Multimodal Interaction
Technische Universitat Berlin
Lecture 1.2.1
Speech Production

This week's topic is speech and hearing.

(Refer Slide Time: 0:23)

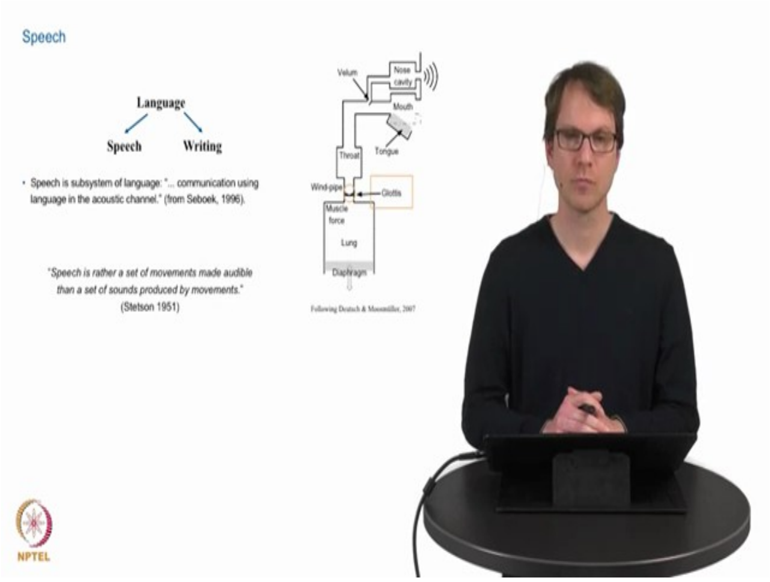
Speech & Hearing

Outline:

- Speech
  - Function
  - Phonetics
- Hearing
  - Ear
  - Psychoacoustics

This is the outline – I will start with Speech, talk a little bit about the functions and phonetics speech production, then I will talk about Hearing that means the Ear, and Psychoacoustics.

(Refer Slide Time: 0:37)



Speech

- Speech is subsystem of language: "... communication using language in the acoustic channel." (from Seboek, 1996).

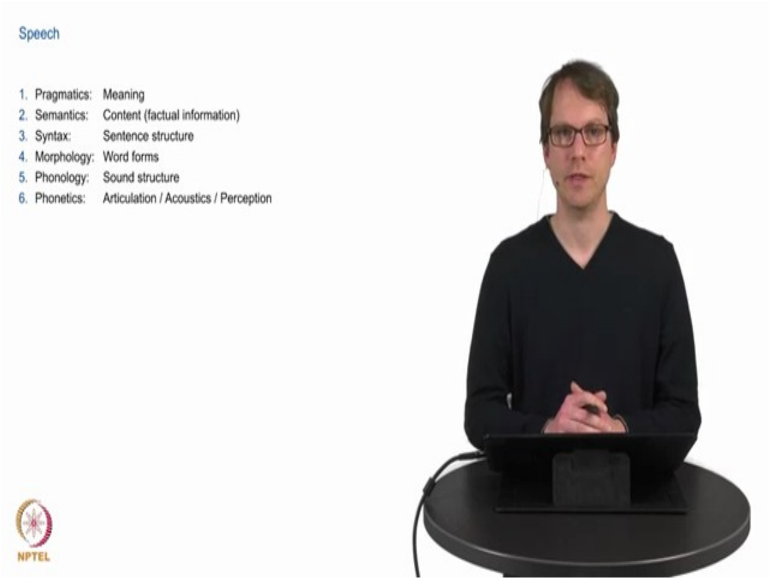
"Speech is rather a set of movements made audible than a set of sounds produced by movements."
(Stetson 1951)

Following Deutsch & Mossmüller, 2007

So, let us start with speech. Actually, speech is just one part of language use it is not the written one but the spoken one.

Here is a citation from Seboek who says, “speech is a subsystem of language, namely a communication using language in the acoustics channel”. There is an alternative way which is dedicated to the articulation of phonetics and speech and this is by Stetson who says, “speech is rather a set of movements made audible than a set of sounds produced by movements.”

(Refer Slide Time: 1:18)

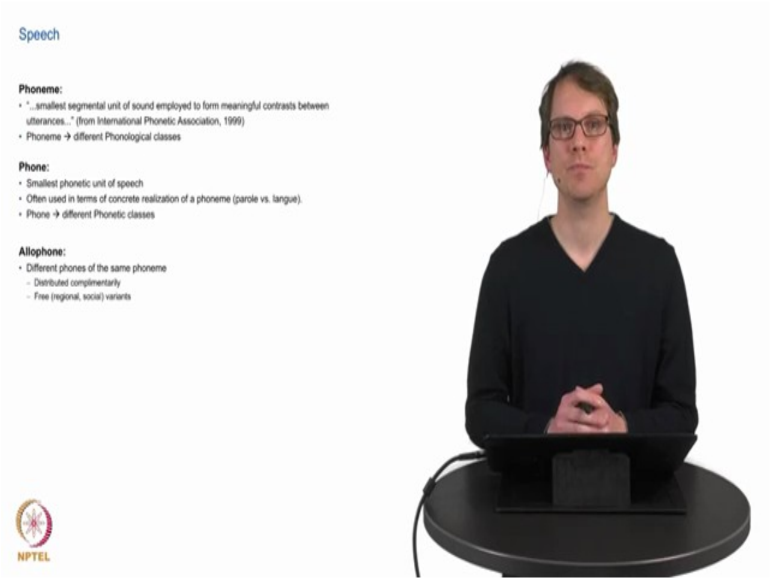


Speech

1. Pragmatics: Meaning
2. Semantics: Content (factual information)
3. Syntax: Sentence structure
4. Morphology: Word forms
5. Phonology: Sound structure
6. Phonetics: Articulation / Acoustics / Perception

So, let’s come to Phonetics what are actually Phonetics? If we have a look at this kind of linguistic hierarchy the first one and the most abstract one is Pragmatics. Pragmatics in linguistics deals with the meaning and intension of the user or the person uttering something. There is also Semantics this is about the meaning of single sentences or the factual information. So, this I about the relationship between sign, the letters for example in written language or sounds and words in spoken language as well and it is dedicated or assigned meaning, you also have syntax which is the laws and the structures of different words and parts of a sentence, there is Morphology what makes actually the different words in which sub groups or sub entities are used for building up words, there is Phonology the structure of sounds of spoken sounds and finally there is Phonetics. And this is how spoken sounds are articulated how they are manifested in acoustics and how they are perceived.

(Refer Slide Time: 2:41)



Speech

Phoneme:

- "...smallest segmental unit of sound employed to form meaningful contrasts between utterances..." (from International Phonetic Association, 1999)
- Phoneme → different Phonological classes

Phone:

- Smallest phonetic unit of speech
- Often used in terms of concrete realization of a phoneme (parole vs. langue).
- Phone → different Phonetic classes

Allophone:

- Different phones of the same phoneme
  - Distributed complimentarily
  - Free (regional, social) variants

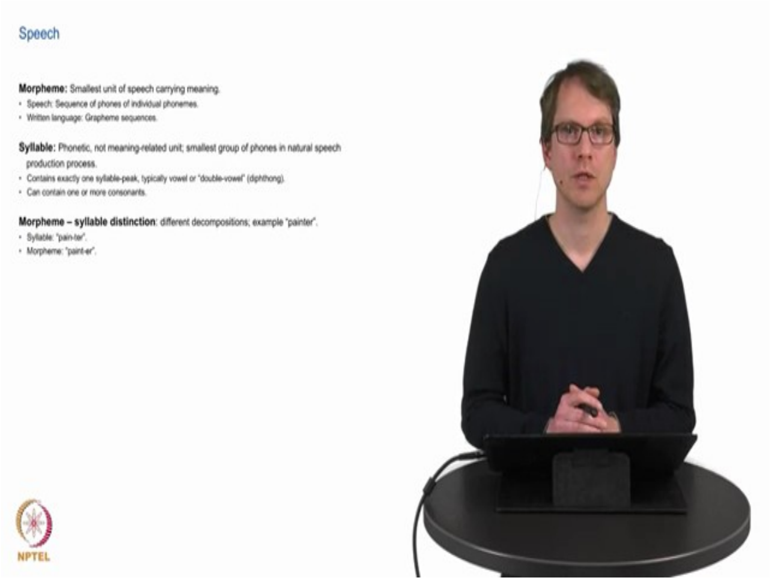
So, in Phonetics there are several important definitions one is Phoneme, so what is a phoneme? It is actually the smallest segmental unit of sound, employed to form a meaningful contrast between utterances' words. So a typical example would be the functional analysis between the two words fit and bit or even kit and hit. So here we have different words and really small acoustic and phonological differences in this, these are the single sounds and these make different phonemes. So, pee and tea in English, fit and hit so p and h would be different phonemes.

So, a phoneme is actually a functional class which is restricted to a given language, in contrast to this what is the phone? A phone is the smallest segment or unit in phonetics and speech. So, Phone is dedicated to or assigned to the phonetics level, not to the phonological level. You could argue that each phoneme has a phone that is used for producing them, this abstract linguistic unit.

But we also have Allophones. So, what are allophones? Allophones are different phonemes, phone classes articulatory class acoustic classes, which differ on a phonetic definition and classification from phoneme classes. So, we have two ways to find allophones. The most simple one is free variants, consider for example in English or German, the R sounds. So typically, there is no functional or meaningful distinction between using the back R or the front R but they are phonetically different which means the acoustics, the perception and the articulation is different but from a functional point of view in phonology they are not different. Therefore, they belong to the sane phoneme class.

There is another version of allophones and these are allophones which are distributed in such a way that you cannot make a functional distinction between them. Typical example would be in German, the ‘ich’ and the ‘ah’ sounds. And the ‘ich’ and ‘ah’ sound. This I also host for Greek. So, these are phonetically similar but not identical, you would say (they are phoneme classes I am sorry) they are different Phone classes. But, because they do not occur in the same position, you will never find a German word that is distinguished by the ‘ah’ and ‘ch’ sound, therefore they functionally belong to the same phoneme class.

(Refer Slide Time: 6:02)



Speech

Morpheme: Smallest unit of speech carrying meaning.

- Speech: Sequence of phones of individual phonemes.
- Written language: Grapheme sequences.

Syllable: Phonetic, not meaning-related unit; smallest group of phones in natural speech production process.

- Contains exactly one syllable-peak, typically vowel or "double-vowel" (diphthong).
- Can contain one or more consonants.

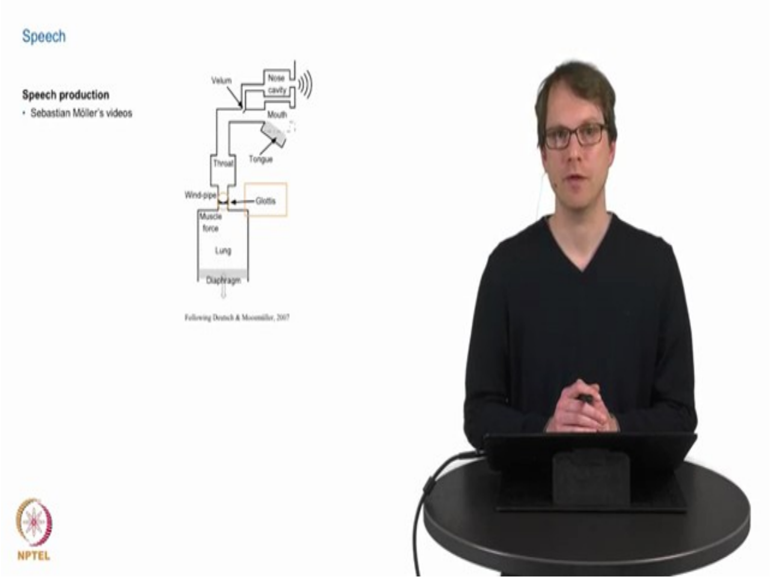
Morpheme – syllable distinction: different decompositions; example "painter".

- Syllable: "pain-ter".
- Morpheme: "paint-er".

What is Morpheme? In contrast to the phoneme, which is a single the smallest segment that is giving a difference in meaning, a phoneme is the smallest unit actually having some meaning. So, word stems are Morphemes. But also, all these declinations or other grammatical markers, which show us plural or gender of a word, all these are morphemes. So, let us consider the example of a painter. So, paint is a thing and painter is the person who uses paint. So, the ‘er’ is the, this small syllable, is actually the morpheme and paint, is a morpheme as well.

And as I used the word syllable, syllable is actually not dedicated to the different linguistic levels. This is a term from phonetics and it just means that it is a unit of articulation. Typically, a syllable has a vowel or something like the vowel in the centre, and may have different consonants or other sounds. Often, a morpheme is a syllable, but not necessarily so. So morphemes can actually have also two syllables, but most have one. And using the same example again, painter you can see the clear difference between the syllable and the morpheme because the border of the syllable would divide painter in ‘pain’ and ‘ter’. But the morpheme border is clearly after paint. So, ‘er’ would be the morpheme.

(Refer Slide Time: 8:14)

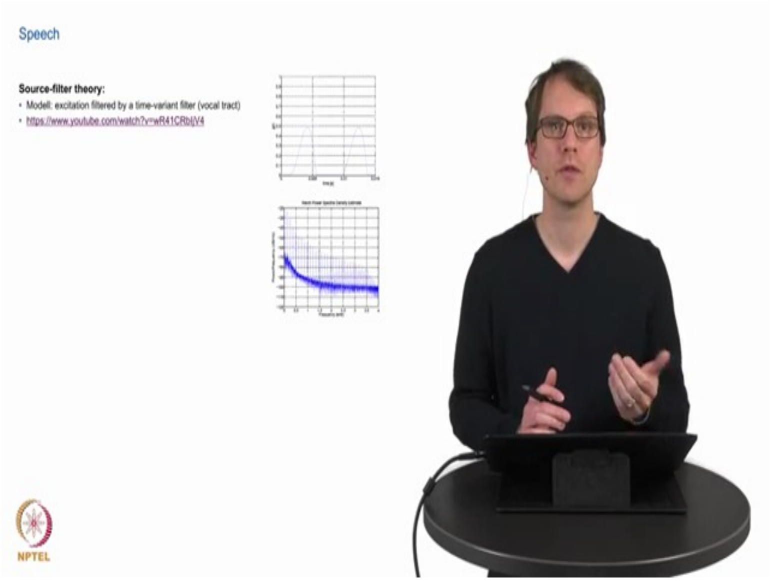


Speech

Speech production

- Sebastian Möller's videos

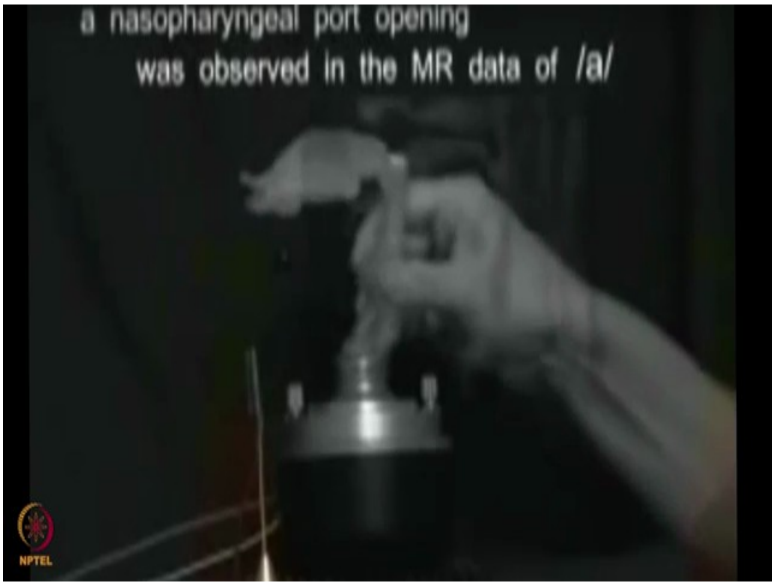
Following Deutsch & Moosmüller, 2007



Speech

Source-filter theory:

- Modell: excitation filtered by a time-variant filter (vocal tract)
- https://www.youtube.com/watch?v=wR41CRbljV4

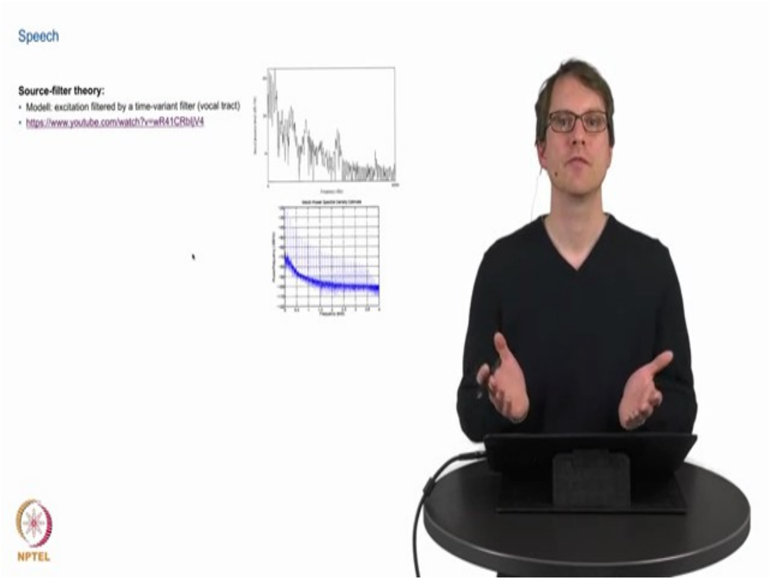

a nasopharyngeal port opening
was observed in the MR data of /a/

So, the whole speech production process is nicely covered by another video by my colleague Sebastian Möller this is there is a link on the ISIS platform. But there are a few things I would like to stress on nevertheless. The important thing in speech production is that we have a so called source filter theory. That means that there is a model that assumes the speech production in humans at least voiced speech production consists of two elements. This is excitation that is happening at the vocal chords, vocal faults where excitation signal is produced and this is formed by kind of filter function and the rest in the upper parts of the articulatory organs like the throat, the mouth (or) and the nasal cavity.

I have a display here a graphic showing on the upper side, the voice source and the excitation function. This is when the vocal faults vibrate passively, rapidly coming together and this is not a sinus this is really more like an abrupt closure of the vocal fault that you can see. The result of this is actually that we have a kind of line spectrum this is the second picture on the bottom of the slides. So, we have the fundamental frequency of the vocal faults coming together, closing and opening again, but we also have multiples of this frequency in the spectral domain and this is now formed by the specific characteristics of the vocal tract it is called so in the upper regions we can actually manipulate by moving our tongue, our jaw and our lips and activating the nasal cavity as well by lowering the vellum.

There is a nice YouTube video showing this source filter theory in action. You will see for the male A sound A vowel an excitation function by loud speaker, and then the person who produces the video will give you or will show a kind of manually produced plastic thing that actually represents the vocal tract so the throat mouth of an articulated ‘a’ and then you can listen to the source the excitation signal and the ‘a’ sound the ‘a’ vowel which comes from the lips.

(Refer Slide Time: 11:35)



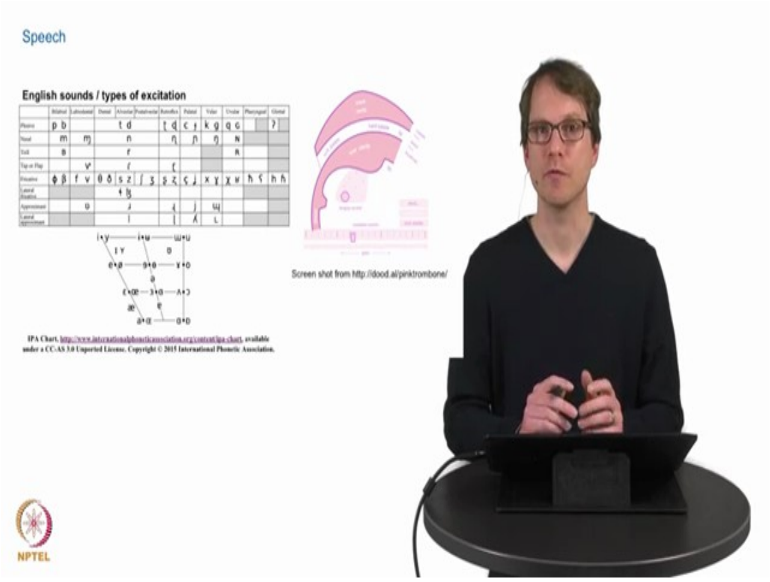
Speech

Source-filter theory:

- Modell: excitation filtered by a time-variant filter (vocal tract)
- https://www.youtube.com/watch?v=wR41CRbIjV4

So, the result of this filtration of the vocal tract signal of the excitation function is then the final spectrum, this is a recording that I made and you can still see the line structure, a little bit, but especially you can see some resonance frequencies so called formants which are quite unique for a particular vowel and this is what is the most important perceptual cue for identifying vowels.

(Refer Slide Time: 12:02)



Speech

English sounds / types of excitation

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

IPA Chart, http://www.internationalphoneticassociation.org/content/ipa-chart, available under a CC-AS 3.0 Unported License. Copyright © 2015 International Phonetic Association.

Screen shot from http://dood.al/pinktrombone/

Here is a table of English sounds. They are the consonants and the vowels. So, the vowels do not have any strong constriction in the mouth, but they have certain pattern by the jaw the opening of the lips and the lowering of the jaw, and the position of the tongue and this is a two dimensional distribution or two dimensional display of the different sounds the different vowels that we have and this according to the opening of the tongue the position of the tongue whether these are the high vowels with more closed mouth and high tongue position like ‘e’ and ‘u’ and low vowels which are more open and rather low position of the tongue the ‘a’.

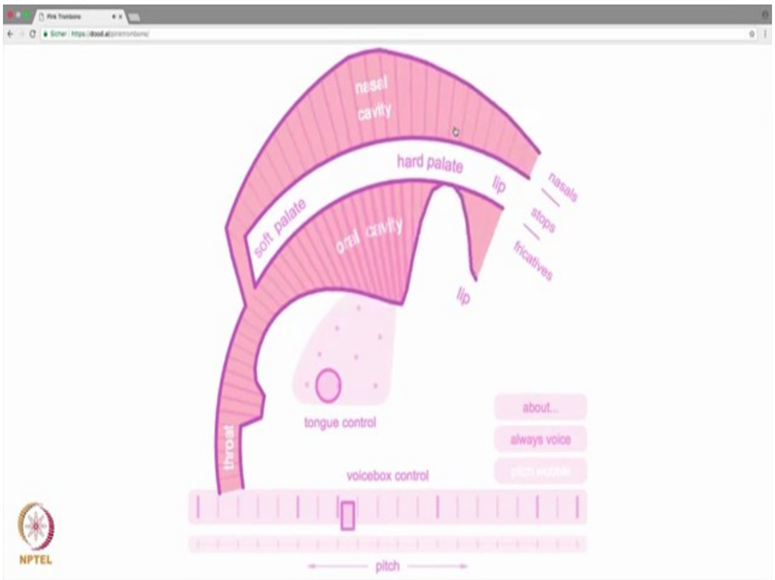
And then we have front and back vowels where the tongue has a certain movement towards the palette in the front or in the back and the front would be ‘e’ or ‘a’ and the back would be ‘u’ or ‘o’. This corresponds, rather loosely but it corresponds to the formant frequencies of the first and the second so called formant which are the first and second resonant frequency. You will have an assignment where you will have to measure your own formants.

The consonants in this table are separated according to the manner and place of articulation. The place of articulation just tells us where is something in the happening in the mouth like at the front, at the lips or at the back with the vellum and then we have different manner of articulation. So, for consonants there are plosives that is there is a real closure somewhere so no air is coming out of the nostrils or the mouth and this closure is then abruptly opened by a so called burst.

So, plosives are ‘p’ or’ t’, or ‘k’ and then there are fricatives where there is no true closure but there is a constriction and this constriction is causing a kind of turbulence or noise like ‘s’, ‘f’ of ‘sh’ and then we have other consonants which are even less strong constrictions like letterals or like approximants. The thing is that all these consonants have two kinds of sounds either being voiced or unvoiced. So whether there is activity, vibration in the excitation function or whether there is excitation or not. So for example there is ‘b’ and ‘p’. The first one was voiced and second there was no activity here at the vocal faults. Same we have with ‘sh’ and ‘y’ second one was voiced.

The only one which is rather difficult to articulate unvoiced are nasals. So nasals are where we have a closed a closing in the mouth, but the air is coming through the nose like ‘m’, ‘n’ and ‘gn’ these are most often or most typically voiced because it is rather difficult to produce any kind of acoustic intensity with all this voicing when the mouth is closed.

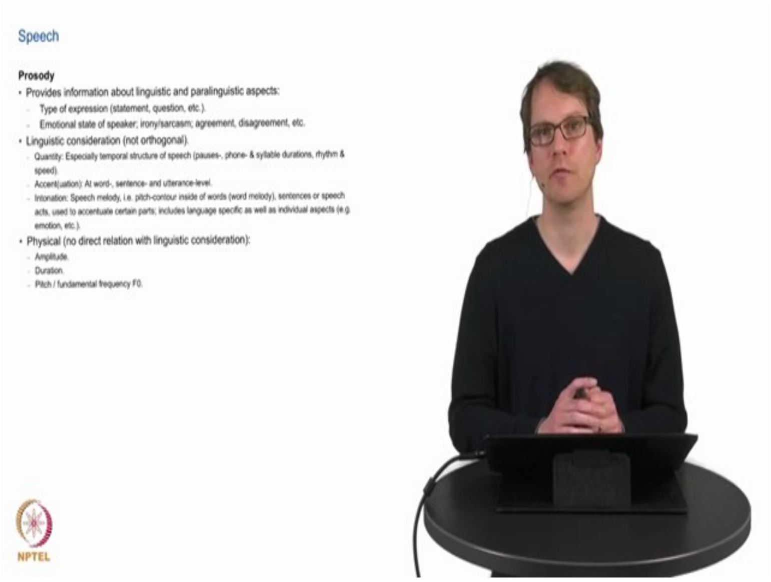
(Refer Slide Time: 16:24)



Now I will show you some examples of some of these consonants, by this online synthesizer. What you can see here is actually a cut through the head and here are the lips and here is the voice source the larynx, which produces the which has the vocal chords vocal faults and produces the excitation function of the excitation signal. And this small synthesizer allows us to move a little bit the tongue here and this is really the restriction that we can manage to build here and of course we have the nasal cavity as well that we can also use. So I will produce now a small sound source. And in general, the vowels are depended on the vocal cord shape and that is where the tongue is located.

Let us come to the consonants, so the consonants they are plosives, where there is a complete closure of the mouth, and plosives have as all the other consonant a certain place of articulation. I am trying to do a ‘ga’ of course with the sound source voice, you always have voice consonants right now. There is also the ‘l’ which should be not a plosive and not a fricative but there is much more air that can pass through this constriction and then we have as well the nasals. Let us go back to the ‘a’ for some other nasals.

(Refer Slide Time: 19:05)



Speech

Prosody

- Provides information about linguistic and paralinguistic aspects:
  - Type of expression (statement, question, etc.).
  - Emotional state of speaker; irony/sarcasm; agreement, disagreement, etc.
- Linguistic consideration (not orthogonal).
  - Quantity: Especially temporal structure of speech (pauses-, phone- & syllable durations, rhythm & speed).
  - Accent(uation): At word-, sentence- and utterance-level.
  - Intonation: Speech melody, i.e. pitch-contour inside of words (word melody), sentences or speech acts, used to accentuate certain parts; includes language specific as well as individual aspects (e.g. emotion, etc.).
- Physical (no direct relation with linguistic consideration):
  - Amplitude.
  - Duration.
  - Pitch / fundamental frequency F0.

The last slide on speech production is about Prosody. So I have talked a lot about sounds and these are rather short terms right? So, these have like durations of let us say to 50 to 200 milli seconds. But of course, when we speak, we produce a lot of other signals and information. We have for example the pitch contour so the frequency the fundamental frequency of our vocal faults and this changes in the kind of intonation and melody. This is typically assumed to be Prosody so all so called supra-segmentals are Prosody. So slowly the varying and aspects of speech production and also aspects which are not restricted to a single segment.

So, for example, how the intensity is Prosody so how loud do we speak, how loud are the sounds and especially how is the sound pattern over time the same with the fundamental frequency of intonation and the same for the duration that means how long are pauses and how long are the segments or the sounds that we produce. Do we speak rather quickly or slowly and this prosody is also relevant for linguistics of course.

For example in German and other languages, Germanic languages we have the distinction of a sentence being a statement “Yesterday it was cold.”, or being a question “Yesterday it was cold?” And this is purely intonation that means the pattern of the fundamental frequency the pitch so this is Prosody. But also prosody plays an important role in nonverbal information. So information which are not a part our spoken language but part of other signals like our emotions or irony or even expressing who we are and where do we come from, I will talk about non-verbal information, in a different week.