DEPARTMENT OF COMPUTER SCIENCE UNIVERSITY OF CRETE

Speech Signal Processing Laboratory

Thesis Report

System for Bandwidth Extension of Narrow-band Speech

Advisors:	Dr. George Kafentzis	kafentz@csd.uoc.gr
	Professor Yannis Stylianou	yannis@csd.uoc.gr
Student:	Georgios Manos	csd4333@csd.uoc.gr

HERAKLION, JUNE 2023

Contents

List of Figures

List of Tables

1 Introduction

In this work, we explore the system[1] invented by David Malah and Richard Vandervoot Cox for extending the bandwidth of a signal. The goal of this work is to implement a version of the system as suggested by the patent in a modular way such that one can modify its components and the system's configuration, while also assessing its performance both qualitative and quantitative. The code is written in Python and is available on GitHub. Figure 1.2 and 1.1 were taken directly from the patent, while some equations and implementation ideas were taken from CS578 Speech Processing course.

1.1 Short System Description

This is a system for extending the bandwidth of a narrowband signal such as a speech signal. The method applies a parametric approach to bandwidth extension but does not require training. The parametric representation relates to a discrete acoustic tube model (DATM). The method comprises computing narrowband linear predictive coefficients (LPCs) from a received narrowband speech signal, computing narrowband partial correlation coefficients (parcors) using recursion, computing M_{nb} area coefficients from the partial correlation coefficient, and extracting M_{wb} area coefficients using interpolation. Wideband parcors are computed from the M_{wb} area coefficients and wideband LPCs are computed from the wideband parcors. The method further comprises synthesizing a wideband signal using the wideband LPCs and a wideband excitation signal, highpass filtering the synthesized wideband signal, and combining the highband signal with the original narrowband signal to generate a wideband signal. As mentioned by the inventors, the preferred variation of the invention is implemented here, where the M_{nb} area coefficients are converted to log-area coefficients for the purpose of extracting, through shifted interpolation, M_{wb} log-area coefficients. The M_{wb} log-area coefficients are then converted to M_{wb} area coefficients before generating the wideband parcors.

The system is described in Figure 1.1, and explained in depth in Section 2

[1] You can find the system patent here

FIG. 8

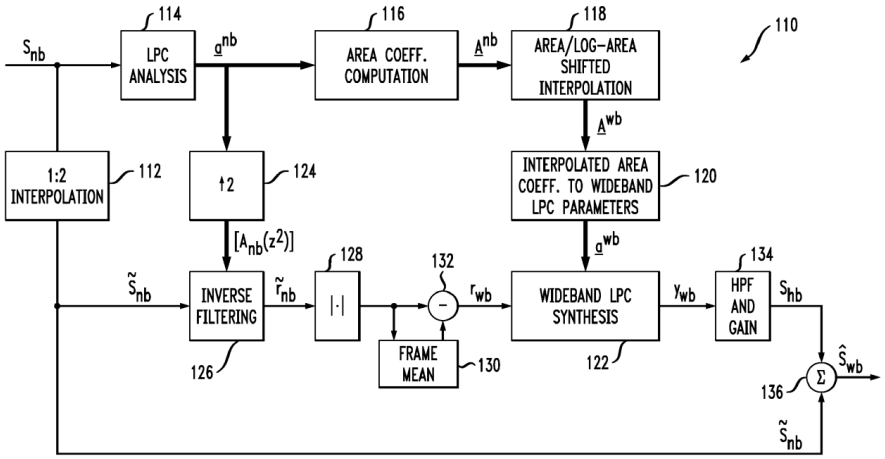


Figure 1.1: System description

1.2 Experiments Setup

The present implementation of the patent follows the flow diagram as shown in Figure 1.2. The bandwidth extension is performed on a frame-by-frame basis. The inventors mention that some of the parameter values discussed are merely default values used in simulation. In this work, we mostly follow their suggested default values, while in section 3, we also discuss the effect of some of those parameters.

Finally, this work uses the student's voice data to perform experiments on as they were recorded on a controlled environment. We will present the results of various recordings under different settings, all of which were originally recorded using Audacity on mono audio settings and 16khz sampling rate. Also, using librosa, we are downsampling the speech signal to 8khz and applying 2x bandwidth expansion (back to 16khz).

The results are evaluated in 3 ways; aurally, visually through waveforms and spectrograms, and using 3 metrics, Mean Squared Error (MSE), Spectral Convergence (SC), L1-norm Mel Spectrogram difference.

FIG. 11

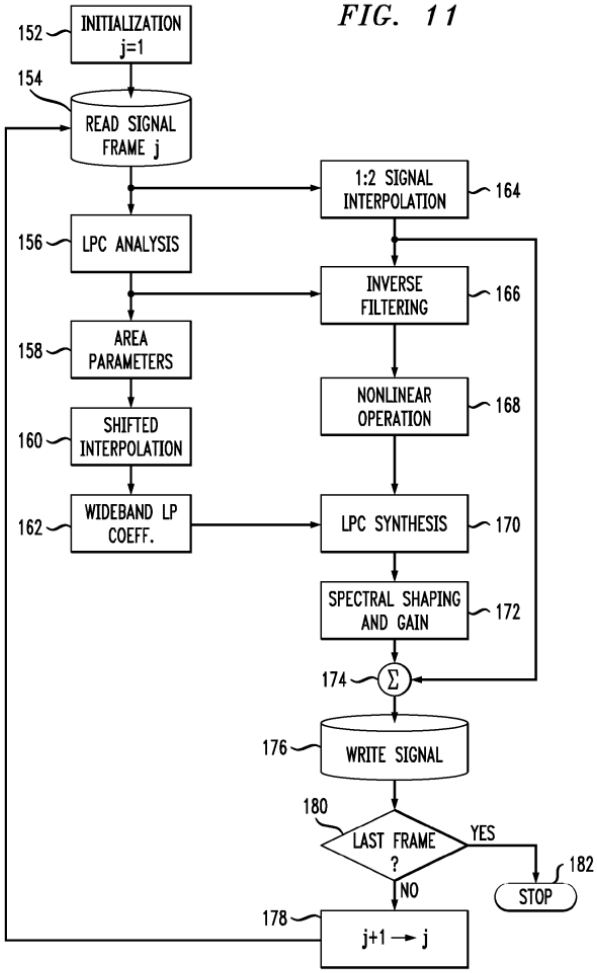


Figure 1.2: Algorithm flowchart

2 Implementation Details

2.1 Parameters

During the initialization (152), the following parameters are established: Input signal frame length = N (256), Frame update step = N / 2, Number of narrowband DATM sections M (8), Sampling Frequency (in Hz)=f_s^{nb} (8000), Input signal upper cutoff frequency in Hz=F_c (3900, as the input was a microphone; inventors suggest using 3600 for MIRS input and 3400 for IRS telephone speech) and R(0) modification parameter δ (0.01). The inventors mention that the values may vary depending on the source characteristics and application.

2.2 Signal Analysis

A speech signal in wav format is read from disk. For frame j, the signal undergoas an LPC analysis (156) that comprises of the following steps: computing a correlation coefficient ρ_1, pre-emphasizing the input signal frame using $(1 - \rho_1 z^{-1}$, windowing of the pre-emphasized signal using a Hann window of length N (256), computing the M+1 autocorrelation coefficients R(0), R(1), ..., R(M), modifying R(0) by a factor $(1 + \delta)$, and applying the Levinson-Durbin recursion[2] to find LP coefficients $\underline{\alpha}^{nb}$ and reflection coeffcients $\underline{k}^{nb}$. Parcors are then computed through reflection coefficients using the relationship $\underline{k}^{nb} = -\underline{r}^{nb}$. Finally, the gain G is computed using equation 2.1

$$G = \sqrt{\sum_{i}^{M+1} \underline{\alpha}_i^{nb} R(i)} \tag{2.1}$$

2.3 Computing Area Parameters

Next, the area parameters are computed (158) according to an important aspect of the invention. Computation of these parameters comprises computing M area coefficients via equation 2.3 that describes the parameters of the discrete acoustic tube model (DATM), and computing M log-area coefficients. The relationship between the LP model parameters and the area parameters of the DATM are given by the backward recursion of equation 2.3, where A_1 corresponds to the cross-section at the lips and $A_{M^{nb+1}}$ corresponds to the cross-section at the glottis opening. Computing the M log-area coefficients is mentioned to be an optional step but a preferred one and thus followed in this implementation. The computed

[2]A thank you to Dr. George Kafentzis for providing a python implementation of the algorithm

area or log-area coefficients are shift-interpolated (160) by a desired factor with a proper sample shift. Cubic spline is applied by default as interpolation method.

$$r_i^{wb} = \frac{A_i^{wb} - A_{i+1}^{Wb}}{A_i^{wb} + A_{i+1}^{Wb}},\ i = 1, 2, ..., M_{wb} \tag{2.2}$$

The next step relates to calculating wideband LP coefficients (162) and comprises computing the wideband parcors from interpolated area coefficients via equation 2.2 and computing wideband LP coefficients, α^{wb}, by applying the Step-Down Recursion to the wideband parcors. As log-area coefficients were used, exponentiation is applied to obtain the interpolated area coefficients.

$$A_i = \frac{1 + r_i}{1 - r_i} A_{i+1};\ i = M_{nb}, M_{nb} - 1, ..., 1 \tag{2.3}$$

2.4 Producing Wideband Signal

Returning now to the branch from the output of step 154, step 164 relates to signal interpolation. Step 164 comprises interpolating the narrowband input signal, S_{nb}, by a factor (i.e. 2, upsampling and lowpass filtering). This step results in a narrowband interpolated signal $\tilde{S_{nb}}$. The signal $\tilde{S_{nb}}$ is inverse filtered (166) using, for example, a transfer function of $A_{nb}(z^2)$ having the coefficients shown in equation 2.4, resulting in a narrow band residual signal $\tilde{r_{nb}}$ sampled at the interpolated-signal rate.

$$\underline{\alpha}^{nb} \uparrow 2 = \{1, 0, \alpha_1^{nb}, 0, \alpha_2^{nb}, 0, ..., \alpha_{M^{nb}-1}^{nb}, 0, \alpha_{M^{nb}}^{nb}\} \tag{2.4}$$

Next, a non-linear operation is applied to the signal output from the inverse filter. The operation comprises fullwave rectification (absolute value) of residual signal $\tilde{r}_{nb}$ (168). The inventors mention that other nonlinear operators may also optionally be applied. Also, other potential elements associated with step 168 may comprise computing frame mean and subtracting it from the rectified signal (as shown in Fig 1.1), generating a zero-mean wideband excitation signal r_{wb}; optional compensation of spectral tilt due to signal rectification via LPC analysis of the rectified signal and inverse filtering. The preferred setting here is no spectral tilt compensation.

Next, the highband signal must be generated before being added (174) to the original narrowband signal. This step comprises exciting a wideband LPC synthesis filter (170) (with coefficients $\underline{\alpha}^{wb}$ by the generated wideband excitation signal r_{wb}, resulting in a wideband signal y_{wb}. Fixed or adaptive de-emphasis are optional, but the default and preferred setting is no de-emphasis. The resulting wideband signal y_{wb} may be used as

the output signal or may undergo further processing. In this implementation, the further processing involves the following steps: the wideband signal y_{wb} is highpass filtered (172) using a butterworth digital filter of order 16 with cutoff frequency $f_c = 3900hz$ to generate a highband signal and the gain is also applied here. The inventors mention using a fixed gain value (e.g. 2) here instead or adaptive gain matching, but we used the previously computed gain value. The resulting signal is S_{hb} (as shown in Fig 1.1). The butterworth filter was applied using *scipy.signal.filtfilt* as it is zero-phase filtering, which doesn't shift the signal as it filters.

The Buttworth filter is presented in Figure 2.1.

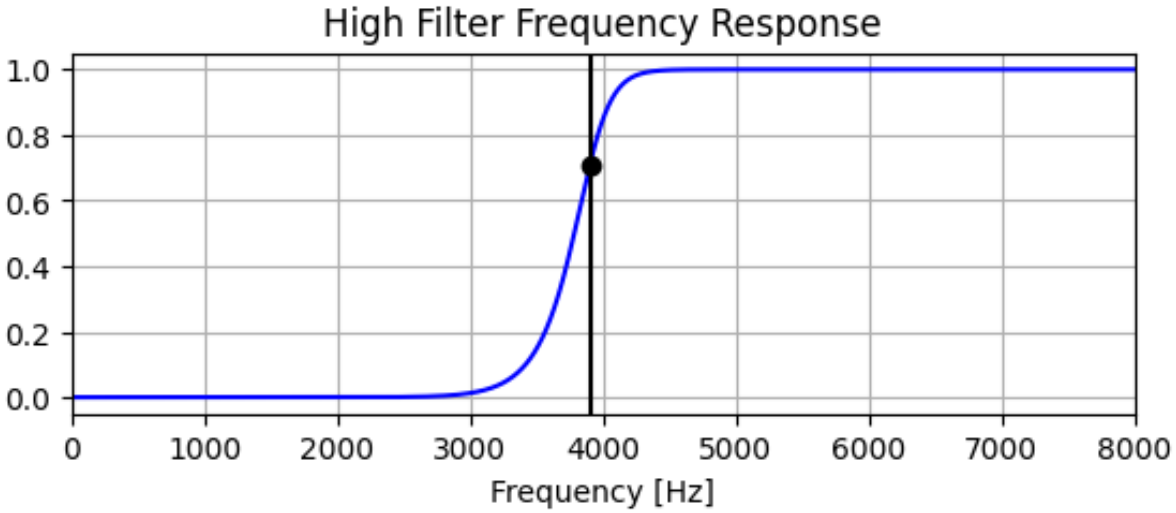


Figure 2.1: Butterworth Filter Frequency Response

2.5 Signal Synthesis

Finally, the output wideband signal is generated. This step comprises generating the output wideband speech signal by summing (174) the generated highband signal S_{hb} with the narrowband interpolated input signal $\tilde{S}_{nb}$. The resulting summed signal is stored in the main buffer. The output signal frame (of 2N samples) can either be overlap-added (with a half-frame shift of N samples) to a signal buffer, or because $\tilde{S}_{nb}$ is an interpolated original signal, the center half-frame (N samples out of 2N) is extracted and concatenated with previous output stored in a signal buffer. The authors mention using the latter simplier option and thus is used in this implementation as well.

3 Experiments

In this section, the system's results are presented for speech file *sample2.wav*. All speech files were originally sampled at 16khz, and were downsampled to 8khz prior to the system's input. The system's output is then compared to the narrowband signal (8khz), the original wideband signal (16khz) - which will be considered as ground truth, as well as a signal which was simply upsampled to 16khz without bandwidth expansion.

This report also includes results for the rest of the speechfiles in the database, presented in section 5.

3.1 Time Domain Waveforms

The signals' waveforms are presented in Figure 3.1. It includes the waveforms for the signal downsampled to 8khz *"Orig (NB)"*, the original signal at 16khz *"Orig (WB)"*, the signal that was upsampled using High-quality FFT-based bandlimited interpolation *"Interpolated"* and finally the bandwidth expansion system output *"Reconstructed"*.

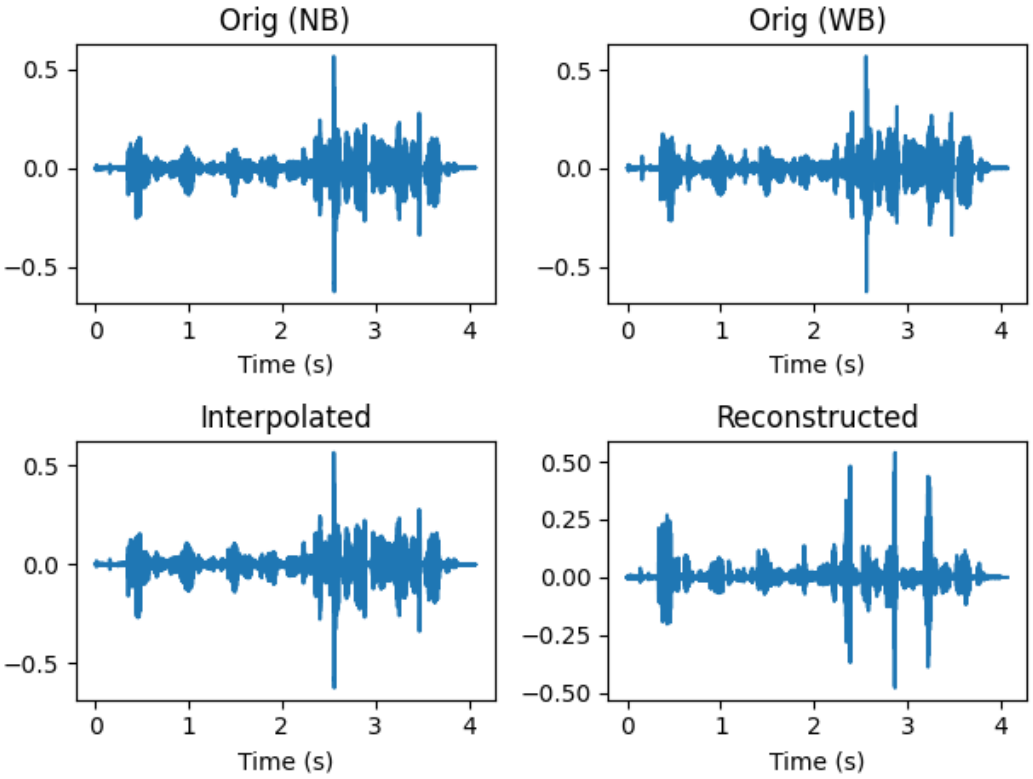


Figure 3.1: Signal Waveforms

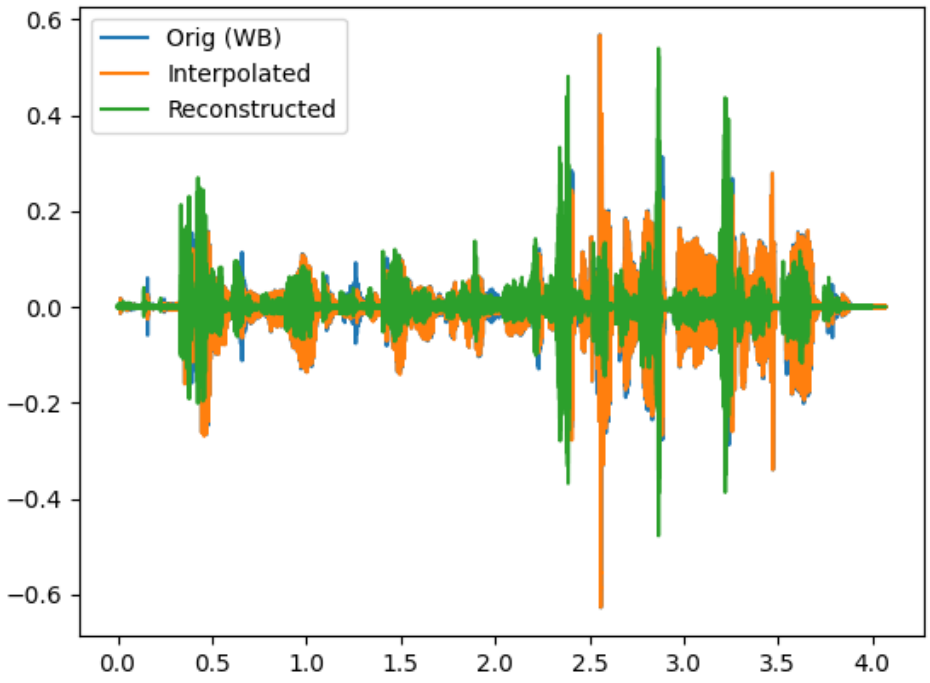


Figure 3.2: Signal Waveforms Merged

We can see that while the first 3 look very much alike, the reconstructed signal appears to be quite different. The energy of the signal after time 2.5s appears to be missing, while the main difference between the other 3 is the peak height at specific points.

Aurally, as per to the author's subjective opinion, the narrowband signal has poor resolution, resulting to a telephone sounding like speech, with small-but-existing impact to the speech intelligibility. The wideband original signal clearly has a higher resolution. One may also observe that aural difference and higher resolution on the SBE signal; however, the voice sounds distorted. The interpolated signal lacks that higher resolution, but the voice is not distorted and the result sounds better than the narrowband signal, yet far from the actual wideband one.

3.2 Metrics

The signals are also compared in 3 different metrics against the ground truth (i.e. 16khz original sampled signal); Mean Squared Error (MSE), Spectral Convergence (SC), and L1-Norm of Mel Spectrogram difference. The results are presented in Table 3.1.

Each metric is presented twice; once comparing the interpolated signal (*Interpolated*) and the second time the bandwidth expanded (*SBE*) signal to the Wideband original signal.

Mean Squared Error quantifies the average difference between 2 signals (i.e. the mean value of the error signal). The mathematical formulation is given in Equation 3.1. The results on the table shows that on every case, the mean squared error of the simply interpolated signal is lower than the bandwidth expanded one, and thus on average, the interpolated signal is closer to the ground truth in the time domain.

$$MSE = \frac{1}{N}\sum_{i=0}^{N-1}(s[i] - \hat{s}[i]])^2 \tag{3.1}$$

Spectral Convergence emphasizes spectral peaks and other spectral components when comparing 2 signals. For this purpose, it became popular as a Loss Function when training Neural Networks on speech signals. The SC function is defined in Equation 3.2 where STFT is the discrete Short Time Fourier Transform of the ground truth speech signal, and $\|A\|_F$ denotes the *Frobenius* norm of a matrix A: $\|A\|_F = \sqrt{\sum_{i=0}^{N-1}\sum_{j=0}^{K-1}|\alpha_{i,j}|^2}$. Once again, the SBE signal has significantly higher score in this metric, thus concluding that the spectral components of the interpolated signal are more similar to the ground truth, compared to the SBE signal.

$$L_{sc}(s[n], \hat{s}[n]) = \frac{\|\,|STFT(s[n])| - |STFT(\hat{s}[n])|\,\|_F}{\|\,|STFT(s[n])|\,\|_F} \tag{3.2}$$

Humans perceive sound in a logarithmic scale rather than a linear scale. The Mel Scale was developed to take this into account by conducting experiments with a large number of listeners. It is a scale of pitches, such that each unit is judged by listeners to be equal in pitch distance from the next.

A Mel Spectrogram makes two important changes relative to a regular Spectrogram that plots Frequency vs Time; It uses the Mel Scale instead of Frequency on the y-axis and it uses the Decibel Scale instead of Amplitude to indicate colors. Therefore, L1-Norm of the difference of 2 Mel Spectrogram quantifies the distance on Mel Scale between the respective 2 signals. Once again, the observations on the table show that the interpolated

Mel difference is significantly lower to the SBE signal.

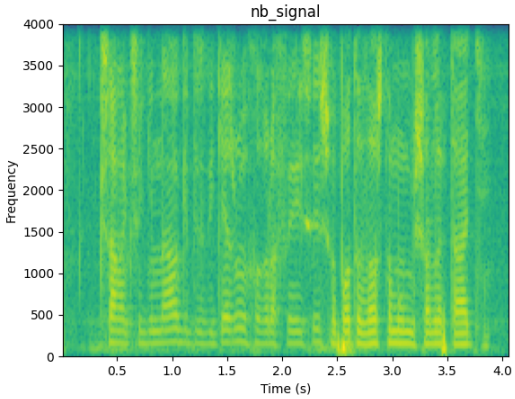
Filename	MSE Interpolated	MSE SBE	SC Interpolated	SC SBE	MEL Diff Interpolated	MEL Diff SBE
sample2	**0.0000583**	0.002903	**0.161815**	0.988587	**8.117**	1706.454
sample3	**0.0005730**	0.018565	**0.189323**	0.966934	**72.457**	11655.512
sample5	**0.0004094**	0.008246	**0.271466**	0.840376	**38.346**	827.613
sample6	**0.0000690**	0.005829	**0.129014**	0.881263	**15.192**	589.839
stars_16k	**0.0000093**	0.000156	**0.303956**	0.963086	**1.288**	25.69

Table 3.1: Metrics Table over all Speech Signals

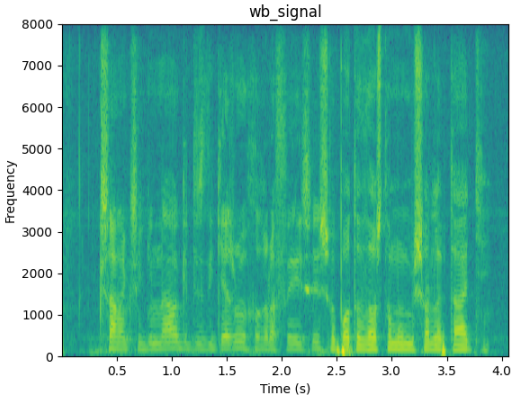
3.3 Spectrograms

The spectrograms of the 4 different signal types are presented in Figure 3.3. As we can see, the narrowband signal has no spectral information above 4khz (i.e. $0.5 \cdot F_s$, where $F_s = 8$khz) whatsoever, while the wideband signal goes up to 8khz. Respectively, while the interpolated signal has a sampling rate $F_s = 16$khz, we see no significant information on the highband zone ($4 - 8$khz) while the signal mostly lives on the lowband zone ($0 - 4$khz). On the Bandwidth Expanded signal, the highband zone has much more spectral information.

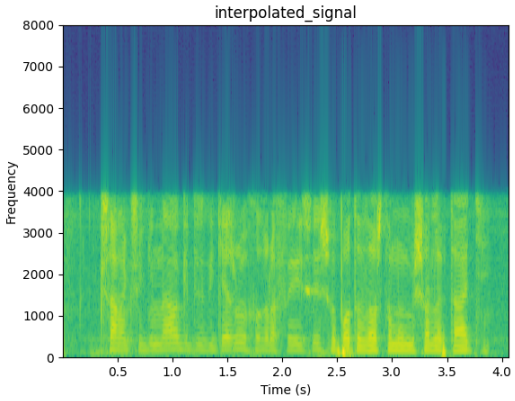
The spectrograms produced are a result of a narrowband analysis, as a long window was used (30ms). Narrowband spectrogram gives good frequency resolution as the harmonics are effectively resolved (horizontal striations on the spectrogram). However, it also gives poor time resolution, because the long analysis window covers several pitch periods and thus is unable to reveal fine periodicity changes over time. Since this work focuses on the frequency resolution of a signal when applying the artifical bandwidth expansion method, narrowband analysis was preferred. It should be noted that colors in spectrograms have a meaning; intense yellow color corresponds to high magnitude values (high energy), whereas green or blue color correspond to lower magnitude values (low energy).



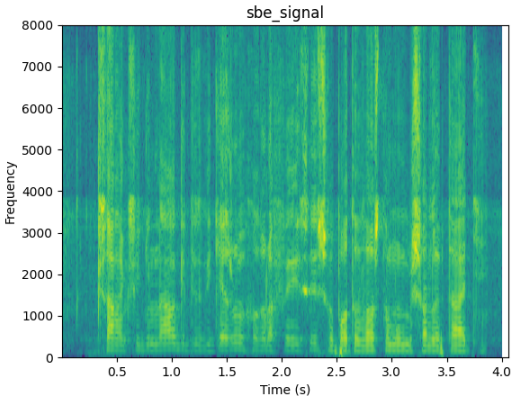
(a) Narrowband Signal Spectrogram



(b) Wideband Signal Spectrogram



(c) Interpolated Signal Spectrogram



(d) SBE Signal Spectrogram

Figure 3.3: Narrowband vs Wideband vs Interpolated vs SBE

4 Conclusion and Future Work

This invention focuses on a novel bandwidth extension approach in the category of parametric methods that do not require training. Almost all parametric techniques use an LPC synthesis filter for wideband signal generation (typically an intermediate wideband signal which is further highpass filtered), by exciting it with an appropriate wideband excitation signal.

In Section 3.2, we compared the resulting signals using various metrics. In Section 3.3, we compared the signals visually through their spectrograms. The metrics showed that the interpolated signal is closer to the wideband signal than the SBE; however, the SBE signal - regardless of the voice distortion - appears to have the wideband resolution, as well as better frequency information on the highband (4-8khz), compared to the interpolated signal.

As a future work, the most important part is to find out where that voice distortion comes from. This may be a result either of redesigning the system's components (i.e. better implementation of specific algorithms) or changing the system's configuration and parameters (i.e. switch to wideband analysis, change cutoff frequency, LPC order).

Also, we only applied a bandwidth expansion of a factor 2 (8khz to 16khz). The system should further be tested for other scales (e.g. 16khz to 32khz, 16khz to 44khz, 8khz to 32khz etc) and compare the results on the same basis.

Reported bandwidth extension methods can be classified into two types - parametric and non-parametric. Non-parametric methods usually convert directly the received narrowband speech signal into a wideband signal using simple techniques like spectral folding and non-linear processing. This invention aims to find a relationship between the narrowband and wideband speech parameters through a parametric method that doesn't require training. Other approaches may even be tried for that goal, such as neural-net-based methods and statistical methods.The inventors mention that the main advantages of a non-parametric approach are its relatively low complexity and its robustness, stemming from the fact that no model needs to be defined and, consequently, no parameters need to be extracted and no training is needed. These characteristics, however, typically result in lower quality when compared with parametric methods.

5 Other Results

5.1 sample3

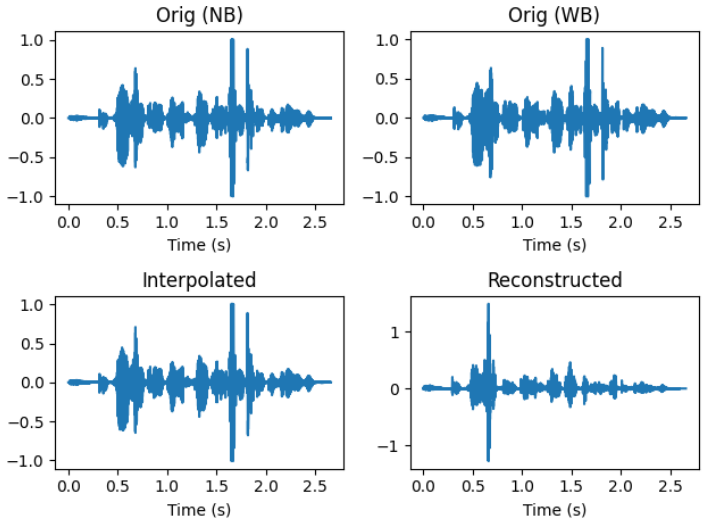


Figure 5.1: sample3 Signal Waveforms

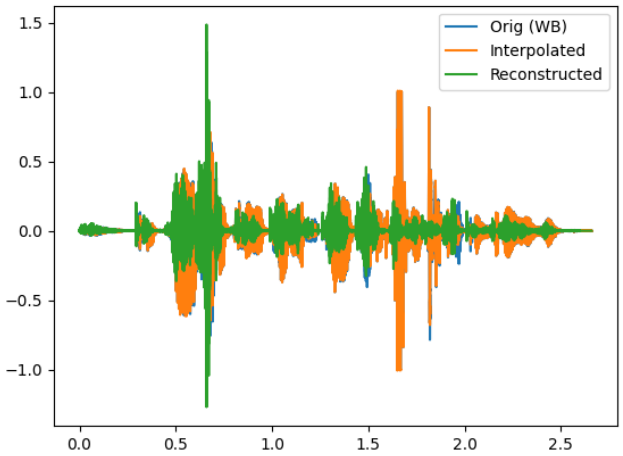
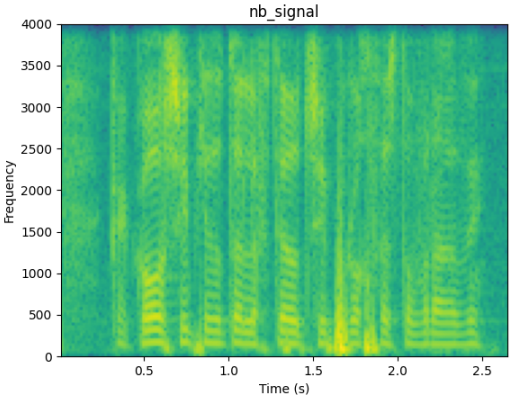
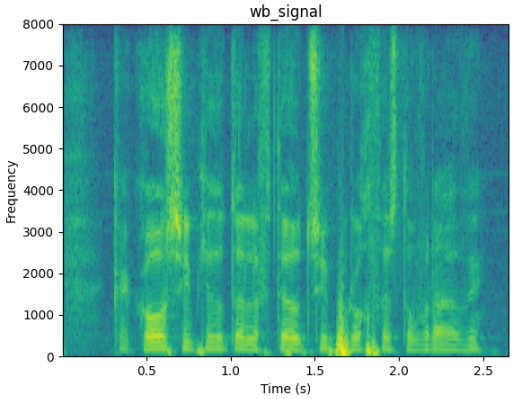


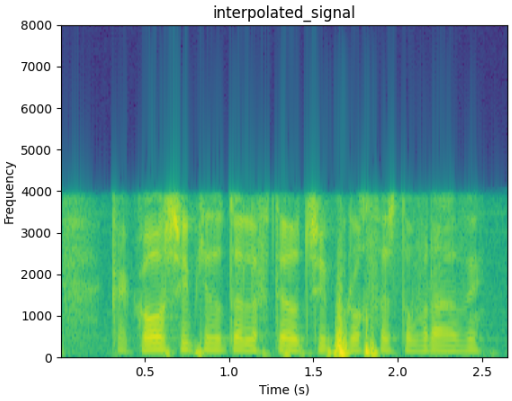
Figure 5.2: sample3 Signal Waveforms Merged



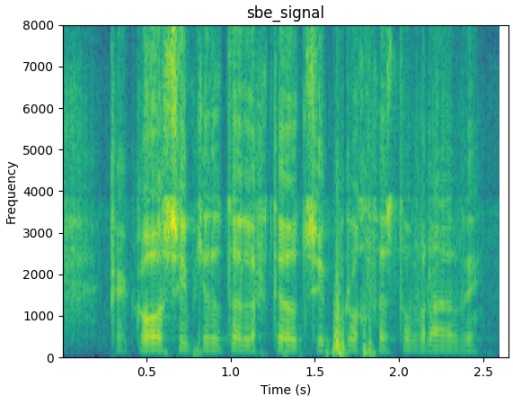
(a) sample3 Narrowband Signal Spectrogram



(b) sample3 Wideband Signal Spectrogram



(c) sample3 Interpolated Signal Spectrogram



(d) sample3 SBE Signal Spectrogram

Figure 5.3: sample3 Narrowband vs Wideband vs Interpolated vs SBE

5.2 sample5

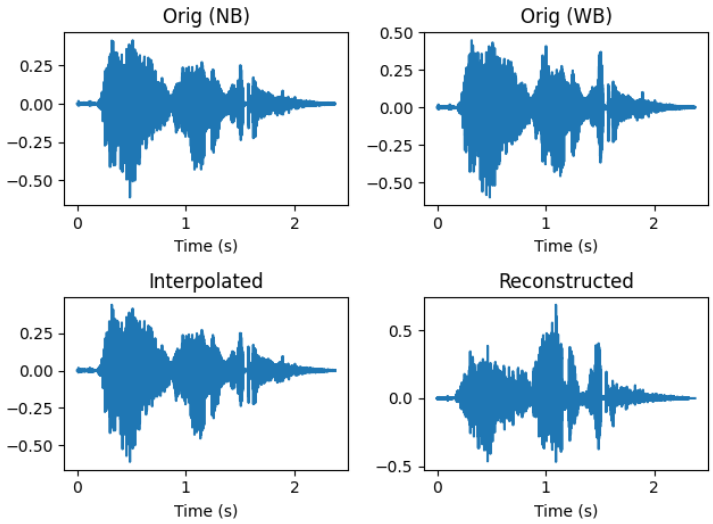


Figure 5.4: sample5 Signal Waveforms

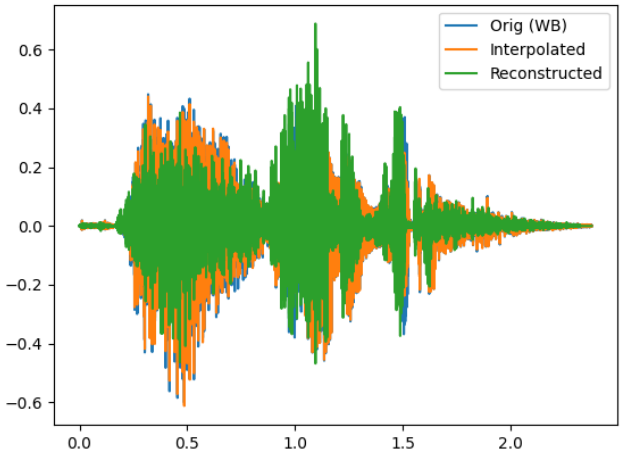
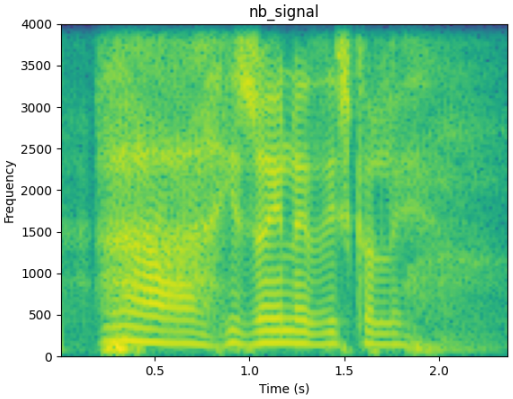
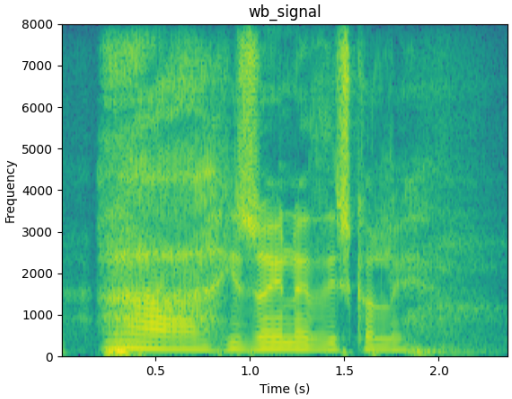


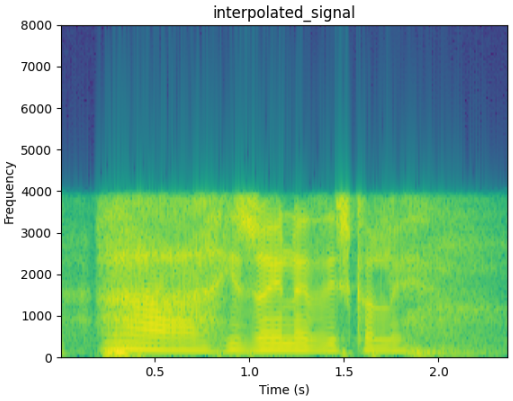
Figure 5.5: sample5 Signal Waveforms Merged



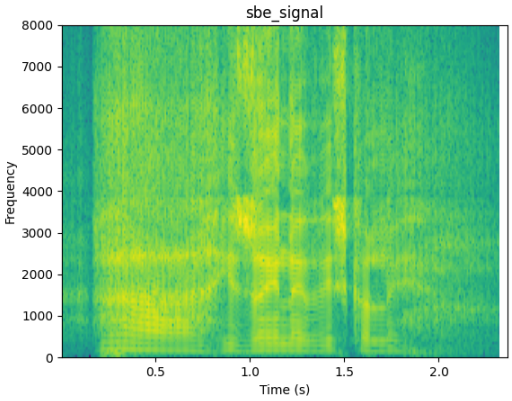
(a) sample5 Narrowband Signal Spectrogram



(b) sample5 Wideband Signal Spectrogram



(c) sample5 Interpolated Signal Spectrogram



(d) sample5 SBE Signal Spectrogram

Figure 5.6: sample5 Narrowband vs Wideband vs Interpolated vs SBE

5.3 sample6

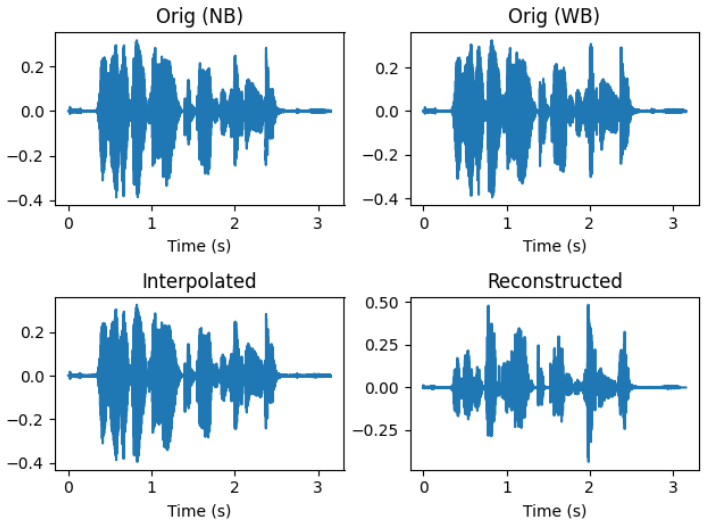


Figure 5.7: sample6 Signal Waveforms

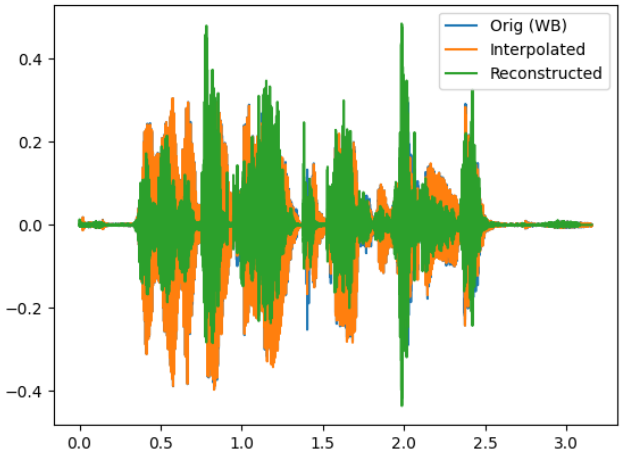
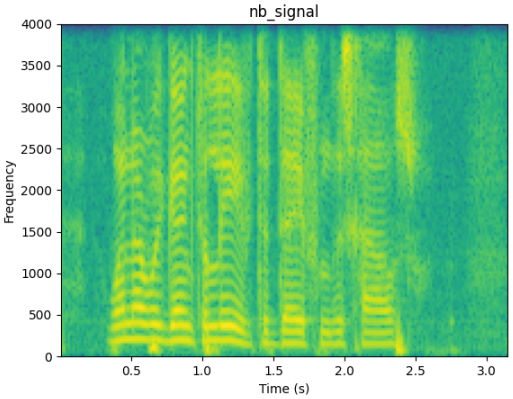
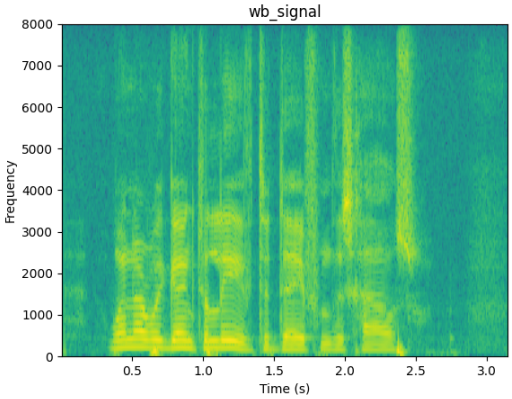


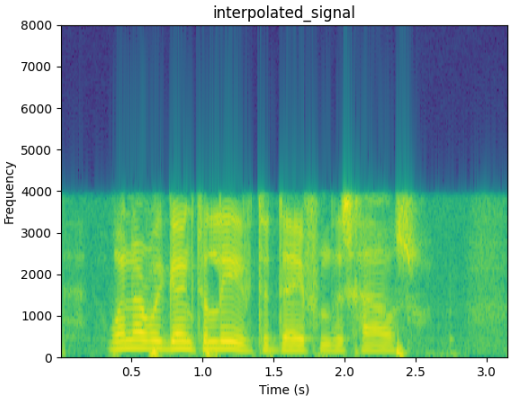
Figure 5.8: sample6 Signal Waveforms Merged



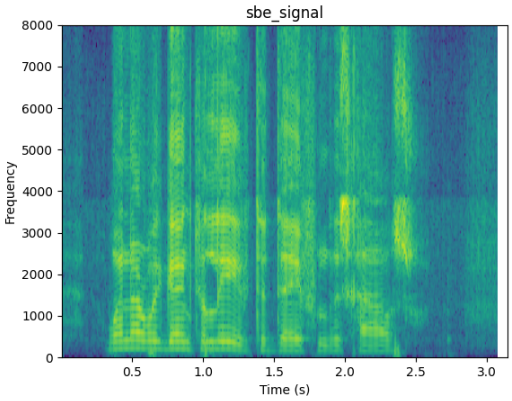
(a) sample6 Narrowband Signal Spectrogram



(b) sample6 Wideband Signal Spectrogram



(c) sample6 Interpolated Signal Spectrogram



(d) sample6 SBE Signal Spectrogram

Figure 5.9: sample6 Narrowband vs Wideband vs Interpolated vs SBE

5.4 stars_16k

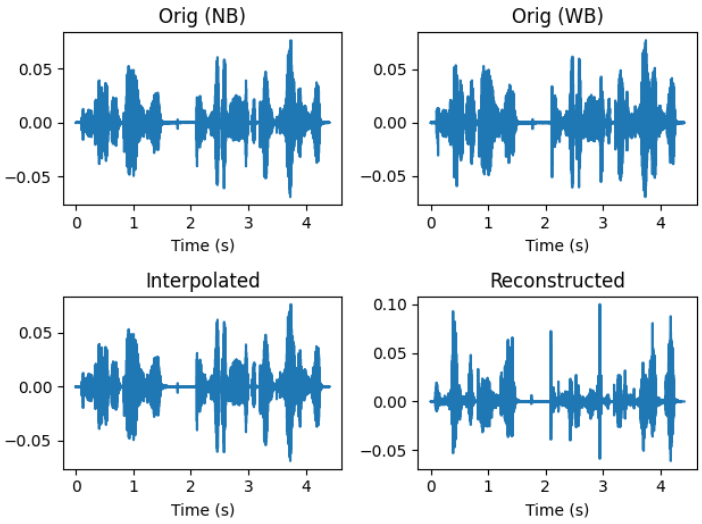


Figure 5.10: stars_16k Signal Waveforms

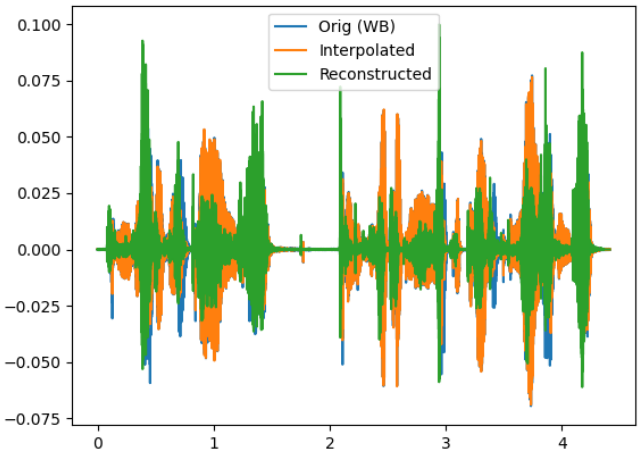
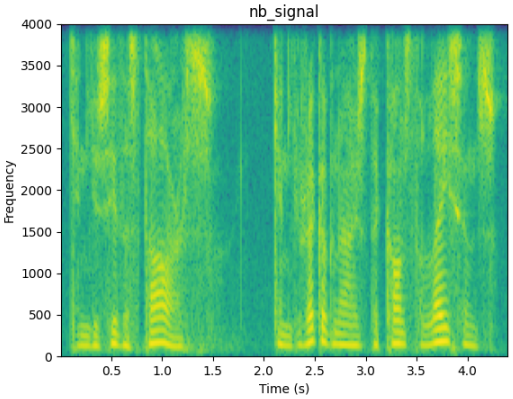
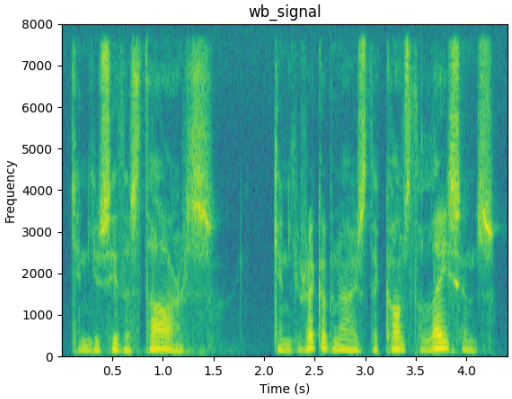


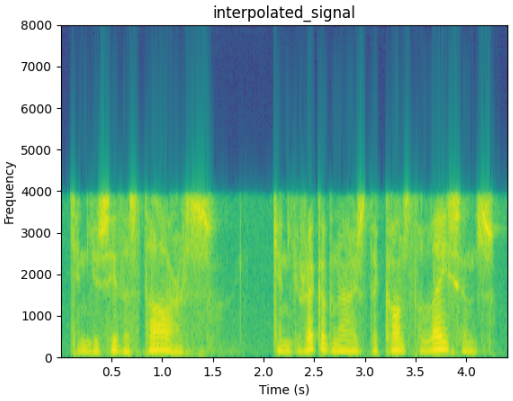
Figure 5.11: stars_16k Signal Waveforms Merged



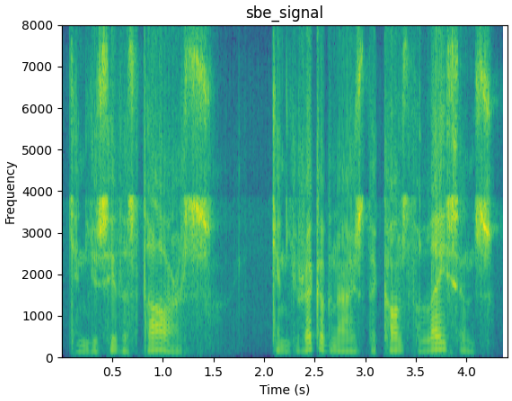
(a) Narrowband Signal Spectrogram



(b) Wideband Signal Spectrogram



(c) Interpolated Signal Spectrogram



(d) SBE Signal Spectrogram

Figure 5.12: stars_16k Narrowband vs Wideband vs Interpolated vs SBE