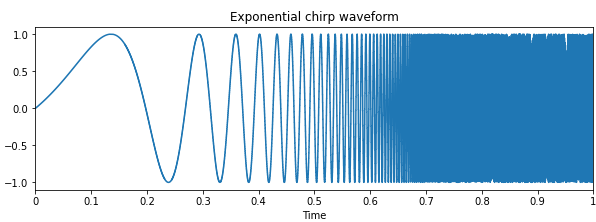
***Speech Lab Report***

**SENANE Zineb, HARKATI Chiraz Rayene**

1. **Spectrogram analysis**

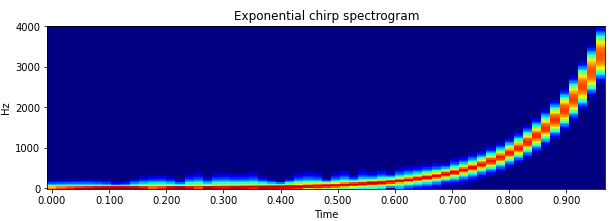
**1.1. Chirp signals**

**Positive exponential chirp**

****

***Explain why the sound seems to last less than 1 second.***

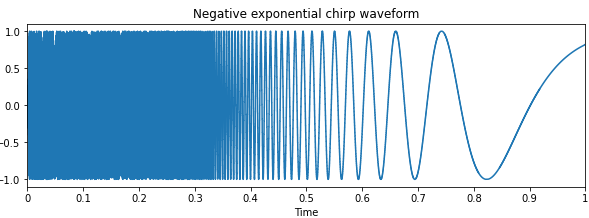
**The signal starts with a frequency of 1 Hz and reaches 4kHz. Because the audible interval for humans is [20Hz,20kHz] which we call audible frequencies, we will not hear anything in the beginning until the frequency reaches 20Hz. As we have an exponential chirp waveform, the part that consists of inaudible frequencies (from 1 Hz to 20 Hz) lasts for an important portion of the chirp signal’s total duration.**

****

***Explain why you see such wide bands or smearing for the chirp at higher frequencies.***

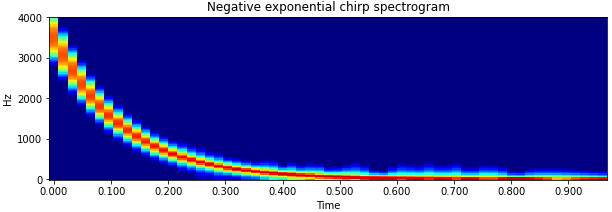
**As the long signal is multiplied by a time window (which is equivalent to a convolution in the frequency domain), we’ll have a smearing effect on the chirp spectrogram. The wideness expresses the energy of what we hear. Higher frequencies reflect higher energy levels and then a wider band. As we can see, in the high frequencies, the frequency changes a lot in one window thus we see more smearing. However, in low frequencies, the frequencies are barely changing and the smearing is too low because of the exponential shape and characteristics.**

**Negative exponential chirp**

****

***Before plotting it, think about what you would expect to see in the spectrogram.***

**Maybe an exponential chirp spectrogram with the same shape but inverted.**

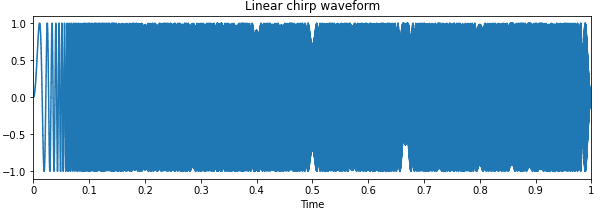
****

***Describe the correspondence between the chirp characteristics, the timewave form, and the spectrogram.***

**As the signal variability is high at the beginning of the timewave form and decreases towards the end of the signal, the high frequencies in the temporal domain will be shown in the first part of the spectrogram.**

**Hence, the high frequencies seen in the temporal domain(timewave form) are expressed by a wider band in the frequency domain(spectrogram). In the high frequencies, the frequency changes a lot in one window; however, in low frequencies, it's barely changing.**

**Positive linear chirp**

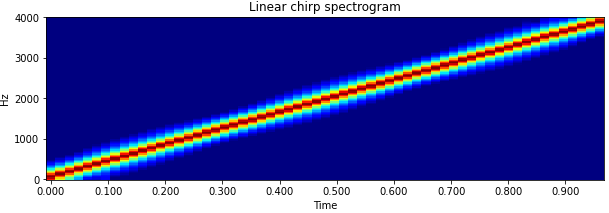
****

***Explain why the sound now seems to last the full 1 second, or at least longer than in the first two examples***

**As we have a linear function, reaching 20Hz will be faster than the first case (exponential evolution in low frequencies is lesser compared to a linear function). The inaudible part in this case lasts only 5ms.**

***Again, before plotting it, think about what you would expect to see in the spectrogram.***

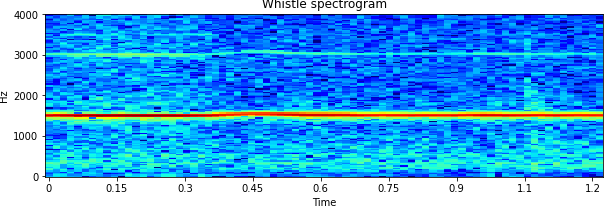
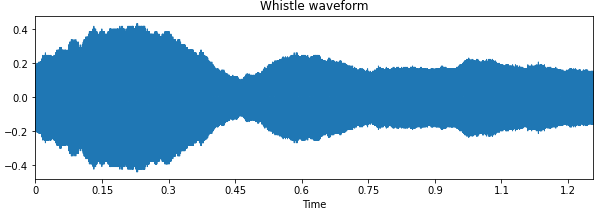
**The spectrogram should be a line with the same width for all bands.**

****

***Describe the spectrogram, the differences between the spectrograms of each chirp signal, and why there is less smearing across frequency for the linear chirp.***

**As we have a linear signal and we’re shifting the window along with it, the frequencies’ evolution this time will be constant in time within the window. Hence the smearing is also constant against frequencies and smaller than in the exponential evolution where there is a fast increase in a time window.**

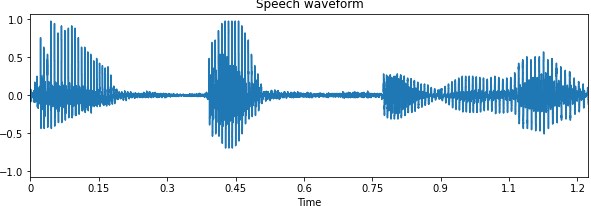
**A whistle**

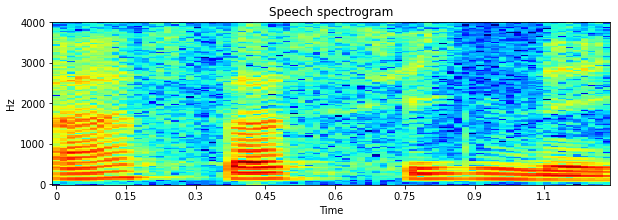
****

***What is the approximate frequency of the whistle? To what corresponds the second visible component? While it is a little difficult to see, describe the correspondence between the amplitude in the waveform and the color variation in the spectrogram.***

**The whistle's approximate frequency is 1.55kHz. The second visible component corresponds to the second formant (~3kHz) and it is twice the approximate frequency of the whistle. lighter colors refer to high energy and then high magnitude (envelope) in the waveform. The second yellow line represents a harmonic.**

**1.2. A speech utterance**

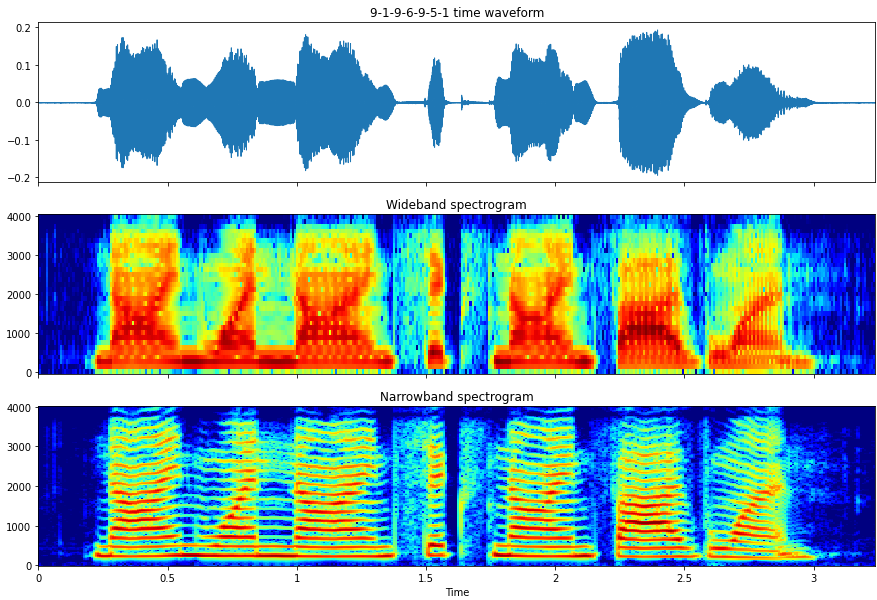
****

****

***Describe the spectrogram and the correspondence to the waveform. Identify the intervals of voiced and unvoiced speech. What are the red, horizontal striations? Estimate the fundamental frequency for the vowel sounds.***

**The voiced speech (e.g. vowel) is represented by high amplitudes in the waveform and the presence of red/yellow striations (pitch harmonics). The unvoiced speech (e.g. consonant) is represented by lower amplitudes in the waveform and the presence of blue striations. The red horizontal striations correspond to the formants. The best way to identify the fundamental frequency for the vowel sound is to compute the number of red lines under 1000Hz and divide 1000Hz by this number. This method gives us a fundamental frequency for the vowel sound of 142 Hz.**

**1.3. Narrowband and wideband speech spectrograms**

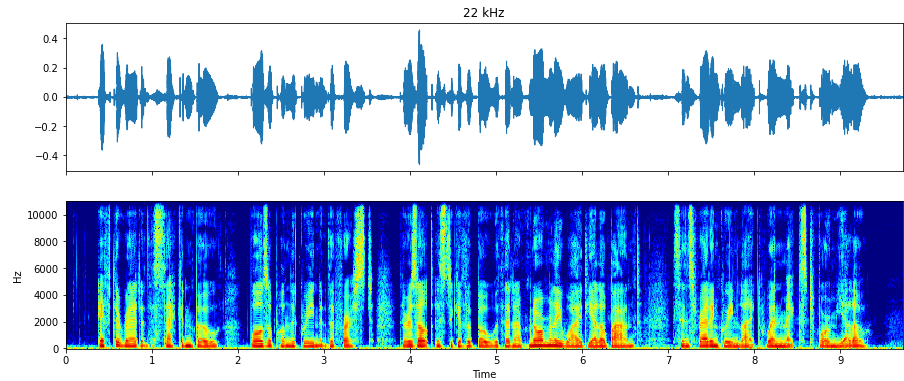
****

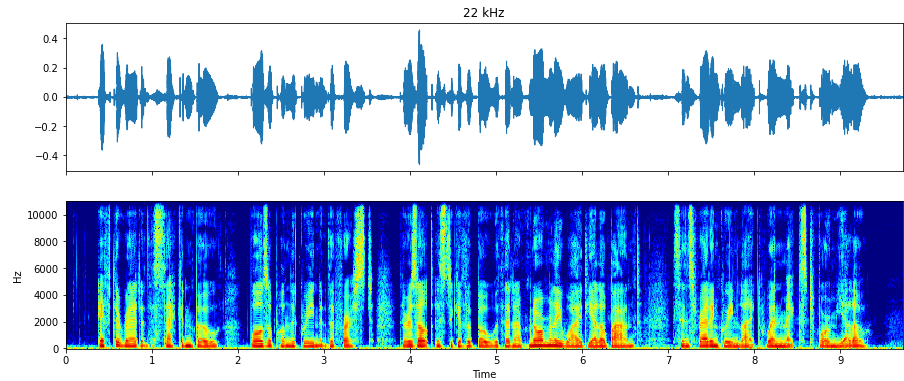
***Going beyond the differences in terms of n\_fft and hop\_length, compare and contrast the two spectrograms.***

**The hop length of the narrowband is larger than in the wide band ( this is because the hop length is set as half of the window size). For this code we set that hop length to 64 then there is no difference as it characterizes the difference /shift between successive windows. The n-fft characterizes the number of samples considered for a window. The number is bigger in narrowband as we take a big window in the timewave form (time-domain). The narrowband spectrogram contains the fundamentals (pitch harmonics), while the wideband capture the envelope of the signal (formants, the red patches). The wideband spectrogram has a low-frequency resolution and high temporal resolution however the narrowband spectrogram has a high-frequency resolution and low temporal resolution since we can the red striations representing the excitation pics.**

**1.4. Speech bandwidth and quality**

**An utterance sampled at 22 kHz**

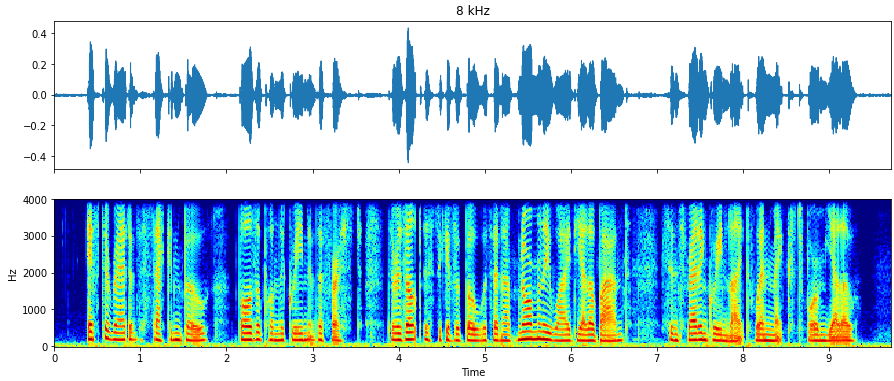
****

****

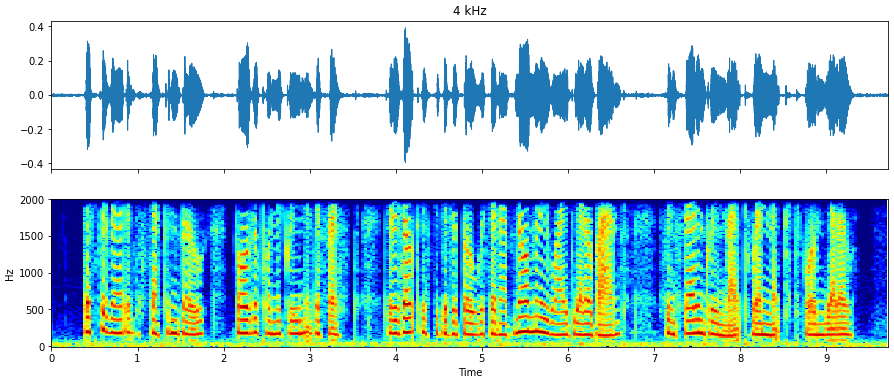
***Question Describe the differences in terms of perceived quality.***

**When we reduced the bandwidth frequency the quality became worse (we cannot hear clearly the sentence as the first one).**

**An utterance sampled at 8 kHz**

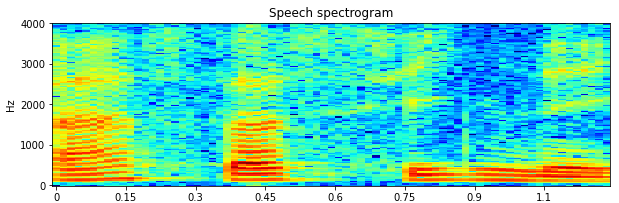
****

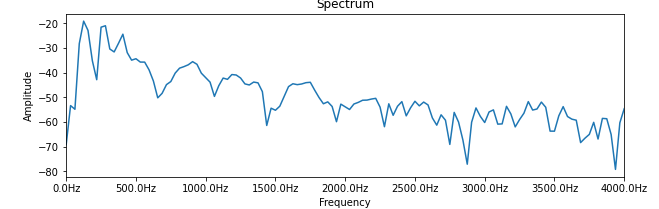
**An utterance sampled at 4 kHz**

****

***Notice the substantial degradation in quality. Describe your observations and explain why the degradation is so much more significant at 4 kHz than at 8 kHz.* When we set a low bandwidth it cancels the higher frequencies that exceed it. That's why we got fewer fundamentals at 4khz than at 8kHz and hence the degradation of the quality. That means with downsampling we lose a lot of frequencies that we can hear (<20kHz) compared to the utterance sampled at 22kHz(the blue light we got in 22kHz spectrogram has disappeared when we downsampled at 4kHz and 8kHz).**

1. **Voiced/ Unvoiced speech and the fundamental frequency**

****

****

***Answer the following questions by plotting the spectrum for different frames:***

***- What characteristics differentiate voiced and unvoiced sounds in the frequency domain?***

**Voiced signals have formants while unvoiced ones don't. Voiced signals have high energy (amplitudes higher) than unvoiced signals. The voiced sounds (e.g. vowel) are represented by high amplitudes in the waveform (high energy) and the presence of formants in spectrograms. The unvoiced sounds (e.g. consonants) are represented by lower amplitudes (lower energy) in the waveform and the absence of formants.**

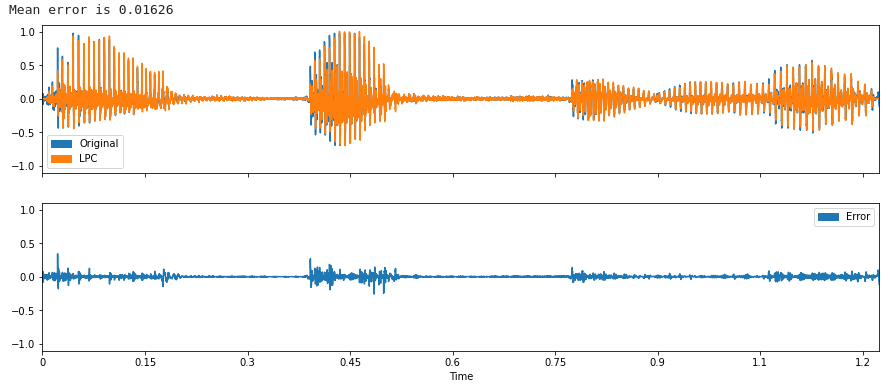
***- What is the approximate fundamental frequency F0 for the first ‘a’ of ‘assassiner’?***

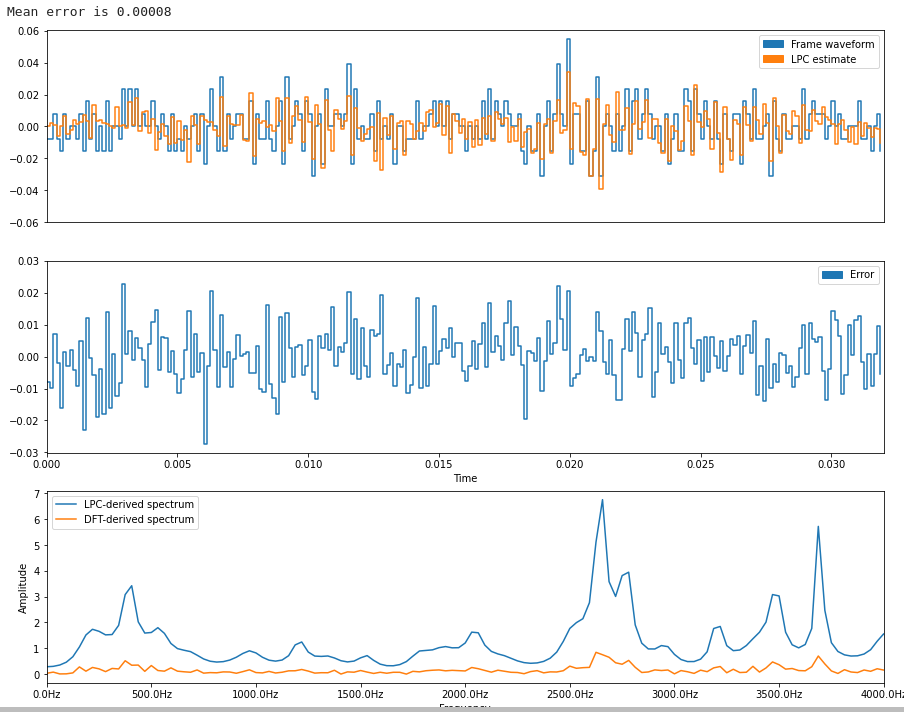
**Under 1000Hz we counted the 7 peaks. 1000Hz/7 = 142.8 Hz (choosing t=0s).**

***- What are the approximate frequencies of the three first formants of the second ‘a’ of ‘assassiner’?***

**At t=0.45s, the first three formants are respectively 500 Hz, 1500 Hz, and 2500 Hz (the frequency of the center of each yellow/red part corresponding to the second ‘a’).**

1. **Linear prediction and linear predictive coding**

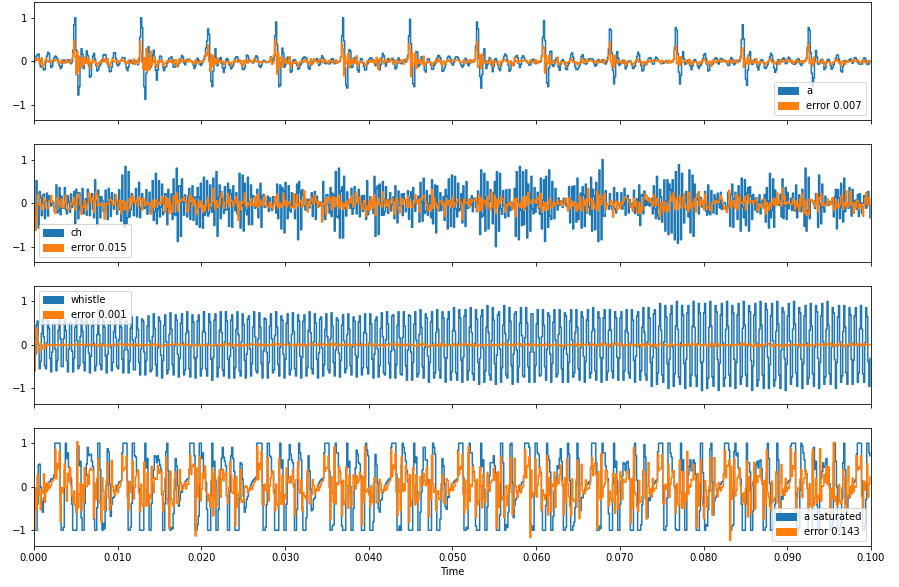
****

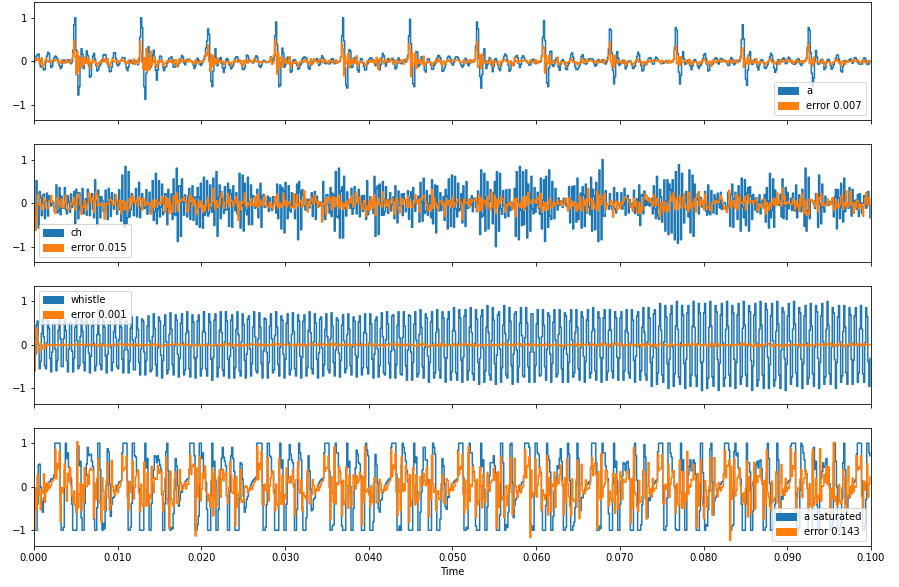
****

***Plot the DFT-derived and LPC-derived spectra for different frames containing either voiced or unvoiced speech and for different prediction orders and then describe and account for the differences you observe. What is the relationship between the number of detected formants and the order of the prediction?***

**We chose t=0.3s for unvoiced and t=0s for voiced speech. The DFT-derived spectrum contains the excitation and the filter. To separate the filter from the source we used LPC. By choosing an adapted order of prediction, the filter is obtained within the LPC-derived spectrum. The LPC-derived spectrum gives us the formants for voiced speech if we’re using an appropriate order of prediction. However, for unvoiced speech, the LPC does not draw the formants. In fact, it tries to predict the next sample using p previous samples, where p is the order chosen, even if they are just noise and fundamentally unpredictable. Hence, for unvoiced speech, the data will not fit the LPC model and the resulting peaks do not have any interesting meaning. The number of detected formants for voiced speech is related to the order of prediction. By setting a low order of prediction, we detect fewer formants and by setting a very high order of prediction, the LPC-derived spectrum imitates in some sort the DFT-derived spectrum and we detect an equal or higher number of formants than the number of samples in a single pitch interval.**

**3.1. Error analysis**

******

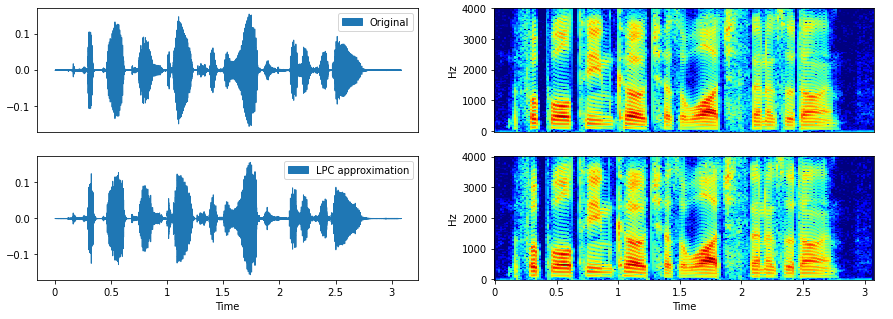
******

***Comment on and compare the results for each of the four waveforms.***

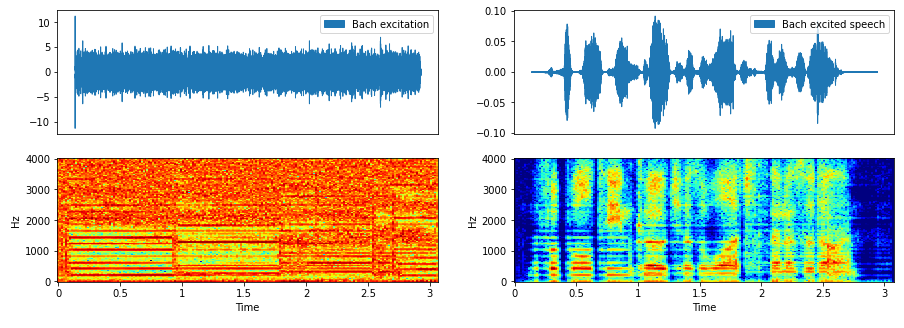
**The errors are really small for the ‘a’ sound and the whistle (less than 0.01), a little bit higher for the ‘ch’ sound but still low (0.015), and quite important for the ‘a’ saturated sound. The whistle signal is quite pure since it's a sinus wave and thus we can easily predict it, which justifies the very small error. Since the ‘a’ sound is voiced, we can predict it with a high order as explained before. By setting the order of prediction to 12, we’ll get accurate predictions for the next sample with the LPC model and hence the low error. However, as the ‘ch’ sound is unvoiced, it is less predictable (explanation given in the previous question) and thus the quite important and higher value we obtained for the mean error than the two other sounds (‘a’ and whistle). Finally, as the saturated ‘a’ contains the amplified original ‘a’ signal after being clipped, it will have some discontinuities that we can consider noise and hence the unpredictability of the saturated signal. This justifies the moderate error value we obtained for the saturated ‘a’ sound.**

**3.2. Cross-synthesis**

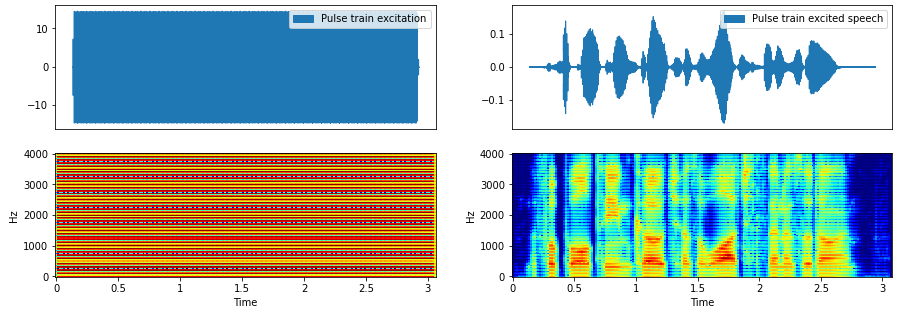
**Waveforms and spectrograms for the original and resynthesised utterances**

****

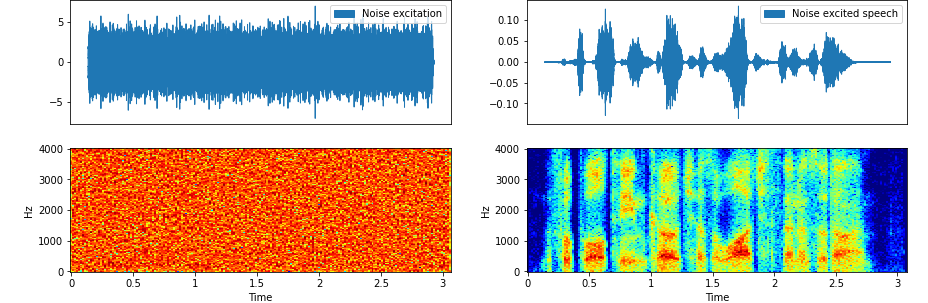
**Waveforms and spectrograms of the piano music excitation and then the cross-synthesis**

****

**Waveforms and spectrograms of the pulse train excitation and then the cross-synthesis**

****

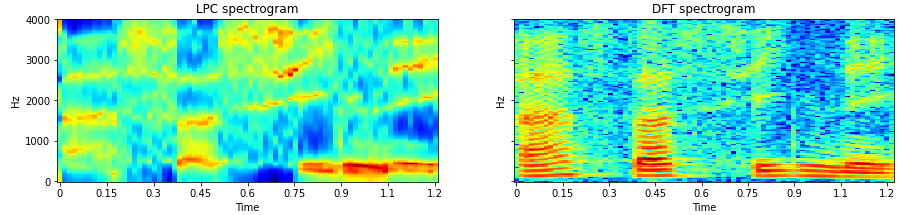
**Waveforms and spectrograms of the white noise signal as excitation and then the cross-synthesis**

****

***Briefly describe what you heard in the examples above and then describe in your own words what are the source and the filter in the source-filter model and their physiological origins.***

**The source is the way of perceiving the speech and it’s what differentiates the old man and the child. The filter is what was said and what makes us understand the words (it’s about the vocal cavity and vocal cords). When we change the excitation, the understanding and the perception of the utterance will not change.**

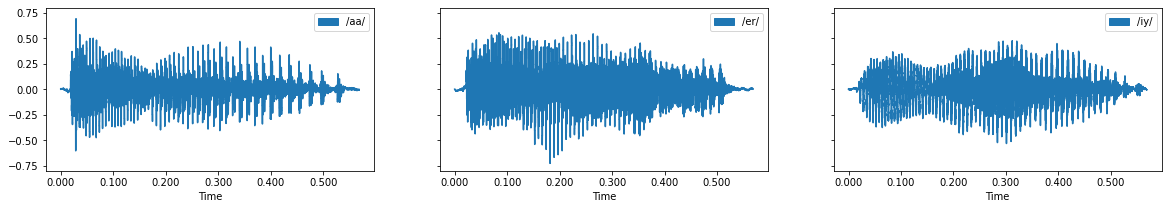
**3.3. LPC spectrograms**

****

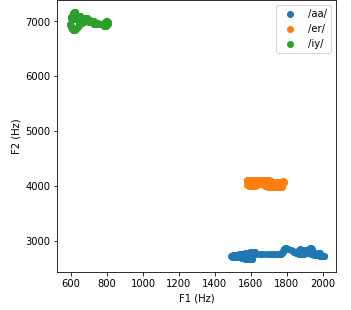
***Explain the principal differences you observe between the two spectrograms and describe their origins as regards both the human speech production process and the tools used for signal analysis.***

**The DFT-derived spectrogram keeps both the source and the filter (all the frequencies of the sound) that’s why we have the striations in the spectrogram whereas, the LPC-derived spectrogram sperate the source and the filter and keeps only the filter that’s why we have a lower resolution of the DFT spectrogram containing only formants for voiced speech.**

**3.4. Modelling**

****

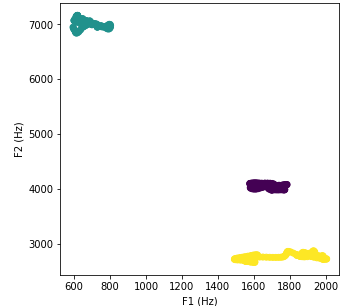
***Frame-level formant estimates for the three sounds***

******

***Comment on the results and the potential for characterizations of the vocal tract or filter configuration can be used for speech recognition.***

**As we have seen in the course, the F1 frequency corresponds to the position in the back and is related to the back pharyngeal cavity while the F2 frequency corresponds to the position in front and it is related to the oral cavity and the tongue. The longer the cavity, the lower the F1 frequency. For ‘iy’ sound, the tongue is on the palate and thus the pharyngeal cavity is longer which means the lowest F1 (~700Hz) while F2 is the highest (7000 Hz). For ‘aa’ sound, the position of the tongue is at the back of the mouth and thus the pharyngeal cavity is short and the oral cavity is big which means a high F1 frequency (~1800 Hz) and a too-small F2 frequency (2800 Hz). For ‘er’ sound, the position of the tongue is in between, so both frequencies F1 (~1650Hz) and F2 (~4100Hz) have values in between the two previous ones.**

***Design a trivial classifier to recognize these different vowel sounds using the k means clustering algorithm***

******

***Describe how you would implement an automatic speech recognition system using such features. How might it perform?***

**We will start by preprocessing each sound to extract the features, this step will be done by estimating F1 and F2 formants from the LPC-derived spectrogram. Next, we will use the K-means classification algorithm to train the ASR system by considering K clusters of an alphabet of known phonemes. The input of our system will be the extracted features (F1 and F2 formants) and it will output the corresponding phoneme.**

**From a theoretical point of view, the performance of this automatic speech recognition using such features seems to be good since the formants for different phonemes will be different. However, a bigger variance of the inter-class variance than intra-class variance might cause a problem and bias the clustering since the clusters will overlap.**

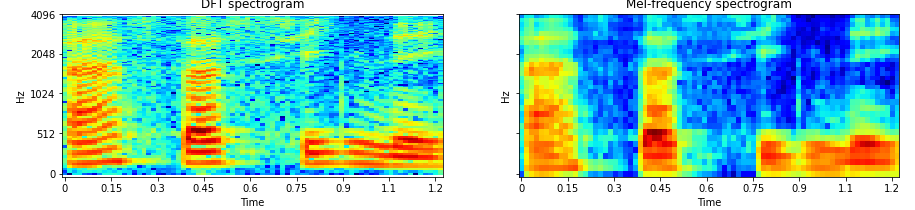
***Repeat this exercise after adjusting the code to plot the same features for the full recordings (not the trimmed versions) and comment on the results. How would an automatic speech recognition system perform now? Do formant estimates make for suitable and sufficient features for speech recognition? Can you think of how you might do better?***

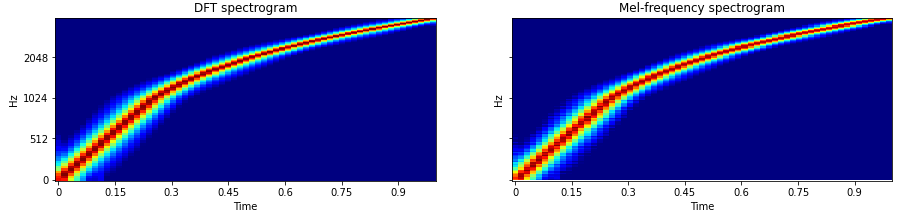
**As mentioned in the question above, for the full recordings, we need to cut them to obtain a trimmed version and apply the model directly. For this purpose, we will use a short sliding window (to be able to separate vowel and consonant sounds) and cut the audio into small parts overlapping with this window. Once this step is done, we will take each sample and use our ASR model to assign the sound to the corresponding cluster. To get the originally said utterance, we can just combine the phonemes. As speaking rate changes within an utterance, and processing made to the audio results in small sounds, the ASR model might have a bad performance. Moreover, increasing the number of phonemes and including similar phonemes will also influence badly the performance of the model (it may merge two phonemes in the same cluster). Hence, using the two formants F1 and F2 estimates for such classification models as they are not sufficient features for speech recognition. We might do better by using acoustic modelling.**

1. **Feature extraction**

**4.1. Mel scaling**

***Plot Mel-spectrograms with different numbers of Mel filters and comment on the results.***

******

******

**The mel-spectrogram with a low number of mel-filters looks like a low resolution of the DFT spectrogram.**

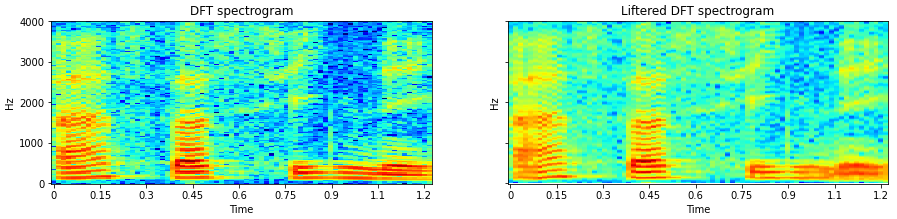
**The mel-spectrogram with more number of mel-filters is closer to the DFT spectrogram. By setting n\_mel to 128 we get the roughly same spectrogram but if we decrease too much the number of mel-filters, we will be able to observe the formants only. We can say that the Mel- spectrogram is a weak representation of the DFT spectrogram**

***Explain the differences and their origins between the two spectrograms.***

**We can see that both plots have the same shape where the frequency increases linearly given the time. By increasing the number of mel-filters, the plot of the chirp signal for the Mel spectrogram looks more similar to the one of the DFT spectrogram as said before.**

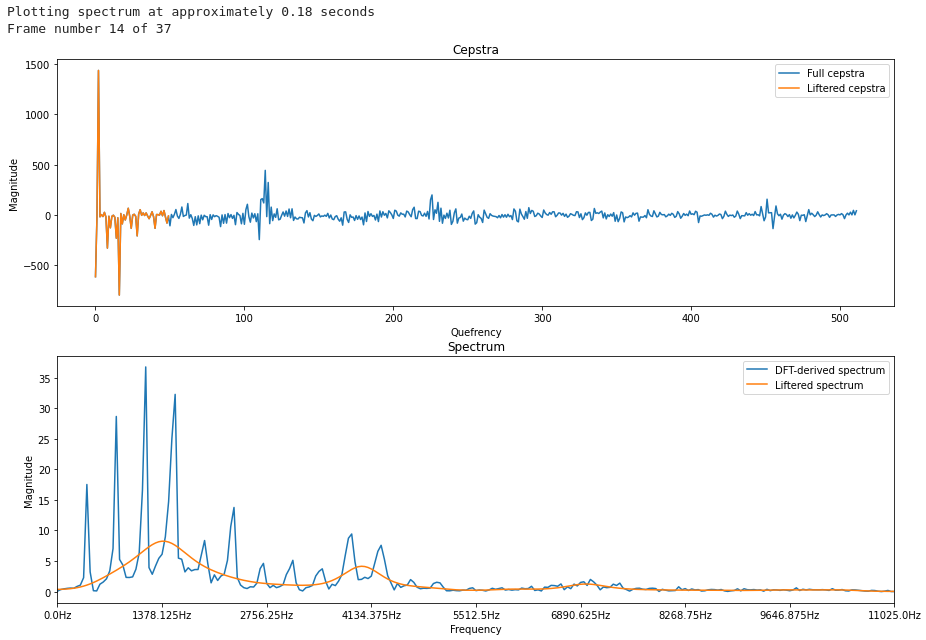
**We can observe also that when we have a lower number of mel-filters, we got more horizontal smearing in the high frequencies than in low frequencies this is because the Mel spectrogram takes into account that humans are better at detecting differences at lower frequencies than at higher frequencies and the, It means that within a large window of time, we have the impression of hearing the same frequencies (while in reality, it is increasing).**

**4.2. Liftering**

****

***Observe and comment on the effect with different liftering parameters, in other words, fewer and fewer cepstral coefficients. At some point, the fundamental frequency harmonics will disappear. At what point (number of retained cepstral coefficients after liftering) does this occur? Can you explain this?***

**We know that the liftered spectrogram is just a derivation from low pass filtering the DFT spectrogram which uses the cepstral processing to separate the filter and source or excitation and keeps only the filter. When we use more retained cepstral coefficients, we get closer to the DFT spectrogram.**

**Since we know that the low pass filtering erases the fundamental frequency harmonics and keeps the formants only, we can get why using 32 cepstral coefficients enables us to recognize the fundamental frequency of harmonics and a lower number of cepstral coefficients did not. ****

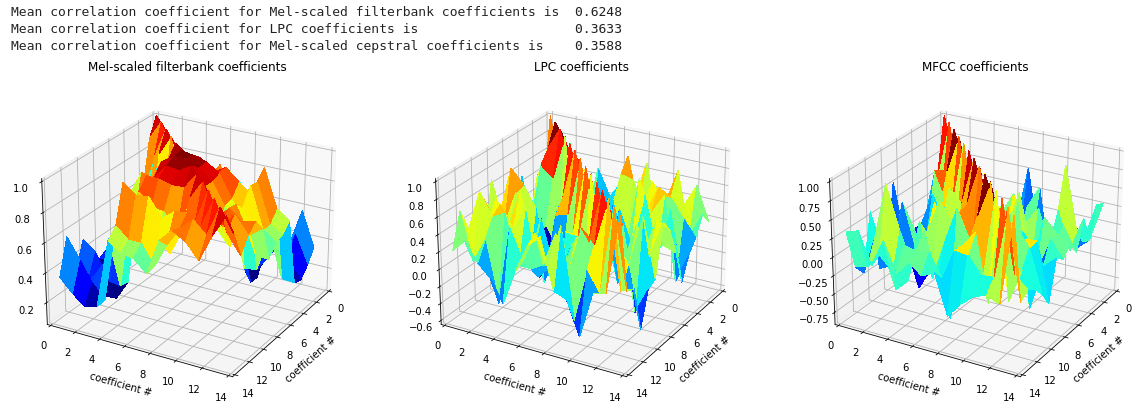
***Explain how cepstral coefficients are extracted and the importance of using a logarithm. Plot the cepstra for an interval of voiced speech and estimate the fundamental frequency. For plots corresponding to intervals of voice speech, you should be able to identify a component that corresponds to the fundamental frequency. You should then be able to set the number of retrained cepstral coefficients to either remove or retain it. Determine the maximum possible number of cepstral coefficients for which the fundamental frequency is still \*removed\*. Then, explain the characteristics which differentiate voiced and unvoiced sounds in the quefrency domain.***

**Cepstral coefficients are extracted through the inverse discrete Fourier transform of the log magnitude spectrum. The logarithm is used in order to observe clearly both low and high amplitude frequencies. The fundamental frequency is the average over 4 periods: F0=1374/4=343.5Hz. The maximum number of cepstral coefficients for which the fundamental frequency is still removed is 110. In the quefrency domain with voiced speech, we can see clearly the bumps that represent the formants of the voice while with unvoiced speech we will have only very small bumps because, as known, unvoiced speech is made mostly of the fundamental frequency and noise.**

***Comment on the similarities and differences between DFT, LPC and liftered spectra***

**In the LPC and liftered spectra we keep only the filter so the spectrograms show low resolution of the DFT. While for the DFT spectrogram, we keep the information from both the source and the filter so we can detect red striations and identify fundamental frequency.**

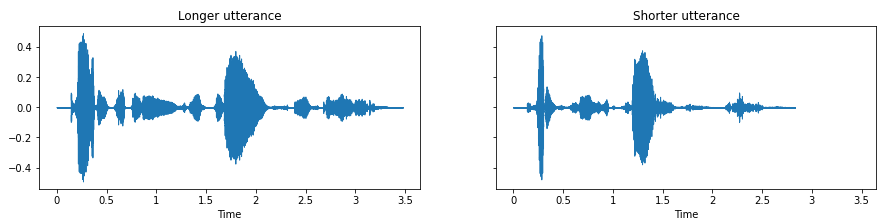
**4.3. Correlation**

****

***Explain which features, among the three choices, might be best for modeling and why.***

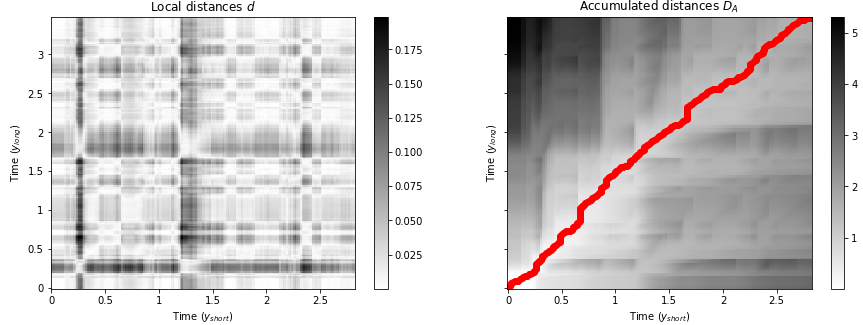
**The features with the lowest mean correlation coefficient will be best for modeling as they are less correlated and keep the most information without any redundancy. Hence, we will choose Mel-scaled cepstral coefficients because they are the least correlated coefficients with a mean correlation coefficient of 0.3588.**

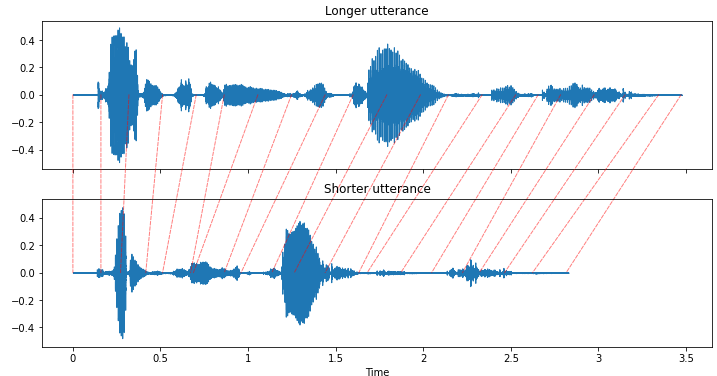
1. **Dynamic time warping (DTW)**

****

**5.1. Addition and linear warping**

**5.2. Dynamic warping**

****

****

The dashed red lines in the above plot show that DTW aligns the two utterances in a dynamic fashion, adjusting for differences in the inter-utterance speaking rates

1. **Hidden Markov modeling**

***Briefly describe how you would design an automatic speech recognition system to distinguish between "On" and "Off" utterances using HMMs***

**To design an automatic speech recognition system to distinguish between “On” and “Off” utterances, we need to use two HMMs, one for the utterance “On” and the second for the utterance “Off”. For each of these HMMs we can use the same structure given in the lab. We will start by training them using two algorithms: forward pass and Viterbi and update the state transitions and probabilities of emissions every time using the corresponding utterances. Once finishing the training phase, we will extract the MFCCs from the test utterance and see the probability of generating the utterance by one of the HMMs. Finally, a comparison of the probabilities and a threshold probability fixed prior to the observations reveal the type of the utterance (“On” or “Off” or none of these).**