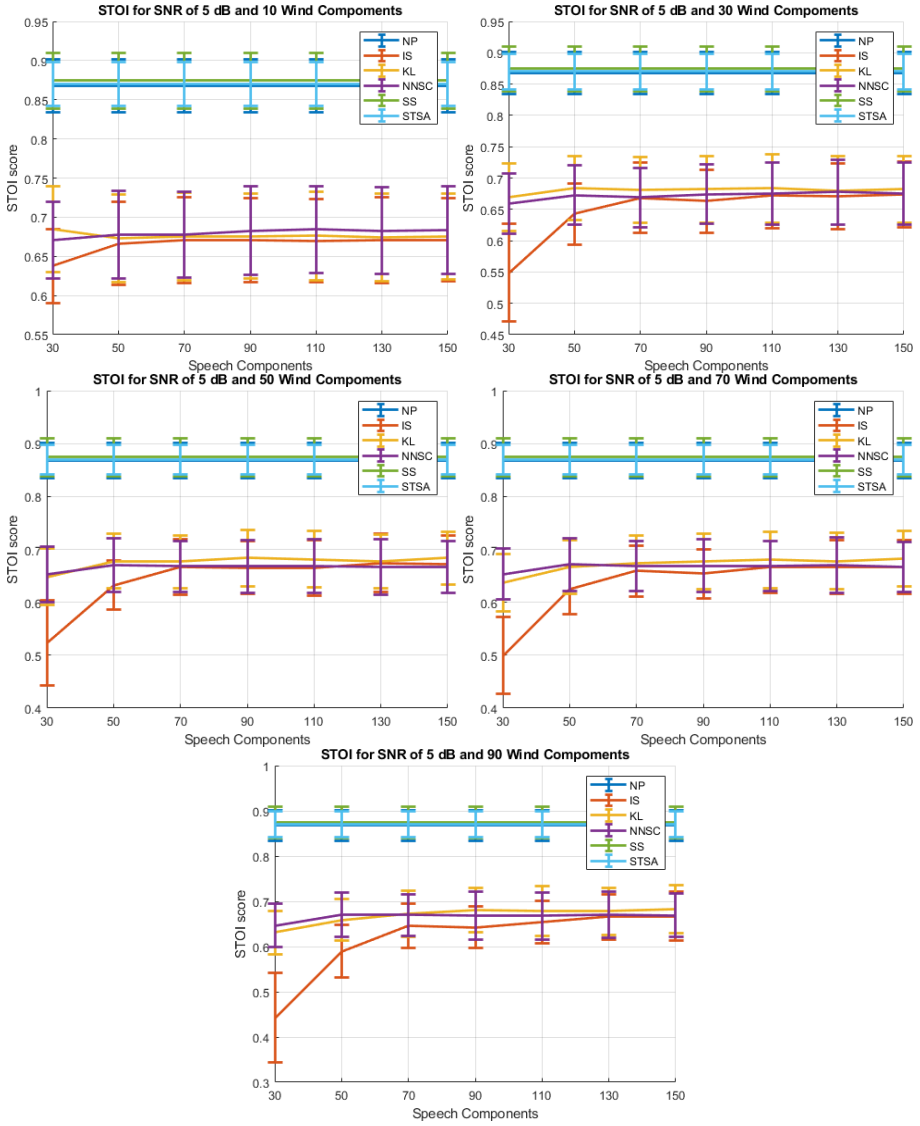

STOI for SNR of -10 dB and 30 Speech Compoments
STOI for SNR of -10 dB and 50 Speech Compoments
STOI for SNR of -10 dB and 70 Speech Compoments
STOI for SNR of -10 dB and 90 Speech Compoments
STOI for SNR of -10 dB and 110 Speech Compoments
STOI for SNR of -10 dB and 130 Speech Compoments
STOI for SNR of -10 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Wind Components

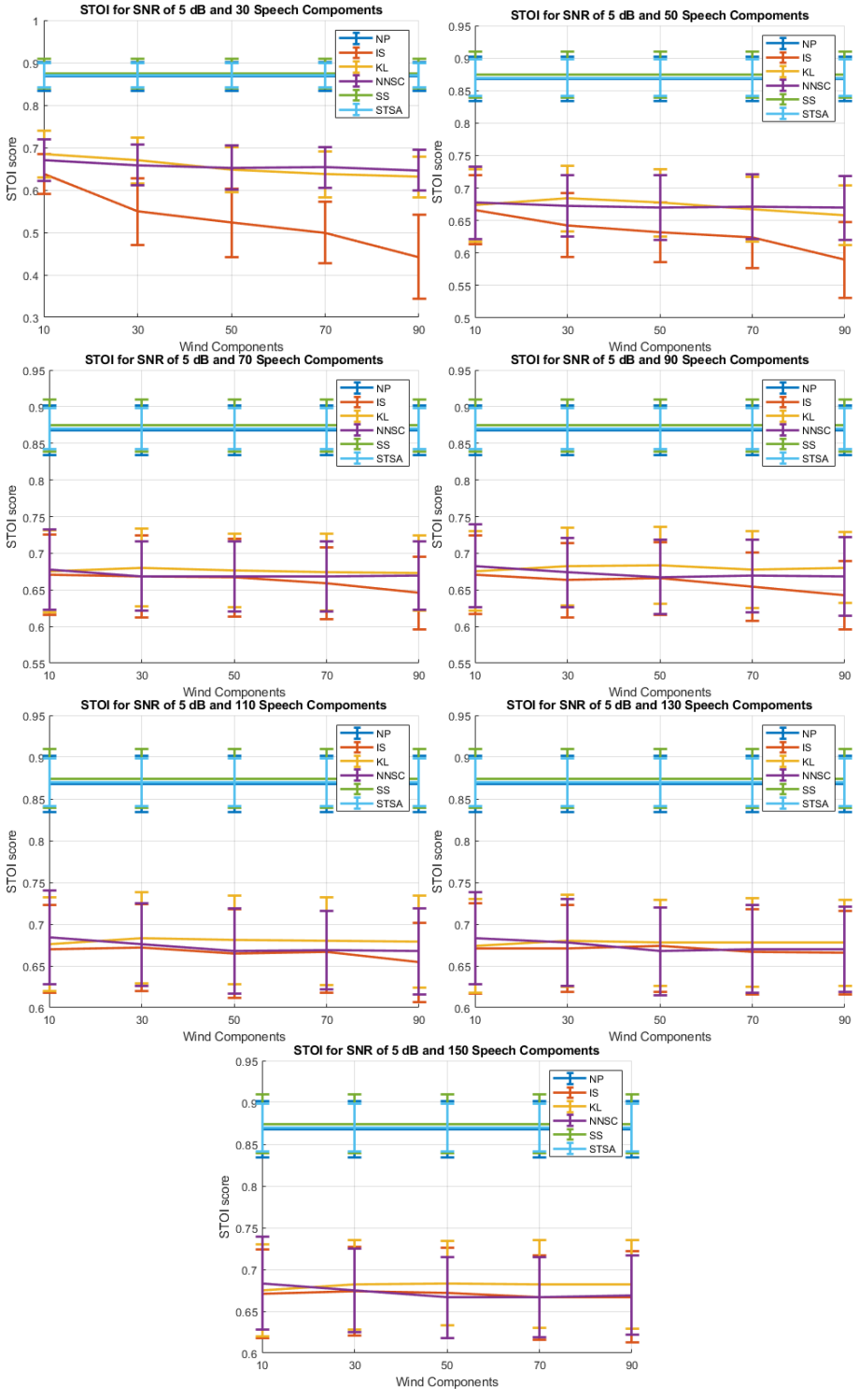

STOI for SNR of -5 dB and 10 Wind Compoments
STOI for SNR of -5 dB and 30 Wind Compoments
STOI for SNR of -5 dB and 50 Wind Compoments
STOI for SNR of -5 dB and 70 Wind Compoments
STOI for SNR of -5 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Speech Components

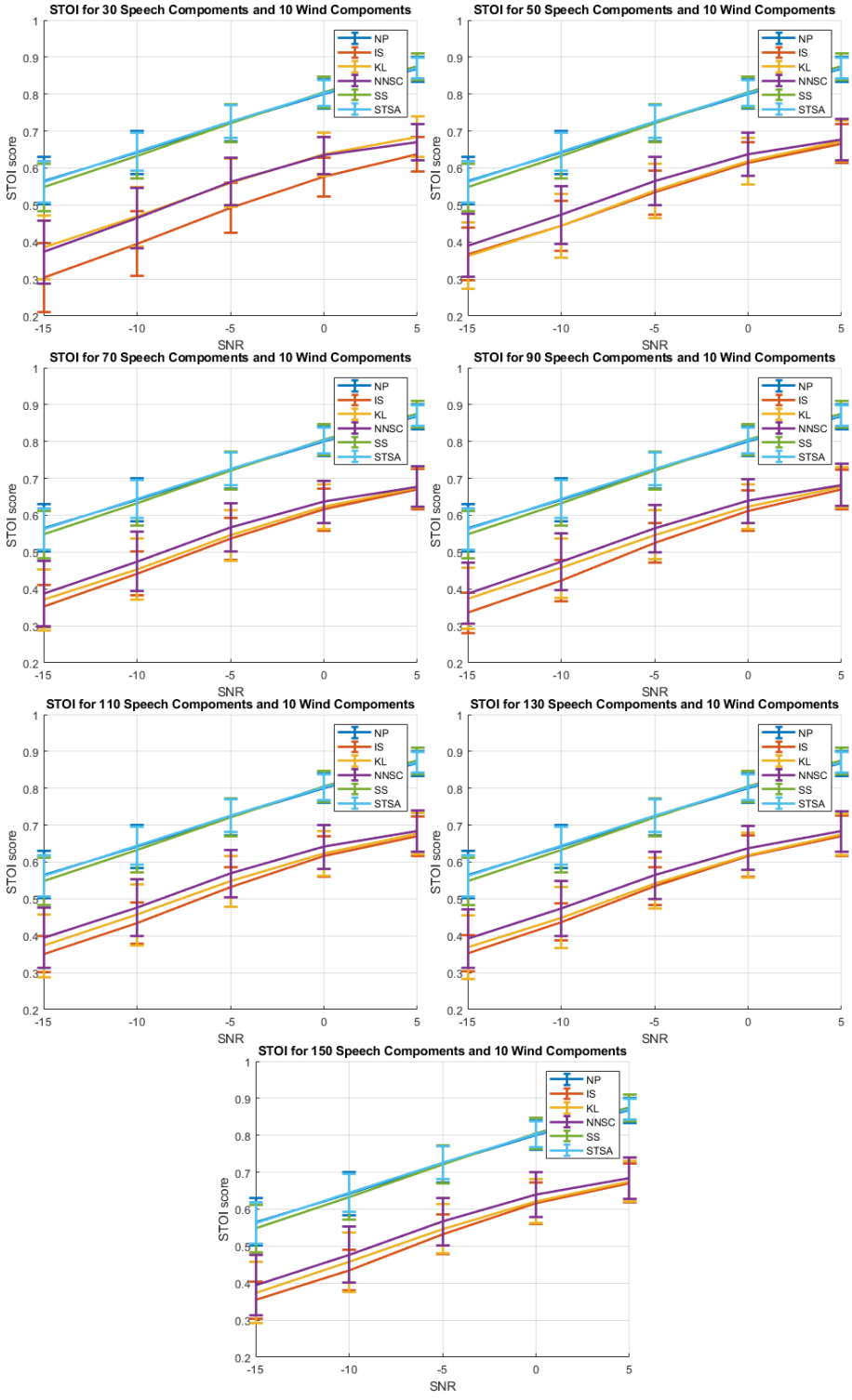

STOI for SNR of -5 dB and 30 Speech Compoments
STOI for SNR of -5 dB and 50 Speech Compoments
STOI for SNR of -5 dB and 70 Speech Compoments
STOI for SNR of -5 dB and 90 Speech Compoments
STOI for SNR of -5 dB and 110 Speech Compoments
STOI for SNR of -5 dB and 130 Speech Compoments
STOI for SNR of -5 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Wind Components

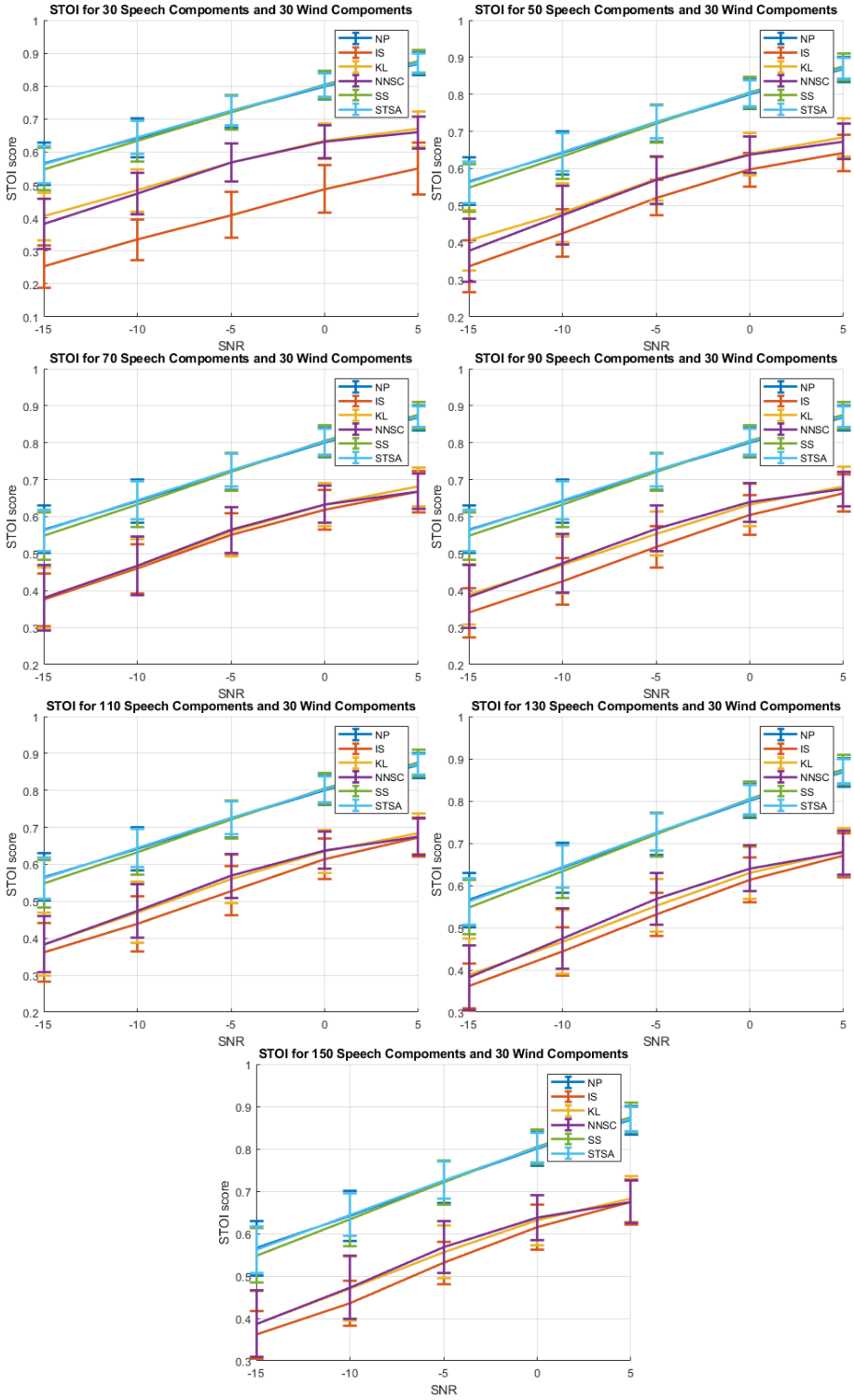

STOI for SNR of 0 dB and 10 Wind Compoments
STOI for SNR of 0 dB and 30 Wind Compoments
STOI for SNR of 0 dB and 50 Wind Compoments
STOI for SNR of 0 dB and 70 Wind Compoments
STOI for SNR of 0 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Speech Components

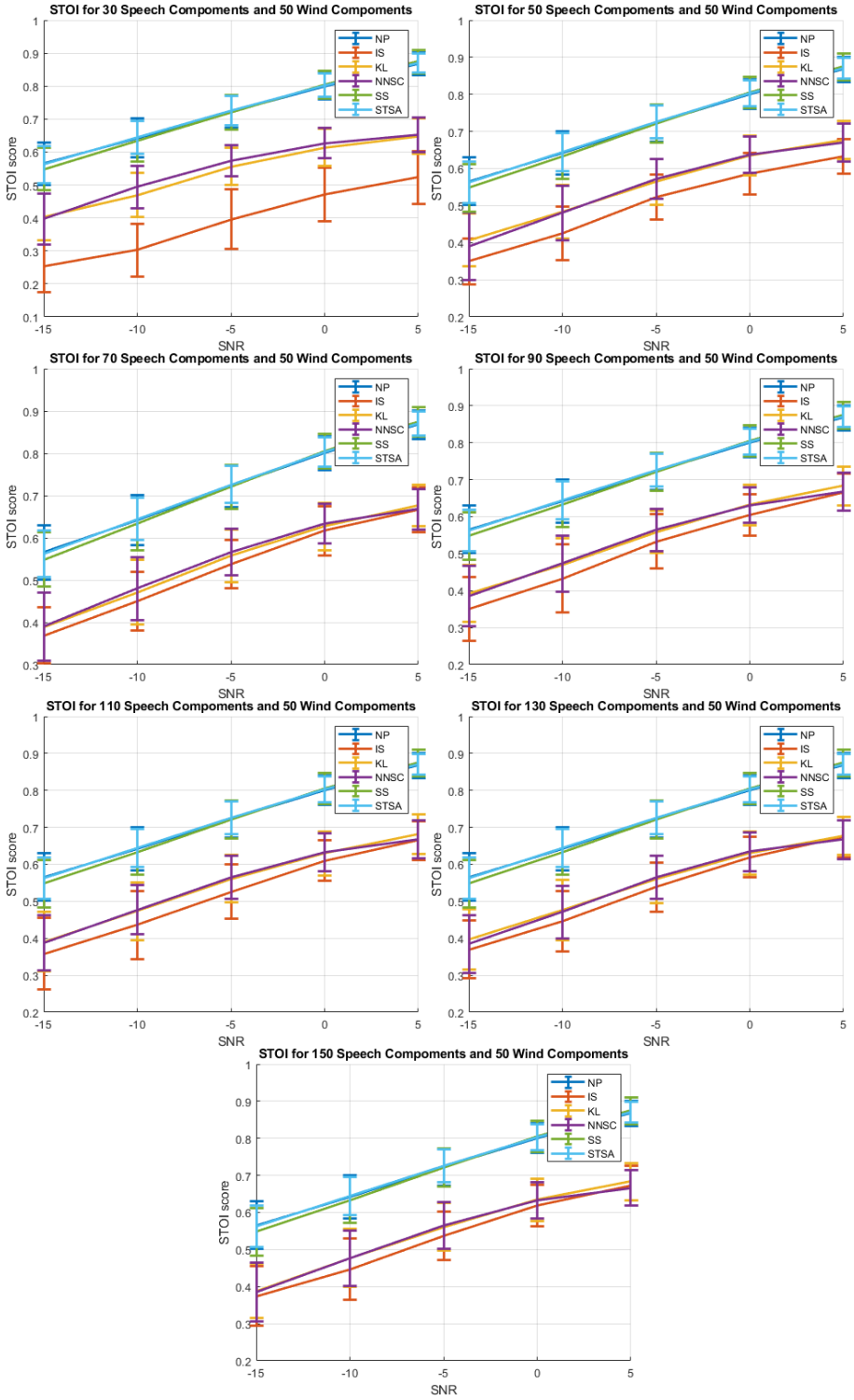

STOI for SNR of 0 dB and 30 Speech Compoments
STOI for SNR of 0 dB and 50 Speech Compoments
STOI for SNR of 0 dB and 70 Speech Compoments
STOI for SNR of 0 dB and 90 Speech Compoments
STOI for SNR of 0 dB and 110 Speech Compoments
STOI for SNR of 0 dB and 130 Speech Compoments
STOI for SNR of 0 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Wind Components

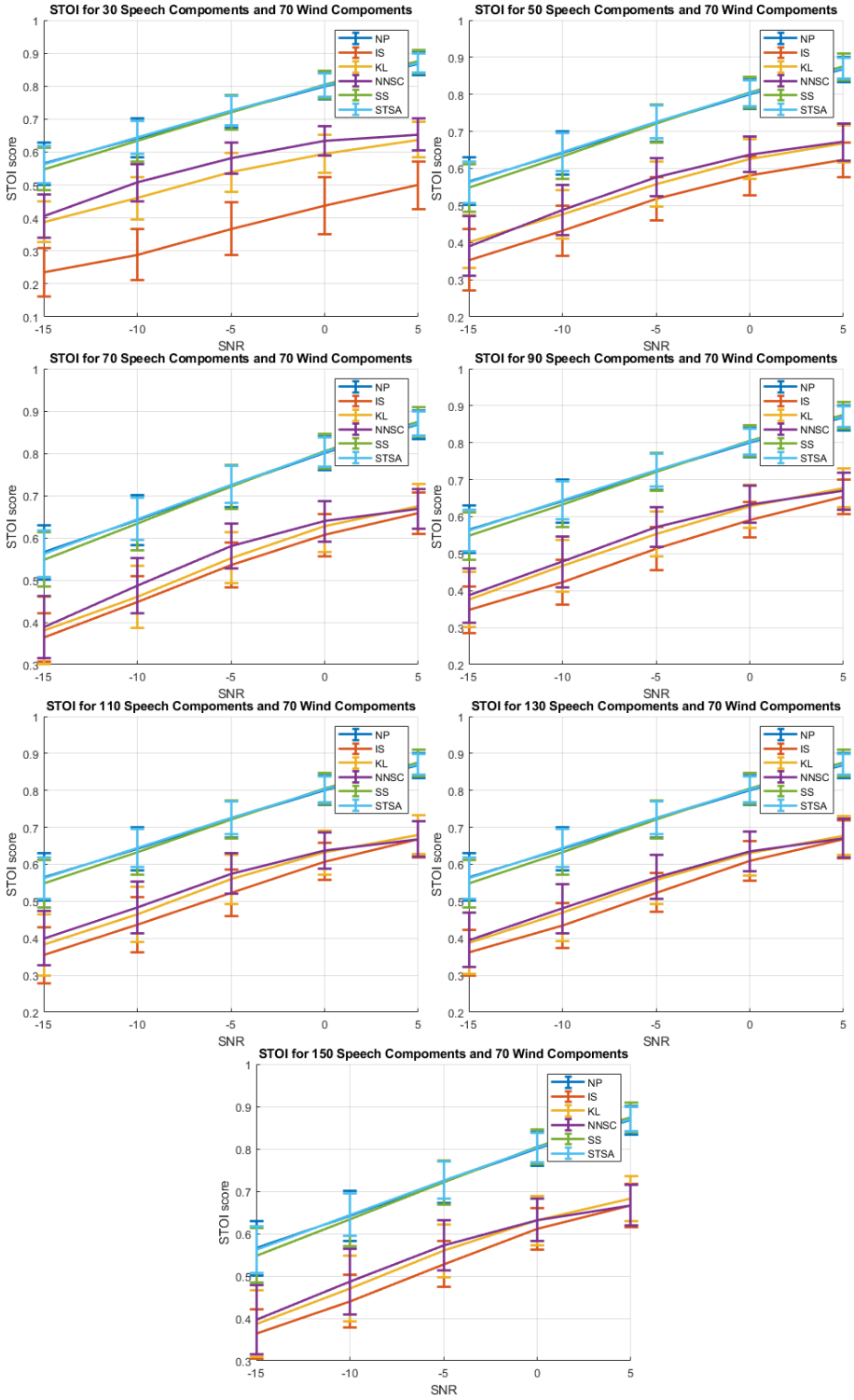

STOI for SNR of 5 dB and 10 Wind Compoments
STOI for SNR of 5 dB and 30 Wind Compoments
STOI for SNR of 5 dB and 50 Wind Compoments
STOI for SNR of 5 dB and 70 Wind Compoments
STOI for SNR of 5 dB and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Speech Components

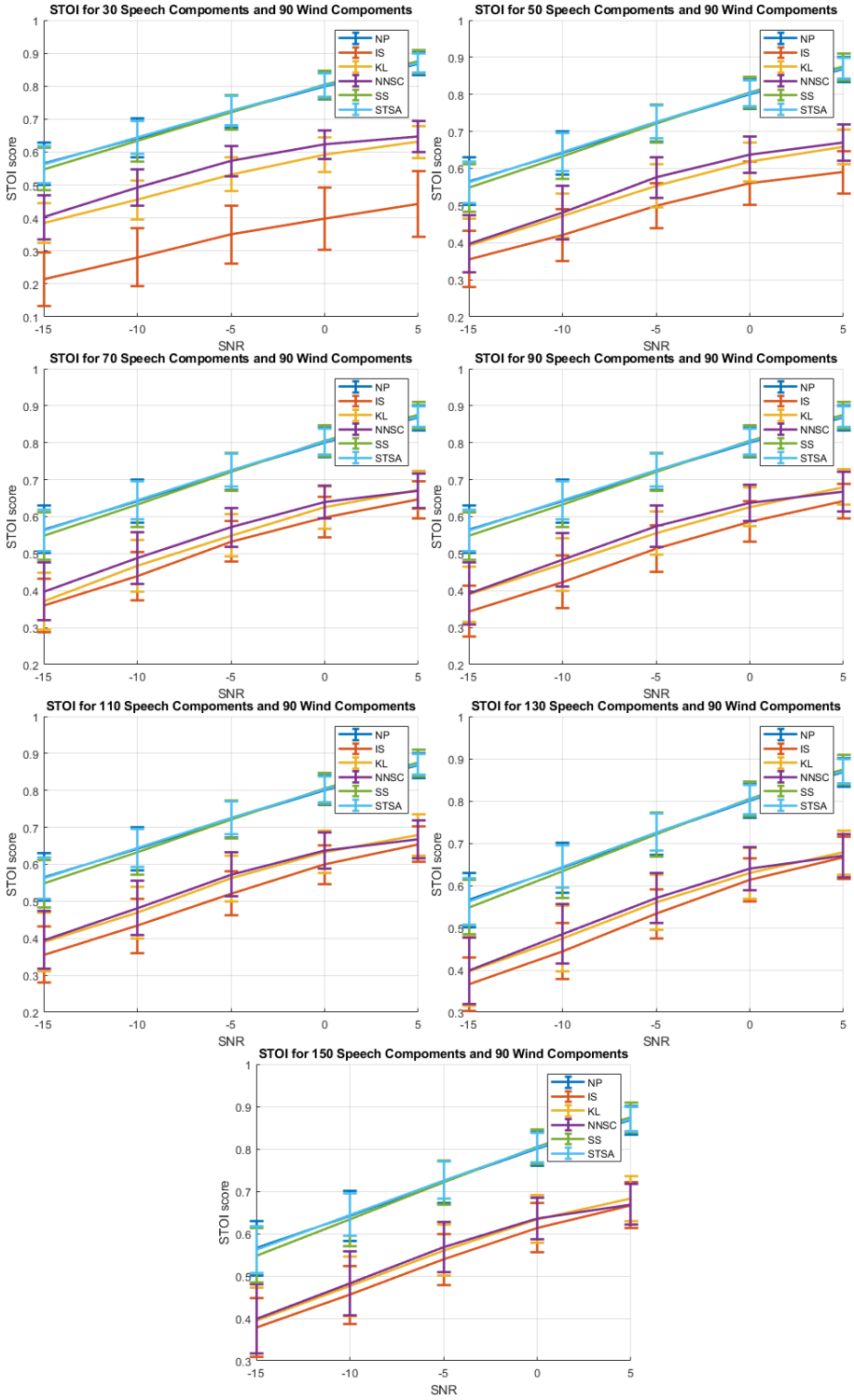

STOI for SNR of 5 dB and 30 Speech Compoments
STOI for SNR of 5 dB and 50 Speech Compoments
STOI for SNR of 5 dB and 70 Speech Compoments
STOI for SNR of 5 dB and 90 Speech Compoments
STOI for SNR of 5 dB and 110 Speech Compoments
STOI for SNR of 5 dB and 130 Speech Compoments
STOI for SNR of 5 dB and 150 Speech Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
Wind Components


STOI for 30 Speech Compoments and 10 Wind Compoments
STOI for 50 Speech Compoments and 10 Wind Compoments
STOI for 70 Speech Compoments and 10 Wind Compoments
STOI for 90 Speech Compoments and 10 Wind Compoments
STOI for 110 Speech Compoments and 10 Wind Compoments
STOI for 130 Speech Compoments and 10 Wind Compoments
STOI for 150 Speech Compoments and 10 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
SNR


STOI for 30 Speech Compoments and 30 Wind Compoments
STOI for 50 Speech Compoments and 30 Wind Compoments
STOI for 70 Speech Compoments and 30 Wind Compoments
STOI for 90 Speech Compoments and 30 Wind Compoments
STOI for 110 Speech Compoments and 30 Wind Compoments
STOI for 130 Speech Compoments and 30 Wind Compoments
STOI for 150 Speech Compoments and 30 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
SNR

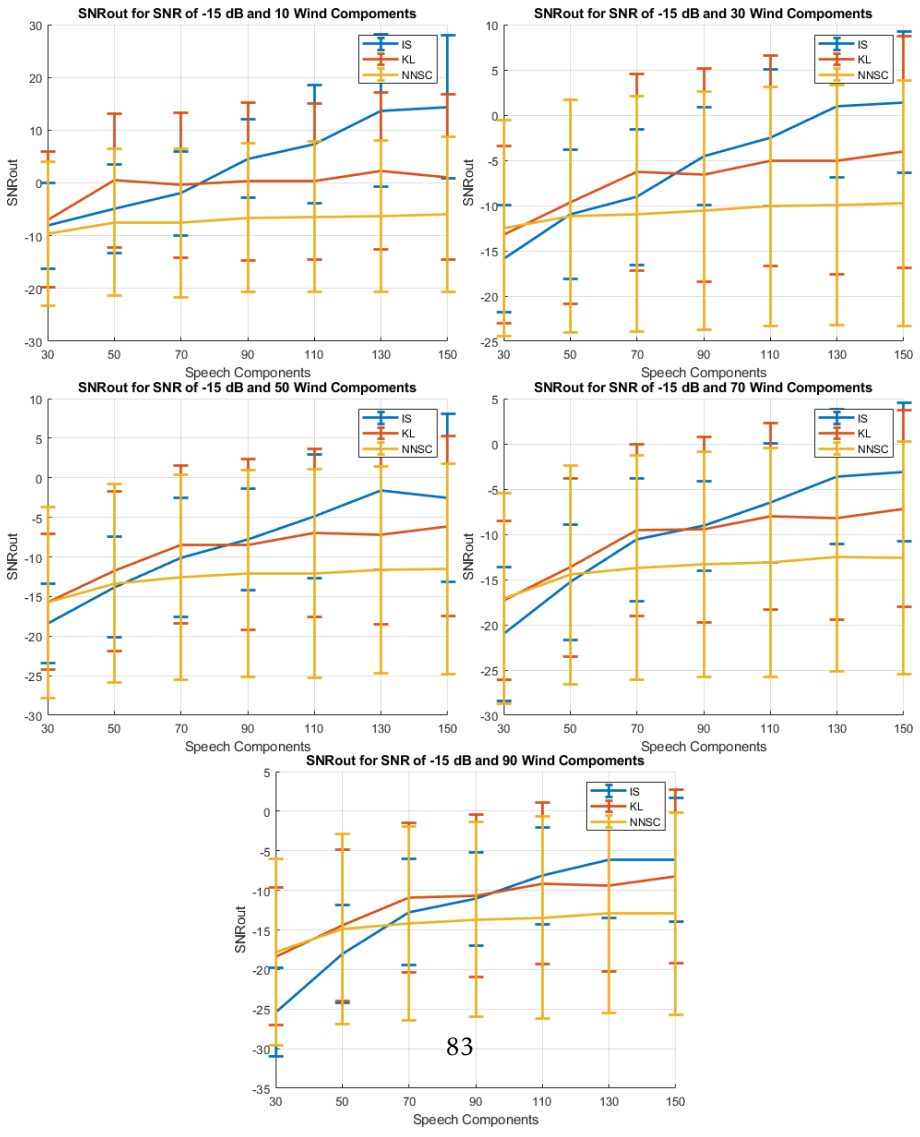

STOI for 30 Speech Compoments and 50 Wind Compoments
STOI for 50 Speech Compoments and 50 Wind Compoments
STOI for 70 Speech Compoments and 50 Wind Compoments
STOI for 90 Speech Compoments and 50 Wind Compoments
STOI for 110 Speech Compoments and 50 Wind Compoments
STOI for 130 Speech Compoments and 50 Wind Compoments
STOI for 150 Speech Compoments and 50 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
SNR

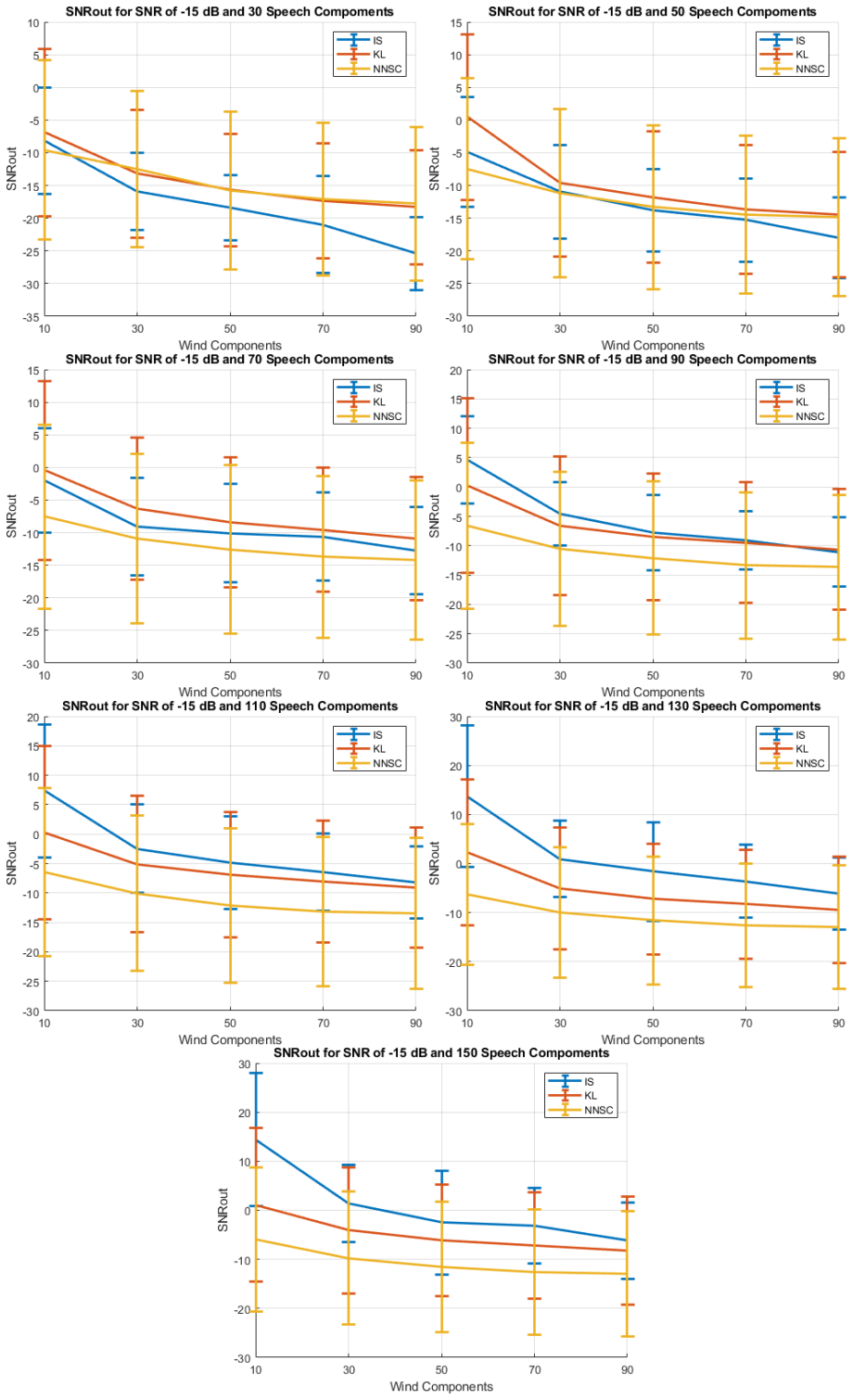

STOI for 30 Speech Compoments and 70 Wind Compoments
STOI for 50 Speech Compoments and 70 Wind Compoments
STOI for 70 Speech Compoments and 70 Wind Compoments
STOI for 90 Speech Compoments and 70 Wind Compoments
STOI for 110 Speech Compoments and 70 Wind Compoments
STOI for 130 Speech Compoments and 70 Wind Compoments
STOI for 150 Speech Compoments and 70 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
SNR

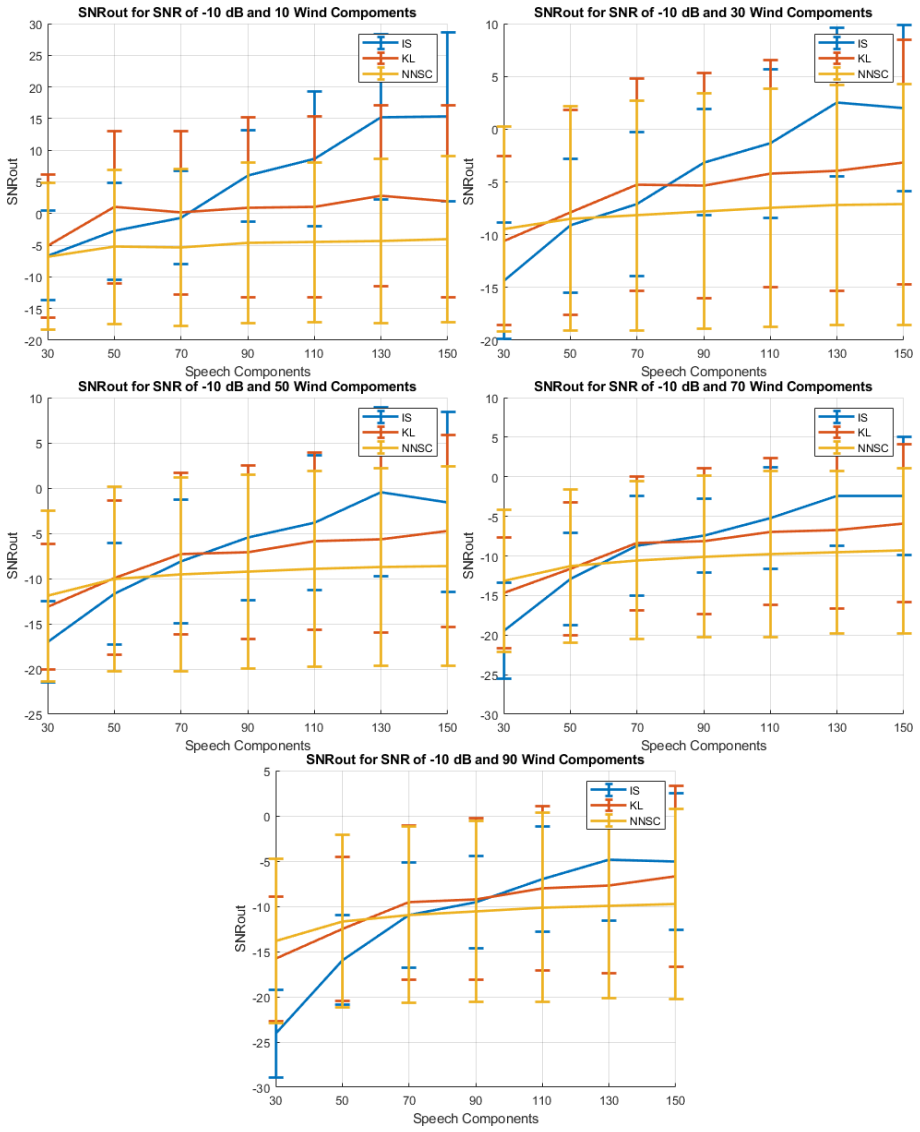

STOI for 30 Speech Compoments and 90 Wind Compoments
STOI for 50 Speech Compoments and 90 Wind Compoments
STOI for 70 Speech Compoments and 90 Wind Compoments
STOI for 90 Speech Compoments and 90 Wind Compoments
STOI for 110 Speech Compoments and 90 Wind Compoments
STOI for 130 Speech Compoments and 90 Wind Compoments
STOI for 150 Speech Compoments and 90 Wind Compoments
NP
IS
KL
NNSC
SS
STSA
STOI score
SNR

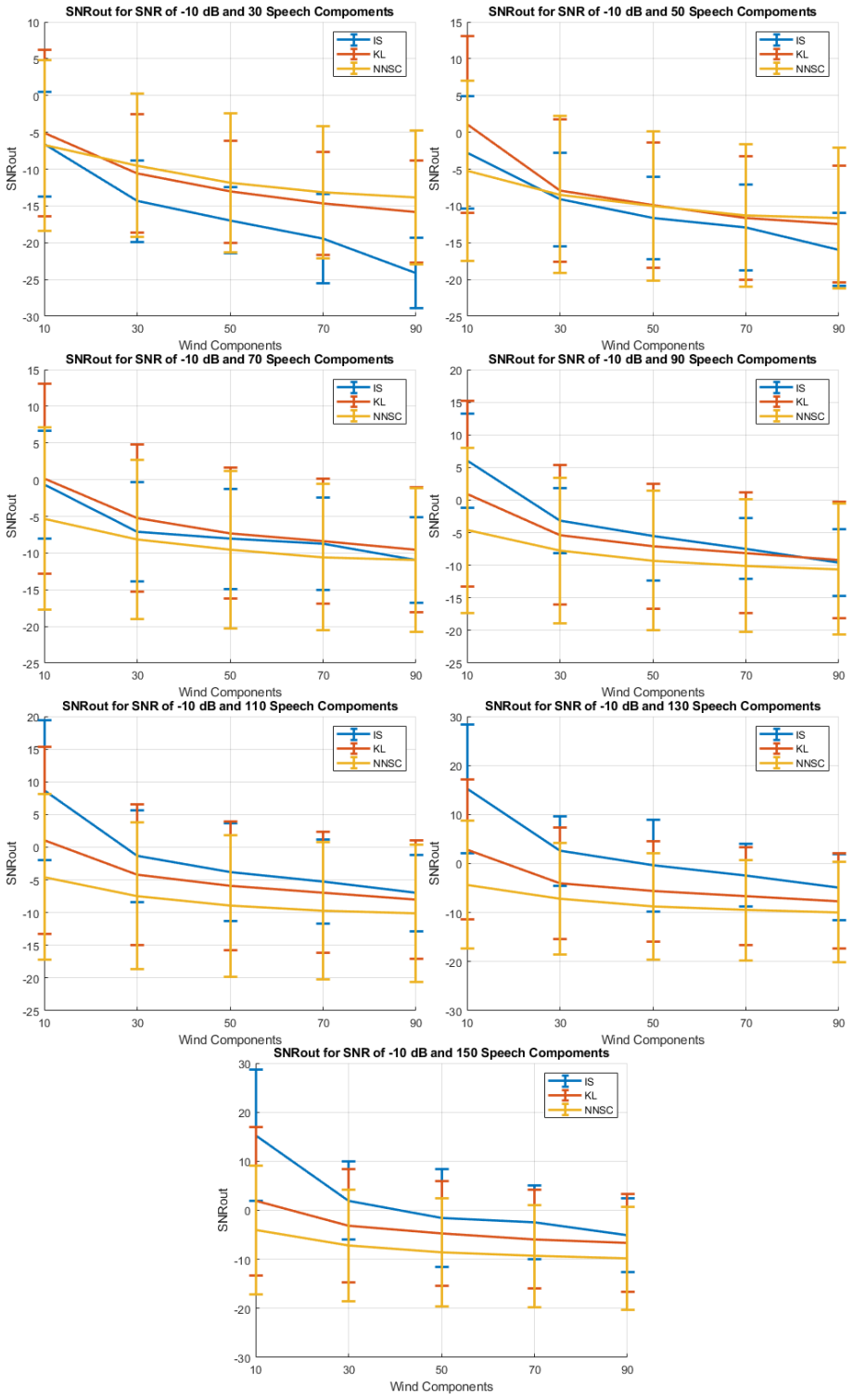
Appendix C

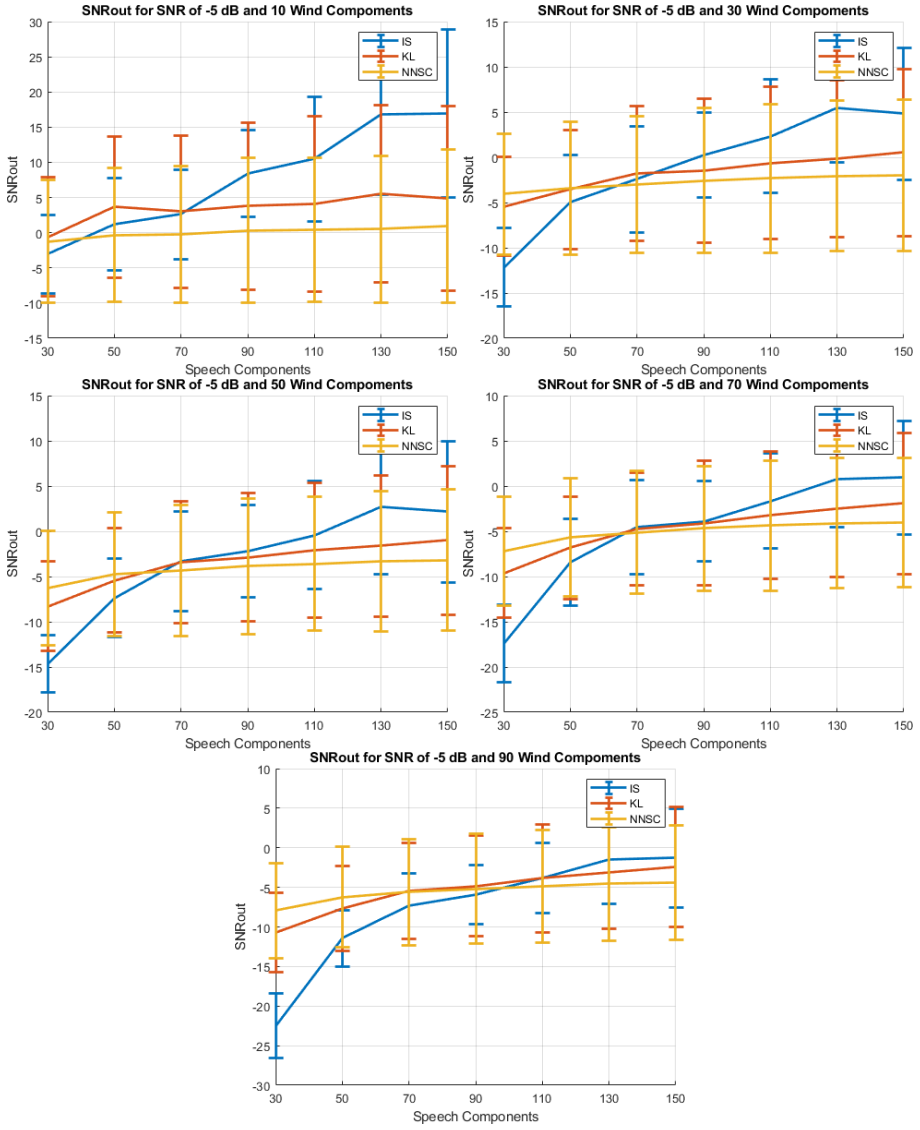
SNR Plots

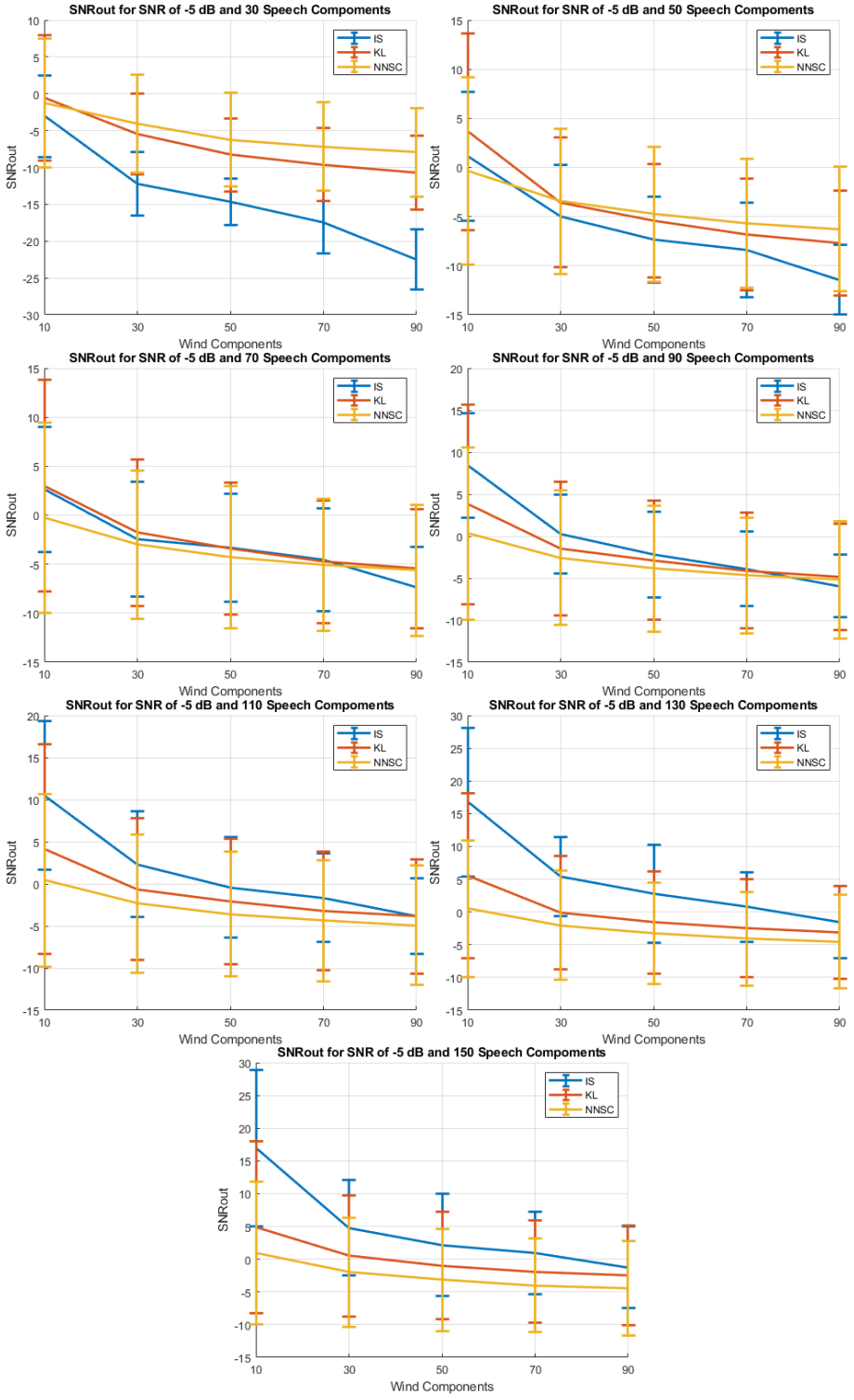


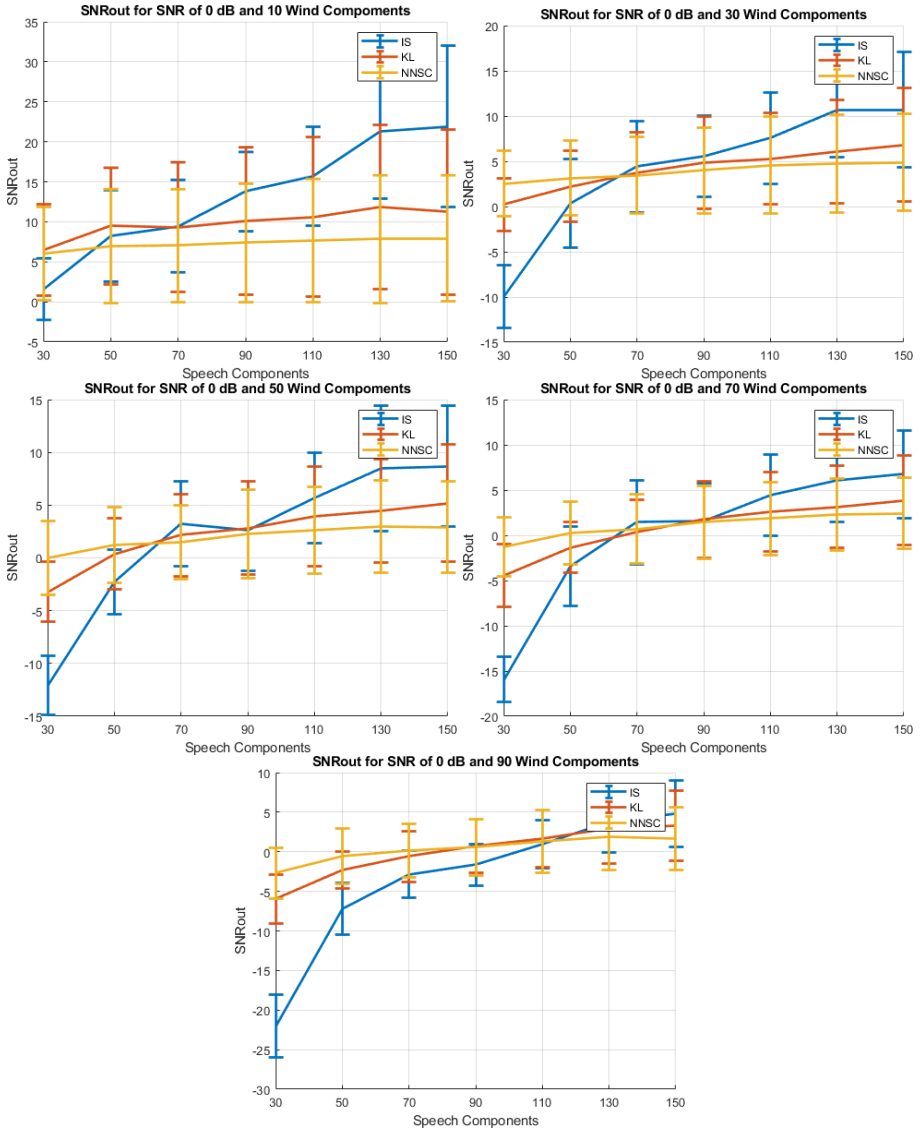

SNRout for SNR of -15 dB and 30 Speech Compoments
SNRout for SNR of -15 dB and 50 Speech Compoments
SNRout for SNR of -15 dB and 70 Speech Compoments
SNRout for SNR of -15 dB and 90 Speech Compoments
SNRout for SNR of -15 dB and 110 Speech Compoments
SNRout for SNR of -15 dB and 130 Speech Compoments
SNRout for SNR of -15 dB and 150 Speech Compoments
IS
KL
NNSC
SNRout
Wind Components

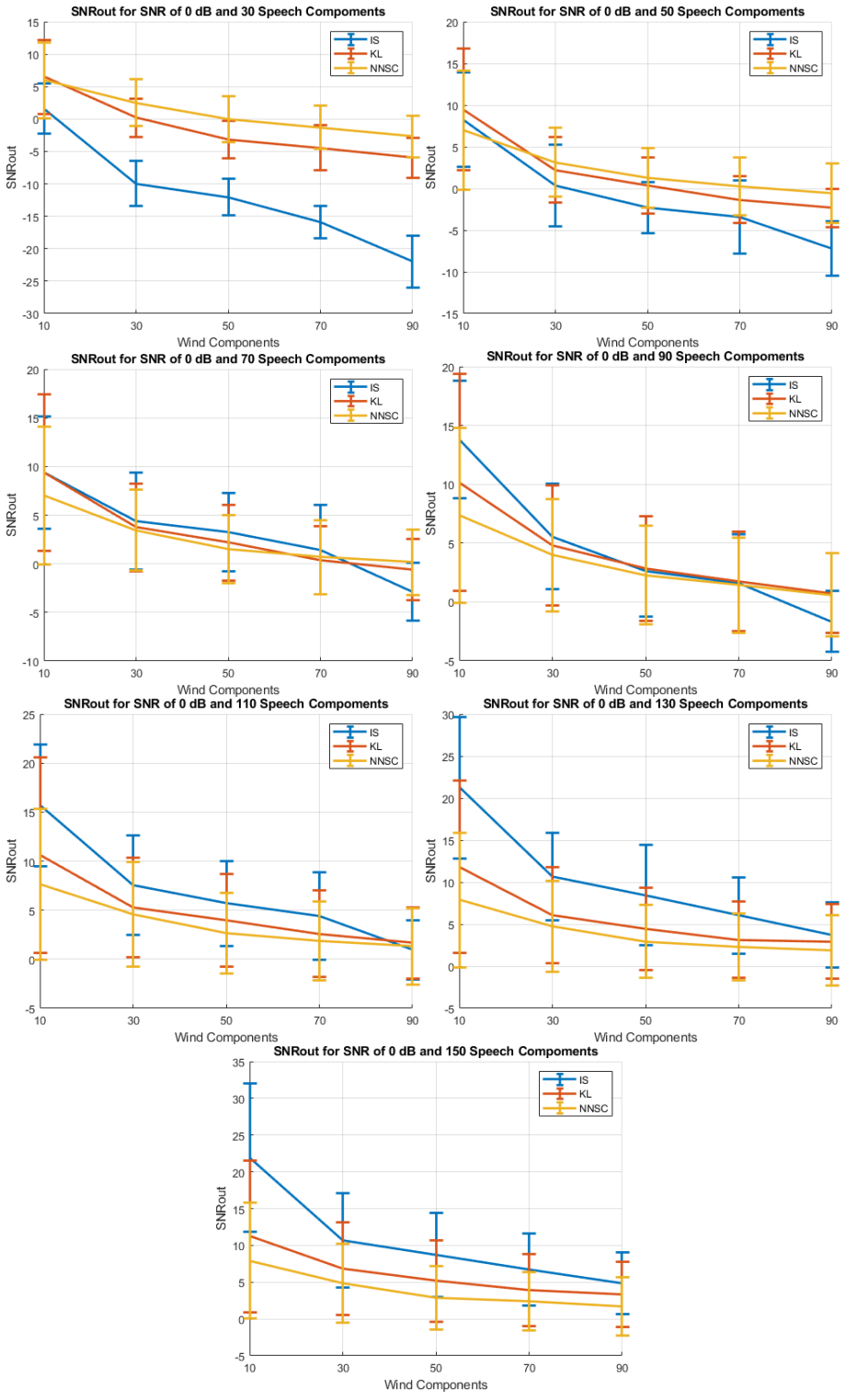

SNRout for SNR of -10 dB and 10 Wind Compoments
SNRout for SNR of -10 dB and 30 Wind Compoments
SNRout for SNR of -10 dB and 50 Wind Compoments
SNRout for SNR of -10 dB and 70 Wind Compoments
SNRout for SNR of -10 dB and 90 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components

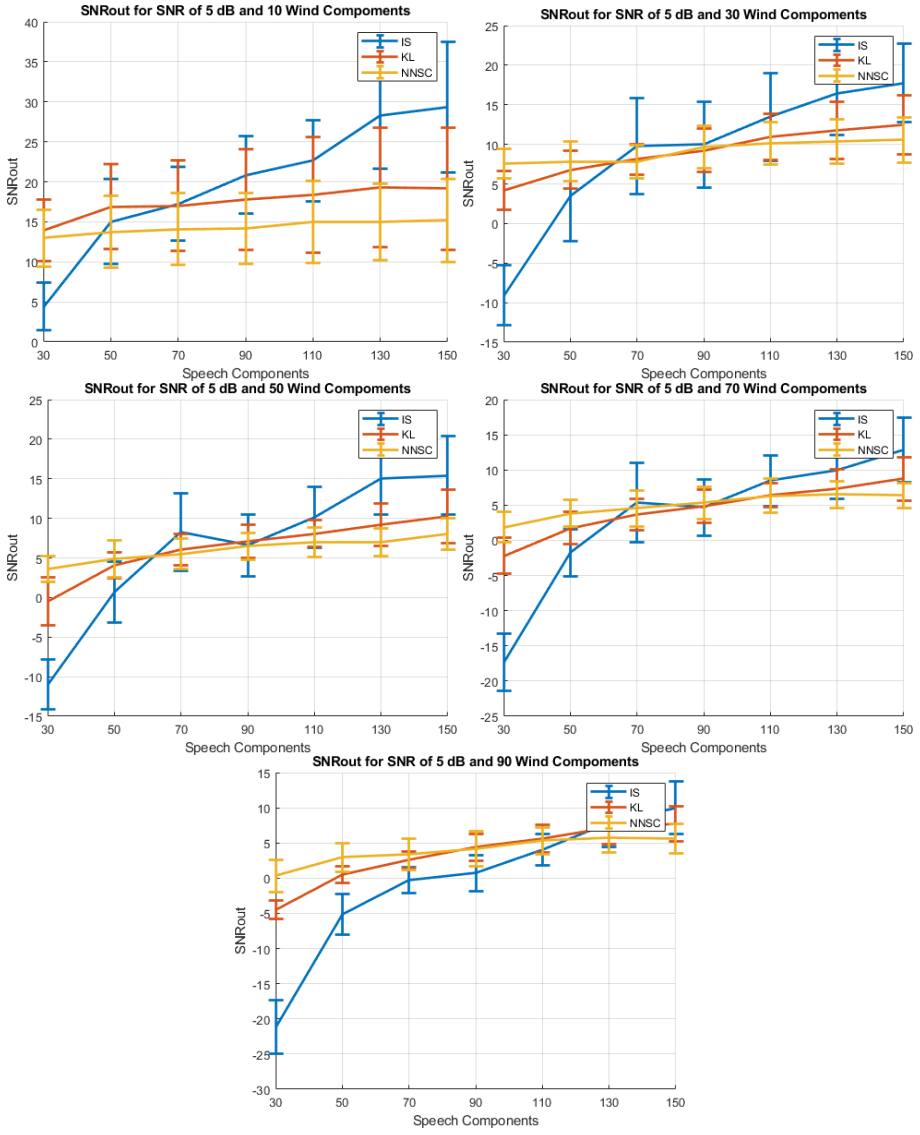

SNRout for SNR of -10 dB and 30 Speech Compoments
SNRout for SNR of -10 dB and 50 Speech Compoments
SNRout for SNR of -10 dB and 70 Speech Compoments
SNRout for SNR of -10 dB and 90 Speech Compoments
SNRout for SNR of -10 dB and 110 Speech Compoments
SNRout for SNR of -10 dB and 130 Speech Compoments
SNRout for SNR of -10 dB and 150 Speech Compoments
IS
KL
NNSC
SNRout
Wind Components

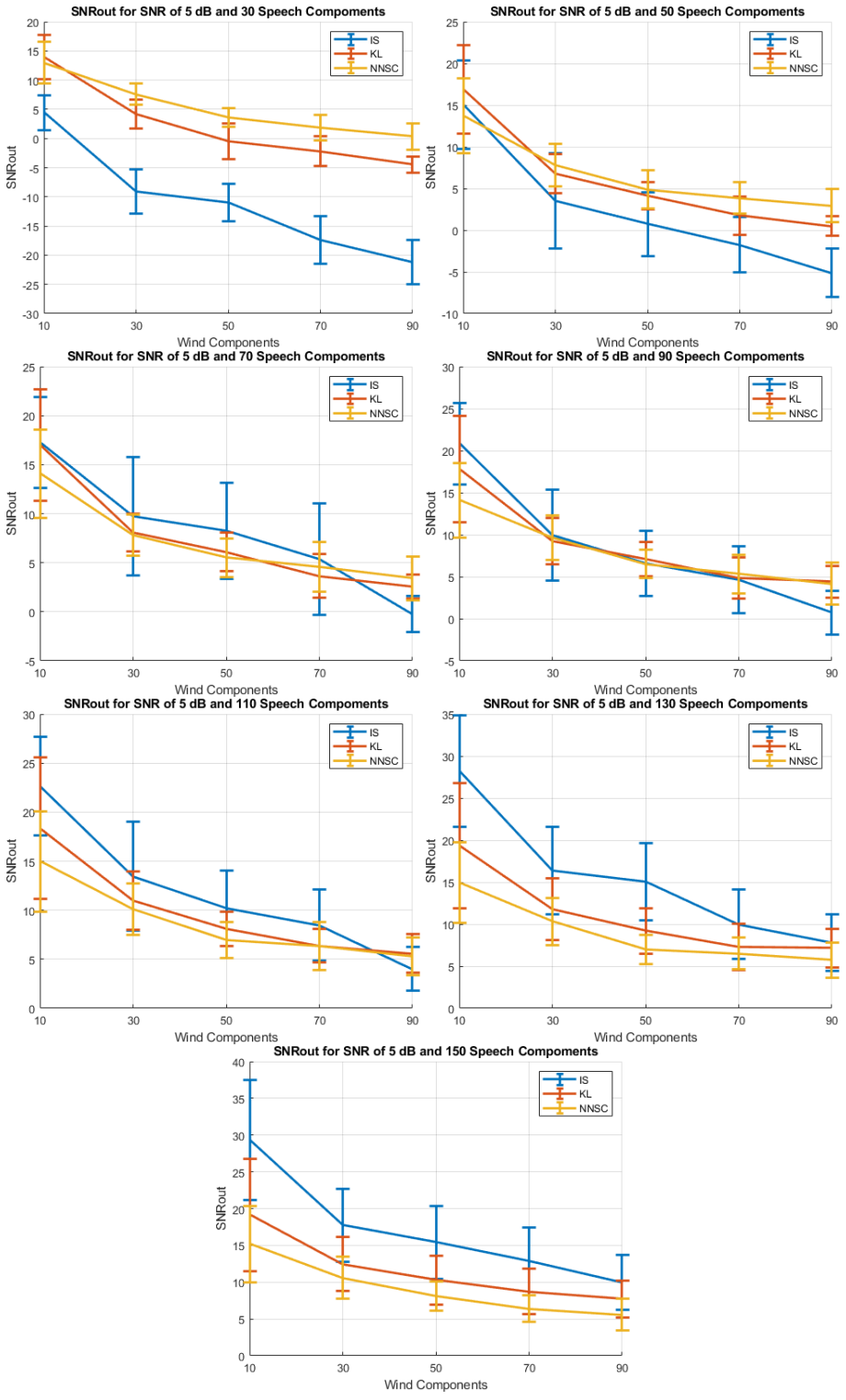

SNRout for SNR of -5 dB and 10 Wind Compoments
SNRout for SNR of -5 dB and 30 Wind Compoments
SNRout for SNR of -5 dB and 50 Wind Compoments
SNRout for SNR of -5 dB and 70 Wind Compoments
SNRout for SNR of -5 dB and 90 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components

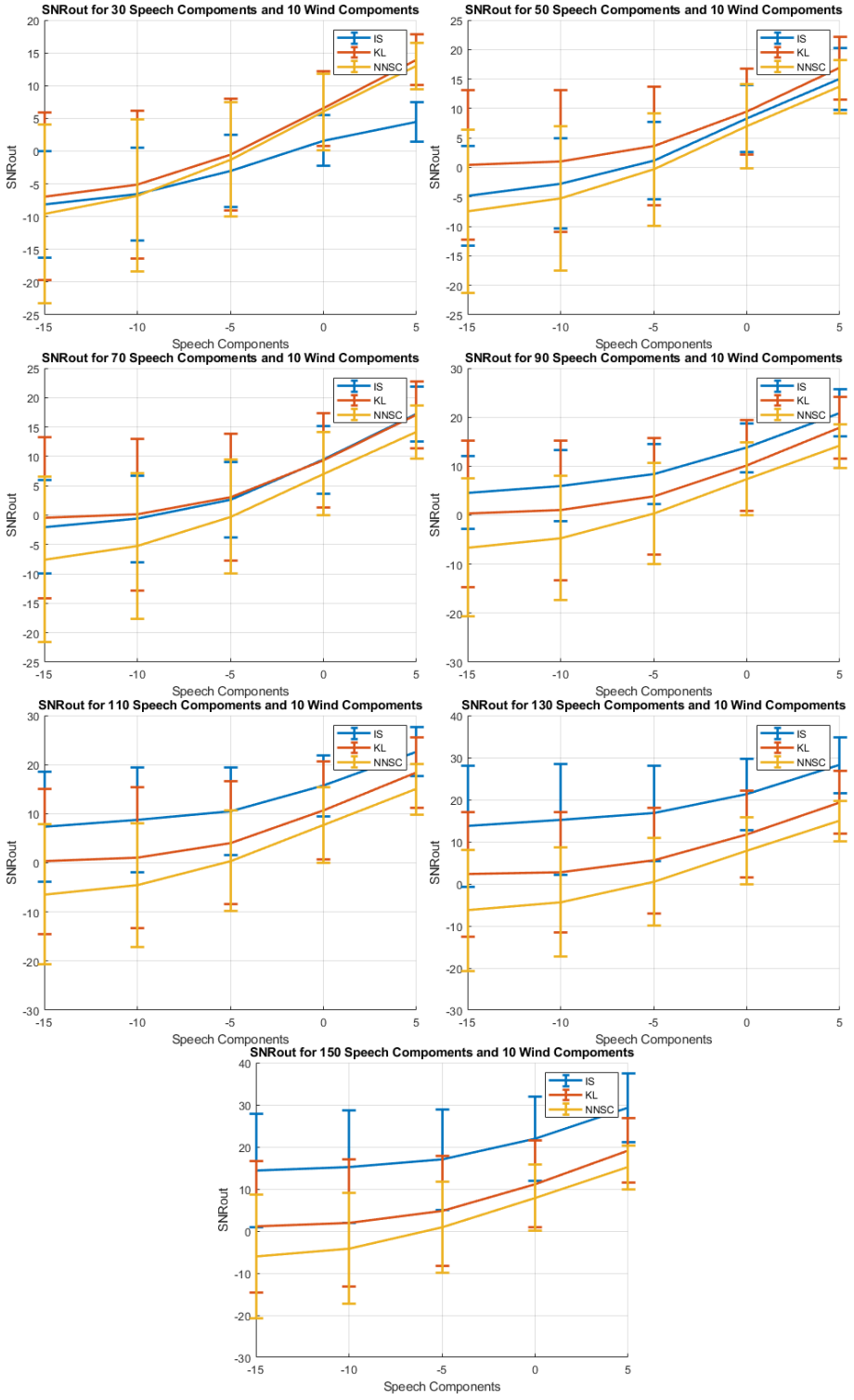

SNRout for SNR of -5 dB and 30 Speech Compoments
SNRout for SNR of -5 dB and 50 Speech Compoments
SNRout for SNR of -5 dB and 70 Speech Compoments
SNRout for SNR of -5 dB and 90 Speech Compoments
SNRout for SNR of -5 dB and 110 Speech Compoments
SNRout for SNR of -5 dB and 130 Speech Compoments
SNRout for SNR of -5 dB and 150 Speech Compoments
IS
KL
NNSC
SNRout
Wind Components

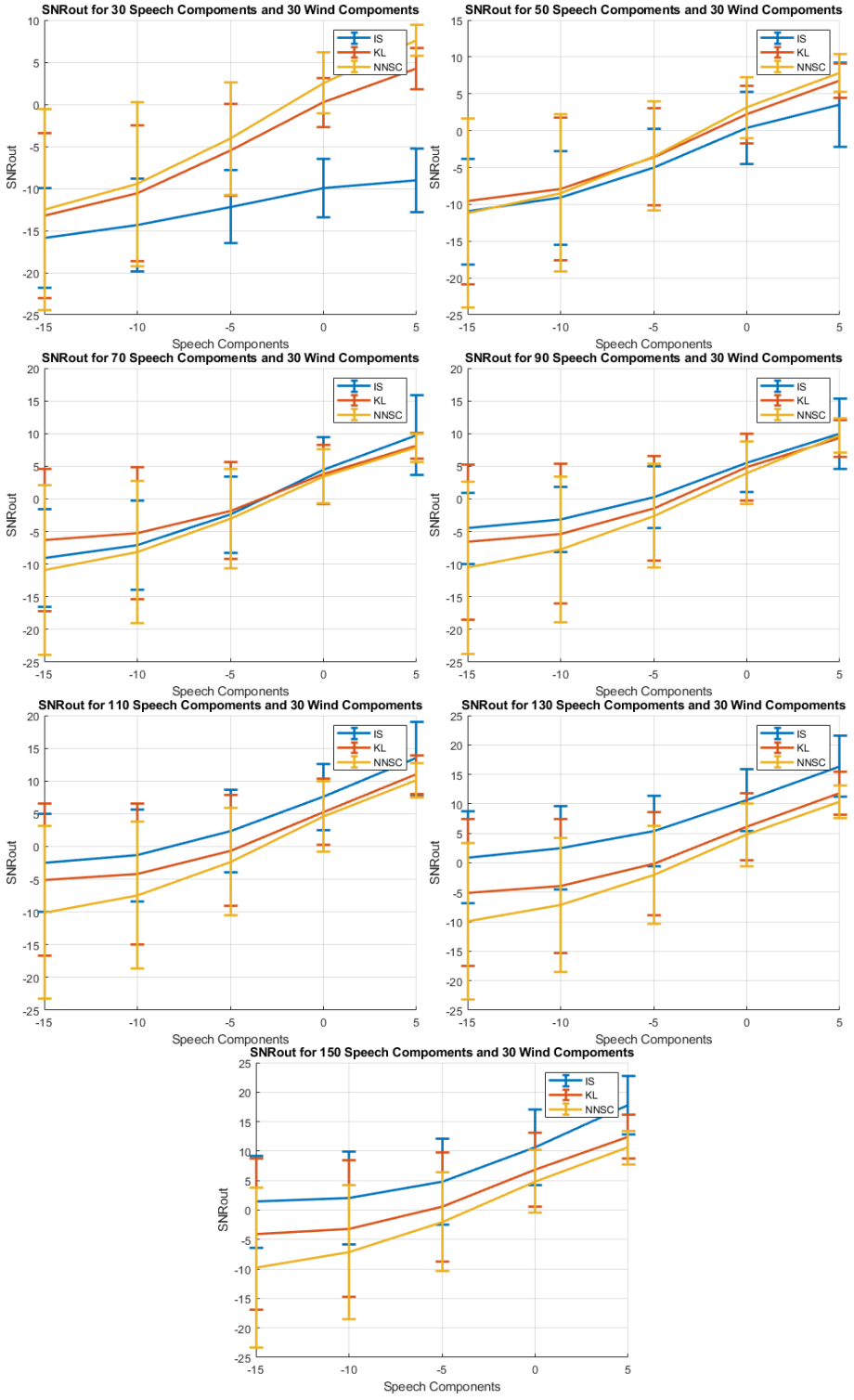

SNRout for SNR of 0 dB and 10 Wind Compoments
SNRout for SNR of 0 dB and 30 Wind Compoments
SNRout for SNR of 0 dB and 50 Wind Compoments
SNRout for SNR of 0 dB and 70 Wind Compoments
SNRout for SNR of 0 dB and 90 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components

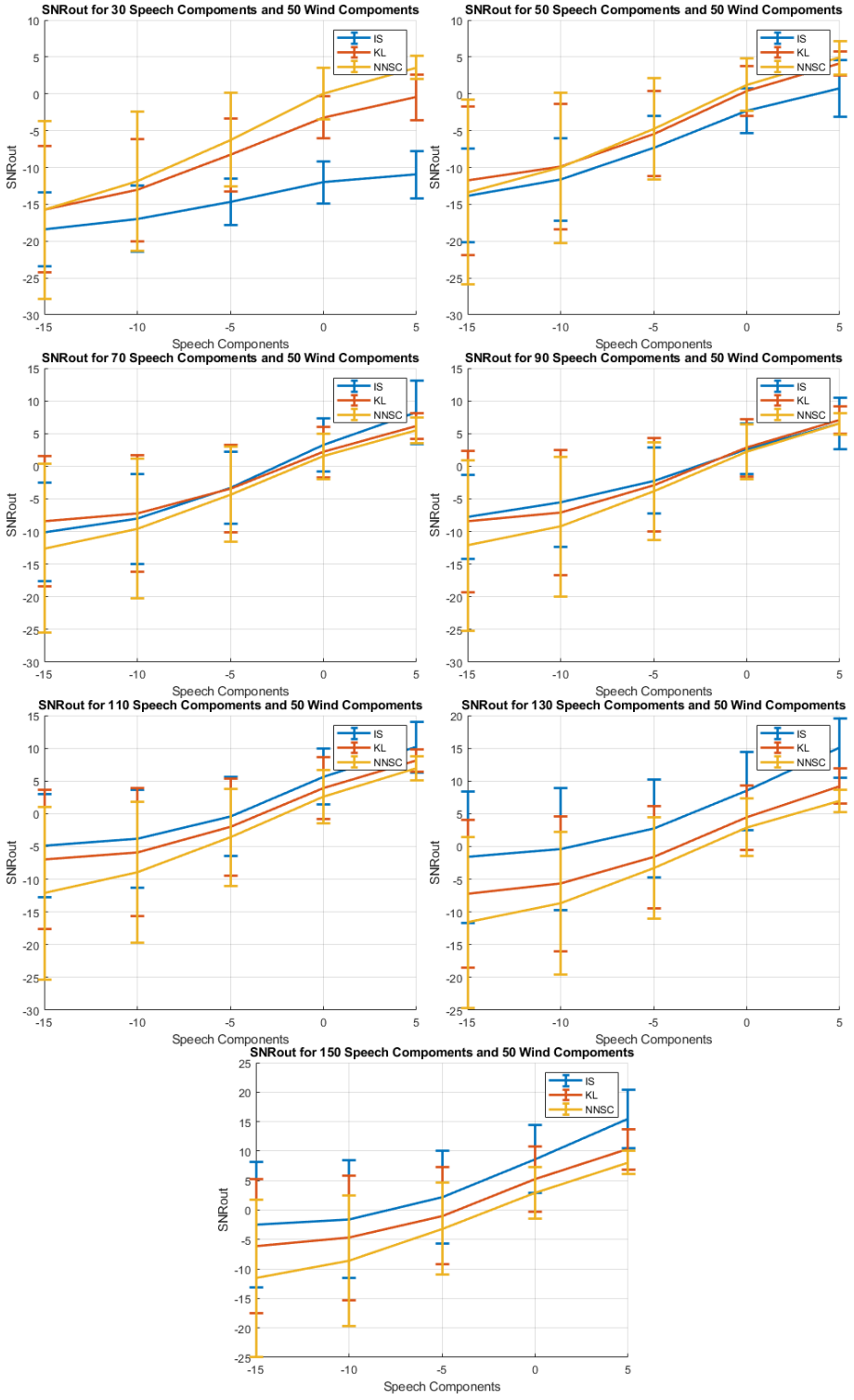

SNRout for SNR of 0 dB and 30 Speech Compoments
SNRout for SNR of 0 dB and 50 Speech Compoments
SNRout for SNR of 0 dB and 70 Speech Compoments
SNRout for SNR of 0 dB and 90 Speech Compoments
SNRout for SNR of 0 dB and 110 Speech Compoments
SNRout for SNR of 0 dB and 130 Speech Compoments
SNRout for SNR of 0 dB and 150 Speech Compoments
IS
KL
NNSC
SNRout
Wind Components

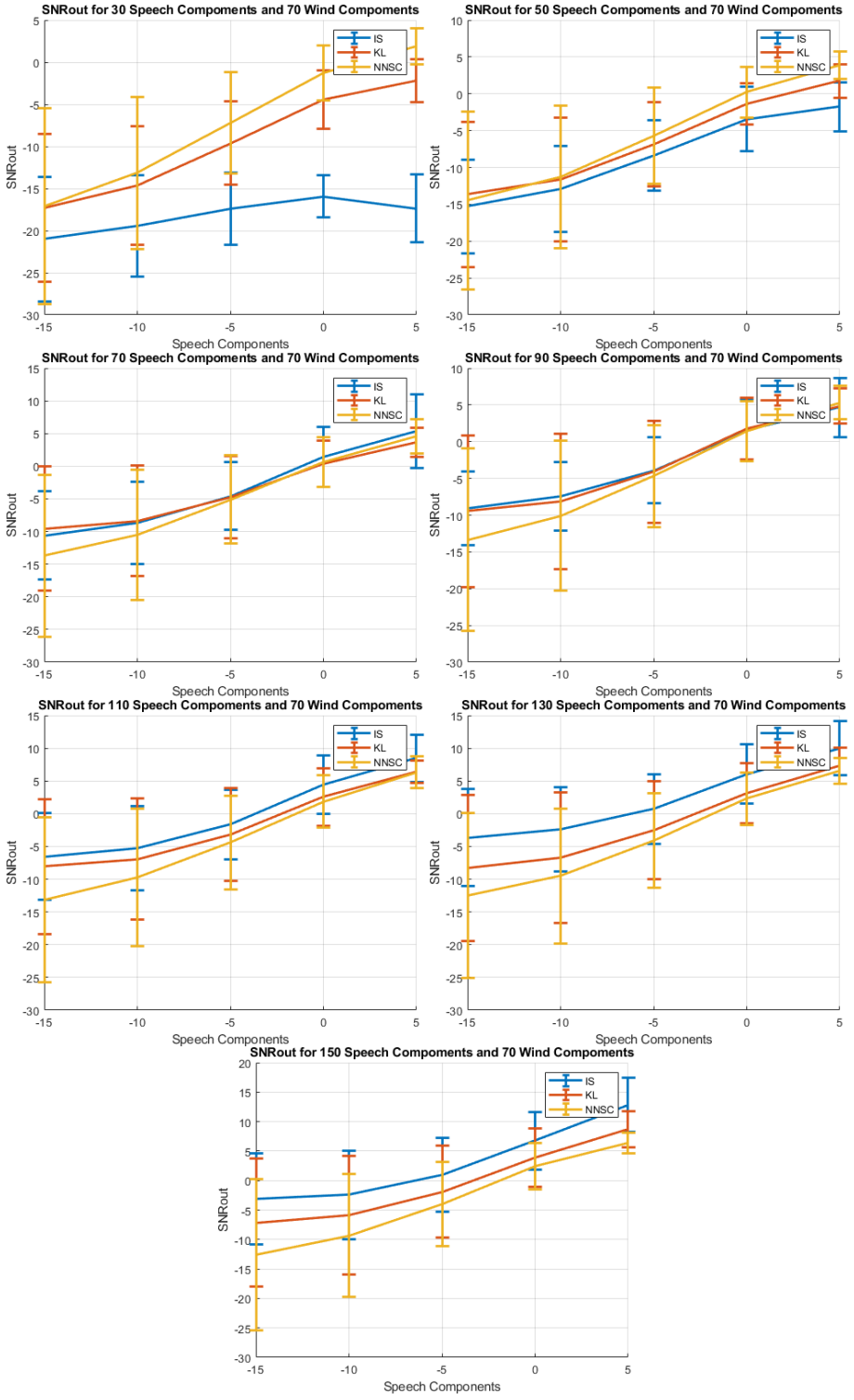

SNRout for SNR of 5 dB and 10 Wind Compoments
SNRout for SNR of 5 dB and 30 Wind Compoments
SNRout for SNR of 5 dB and 50 Wind Compoments
SNRout for SNR of 5 dB and 70 Wind Compoments
SNRout for SNR of 5 dB and 90 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components

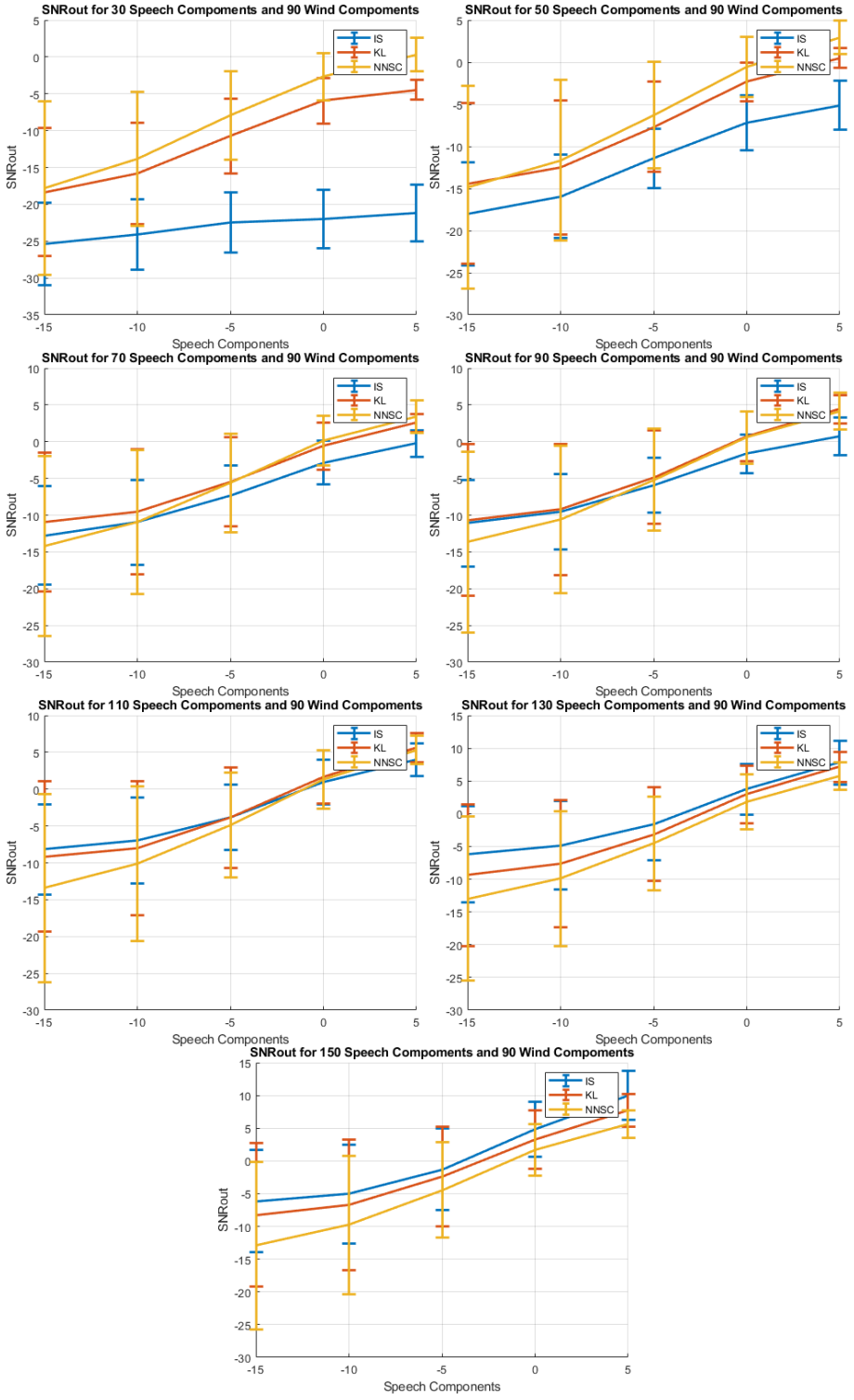

SNRout for SNR of 5 dB and 30 Speech Compoments
SNRout for SNR of 5 dB and 50 Speech Compoments
SNRout for SNR of 5 dB and 70 Speech Compoments
SNRout for SNR of 5 dB and 90 Speech Compoments
SNRout for SNR of 5 dB and 110 Speech Compoments
SNRout for SNR of 5 dB and 130 Speech Compoments
SNRout for SNR of 5 dB and 150 Speech Compoments
IS
KL
NNSC
SNRout
Wind Components


SNRout for 30 Speech Compoments and 10 Wind Compoments
SNRout for 50 Speech Compoments and 10 Wind Compoments
SNRout for 70 Speech Compoments and 10 Wind Compoments
SNRout for 90 Speech Compoments and 10 Wind Compoments
SNRout for 110 Speech Compoments and 10 Wind Compoments
SNRout for 130 Speech Compoments and 10 Wind Compoments
SNRout for 150 Speech Compoments and 10 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components


SNRout for 30 Speech Compoments and 30 Wind Compoments
SNRout for 50 Speech Compoments and 30 Wind Compoments
SNRout for 70 Speech Compoments and 30 Wind Compoments
SNRout for 90 Speech Compoments and 30 Wind Compoments
SNRout for 110 Speech Compoments and 30 Wind Compoments
SNRout for 130 Speech Compoments and 30 Wind Compoments
SNRout for 150 Speech Compoments and 30 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components


SNRout for 30 Speech Compoments and 50 Wind Compoments
SNRout for 50 Speech Compoments and 50 Wind Compoments
SNRout for 70 Speech Compoments and 50 Wind Compoments
SNRout for 90 Speech Compoments and 50 Wind Compoments
SNRout for 110 Speech Compoments and 50 Wind Compoments
SNRout for 130 Speech Compoments and 50 Wind Compoments
SNRout for 150 Speech Compoments and 50 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components


SNRout for 30 Speech Compoments and 70 Wind Compoments
SNRout for 50 Speech Compoments and 70 Wind Compoments
SNRout for 70 Speech Compoments and 70 Wind Compoments
SNRout for 90 Speech Compoments and 70 Wind Compoments
SNRout for 110 Speech Compoments and 70 Wind Compoments
SNRout for 130 Speech Compoments and 70 Wind Compoments
SNRout for 150 Speech Compoments and 70 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components


SNRout for 30 Speech Compoments and 90 Wind Compoments
SNRout for 50 Speech Compoments and 90 Wind Compoments
SNRout for 70 Speech Compoments and 90 Wind Compoments
SNRout for 90 Speech Compoments and 90 Wind Compoments
SNRout for 110 Speech Compoments and 90 Wind Compoments
SNRout for 130 Speech Compoments and 90 Wind Compoments
SNRout for 150 Speech Compoments and 90 Wind Compoments
IS
KL
NNSC
SNRout
Speech Components