SPEECH PROCESSING

Introduction

Inma.hernaez@ehu.eus (1st teacher)

Ibon.saratxaga@ehu.eus

Goals:

- Learn the **basic principles** of Speech Processing and Speech Technologies.
- Learn the fundamentals of the **speech production** and **speech perception** mechanisms.
- To be able to **correctly use the basic tools to analyse, visualize, and process audio signals** in general and speech signals in particular.

Syllabus

1. **Lesson 1**: Speech production and perception
2. **Lesson 2**: Basic concepts about signals and systems (Part I and II).
3. **Lesson 3**: Speech signal: representations (Part I and II).

Practices

- Short tasks and exercises
- Practice 1: Introduction to audio managing software
- Practice 2: Basic speech signal analysis.
    - Part I
    - Part II

Programs to download: **MATLAB**, Speech Analyzer, **Praat**, **Audacity**

Introduction: information in the speech signal

- What kind of information can humans extract from a speech signal?
    - **Message or linguistic information**:
    - **Language recognition**: we can extract from the speech signal information or knowledge about the speaker.
    - **Keyword spotting**: listening to speech signal and try to extract the main information through keywords. Instead of having an automatic colour system that says: say 1 if you want to get an appointment for vaccination or say 2 if you do not. This is very simple, it is forcing you to say 1 or 2. Another way would be "how can I help you?" and you say "get vaccine" and 'vaccine' is a keyword in

the system. Other keywords could be in this case: appointment, morning, afternoon, etc.
    - **Speech synthesis**: the technology that aims at generating the signal from the test. Speech recognition is speech to text, and speech synthesis is text to speech. Where is the information in a speech signal and what can we do with speech?
- Another part of information is the identity of the speaker:
    - **Biometric verification**: you are telling me that you are Inma, prove it with your speech. You talk to a system and the system validates your identity.
    - **Voice conversion**: voice cloning is another very similar way to call it. The system that converts one speech from one person into another. I mean, the speech of another person. Phonetics research: research language evolution, dialects from different regions. *Research on young Basque speakers, the differences between adults and youths*.
    - **Speaker identification**: we have a set of speakers and you come, and I must identify you. This is useful for forensics linguistics.
- Paralinguistic information: emotional state, etc. Parameters that are over linguistic information.
    - **Emotion recognition**:
    - **Speaker's state recognition**: detection if the speaker was drunk. It could be useful for cars, to detect if they driver is drunk, and if that's true you are not allowed to drive. Also: asleep detection. For example in call-centres they have a system for state recognition in order to know if the client is angry or happy.
    - **Expressive synthesis**:

The acoustic realization of this high-level info involves a large variety of measurable low-level features.

Where all this come from? How is this information realized acoustically? How this info appears in the signal? How can we measure all this information and convert it into low-level features?

We are going to talk mostly about low-level features.

Index

c. **Loudness and phons**
d. **Masking effect**
e. **Temporal and frequency discrimination**
f. **Other speech perception effects**

1. Signals and sounds

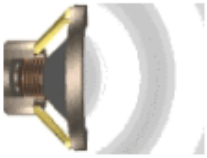
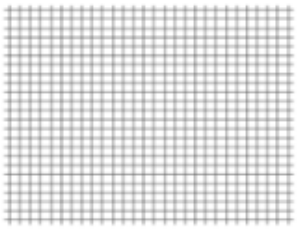
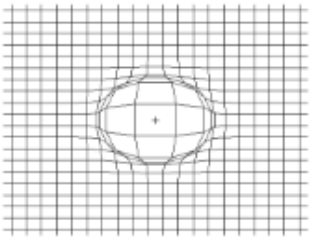
a. Basic concepts.

The sound is a pressured wave generated by a mechanical vibration of some object. Our vocal chords for example. But it can also be a speaker, which modifies the pressure in the air. The variation of this pressure propagates physically in the space.

If we take at a certain point in space, in the air, we measured this pressure with a microphone, for instance. The **microphone** is the opposite of the speaker. And if it is able to measure this pressure, this value or measure it is what is called the speech signal or the sound signal.

The signal is the electric signal, or measure, that we get from the acoustic wave. When we see the speech signal, what are we looking at? It is the variation of the pressure of a (physical) point () in the air. Nothing more than that. If there is a variation of the pressure, you measure it and you get the signal. This is the way it propagates.

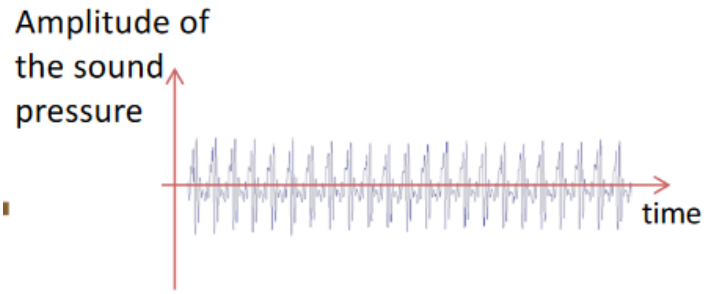
So, if we put a microphone in the middle of the room, we will get the measure in 3d.

With the microphone we convert an acoustic wave/mechanical pressure to an electric signal.

The **electrical signal is proportional to the variation of the pressure**. These electrical signals, we will digitalize them to put them into the computer and process them digitally. Once it is digitalized, it can be stored, processed, recognized, etc.

Here, this representation we are going to see it very frequently:



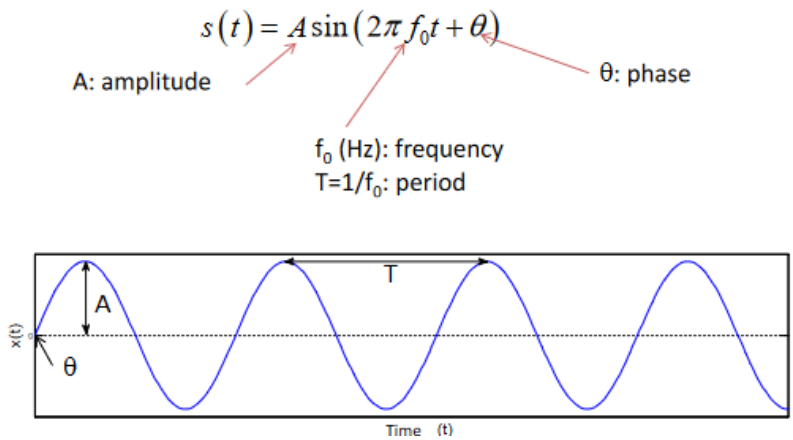
The microphones are the physical mechanical sensibility of capturing sounds. Of course, the sensor (microphone) could be very good or very bad, many things intervene here: the membrane and how it is isolated, how good is the magnet…

b. Time and frequency.

The simplest vibration in nature is the sinusoid. The first mathematics part: we should start with the simpler sound that exists. Which is the simplest sound we can listen in nature? It is a movement, variation of the pressure, of the waves which can be described with this mathematical expression. Get used to this math expression.

How is this movement? There is a particle in the air and moves (we are describing in one dimension). The curves mean that the pressure goes up and then down, and then up again.

This placement from the original point to the final point, is what we are seeing in this signal (look at the pic). For example, you throw a stone to a lake, and you see the movement of the waves. If this movement stays forever, ideally, this function is called sinusoid. This math function describes the simplest vibration in nature.



A is **amplitude**. It is one of the parameters of the sinusoid. We can use "s" for sinusoid but also referring to signal. **S(t)** means how this particle moves with time, so, it is function. Pi is the relationship between the length of a circle and the diameter. **Pi** is a constant. Now we have "t", the variable of time, we can know the value of the signal at a certain time. If we give values to "t" we get the value of the signal at a certain point of time. Then, we have the f_0 **(frequency sub zero**), this is the speed of the movement. this movement could be tuuuuu-tuuuuuu-tuuuuu or tututututu (second is faster), that is what says the frequency. If the movement is fast, the frequency will be high. They are directly proportional.

Another thing that we can see here **is the t, is the period;** it repeats itself all along the signal. For example, in the previous example we can see it is periodic because it repeats itself. The shorter the phase, the faster the movement.

If one period lasts "t" (let's say 1 second), the inverse of that value (1 divided by frequency), then T is measured in seconds. **How many periods do we have per second?** 1 second, 1 period.

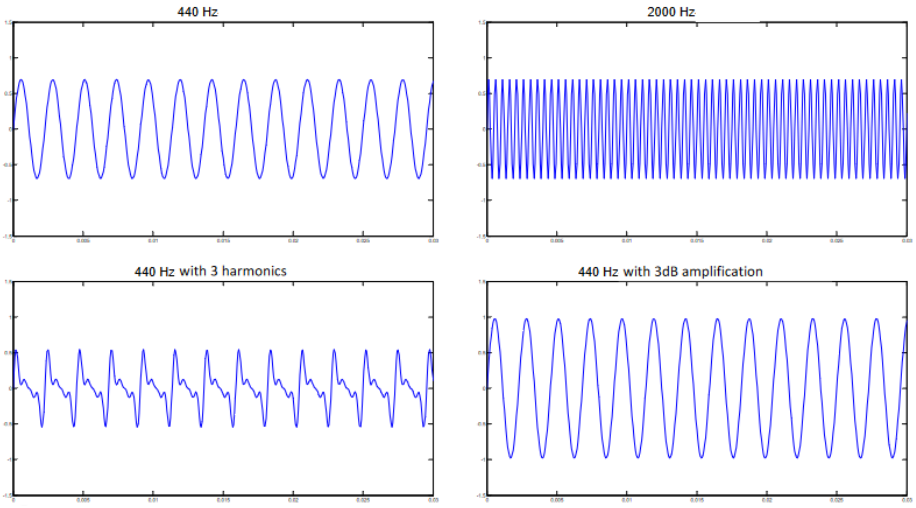
If we have t=1 second, we have 1 period per second. Now suppose that t is half a second (0,5 sec or ½ sec), how many periods do you have in 1 second? Two periods.

The **inverse of the period is called frequency**. The frequency measures the number of periods per second. And this is what I have called it periods and/or cycles. **Inverse of second: 1/s or s^{-1}. And this measure s^{-1} is called Hertz (Hz).**

The speed is measured in km/hour.

The Θ, the phase refers to where we start. But it could mean any point of the wave. This value indicates the starting point of the sinusoid. And we are going to generate these functions in MATLAB. We need to identify these features of the equation.

Relation of sinusoids and sound in time:



Listening to different sinusoids and see what is the higher frequency that you can hear.

440Hz: this is the standard B, musical notation: la. (?) all instruments are tuned at the same frequency. This is a simple sinusoid. It has one only frequency. And it is periodic.

440Hz with 3 harmonics: it sounds more human-made. This is a combination of simple sinusoids. It has more than one frequency, but it is periodic as well. The 1st and this one, they have the same period, but different frequencies.

2000Hz: the wave is faster, and this gives a more acute sound.

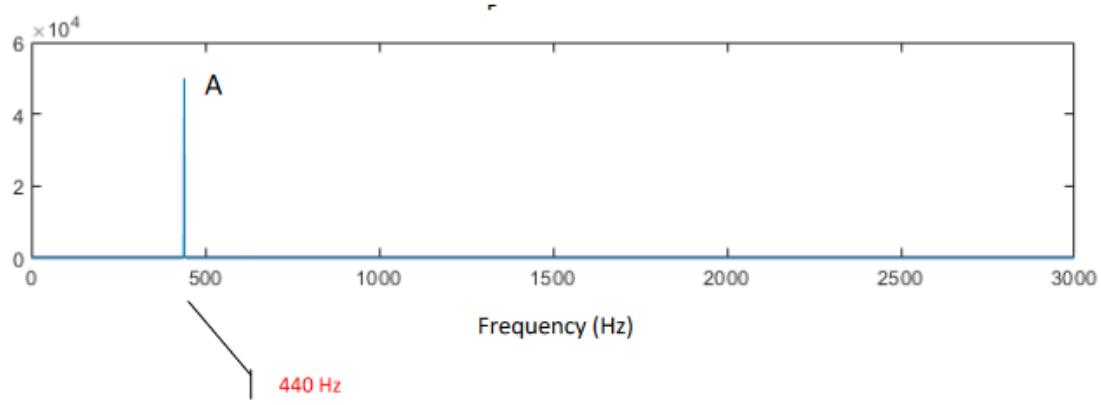
440 Hz with 3dB amplification: double the amplitude.

What do they have in common? The period. This is what we hear, for us is the same note. We perceive the same note. The notes have the same period, some notes with the same period are same notes.

Spectrum: a representation of the signal using the parameters that define it (Amplitude + Frequency)

If we can define a sinusoid with these three parameters, we put the phase to zero, because right now it is irrelevant. If we can differentiate with only two parameters, maybe we can find a better representation instead of a waveform.

Instead of saving this function, or all this data, I have the sinusoid with frequency 440Hz and Amplitude 5. Like in the example in the graph: the line represents one frequency. And we can put more frequencies in order to store them.



Only one line represents and only one sinusoid. This is the representation of the signal. And this representation is called spectrum.

This is the alternative representation of the signal that we have already seen. The 440Hz with 3 harmonics, the important thing is that all these frequencies are multiple one to the other. This is the magic of periodic signals. The y axis is the amplitude. The second line in the 440Hz+3 harmonics is the double of the first one, the third line is the triple and so on.

What is the sampling frequency?

We have the signal and then we take samples to store the signal in the computer. This process is called sampling. How many samples in one second? If the sampling frequency is 44100, that means that in 1 seconds, we have taken 44100 samples.

The distance between two samples, would be 1 divided by 44100.

c. Sound Pressure Level

The intensity of the sound is usually measured in logarithmic units dB SPL (Sound Pressure Level) (dB = decibels).

What is the relationship between the amplitude and the dBs? Decibels means like ten parts of bels. This measure is a relative measure, it is not absolute. Can never be taken as an absolute.

An absolute measure is that I am 1 meter 53 cm tall. And a relative measure could be like I am 10 centimetres smaller than you, because I am doing it relative.

It is a logarithmic function. It is a conversion or a transformation of the number. It means like the inverse; we can go from the logarithm to a unit and then backwards.

We say it is relative because we measure the pressure in reference to a point, or reference level. We do not measure the pressure absolutely, but in reference to a reference value. This reference value is taken empirically. They took this decision long time ago and it is like that.

This reference level is 20 micro-Pascal. It is a very small pressure, and it is the *minimum minimum minimum* pressure measured here. This is just convention, nothing else.

We measured in reference to that point. If this is the minimum that we can measure, then we are always taking this. Then, what we do is divide.

P_1 times bigger than P_{ref}	PdB (SPL)
1	0
2	6
10	20
100	40
1000	60
0,1	-20
0,5	-6
0,01	-40

If we are in 2, we are ten times bigger than the reference level. This is the number that we take that is P1 divided by the reference.

Logarithm base 10 of 100 is 2. Logarithm of 1000 is 3, 10.000 will be 4. And so on.

In the same slide as the graph, we can see that even we say logarithm of 0 is 1. Instead of going multiplying by ten and growing a lot and exponentially we grow linearly thanking to the logarithm.

With the logarithmic expression we convert these multiples by ten, we go from 20 dB to 20dBs. Always adding twenty.

The perception of incrementing sound is linear.

P_{ref} is always = 20 micro pascal = $20 \cdot 10^{-6}$ = 0'00002

And SPL is P1 divided by P_{ref} could be 10^0, 10^1, 10^2…

Now, we have a signal, and we are going to talk about the amplitude. When we use this amplitude, we don't use micro pascal anymore. All the signals are in a range, from -1 to +1. Why do we need the dBs? We amplify the signal for example 20 dBs (also means multiply this by 10), what happens with the A? for example mi Amplitude is in the level x10^2 and I want to amplify ten times, so I go to x10^3, I have amplified 20 dBs.

2. Speech production

a. The human phonation system

We are going to see how speech is produced by the production system. We are not going to enter very deep in characteristics of speech right now.

Speech → the sound wave produced when the air is expelled from the lungs and reaches the air through the vocal tract.

Speech is generated when there are pressure variations in the air. These variations are produced by our human producing system. It starts in the lungs. We take air and full our lungs and then the air goes through the trachea, it gets accelerated and goes through the glottis, pharynx, and it flows through the cavity. And goes out through the nasals and mouth.

Everything that is under the glottis (Adam's apple) it is called the sub-glottal system. We differentiate very much between the sub-glottal system and the upper-glottal system.

The sub-glottal system is made with the lungs, bronchial and trachea. The trachea is very important because the air gets speed there. And then it goes to the vocal tract. The air starts in the lungs, then through the trachea, then the voice box (or larynx, Adam's apple). Afterwards, through the vocal tract.

In the vocal tract we find the glottis (vocal cords), then the pharynx, then the vellum (to close the air that goes to the cavity. Then we have the mouth.

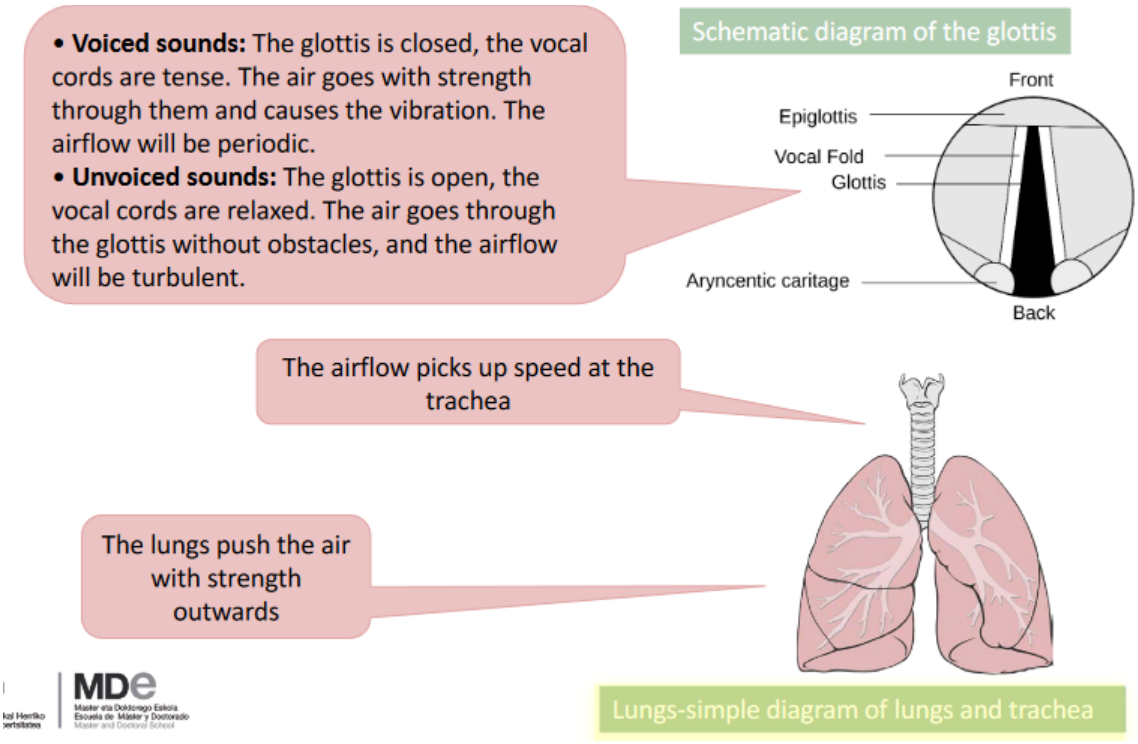
In the oral cavity, that is the mouth, or surrounded by the mouth. And we also have the nasal cavity, the air can go over there as well, for example with the phoneme [m] or [n], nasal consonants.

These limits (Section: 0-20cm) are the values of the frequencies that we are able to generate, it is because of my vocal cords.

In the voice-box we have the vocal cords, we can control them with some muscles. I can control this length, or size, and there are two main situations that we consider:

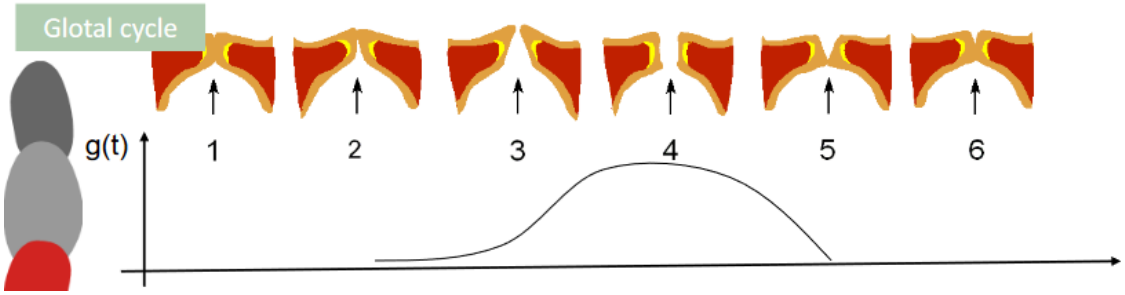
1) vocal cords are relaxed, loose, in that case the glottis is open and the air goes through without opposition. For example, when we do: shhh.

2) vocal cords are not relaxed, the vocal cords vibrate, and we produce voiced sounds.

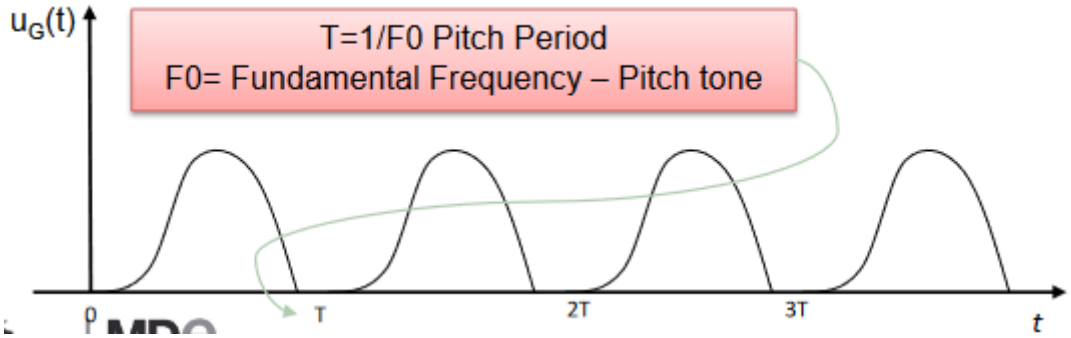


Vocal cords are going to open and close, and it is a periodic movement. so, each period is one cycle of the movement. So, the glottal cycle is the interval of time that includes the whole cycle: opening and closing. We use some models of this glottal opening. If we want to put a microphone in the glottis, what would we see? We would measure the pressure of the air there. Of course, they do not use a microphone, but other instruments.

When the glottis is closed, there is no air (see in the graphic from the 1st to the 3rd pic).



But we can see a glottal signal. And it is going to be something like this:



And it opens and it closes, and then again and so. This is what is called the glottal pause, when the glottal signals are made by consecutive pauses.

The most important thing here is the pitch period, or fundamental frequency. It is the fundamental frequency, but also, we can call it pitch or tone (colloquially). Fundamental frequency is changing in speech. We do not have a constant pitch, tone or fundamental frequency. We will learn to estimate this fundamental frequency in speech.

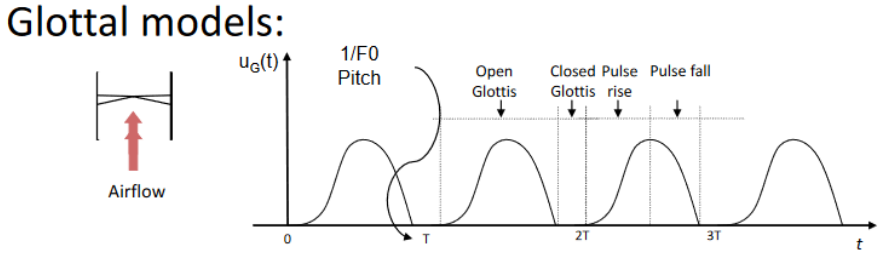
Video the vocal cords in motion:

In the first part, he had the device just directly in the cavity and we could not see the periodic movement. because the movement is faster than we can see or perceive with human eyes. We cannot see movements faster that 15 cycles per seconds.

In the second part we can see the movement. he is using stroboscope light, this light samples the image and we get a copy of the movement. we will see this in the signals lesson and how this works.

b. Voiced and unvoiced sounds: the glottal pulse and the glottal cycle

There are several glottal models, as you can imagine this glottal pause and signal is very important for us. They try to imitate mathematically this movement:



This Rosenberg glottal model use these two parameters: the fundamental frequency, the period is the tone of the signal. Different kinds of the same sound.

What is the ***relative close time***? How much time is opened and closed. The proportion of the opening quotient. So, the pulse duration divided by the pitch period gives us the proportion of time that the glottis is opened.

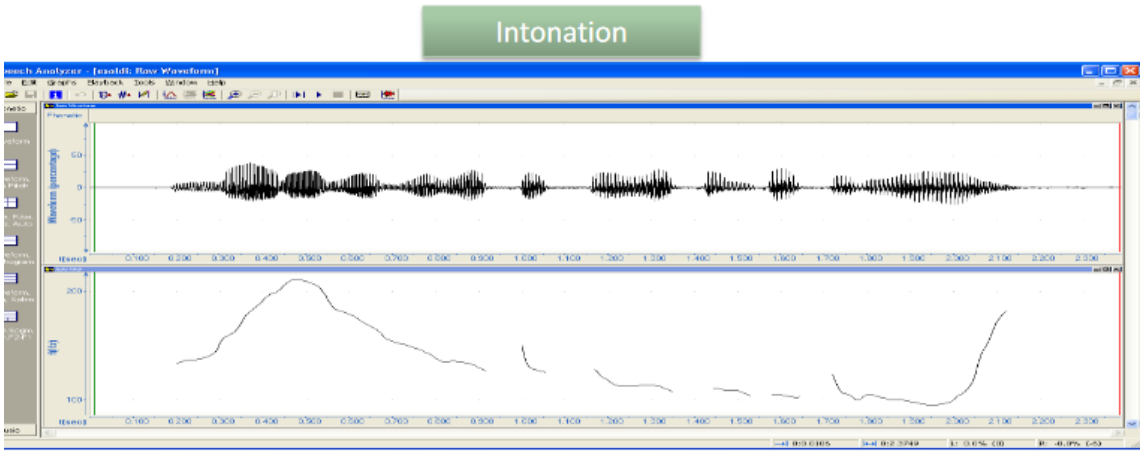
We can see that first the pulse goes up and then goes down, goes to be closed. For example, we can have something like this.

Speed quotient: the time for going up divided by the time for going down/close.

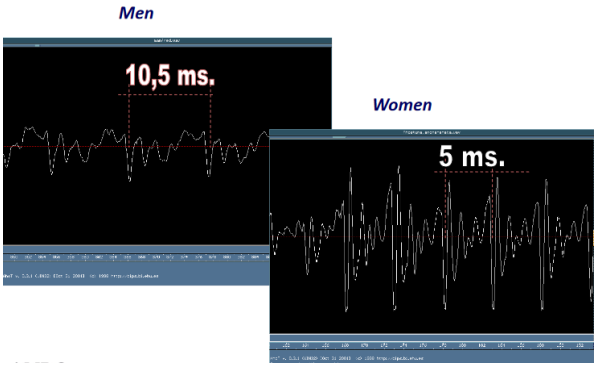
Of course, the pitch is the fundamental frequency. If I say that my ff is about 110Hz, I talk about the average of my speech. In average, men have lower fundamental frequency compared to women and children (they have higher).

Typical values are these values: men (110-130) and women and children (200-230Hz)

If we measure intonation in short segments of time, we can see it changes. Because, as we said we are not talking with the same intonation all the time. And therefore, we change the fundamental frequency:



We see the change of fundamental frequency in speech through time.



First, we must select a voiced segment in order to find the pitch. As for example:

Vowels are the perfect sounds to measure fundamental frequency.

There are variations because we are humans. Not only because we change it on purpose, inside the vowel, but also even if I try to keep my ff constant… there is no exact measure. The measure for each period is going to be randomly different from one period to the next one. The signal that we see

in the voice segments of speech is going to be periodic. It is pseudo-periodic, because periodic mathematically means that is **exactly** the same period.

Voiced sounds:

The vocal cords interrupt the air flow coming from the lungs and then produce the vibration.

The airflow will be periodic, corresponding to the opening-closing period. If this is the signal, if the period is T we will call it T0, it is the inverse of F0 is the fundamental frequency and T0 = 1/F0 is going to be the period. Fundamental frequency = fundamental period = period.

Unvoiced sounds:

We will have a wave form, very noisy thing, because there is no modulation of the signal. During the production of the unvoiced sounds, there could be also obstacles, in the nasal tract we cannot put obstacles because we do not have articulations

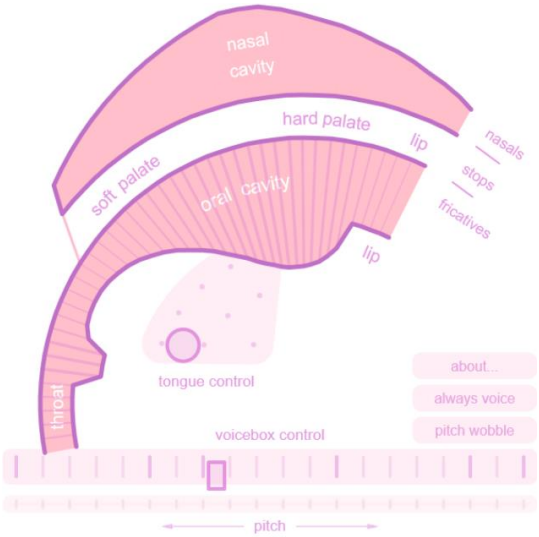
Vocal and nasal tract:

They are cavities. What is a cavity for a sound? When a sound, if we have a box and we make a whole in one side and put a speaker there, the music or the sound will reflect on the walls of the box, inside the cavity.

In the cavity, certain frequencies get reinforced. Due to mainly defined by the size of the cavity. Remember that the **air pressure has a frequency**, the pressure is varying, so some frequencies will be reinforced, and some others will die. We are not able to go through all of them because some do not propagate.

The resonance cavities means that inside these cavities some frequencies will be reinforced. How many? Considering 4kHz, 3 or 4 resonances will only be reinforced. And resonance frequencies are called formants.

The **values of the formants vary with the dimensions of the vocal and nasal tract**. They change also with the shape; I change the positions or values of these resonance frequencies. If I do “aaaeeeiiiooooooaaa” I am changing the formants, the resonance frequencies with the shape of my cavities.



If I do “aaAAAAAAaaaaAAA” I change the pitch, not the resonance frequency.

https://dood.al/pinktrombone/

We can play with this in order to see which things change because of what.

How do we see the formant in the speech signal? What do we see to notice the different formants?

We **cannot measure the formants directly in the speech wave**. We need more tools. We can see that we have different vowels or that the speech signal we have has different vowels. What is the tool that is going to allow me to see the spectrum? We have already talked about the **spectrum** which is the representation of the frequencies the signal has. This software calculates the spectrum (in Cool Edit Pro – Frequency Analysis)

Now, the formants are the resonances of the oral cavity. They are frequencies which are favoured in the transmission. They are the frequencies that are going to be seen as having more power, more volume, so are going to be represented higher.

It is not always ease, to see them with our eyes, we need to have stable segments of speech. The formants are not high peaks. **The peaks in the spectrum are harmonics**. A harmonic is if I have a fundamental frequency of Fa = 100Hz, the harmonics are the frequencies that are multiple to this value. The first harmonic is Fa (100Hz), the second harmonic is 2·Fa (200Hz), the third is 3·Fa (300Hz) … the signal that is periodic with fundamental frequency 100Hz contains frequencies which are multiples, harmonics.

Now, I can see that if I measure a peak, the value is the inverse of the period in the wave signal. The fundamental frequency is the inverse of the period.

I can see the formants because the frequency analysis is not flat, if we do not have fundamental frequencies, we would see it flat. We look at the envelope, not the details (that are the harmonics). To find the formants, we look at the envelope. So, we see which parts are going out with more energy. Luckily, we have software that calculates this for us.

For example, Praat or Speech Analyzer.

With an unvoiced sound, we find a random structure and the spectrum is flat, with very low energy.

We call it F0 to differentiate it from the formants that could be: F1, F2, F3, F4, F5… (Normally we use 3 or 4).

Coarticulation: the influence of the surrounding sounds on the actual sound.

Nasal sounds

It is when the **airflow goes through the nose, or nasal cavity**. We can have nasalization when there is a nasal very close to the vowel. A coarticulation affecting the vowel. Nasal sounds are produced when we close the mouth, there is no way for the air to go through the mouth, so the air goes through the nose. Also, there are some coder and communication devices that reproduce very badly the nasal spectrum. Let's say badly designed for nasals. Some satellite communication, some reporters use at work. Through the nasal cavity. The nasal is also the cavity. It also has resonances. What happens with these frequencies of the oral cavity that now they get trapped (these resonance frequencies) in the oral cavity and cannot go outside anymore. These frequencies are not in the output. But with the nasal sounds, it is the opposite. With the nasal sounds we see anti-resonances. The characteristics are not the maximums but the minimums.

Acoustic phonetics

a. The alphabets

It is an area of study in the Phonetics science. They study the relationship between these acoustics' physical magnitudes and the linguistic concepts. Sounds are represented with an alphabet.

For example, IPA, International phonetic alphabet and it is very common, and all linguist use.

IPA Chart with Sounds | International Phonetic Alphabet Sounds

There are several proposals, and we have one in Europe that is SAMPA, inside this project they developed protocols and standard for databases, recognition databases, recording, etc. So, they developed the SAMPA code. The most extended in informatics. Code in ASCII and IPA symbol combined.

SAMPA computer readable phonetic alphabet (ucl.ac.uk)

X-SAMPA is the same but more extended (diacritics of the IPA)

b. Articulatory mode and place

In these alphabets the sounds are described according to different dimensions. For example, they differentiate between vowels and diphthongs versus consonants. Vowels and diphthongs are sounds that do not find an obstruction. This is the main difference between vowels and consonants. Diphthongs are just two vowels one after the other (linguists do not pay too much attention). In diphthongs there is also no obstruction in the vocal tract.

Another dimension is voiced vs unvoiced. Although all vowels and diphthongs are voiced. If we whisper, we will make the vowels unvoiced. Vowels are voiced by definition, though. Consonants could be voiced or unvoiced.

- /p/ is unvoiced. The obstacle is because of the lips, but not in the vocal tract.
- /b/ is voiced.

<table>
<tr><td>(Basque)</td><td></td><td colspan="3">Closing</td></tr>
<tr><td></td><td></td><td>No</td><td>Yes - Total</td><td>Yes - Partial</td></tr>
<tr><td rowspan="3">Voicing</td><td rowspan="2">Yes</td><td rowspan="2">Vowels
/a/ /e/ /i/ /o/ /u/</td><td>Nasals
/m/ /n/ /J/
Plosives
/b/ /d/ /g/</td><td>Fricatives
/jj/
Liquids
/l/ /L/ /r/ /rr/
Laterals (approximants)
/B/ /D/ /G/</td></tr>
<tr><td colspan="2">Affricates
/gj/</td></tr>
<tr><td>No</td><td></td><td>Plosives
/p/ /t/ /k/ /c/</td><td>Fricatives
/f/ /T/ /s/ /s'/ /S/ /x/</td></tr>
<tr><td colspan="3"></td><td>Affricates
/tS/ /ts/ /ts'/</td></tr>
</table>

Another dimension is if it is oral or nasal. We distinguish sounds being oral or nasal. For example, in Basque there are nasal vowels. For European we find m, n and ñ as nasals.

Another one is place of articulation: where is the obstruction found?

- Bilabial: p,b,m
- Labiodental: m, n, f

Manner of articulation: the way the articulators are placed (stops or plosives, fricatives, affricates…). In plosives, the articulators stop the airflow.

Also check this webpage: http://smu-facweb.smu.ca/~s0949176/sammy/

Acoustic phonetics: articulatory place

The vowels can be defined by 5 standard points. Position of the tongue: forward, middle, backward. And then is the opening of the cavity: minimum, medium, maximum. Then, the opening of the mouth: non-rounded, rounded.

The smaller the opening, the smaller is also the value for the vowel. This resonance F1 varies according to this opening (minimum, medium, maximum). This gives a value to the first format. The smaller the opening, the smaller the threshold.

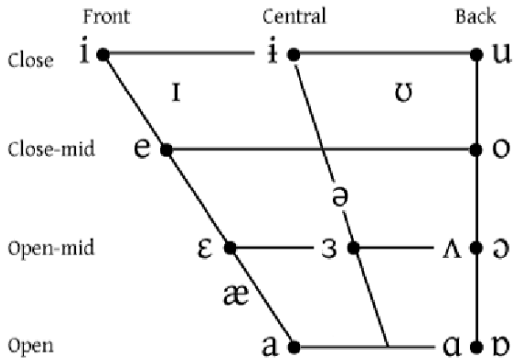
With the position of the tongue, we have the second formant. In a backward position of the tongue, the formant is low. With /u/ the second formant is low. And then it goes to /i/ and has a higher value according to F2, the second formant.

Diphthongs are arrows going from one position to another. And vowels are points in this map.

Differences in pitch do not make differences in formant. /a/ will always have the same values for the formant.

We simplify to determine a standard.

Formant (Hz)	F1	F2
/i/	284	2430
/e/	527	2025
/a/	689	1458
/o/	608	1215
/u/	243	770



4. Speech perception

a. The human auditory system

We perceived messages from our auditory system. We have some different auditory systems: peripheral, intermediate and central.

Something we can see in this pic of the brain is the stereo processing. Meaning we receive audio from one part and here (intermediate auditory system) the perception goes to the other part of the brain. They exchange the message.

The peripheral auditory system. We have the outer-ear that foes to the tympanums. The middle ear that transmits the sound to the cochlea and the inner ear (auditory nerve).

The auricle focuses sound waves through the ear canal. We gain in power with the sounds we perceive

The ear canal which the sound gets speed. We have some frequencies and some difficulties to go through the canal. Then, we have tympanums which is the end of the outer ear. Tympanums is a membrane, like in a drum.

The bones in the middle-ear are transmitting these movements. They do not move the same dimensions. There is an adaptation, because it is transmitted by a liquid. A change of impedance. The oval window is the entrance to the cochlea. The cochlea is a like a snail. Thanks to the semi-circular canals we have balance.

Task1 pdf:

1. Mistake: in the pdf the last point: "to obtain sinusoids of amplitude 0,25 and 1".
   20 log 1 = 0 dB
   20 log ((10 x Pref) / Pref) = 20 log 10 = 20 dB
   What happens if I double the amplitude? B is going to be 2 x A (two times A), in the formula will be 20 log 2 = 6 dB.
   If I amplify 6dBs, I am doubling the amplitude.
   If I go from 10^4 to 10^3 I have to divide by 10. Or substracting 20 dBs.
   If I want half the amplitude (multiply by 0.5) I divide by 2, I substract 6 dBs.

Pitch perception

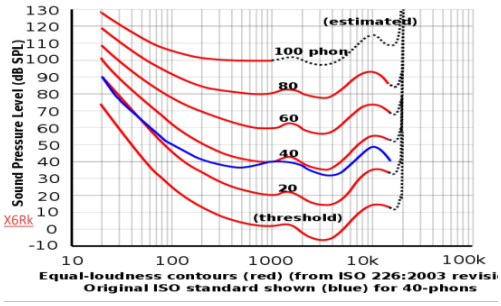
The pitch is the psychoacoustic perception of the fundamental frequency.

- It is based on the maximum excitation point in the Basilar Membrane.
- It depends on the frequency, intensity and the waveform of the sound.
- The pitch scale for individual pure tones is derived by adjusting the frequency of a tone until it sounds, as determined by the listener, half or twice as high as a second tone.
- The perceived pitch value can change if the intensity of the sound increases, even if the frequency keeps constant.
    - F<300Hz if the intensity increases, the perceived pitch decreases.
    - F>400Hz if the intensity increases, the perceived pitch increases.

There are different scales that are used for perception of pitch.

The **phon** is the level of loudness that you perceive. If from the starting 1 KhZ (standard ISO) you change the frequency, for instance putting a tone at a lower frequency, you will hear *AT A HIGHER VOLUME* than at a superior frequency. The more frequency you have, the *LESS* you'll be able to hear. That's why in the Audacity exercise, when increasing the frequency yet keeping the rest of the variables the same, you heard the sound less and less.

To manage these curves, if I play a tone of 2,000 Hz and if you're perceiving a volume of 40 phons, what volume would you perceive at another frequency? To do that, you follow the red line for the 40 phons to the right or to the left.



To know how many phons I would hear with the same dB as 40 phons, but this time at 200 Hz, I follow the 40 phons red line until I reach the 200 Hz in the x axis. I then check at which y axis position the 40 phons red line is at: in this case, 50 dB SPL. Therefore,

Auditory masking

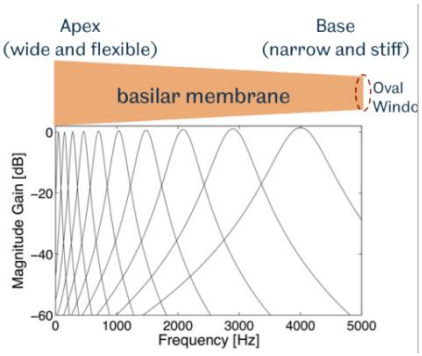
Frequency Discrimination:

It is different depending on which frequencies we are. At lower frequencies, we're able to discriminate better than at higher frequencies.
We're now gonna explain what this is about.
At home we can play with the link at slide 62.
If I can't discriminate between two frequencies, that means that those frequencies are equal to me.

There is also temporal discrimination. If two tones change very fast, it might be that we don't see the difference between the two tones (for instance, if I clapped my hands too fast, clapping two times but with barely seconds of difference, I might not be able to determine that I heard two claps; I might think that I heard a single clap, simply because of biological limitations of our hearing organs).

Critical bandwidths

Critical bands means that two tones (two sounds; we talk about "tones" because they're simple sounds that only affect one point in the basilar membrane). If you remember, in the basilar membrane the higher frequencies are processed first. Critical band is a range of frequencies where two tones cannot be distinguished: they are heard as the same tone, despite having different frequencies, because our basilar membranes are not capable of identifying the difference. The following graphic represents a series of bands, as lines, each of which represents a critical band:



Notice that the lines are closer together at lower frequencies and more separated at higher frequencies. This is because we can more easily discriminate tones of different frequencies if we're at a higher frequency.

[Slide 64; missing in eGela :-)]
In slide 64, we can see a table where an engineer defined the critical bands: that table is called the Bark Scale. In this table we can see the whole ¿audible? spectrum, starting from 50 Hz to

13,500 Hz, we divide this whole range of frequencies into bands. And, inside each band, there is a range of frequencies that we cannot differentiate one from the other.

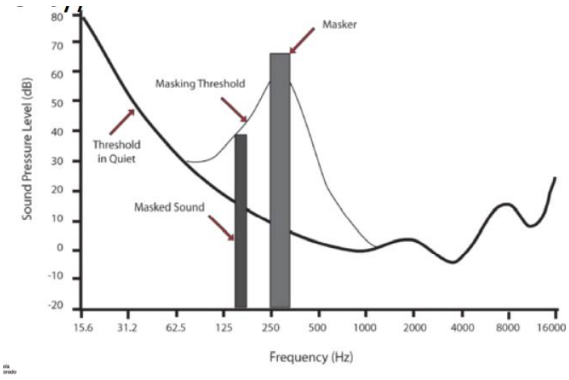
Auditory masking

Auditory masking occurs when the perception of a sound is affected by the presence of another sound.

- Simultaneous masking (*frequency masking*): two sounds of two different frequencies occur simultaneously, and one of them (*masker*) masks the other (*maskee*).
- Temporal masking: the *masker* and the *maskee* occur one after the other.

Simultaneous masking

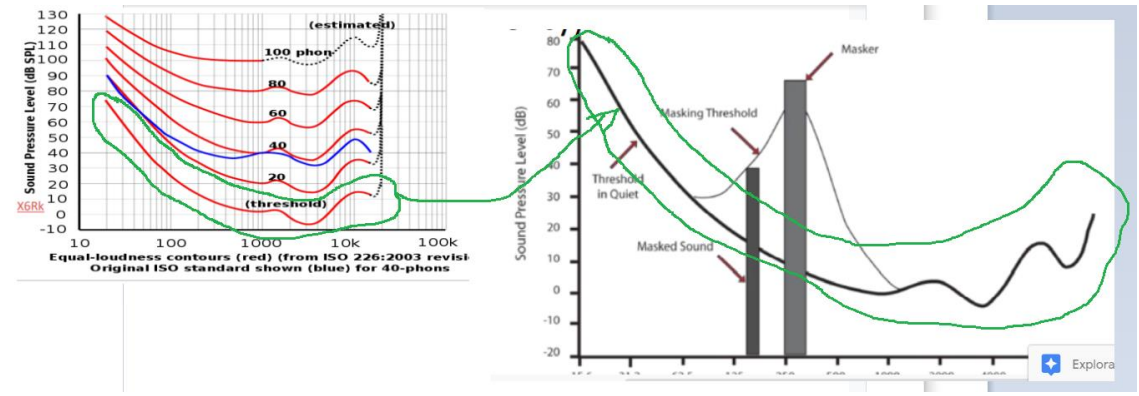
In simultaneous masking, we have two tones.



One of the tones generates a mask, called the "Masking Threshold", that makes it so any sounds that are below it are not heard. On the other hand, we have the "Threshold in Quiet", which is another mask under which NO sound can be heard at all. Notice that the mask changes as the frequency changes.
Any tone that is located under the "Masking Threshold" cannot be perceived. Therefore, any tone located under that mask, we will not be able to hear it. Therefore, in order to be heard it needs to have a higher amplitude, grow beyond the mask.
We cannot hear below 0 dB SPL, nobody in the world is capable of doing that. However, even if a tone is not at 0 dB SPL, it might be possible that it would still not be heard, due to the threshold in quiet, which changes depending on the frequency.

In MP3, the file format eliminates all tones that are NOT going to be heard, due to the variables of the graph above.

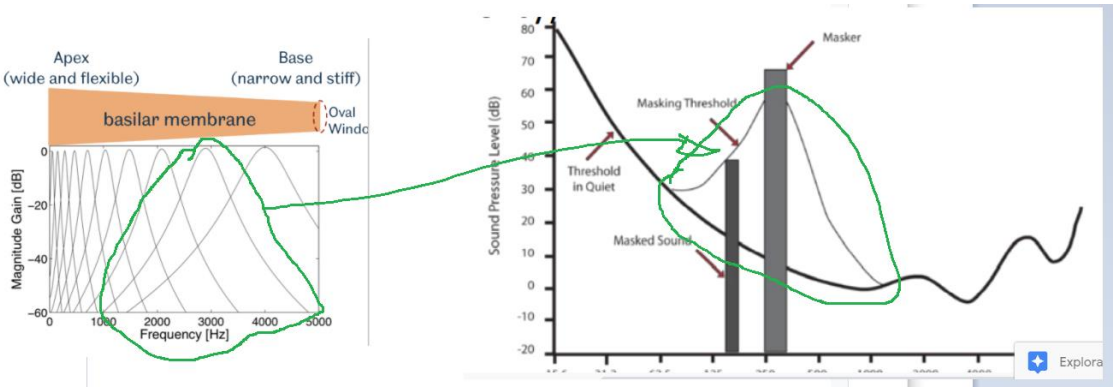
NOTICE, AND THIS IS IMPORTANT, that the "Threshold in Quiet" is the same one as this graph below:



1

[1] Dibujos proporcionados por el artista Miguel López, que muy amablemente ha decidido sus derechos de autor para que la edición de este documento sea mejor ejemplificada. Gracias, Miguel.

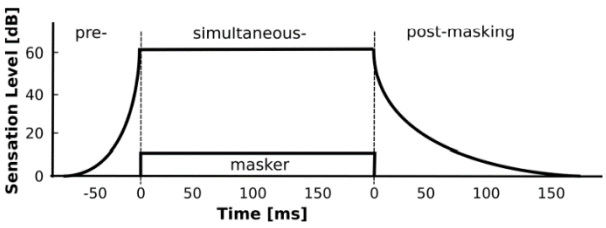
And that the “Masking Threshold” is the same as this other graph below:



So this **is really a combination of these two graphs**![2]

Temporal masking

Temporal Masking: a masker makes inaudible sounds that occur before and after the sound.

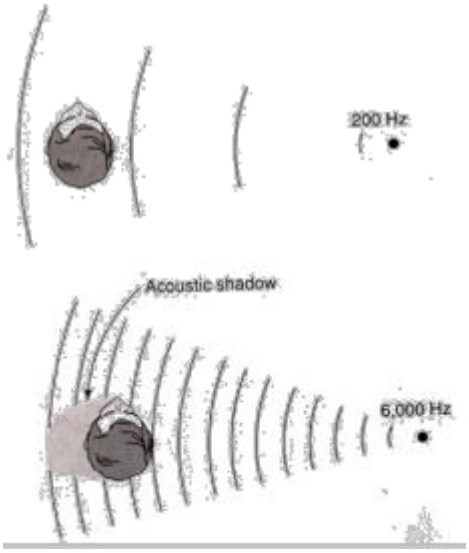


[I DIDN’T UNDERSTAND THIS PART!!!!][3]

Temporal masking is independent of the frequency.

Sound localization

Another important thing is stereo processing. Stereo means that we can hear from both ears, and the ear flaps (*orejas*) act as antennas with the acoustic waves. What do we need to know? We use both air flaps to locate a sound, to know the origin of a sound. We use different strategies depending on the frequency of the sound.
So, if we have a low frequency, we have low waves spreading to us.
The length of this wave is called the Wavelength. At lower frequencies, the wavelength is longer. At higher frequencies, the wavelength is shorter. The waves are smaller.



[2] Notas añadidas en el manuscrito original del artista Miguel López que se mantienen en esta producción, en relación a sus creaciones artísticas.
[3] El autor muestra su preocupación y desesperación ante el concepto de *temporal masking*. Esperamos que haya sido realmente temporal y actualmente se encuentre en mejores condiciones. Bendiciones, Miguel.

- If we have a high frequency, we don't know where it comes from. In order to locate the origin of the sound, we do a difference between the intensity of the sound in one of the ear flaps versus the other. This creates an acoustic shadow ("una sombra de onda").
- With lower frequencies, we don't have to do that, because in shorter wavelenghts the waves are located more or less in the same position in both ears. (Further description in slide)

We use both strategies, simultaneously, to locate the origin of sounds.

Other perception effects

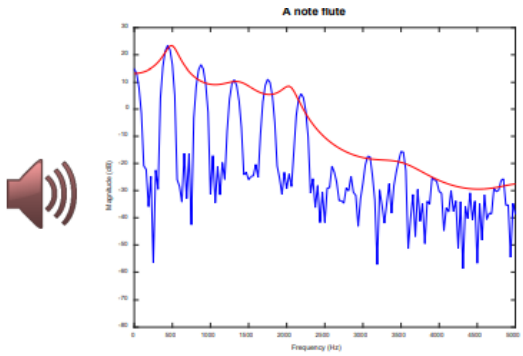
There are many perception effects, due to the limitations of our audition system and our brain. This slide contains some examples of it.
For instance, the "McGurk Effect" tells us that what we see can influence what we hear. So, we can identify a "p" sound DESPITE, in the description of the sound, its frequencies telling us that it is actually a "b".

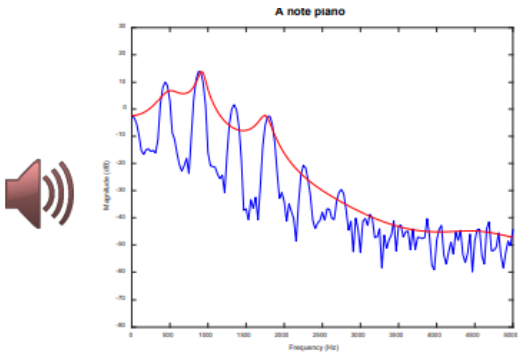
Speech perception

Physical Magnitude	*Perceptual properties*
Intensity (sound pressure level)	Sonority
Fundamental Frequency	Pitch
Spectrum	Timbre
Onset/offset	Temporal control
Phase difference	Position

The fact that physical magnitude and perceptual properties are so mixed up means that the perception of sounds is influenced by it. These factors are related to each other, but its relationship is extremely complex.

These two sounds below are both the same musical note, but played by different instruments (a flute and a piano). The difference in their sounds lie in their timbre:





[END OF THE PRESENTATION]

[3-minute break]

[WE CAN DO TEST-Lesson1 in eGela, AS WAY OF TESTING HOW WELL WE DO IN THE FINAL TEST]

[We seen an evaluation rubric for laboratory practices]

[We start with MATLAB, practice in eGela: Missing Fundamental Effect with Matlab]

[We start with "Basic concepts about signals & systems"]

- Difference between Signals and Systems: explained on the slides.

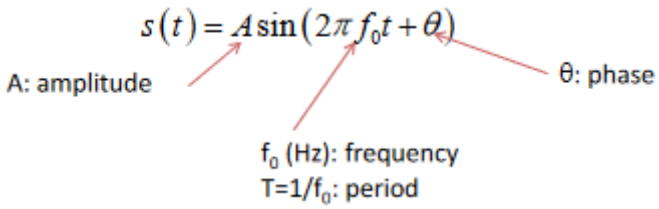
- Analog to digital signals → Explained in slides, more or less, if you need more info please let me know, my mind is too exhausted for now hehe
    - Sampling + Quantization = Digitization
    - A sound wave, containing speech, is a collection of samples
        - There are x samples for each second (called Sampling Frequency). The more samples, the more realistic the sound will be, BUT THE MORE STORAGE IT WILL TAKE.
        - Each individual sample is a collection of points that has an integer value. Each sample is a data value that has a number of bits to be represented with. Usual numbers are 16 bits used per sample, or 24 bits used per sample. This value is called the Bitrate.
    - If we have **Stereo sound** (2 channels, same sound twice!!!), we have a **Sampling Frequency of 44,100** samples per second, and 16 bits per sample. The sound we want to store is 1 hour long. (This is the Standard CD format storage) ***HOW MUCH DATA STORAGE, IN BITS AND IN MEGABYTES, DO YOU NEED TO STORE ALL THIS SOUND***?
        - (16 * 44100) * 3600 * 2 = 5.080.320.000 bits for 1 hour of audio
        - 5.080.320.000 / 8 = 635040000 bytes / 1024 = 620156,25 KB / 1024 = 605,621337890625 MB

How to calculate the Megabits per second.
One thing is Megabits, and another is Megabytes. Do not confuse:

1 byte	8 bits
1 kilobyte	1000 byte
1 Kib (kibibyte)	1024 byte
1 Megabyte (MB)	10^3 Kbyte 10^6 bytes

Basic concepts about signals and systems

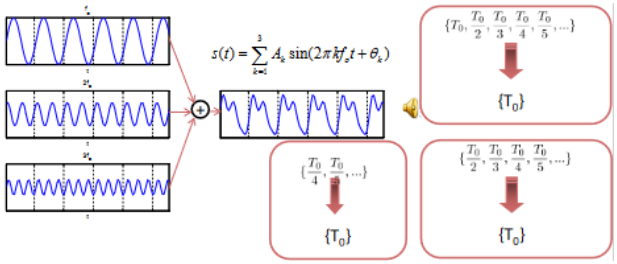
Basic signals: Periodic signals



One of the most important ones are periodic signals. Because the speech wave have many segments that are semi-periodic. The basic one is sinusoid. We already know the sinusoid and its characteristics: amplitude, phase, frequency.

Sometimes we are going to see sinus and other cosines, they are only different in the phase. The period of the signal: the higher the frequency, the smaller the period. They are inverse.

This is not a sinusoid, but it is periodic. We define the minimum period as the piece it repeats itself: fundamental period. The inverse of this fundamental period is the fundamental frequency.



We have here three sinusoids and we add the three sinusoids. We sum them and this is the mathematical expression of this sum. Sigma means addition. We have the amplitude (Ak) and this k can take the values: 1, 2, 3. So, from 1 to 3. And then the sinusoid that has the fundamental frequency of f0. And the phase from 1 to 3 sinusoids. The frequency f0 it is multiplicated by the k. first signal k =1, the second k=2. So, you have to multiply by 2.

The fundamental period is T0, in the first signal we can see it. In the second signal the fundamental period is doubled. However, the fundamental frequency is half the previous one.

We can see that if we do addition from the first and the second signal the result is: the period will be added. It is going to be periodic mixed or blended. Then with the third signal: if this is three times this frequency, the period in the 3rd is going to be three times the period of the first signal. I have added all of them the fundamental period is going to be the smaller one, the smaller frequency. [4]

[4] I think that here the teacher got confused.

We have added signals with the fundamental period of the first one. We take the first one as reference. All of them are multiple from the first. This also is valid when I remove any of the other signals.

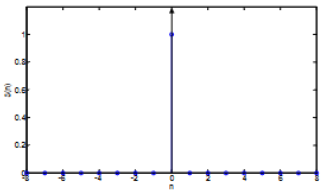
For example, if in this example I add the first one and the third one, they have fundamental period T0 anyway. If I ignore the first one, it does not matter, because it has fundamental period T0, because that's how we have calculated it and it still has the same fundamental period.

The least common multiple of the frequency components is perceived as the 'tone' of the signal. The least common multiple is in this example the third signal. It is multiple of the two periods. And this is the lowest (?)

Basic signals: the unit impulse signal

0 where n≠0

$$\delta[n] = \begin{cases} 1 & n = 0 \\ 0 & rest \end{cases}$$



The signal is always a function. So, we have one independent variable, and it is an integer number. So, this variable takes the values: -2,-1,0,1,2…

This signal, the unit impulse, is a very simple signal. Only takes one value and this value is in the origin. And it is called Delta. Delta of 'n' is a signal that is equal to 1 when n is equal 0 and it is zero for any other value. We will see the usefulness later.

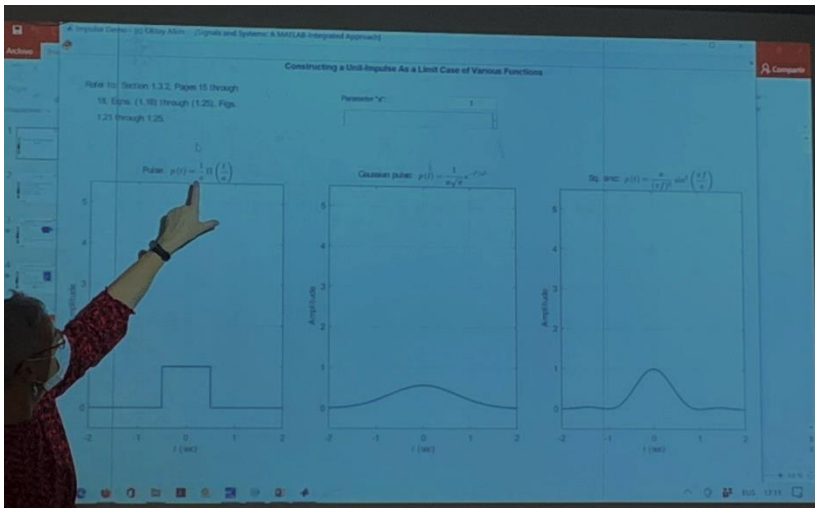
The slides that have a STAR in the title are optional. I am not going to ask these slides. I do not recommend ignoring them, but it is not going to be compulsory to understand them.

What is the particularity of this discrete signal? N is a discrete number, but we know that this will come usually from another analogue signal that has been samples. Come from sampling an analogue continuous signal.

What is the equivalent to this impulse signal? We design a signal that only takes one value for one instance of time. Because this is a continuous value, we define a function that only can take one value. t=0. Outside this value, it is zero. And at that point it is infinite value.

But with the particularity that the area of this function is 1.

Now we open Matlab:



And we can see that the first that it has area 1. A is value 1. The base is a, and the high is the value divided by a.

This rectangle becomes high when we change A. when this value A is *very very* small, the rectangle is *super super* high or infinite.

This is a way to say that I can define the function as the limit when t tends to zero of the function that follows. There are many functions that could be used to get this delta function.

For example, we can use as well delta functions and the third one that I do not know the name.

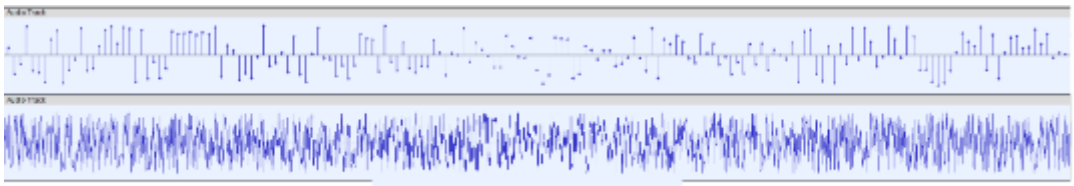
You should be able to see the equivalence between the complex mathematical explanation of the signal and the "easy" one.

Another basic signal: noise

It is very different. It is a signal with a random nature. This means that it is not deterministic, the values of the signal of noise cannot exactly be determined by a formula, like in the case of the sinusoid. If I write the sinusoid formula and I give a value to 't', I know the result. I give 1 to A and the f0 with 100 and a value for t and we know the result. That's why it is deterministic.

But not noise. These signals are called random or stochastic. It is the same as gamble, more like elegant, but the same.

This is a set of samples, random samples:



Segment of white gaussian noise

The only thing that changes is that I have put zoom, they are the same. Noise signals exist both for discrete signals and continuous signals[5]. These signals, what can we do with them? How do we work with them? We model the signals statistically. One of the most important noise is the Gaussian noise. What does this means? These values (that the signal takes) have the Gaussian probability function. The values have a Gaussian distribution. Means that all the values closer to the mean are more probable than values closer to the edges (-infinite or +infinite). This is what Gaussian means.

White Gaussian noise is the glottal signal. For example, when we do "aaaah" but whispering, it is white gaussian. Try it with MATLAB. Noises differ also in their frequency components.

Basic operations with signals

[5] When we use 't' we have continuous time signals. We differentiate these cases because they are different. This could be continuous signal. For example, the first picture of signal (on top) of the slide is discrete, a sequence of noise. The second is represented as continuous "because it looks continuous" (Inmaculada Hernaez,2021).

Amplify: we can do this operation in both domains. When we amplify a signal, we change the volume. But the mathematical representation is the one in the slide.

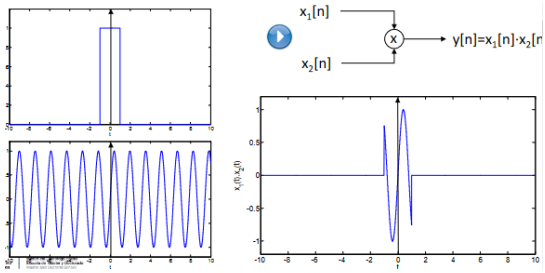
We can also do addition of signals. How do we do addition? We synchronize the signals in time, we draw them, align them with the time. And then we are going to do the addition of signal1 and 2, this one is 0 from -2 to 0 and the second the same, so first part of the resulting third signals is going to be 0. But if you have 1 in the first signal and 0 in the second signal for the same slice of time, the resulting signal in this period would be 1. And so on.

The **addition** of signals occurs very frequently in real life. For example, where are going to find it? The most common situation is a noisy signal. When we get in one hand speech, and there are noises in the room (electricity, people talking, interferences… think about a restaurant). We record it in a microphone, for example. And we can see:

$s_n(t)=s(t)+n(t)$
Noisy signal

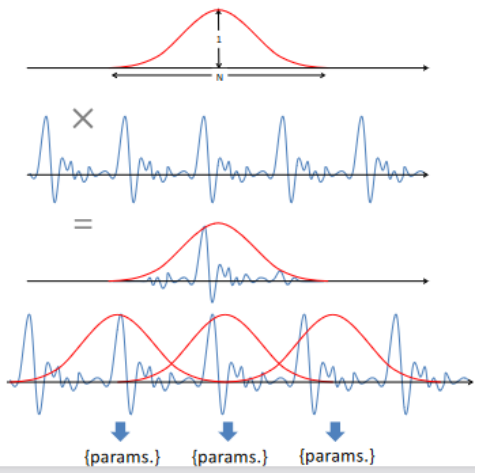
MODEL: an original signal + noises, interferences or whatever. For example, we have echoes or reverberations. Or some additions that are made on purpose like stereo effect, we have an equipment that generates different signals.

Multiplication



We align the two signals and once they are aligned in the zero point of time. We see the picture of the sinusoid. We multiply point by point: 0 up to -1, so everything multiplied by 0 is 0. So, we get a 0. From 1 to infinite, the signal is 0, we multiply the two of them and we also have 0. And from -1 to 1 we have one. So, when we multiply, we get this piece of the sinusoid.

I have this **window** (the one we can see in the picture that is a big rectangle): from -1 to 1. That's the only part where you can see (or multiply) because you have values. I am in this window, and I see things passing by, a car, a sinusoid, whatever you can imagine. I have this window with length = n, it can be expressed in time (like milliseconds), and I multiply this to my signal in order to select a segment of my signal: My window could be **rectangular or another shape**. So, we can use different kind of windows to **remove border effects**.

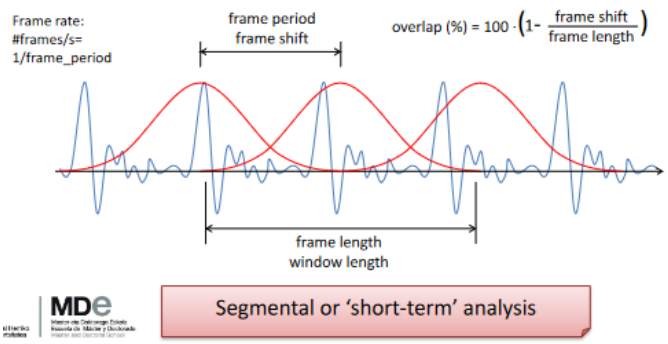


You can see that in the previous case does not seem like a sinusoid anymore, it had discontinuity because of the borders. But look at this one, it does look like a sinusoid. Window is a very intrinsic distortion element.

So, when we select the window, we select the segment of the sinusoid or the signal. Why do we select the segment? We do it in order to **extract information**. If I want to analyse a piece of speech, it is clear that we cannot mix different parts. Because the characteristics are going to be mixed. For example, I want to know if it is you talking or me, perhaps I want to mix the characteristics. But in many other situations, I do not want to mix the sound for speech recognition. We do **windowing** because we want to see the characteristics of the signals segment by segment. This is called **short-time/term analysis**. Very short segments where in these segments the signal is stationary, and it keeps its characteristics.

Coming back to the last picture: I move my window and then I get a new set of parameters. And I can do it again and again and do the called short-term analysis.

Every time we move the window, we get a different set of parameters, this is called the frame shift. Also frame period. I have the window and I move it from here to here. This segment of time, and then it is called frame shift. When I move, I can or not overlap the windows. Look the first window in red and the second, they are overlapped. The second window takes a little bit of the space of the segment taken in the first window. Also, we could not overlap, it is our choice. I can choose how many frames per second. The closer the frames they are, the more frame rate they have. In speech we have a usual value: 30 milliseconds. So, I move these 10 milliseconds and the next window is 30 milliseconds, and I am going to move 10 milliseconds.



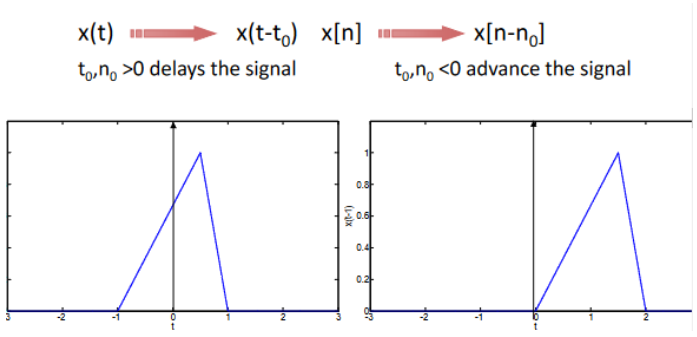
In speech, 30 milliseconds are going to like 3 or 4 periods, more or less. I move and the usual overlap is 66%. I have 30 milliseconds and I move 10 milliseconds over and over again, so I do not lose anything. If my window is so small, I am missing the period. If I only see just a small piece like this, then my analysis will not give me a periodic signal.

Data of the example:

- Frame shift: 10 ms
- Frame rate: how many frames per second. 1/FrameShift = 1/10ms = 1/10^-3 = 1/0,01 = 100 Frame/s.
-

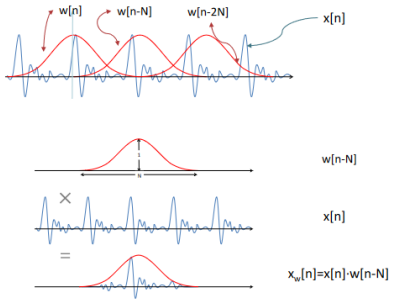
1 divided by the frame shift is the frame rate: how many frames per seconds.

More operations: delay



We need to know this representation, we have *again* a window (well or not, it is just an example). I have the first signal, and I want to represent it but in another time instance. How do I represent this mathematically?

If the first one is x of t, the second is x of t-1. This parameter is the delay. It is written in the way that it means delay. If we have x(t+t0), the signal is advanced. But x(t-t0) is delayed, here it happens at t=1, so it is delayed. If the parameter is positive, so t+t0, moving the signal before in time, instead of beginning at -1, it began at -2, for example.

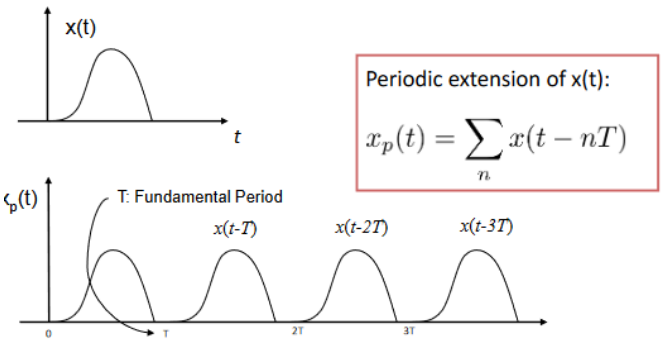


We have that w[n] is the first window and then x[n-N], it is delayed. Xw[n] this is the expression of my windowed signal:

Xw[n] = x[n]·w[n-N]

So, I have my signal and I want to select the point t0 and from that point on. I have my signal x(t) and I have my window w(t). but I have to put my window in the point in time I want to. Simply I just make w(t-t0), because this is where I want to put my window and then I multiply them. So, I have selected the new window of my signal. Also, it is my windowed segment. Of course, it is only the selection. The window of the signal depends on the signal and on time. I want to select another segment, so t – t1, t – t2… I will get any windowed of my signal.

Periodic extension



We take a basic function, or signal, could be for example a rectangle or triangle, but the one in the slide is more realistic. Periodic extension: a periodic signal has a period that repeats itself, so we extend the period. We delay the signal by T, by 2T, by 3T… and also -T, -2T, -3T… and we generate the periodic extension of x(t).

Addition of 'n' that is an integer number. What are we adding? X(t) plus minus 1 t, then xt minus 2t… and so on. Which is the period of this periodic signal? T. we build a periodic signal with fundamental period T from the basic signal which is not periodic.

Suppose we are programming, and we want to generate a periodic signal. We take a period, a basic signal, basic function, and we do like from zero to one thousand. And we use the formula of periodic extension of x(t).

Glottal signal

To model the glottal signal, output of the vocal cords, we use the basic glottal pulse. In the model we have a mathematical function. We have a vowel of 3 hundred milliseconds. We want to specify, we put 100Hz as fundamental frequency, it gives a fundamental period of the inverse of the fundamental frequency: 1/100. So, fundamental frequency is 0,01 s → 10 ms. I am going to take the formula of the glottal pulse and I am going to multiply the period for 10ms and it is going to be how we model the resulting signal.

Properties

Sampling is like windowing, but I take only one sample. With windowing I have the whole signal and I take a piece. With sampling I take one point in the signal.

X[n]delta[n] = x[0]delta[n]

Because the delta function only has one sample, I multiply the two of them and then I get the point which at delta is defined. Imagine delta function is defined only in zero. We also have x[n] that is defined discretely. And we only have one point, when n is zero in both functions.

When n is zero in x[n] it has the value 3. When n is zero in delta[n] is 1. So the result is:

X[n]·delta[n] = 3·delta[n]

This is done with discrete sequences. The same could be done with continuous time signals, any of them can be written as a combination (it is not addition, careful) of the function of delta functions:

X(lambda)delta(t-lambda)d lambda

These rectangles are always the same function, and they are delayed one from another. The rectangle will become ?? in the limit. It will give us a stair, it is an approximation. In the limit, when I reduce the pulse width, the ??. it allows us to get the combination of all signals.

Decomposing a signal into very basic signals.

3. Fourier transform

$$X(f) = F\{x(t)\} = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \qquad x(t) \xrightarrow{F} X(f)$$

Fourier transform of x(t)
Analysis equation

Time (s) — Frequency (Hz)

$$x(t) = F^{-1}\{X(f)\} = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df \qquad x(t) \xleftarrow{F^{-1}} X(f)$$

Inverse Fourier transform of X(f)
Synthesis equation

Time (s) — Frequency (Hz)

Why treat signals and systems in frequency? Easier to be analysed. The Fourier transform has this expression. The first thing to know is called Fourier because of a mathematician. It is called transform because it transforms one signal into another, one function into another function (mathematically). The most important part is that this transformation changes the domain, the variable. In the original one the variable is t, and after the transformation the variable is going to be F, frequency. A very important property is that we can go back to the original signal. It is a transformation that could be reversed. Why do we change the domain? Because it is easier. So the right part of the first

transformation that we can see in the picture: does not exist. It is created by us, that is why also the reason why we can go back to the original signal and we do not lose any information.

The Fourier transform

Is a function of frequency (Hz): X(f). for example if the function is y, we use capital letter y. for each value of f (frequency), measures how similar is a signal x(t) to a sinusoid of frequency f. and it is function of f. I can interpret the value of X(f) at 100Hz

The amount of frequency of 100 Hz is this in the signal, there is a sinusoid inside the signal. Also there is a sinusoid of ??

If the Fourier transform is zero at one frequency it means it does not have values for that frequency.

Recover: this similarity is not very mathematically strict. This is just colloquial language. Strict language is: the square of module of the mathematical function is the measure of the amount of energy/power in the signal at a specific frequency f. Spectral density: power of the signal.

The module of the Fourier transform: is called the spectrum. When we do it with Speech Analyzer for example, checking the spectrum, it is indeed the Fourier transform.

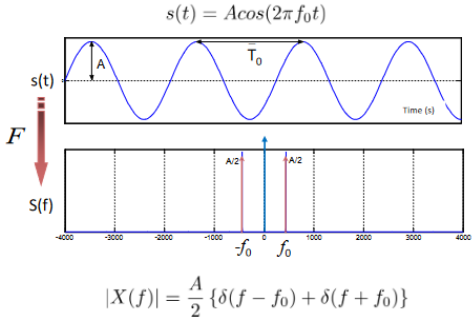
This has a lot to do with the sinusoid I told you, we are going to ignore the phases. The phase is the one that appears here in the Fourier transform. And this causes that the result of this is a complex number. We only are going to work with the module, okay?

Fourier transform for continuous time signals

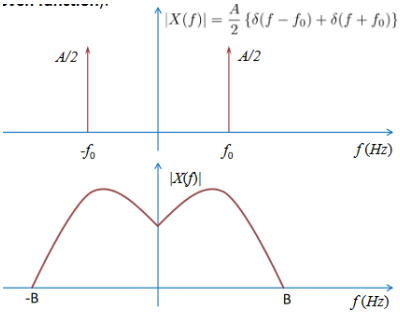
The first signal is a cosine function. Now, I have a signal and the frequency is zero. How many frequencies are in this sinusoid? 1. And the rest of the frequencies are going to be zero.

The Fourier transform is a delta function at that value. I have the delta function at this frequency (that is zero), and another one delayed(-f0). And then I have another delta function advanced (+f0). The Fourier transform eliminates the negative frequencies.

It measures the similarity with a complex exponential. So the cosine have ??



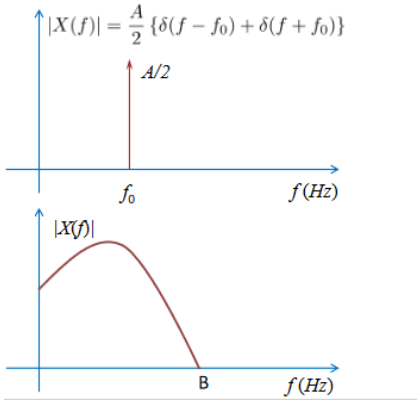
$$|X(f)| = \frac{A}{2}\{\delta(f - f_0) + \delta(f + f_0)\}$$



In a real signal, there will always be symmetry. If x(t) is real, because all speech signals are real, the Fourier transform will have always a symmetry. And in this the module will be a symmetric as well. Symmetry with respect to the axis. It is an even function. If it is symmetric why bother us about what is happening in negative values? We only take the positive. But in MATLAB for example you will get the whole graphic as well as in the picture on the right.

But the following image is what we are going to do: take only the positive values. Only the positive part.

How similar is this signal? The second contains from zero to B, that could 60 thousand Herz, for example.

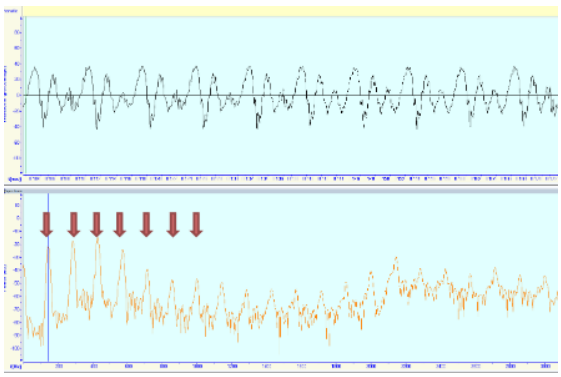


Fourier transform for continuous time signals: linearity of the Fourier transform

It is linear, it means that x(t) → X(f). If I amplify my signal, I get the same transform but amplified, bigger. For example, I have my sinusoid and I scale it by 3. Scaling linearity. Addition of two signals, I do the Fourier transform with both and add them together.

So, everything is equal in one side and in the other, with the basic original signal and with the Fourier transform.

How do we apply this? Now, we can do it more precise. The Fourier transform of this more complex signal (3 signal) gives me the original signal of the transformed.



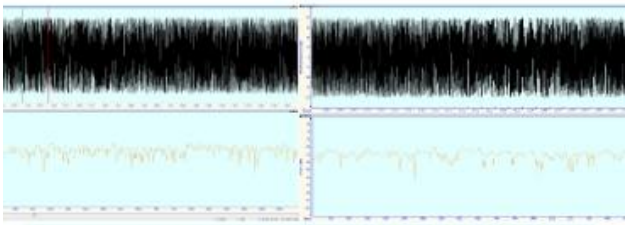
A more realistic example: I think this is a vowel. I do spectrum and we get this picture. This piece in speech Analyzer. In the parameters a window has been specified. We need to know which window has been applied. It draws the model of the FourierT. what do the arrows mean? These are the impulses, the delta functions we were seeing before.

The effect of the window it, and if we do not have an infinite signal, the effect is this picture. Delta function becomes this not so big delta. The arrow is the delta, the first arrow, the fundamental frequency is the first peak. The second arrow marks the 2 times the fundamental frequency and so on. Not all the components have the same weight or energy, but could you distinguish here that if I only see the first part of the arrows, I see multiple values of the fundamental, probably harmonics (okay, do not judge me, I am getting crazy now and I do not understand what the teacher is explaining).

There are several ways that I can measure the fundamental frequency. I know some harmonics are not there. Nasalization for example could have lower energy.

FourierT of noise

Being noise of random nature, its FourierT will strongly depend on the selected segment. I will also look very random.



FT of two different frames of the same wgn signal

With noise we have a problem because it is not periodic. If I do the FourierT in some points, the values are going to be random.

This is taking a window with defined parameters, so if I move the cursor, it calculates again the FourierT. It looks kind of random. When I see the spectrum like noisy, it is because the signal is noisy[6].

Why is it call white noise? White means all the colours are in the same amplitude in width light. White noise has the same amount in all the spectrum.

For noisy signal we do the Scan, averaging the spectrum of several frames. We should do this in order to get a good estimate.

Power Spectral Density: the spectrums calculated from consecutive frame…

Exercise:

- 1 sec
- Parameters: voiced or unvoiced, FourierT, the envelope of the spectrum… whatever.
- Frame rate: 100 frames per second.
- If we have N parameters per frame. We have calculated the FourierT which is usually calculated with the algorithm FT{1024]. For example, we have 1024 values in one N, and then multiplied by 100.

OUTLINE

4. Linear Time Invariant Systems

5. Filters and resonators

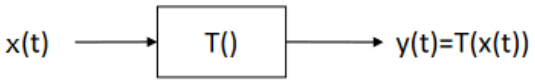
6. The source-filter model

[6] Momento a lo Mariano Rajoy.

LINEAR TIME INVARIANT SYSTEMS

We can calculate the FourierT of these signals. If I send a signal from one point to a another point, the path, the trajectory, can be modelled as a system. And it has some influence into the signal. Any device or any process applied over a signal, can be seen as a system.

We say that one system has one input and one output. The input signal is usually x(t) and the output y(t). we can also have a discrete: x(n) and output y(n).



The effect of the system is modelled from this 'T'.

Not any system can be modelled through the FourierT simply. We apply all the parameters and all the analysis that we are going to do. It is going to be applied over a particular kind of systems. That is Linear Time invariant System.

Let's see what this means. There are limitations. We apply our knowledge to linear systems and time invariant systems. We forced the system to be linear and time invariant.

2) Scaling property

– $T\{a\,x_1[n]\} = a\,T\{x_1[n]\} = a\,y[n]$

So, what is this? A FourierT is a linear operator, because the FT of the scale version of a signal, is the **scale version of the FourierT**. For example, the scaling property. If I have an output y(t) of the system x(t). if the input is scaled, the output will also be scaled. This is a linear system. For example, I go and buy some bread, give one piece, one piece is 1€, 2 pieces 2€, 3 pieces 3€… 10 pieces is 9€… OH, it is not linear!

1) Additive property

– $T\{x_1[n] + x_2[n]\} = T\{x_1[n]\} + T\{x_2[n]\} = y_1[n] + y_2[n]$

And the other operation related to linearity is **adding signals**. If the output to the signal is y1[n] and the output of other signal is y2[n]. If the input the sum of the two tones, the output will be the sum of the two outputs. This must be interpreted as I have 2 inputs and each of them has a corresponding output. Then I add the two signals before inputting the system, then the sum would be the sum of the two outputs.

- Expressing both at the same time:

$T\{ax_1[n] + bx_2[n]\} = a\,T\{x_1[n]\} + b\,T\{x_2[n]\}$

If I have an input very complex which in any way I can decompose in individual inputs. If I have one very complex input, a signal which is very complex, but through some analysis I am able to separate these components. Then, I can go let's say to the same system. These systems will apply some process, or operations. Then, I can input the first and I will get y1[n] and with the second input I get y2[n] and the final output will be the sum of the individual outputs. It is interesting in order to simplify a process.

- In a more general way

It is the same but generalised.

TIME INVARIANCE

Conceptually very simple: I will ask the system keep the way you are, do not change your properties while I am analysing. Then, otherwise suppose that I input the signal to the system and I get an output, and now later in time, I go to the same system and I expect to get the same output. I don't want a system like a neural network, I do not want a different output. We will ask the system to be time invariant. From one input there is only one output, invariant. Always the same.

Apart from stochastic systems or neural network (initialled with random values), in deterministic sytems we have time variance. It is probable that there is a different response, the answer can change with time. We do not want the system changed, its property, its response. Mathematically how do we express time invariance?

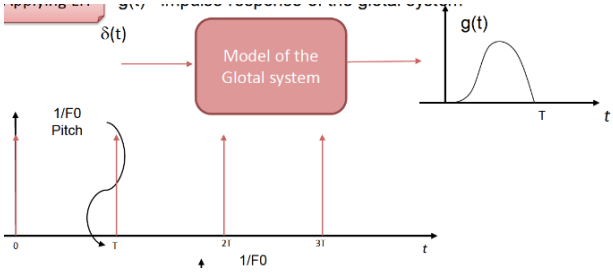
$$x_1[n] = x[n-n_0] \text{ produces } y_1[n] = y[n-n_0]$$

If I have this input x1[n] and then I get one output. Then I put the same input delayed at another instant, x[n-n0] this means delay. If I enter the system with an input that is equal but delayed, what must the output be? The same output as in the original but delayed. And this is time invariance.

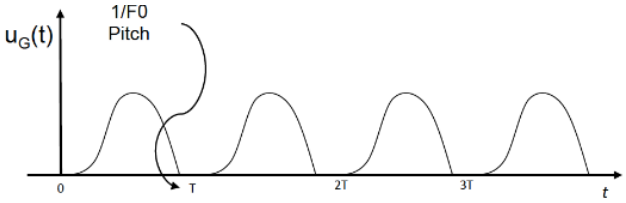
The response of the delayed version would be y[n-n0] = T(x[n-n0]). We expect this!

APPLYING LTI

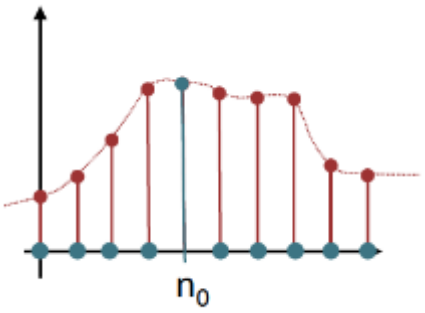
Suppose we have a model a glottal system. What is a model of the glottal system? Suppose we input our model with one very simple signal, one signal, one input. This is a mathematical function, but a model is also a mathematical function. So, we input our model with this impulse S(t) ((No es una 's' es una letra griega)). If I want to measure the acoustic response of this room, and then for measure it I try some impulse. For example, I clap and if there is reverberation, I will hear the clap as echo. The clap is an impulse and the effect is what I measure, I measure the response.



Now, if instead of having only one impulse… If another impulse comes t seconds later, I will get the same output t seconds later. For every impulse, being the impulse at any time, I will get the pulse g(t) at that time instant. So something like this picture. I can model this in a model of the Glottal system. The only thing I need to mathematically express this signal is this function g(t) (the pulse, the glottal pulse modelled this mathematical function) and of course the value of pitch, the tone. Can you see the power of the model? Or the mathematical model? Yes. I have applied in this last picture, I am getting the sum of the output and I am applying that it is time invariant, because the output of the delayed version of the input, is the delayed version of the output.



This is one of the slides I used, and it has a STAR. This is just to remember. If we have a signal x[n], this would be the red little balls. I could say this signal is the result of combination of outputs from different signals. This means only that a signal could be decomposed into very simple signals, and the inputs scale. I can decompose any input into a system like a combination of deltas, impulses. So, applying linearity if I know that the response of the system to a one impulse, I can calculate it as a combination of simple outputs.



$$x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau$$

This picture means continuous signal. With the integral, tao is delayed. Equivalent expression of continuous time signal. Remember the integral.

Suppose that I have one system and I can calculate like before, I clap, and I measure the output. The input is an impulse, and I will get as output some signals that are going lower and lower. These represented arrows, the effect would be h(t), discrete system. I can estimate, suppose I can calculate the output to an impulse. Then, for any signal because any signal can be expressed as a combination of impulses, if I calculate the effect transformation of x, now I have the input decomposed in impulses. I apply linearity, the output of the sum, is the sum of the output. Now I scale, x[k] scaling my input. Now I say okay, the summation of the transformation for this is the trans---

And now I apply invariance, the delayed delta is the same output but delayed. What does this mean? If I know the output of a system to an impulse. I can calculate through this operation, the output to any input. I go to my room that I am modelling, and I clap again, and I measure my impulse response. I can calculate the response.

$y[n] = x[n] * h[n]$ ← ***Convolution*** operation

This operation is called convolution and the star is a symbol.

Continuous time systems

$x(t) * h(t)$ ← ***Convolution*** operation

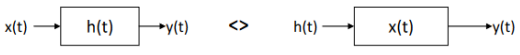
this is symbolically represented with the star. You do not need to know this formula to pass the subject, but this you have to identify. It is not sum, product, just CONVOLUTION. This operation allows us to calculate the output to any input knowing just the impulse response.

LINEAR TIME INVARIANT SYSTEMS: PROPERTIES OF CONVOLUTION (FALTAN PICTURES HERE)

From convolution to product thanks to FourierT. Now I know that I have one input, the system characterised by system response, and the output. The relation is that the output is the relation...

Commutative

$$x(t) * h(t) = h(t) * x(t)$$



it is commutative. A+b=b+a. the importance is that input and input response could be changed.

Associate: I can take the 2 first things, sum them and then go to the third one and so. Same with convolution, we can apply this property. It has very important properties: suppose I have one input to first system and the output goes to the second system. This is the same as having one only system which has an impulse response that is the convolution of the two. This also can be interpreted the other way around, if I am able to put the output as the convolution of the two systems. Imagine we have to communicate Bilbao to Madrid. It is very difficult to put one only wire. It was not possible to do one only path, frame, piece of communication system. We had to use repeaters, also today we have repeaters in radio communication systems. The input to the second system is the output from the first one. I can model the whole system by concatenating all the systems. I have to convolve all the responses of all the systems.

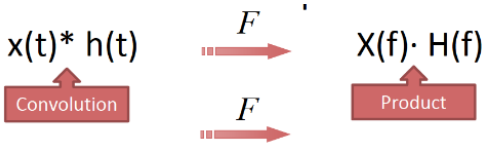
Imagine I have my glottal signal (from my vocal cords) and then I have my mouth, my mouth is another system, I concatenated them. Think about this in this way. Then, we have the effect of the lips, every piece is one system, and they get concatenated together.

Distributive with respect to the sound: this is another thing; I have one input in common with two systems. Two microphones. One only input but with two independent microphones. We sum them up, then this is the same as having one only system whose impulse response is the sum of the two. For those who love maths, we have the commutative property, associative and distributive and now: the neutral element.

The neutral element is the convolution.¿¿??

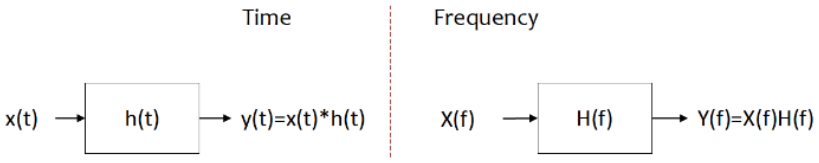
LINEAR TIME INVARIANT SYSTEMS: PROPERTIES OF CONVOLUTION

- FT of the convolution operation: Product!



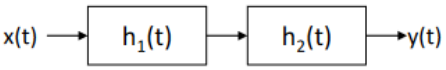
If we do the FourierT of the convolution, so we have these two signals, we apply the FT and each of them has its own FT. if I do the fourier transform of the input I get the product of the FT of the results, outputs.

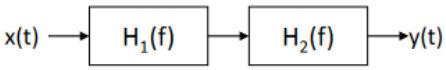
We have a system; we can model it with the impulse response. We can calculate the output through convolution if we were with the signals in the time domain. But if we go to the frequency domain, we can calculate the product of the FourierT and we can go back to time. We do not need to use the convolution, this is why we do not ask you how to do it, you just need to understand it. Remember that the FT is reversible, we can always go back from time to frequency and form frequency to time.



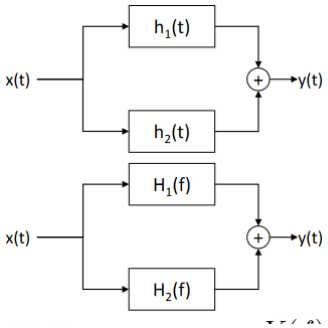
How does apply in the domain?

Concatenated systems also can be called cascades or series. If we connect two system in series, the FT of the impulse responses get multiplied. This is called the impulse response and the bottom one is the frequency response. And it will be the main characteristic of the system. Same as the impulse response but much more understandable.





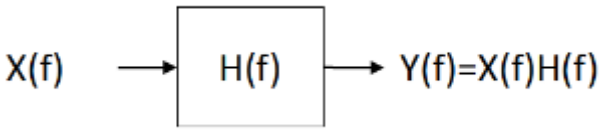
This is called connection in parallel is that we have one common input to all the systems, we can have two or more systems. And then the output are added.



FILTERS AND RESONATORS

Resonators amplify one specific frequency of the sounds (?).

FILTERS

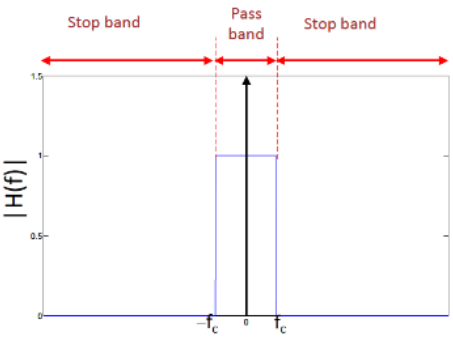


It is the way we use this property that we have just seen. That now we interpret the system fully in the frequency domain. Completely we are going to work only in the frequency domain. We have one input and we have a system, with a frequency response. So, a system that is designed to act to generate effects in the frequency response of the input signal, this is a filter.

Filter is a system that will keep some frequencies on the frequency components in the input signals and will remove some others. What is a filter for us? Something that leaves going through the systems some components and reject some others.

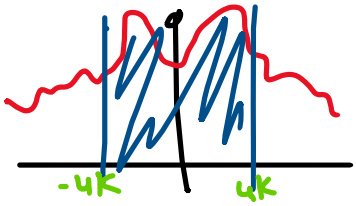
Think about the sun cream, it will leave some components of the light go through the cream and others not. Some components are rejected in this filter in order to protect our skin.

What is a filter in this context signal processing? It is a system, the FourierT is multiplied by this impulse response. Now we work on the frequency domain, we reject some frequencies. How do we decide? We have some components in the input and we want to reject this components and we do not want them in the output.

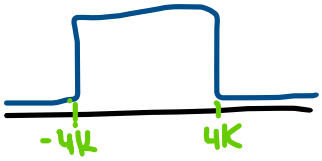


Suppose I have a signal because I am working in the frequency domain, I must obtain the FourierT the transformation that gives me the frequency components of this signal.

I want to eliminate the sides out from the area I have marked in the picture. I have to multiply by zero so I eliminate what is not blue.



When I do H(f) to the signal I want to filter I get outside of the blue part 0, and inside 1. I would be something like this:



This frequency is called cut-off frequency (*frecuencia de corte*). We can now see the usefulness of the FourierT and the power of the FourierT.

In fact, we can implement this filter in the time domain, but the design is in the frequency domain. In the time domain there are much better implementations.

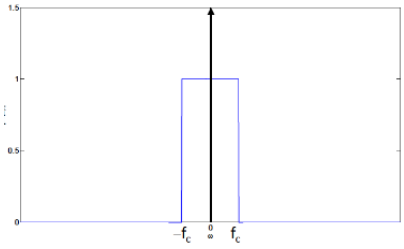
Pass band → the band of frequencies (or range) where the input frequencies go through. And then we have the stop band, they do not go through (see the paint that before -4k is zero and after 4k is zero again, those would be stop bands).

IDEAL FILTERS

We ignore the phase, but in filters is really important. 1 is pass band and 0 is stop band.

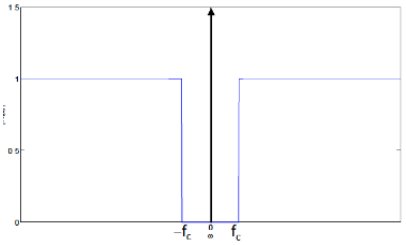
BASIC IDEAL FILTERS

This is a low pass filter, and the bottom picture would be high pass filter. They are the opposite. We want jus the high frequencies and we do not want the low frequencies. The cut-off frequency this time is going to be the frequency from 'fc' on and that would be the pass band.

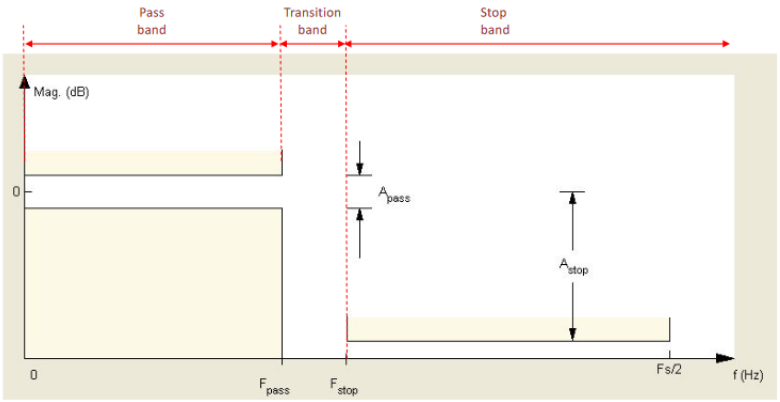


In band pass filter we let a band of frequencies pass through. (pictures)

Band-stop filters that they remove one part. (pictures)



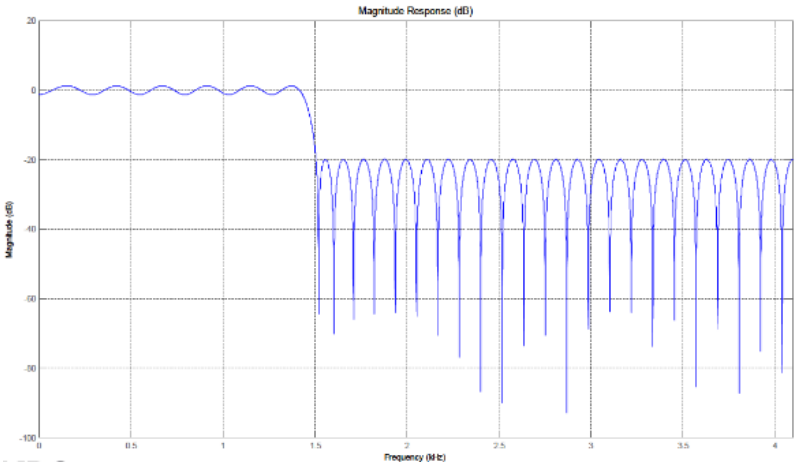
REAL FILTERS



It will be very difficult to get a zero in the stop band. So, we will allow some small signal there. And this is expressed here like attenuation. Okay, you can multiply not by zero, but by 0,00001. Not zero but almost. So, we could see a little bit higher than 0.

The same in the pass band, we want a constant signal, but this is going to be very difficult. We will have some distortion; we will specify also how big the distortion could be. And also, this change from 4000 (or 4K) for example at 1999,99 it cannot pass but in 2000 could pass. We must give some flexibility to these bands or limitations.

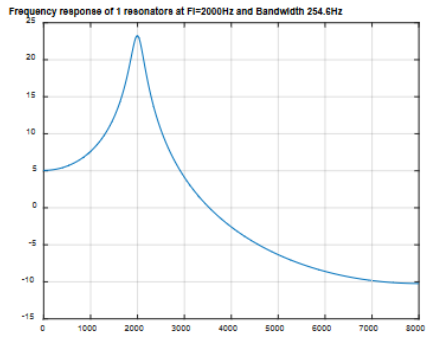
There is an App in MATLAB that you can use to design filters. This is how a frequency response to a real filter responds:



It is a low pass. This is a logarithmic scale. Usually they are expressed in dBs, this is way it does like drops. What amplification is 0 dB? 1. Logarithm of 1 is 0. So, 0 dBs (6dBs was double the dBs). Remember when we did 20·log(P2/P1) if they are equal is 0. The left part of the picture is the pass band, no amplification there. Then in -20dBs related to the first range. That range is going to be multiplied by a number, 20 dB under the input.

$$20 \cdot \log 10 = 20 \cdot 1 = 20$$

So the drops are 20dBs smaller than the input, ten times smaller in amplitude than the input.



Frequency response of 2nd order resonator

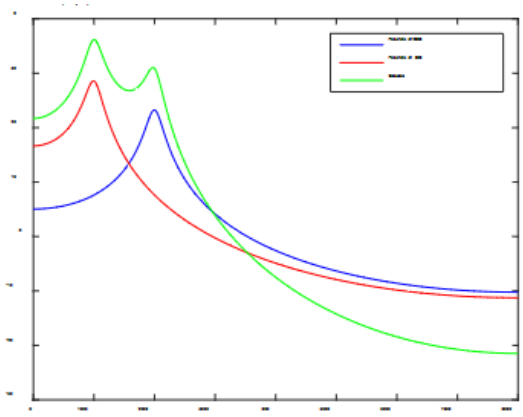
This is how a low-pass filter looks like.

RESONATORS

This is the frequency of one resonator. And the band could be more or less selected. A resonator is defined centred in one specific frequency that is resonator frequency. This is very discriminative filter.

Always centred on the frequency of resonance. Resonators are systems which favour a lot the transfer of energy at one frequency. The others are not favoured. There is always one frequency that has preference. We say to the other frequencies: "*You shall not pass!*" like Gandalf in Lord of The Rings.

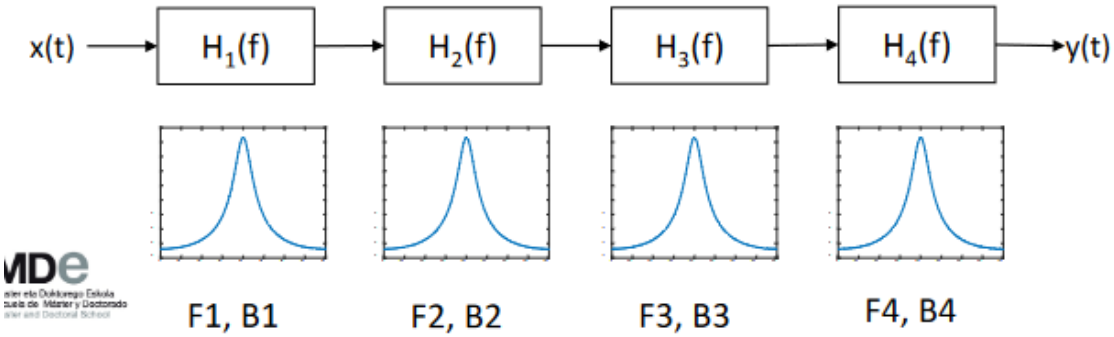
The bandwidth determines how discriminative is: 25.46Hz is more discriminative than 1248Hz.



Three resonators concatenated in series.

Suppose the red resonator has a frequency of 1000. And then the blue frequency of 2000. We concatenate them and the combination of the two would be the green line.

This is in logarithmic scale, when we multiply in the linear scale, in logarithmic scale we sum. Sum the two functions (blue and red) to obtain the green one. Does this bring something to your mind? The formants chart. For example, two formants. In fact, this is why we study these resonators, because the resonance formants are in the oral cavity. We can model easily the oral cavity as having frequency of four resonance. The position and values of these frequencies depend on the form of the cavity. The formants are in one position and this frequency of resonance are in another position. (?)



Chain of resonators. The frequency of the 1st formant, the frequency of the 2nd formant and so on. We will need to specify also the bandwidth, how discriminative are these formants. The resonators have two parameters: frequency and bandwidth. We could also use a model for the oral cavity with these resonators in parallel. Both models have been used for oral cavities. The difference is in the implementation. This way is easier than in parallel.

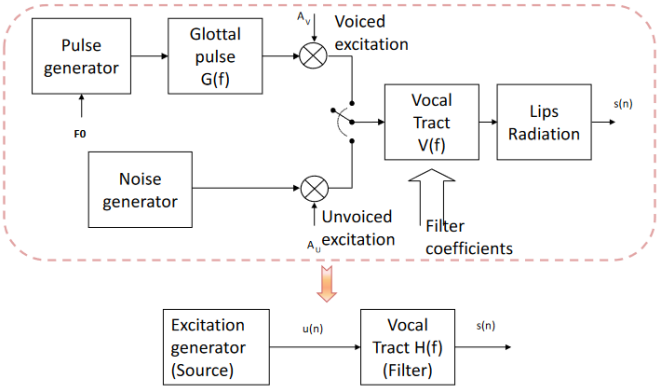
ANTIRESONATORS

You must know they exist. Anti resonators are a move of 180 degrees. It would be like an eliminated band filter. But it is not a band, it is only one frequency. Equal to the resonator but opposite. The frequency cannot go through, we use it for nasals. When we produce a nasal sound, we close our mouth, so some frequencies get trapped there and they cannot go out. They are 0. So, they are antiresonances.

THE SOURCE FILTER MODEL

SPEECH PRODUCTION: THE SOURCE-FILTER MODEL

This is the basic source-filter model. The most popular model for speech. What can we see here?



A simplify, we can see two big parts:

Excitation generator, used to genersate the excitation. The excitation tries to model the output of the vocal cords.

And the part of the vocal tract is from the glottis to the lips. Then lips radiation is just an effect, it is a very simple operation, we will not enter there.

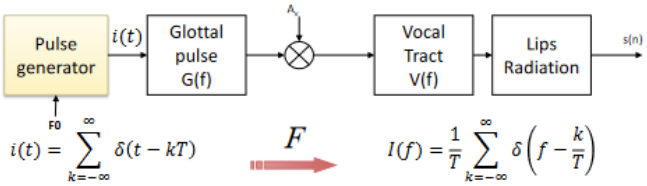
If you take the source sound and you put it in the middle of the room and if you put the speaker next of the wall, the sounds are going to be different. Lips radiation is usually into other models.

Vocal tract: which are the most important parameters for the vocal tract? What does the vocal tract do? Which are the parameters that define the form of the vocal tract? The formants. The formants are the values given to specify this filter. These frequencies will be determined in the filter coefficients.

The excitation: how is the signal that is the output of the vocal cords? We must differentiate two kind of signals: voiced (vocal chords are vibrating) and unvoiced (vocal cords are relaxed and the air flow goes through). The part for voiced is on top, the part for unvoiced is on the bottom.

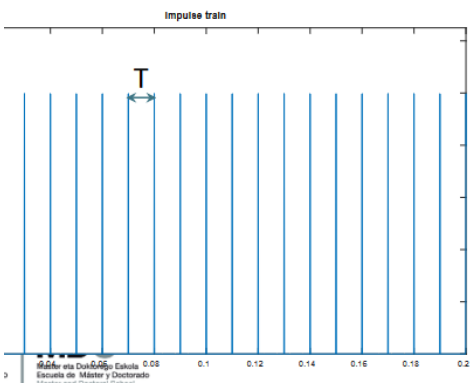
VOICED SOUNDS

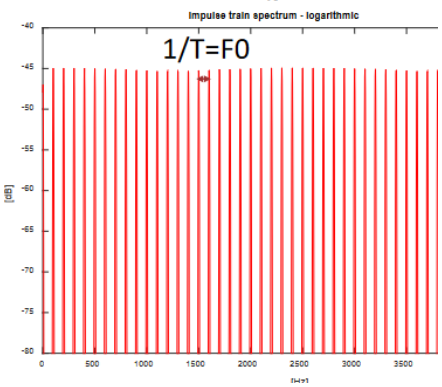
We have here, two systems put in cascade, in series, one after the other. So, this one is a generator, no input, it is configured with only one parameter that is the fundamental frequency.



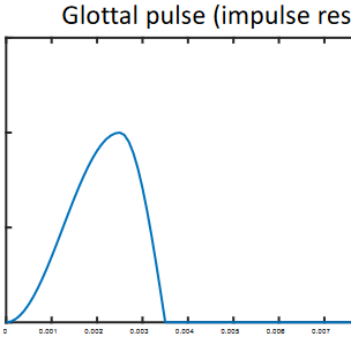
This first step should be called "impulse generator" because the output of this system is an impulse. One impulse every fundamental period. We want to model a segment of speech with a fundamental period of 1 millisecond. Then, we will configure the F0 of the pulse generator and this will generate every one millisecond, one impulse. We generate impulses which are space T, the inverse of this value, and forever (in theory), in practice it will not be until infinity, it will be until the time we choose, or we model. We get the train of impulses.

We can analyse the sounds in time (the waveform) and also in the frequency domain through the Fourier T. if we do the FT of the first part (train of impulses) gets the train of impulses.





GLOTTAL PULSE



Any linear time invariant system can be characterized by the impulse response, and then if we know this we can calculate the output to any input. What is the operation that allow us to get the output? Convolution of the input and the impulse response. We characterized this module as glottal pulse.

Mathematical formula allows us: to see the opening phase, the closing phase, and the closed phase:

This g(t) is the impulse response to this system.

When the input is an impulse the output will be the curve we see in the picture.

If this impulse is at another point we will get the same but in another moment in time. A train of pulses, of glottal pulses.

We have here the i(t) the impulse signal, then this is convolved with the impulse response g(t) and we get as a result the glottal signal as a pulse train sg(t)

We can analyse this in the frequency domain.

VOCAL TRACT

Here this parameter is really important (Av), the amplifier. We are multiplying the signal by this constant. The constant is Av (v from voiced). It is the amplitude for this segment that we are modelling or synthetizing when the sequence is voiced. Here we have the signal and the vocal tract is usually specified is the frequency domain with the formants. V(t) which is the inverse fourier transform from V(f). we must know: convolution in time, product in frequency.

Do not mix them. We have again impulse, impulse response and output.

At the input we have the previous signal: the glottal signal. Our face acts like a high speaker. S(t) always looks like speech.

UNVOICED SOUNDS

NOISE GENERATOR

That generates noise. Here is a segment of noise, and we a FT and we get a very noisy signal as well. If I were able to put a microphone in my phone: which nasal frequency is flat. I can put formants to a noisy signal.