

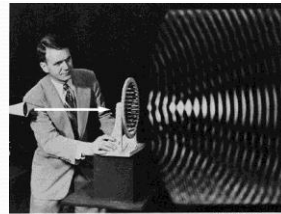
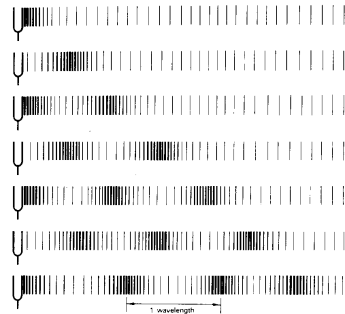
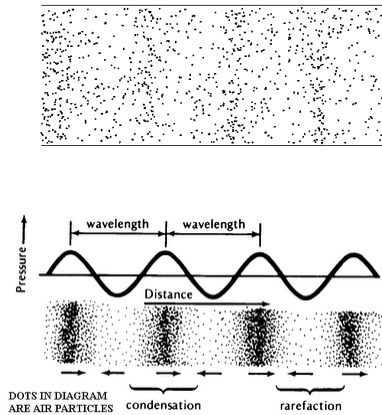
Sound Production and Perception

E.M. Bakker

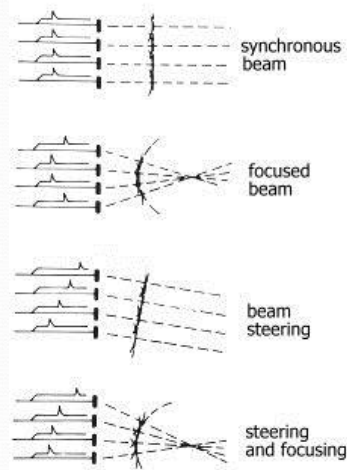
Overview

- The Physics of Sound
- The Production of Speech
- Phonetics and Phonology
- The Perception of Speech and Audio
 - Frequency Masking
 - Noise Masking
 - Temporal Masking
- Vocal Tract Workshop

The Physics of Sound



Phased Arrays



Refs.: <http://www.ob-ultrasound.net/somers.html>

Soundlazer



<https://www.kickstarter.com/projects/richardhaberkern/soundlazer> (www.soundlazer.com (2022: suspended)

LML Audio Processing and Indexing

5

The Physics of Sound

Speed of Sound		
Air (at sea level, 20 C)	343 m/s (1235 km/h)	$V=(331+0.6T)$ m/s
Water (at 20 C)	1482 m/s (5335 km/h)	
Steel	5960 m/s (21460 km/h)	
Tendon	1650 m/s	
Wood hard vs soft	4267 m/s vs 3353 m/s	

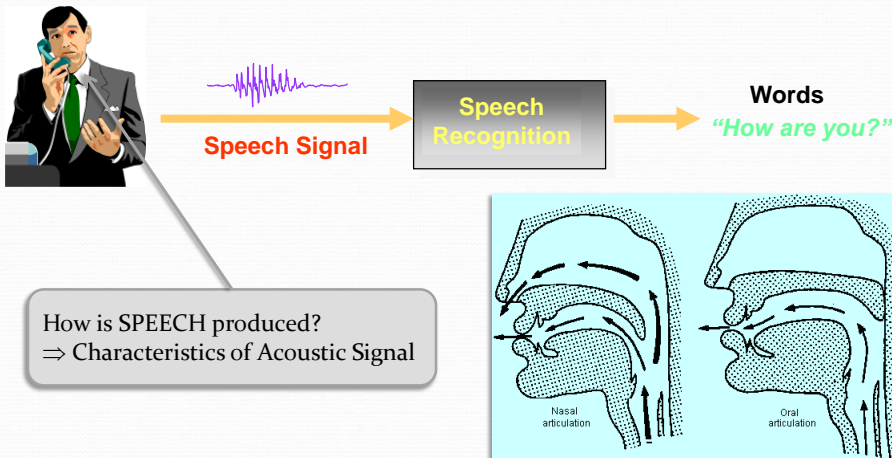
Speed of Sound

- Proportional to $\sqrt{\text{elastic modulus/density}}$
- Second order dependency on the amplitude of the sound => **nonlinear propagation effects**

LML Audio Processing and Indexing

6

The Production of Speech

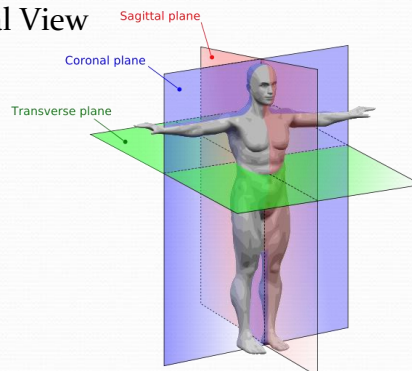


LML Audio Processing and Indexing

7

Human Speech Production

- Physiology
 - Schematic and X-ray Sagittal View
 - Vocal Cords at Work
 - Transduction
 - Spectrogram
- Acoustics
 - Acoustic Theory
 - Wave Propagation

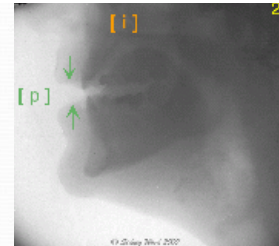
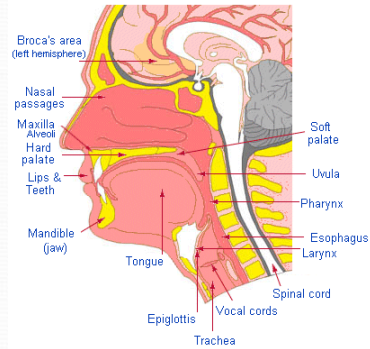
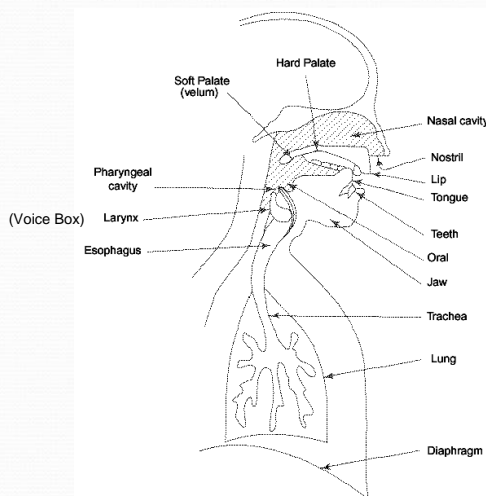


(Picture by Y.Mrabet from Wikipedia)

LML Audio Processing and Indexing

8

Sagittal Plane View of the Human Vocal Apparatus

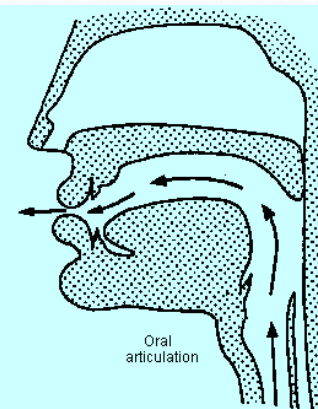
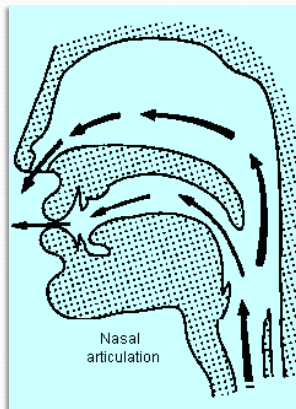
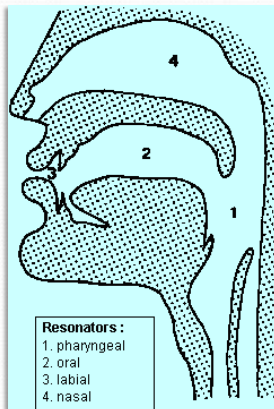


LML Audio Processing and Indexing

https://linguistics.berkeley.edu/aci/p/appendix/vocal_tracts/KNS.html

9

Sagittal Plane View of the Human Vocal Apparatus



LML Audio Processing and Indexing

10

Characterization of English Phonemes

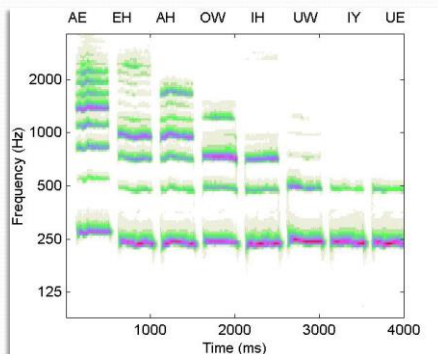
/f/	Labio-dental
/v/	Labio-dental
/θ/	Tip-dental
/ð/	Tip-dental
/s/	Blade-alveolar
/z/	Blade-alveolar
/ʃ/	Blade/front-palato-alveolar
/ʒ/	Blade/front-palato-alveolar
/h/	Glottal
/l/	Tip-alveolar
/r/	Blade-postalveolar
/w/	Bilabial back-velar
/j/	Front-palatal

Consonants	Place
/p/	Bilabial
/b/	Bilabial
/t/	Tip-alveolar
/d/	Tip-alveolar
/k/	Back-velar
/g/	Back-velar
/tʃ/	Blade/front-palato-alveolar
/dʒ/	Blade/front-palato-alveolar
/m/	Bilabial
/n/	Tip-alveolar
/ŋ/	Back-velar

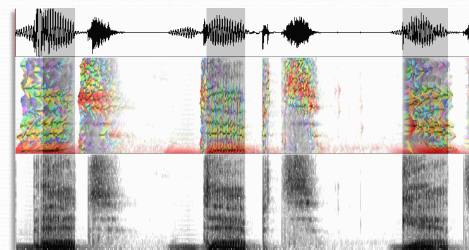
LML Audio Processing and Indexing

11

English Phonemes



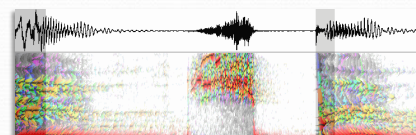
Vowels



Bet

Debt

Get



Pin

Sp

i

n

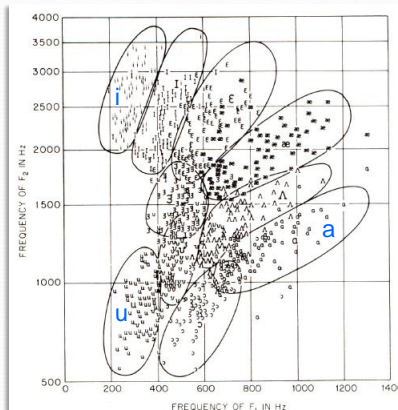
Allophone p

LML Audio Processing and Indexing

12

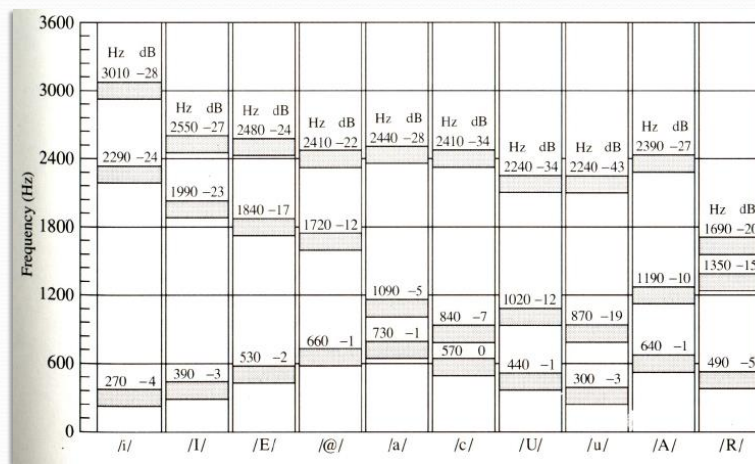
The Vowel Space

- We can characterize a vowel sound by the locations of **the first and second spectral resonances**, known as **formant frequencies**.
- Some voiced sounds, such as **diphthongs** (e.g. **air**), are transitional sounds that move from one vowel location to another.

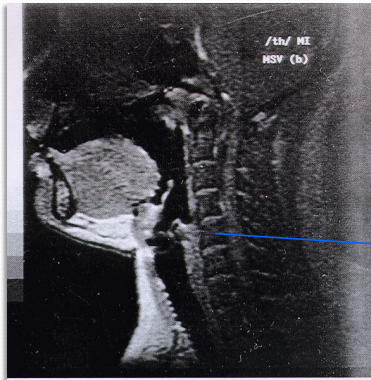


Phonetics

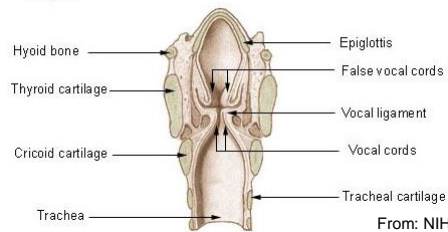
Formant Frequency Ranges



Vocal Cords



Larynx



From: NIH.

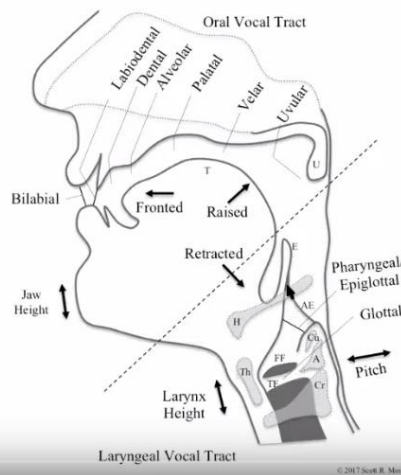


- High velocity air from the air pressure generated by the lungs opens the cords.
- The **Bernoulli effect** of the high velocity air results in a lowered pressure and the closing of the cords.
- This happens with the **resonance frequency of the cords**.
- The resonance frequency depends on the length (12.5 mm - 17 mm - 23 mm) and tension of the cords

LML Audio Processing and Indexing

15

Laryngeal Articulator Model



(Esling, Moisik, Benner & Crevier-Buchman, *Voice Quality: The Laryngeal Articulator Model*, 2019)

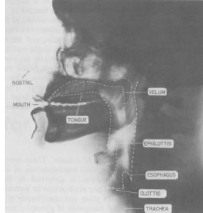
See [3] J. Esling

LML Audio Processing and Indexing

16

Models for Speech Production

Fundamentals of Speech Recognition by Lawrence Rabiner, and Biing-Hwang Juang

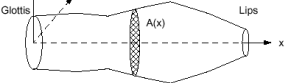


A detailed acoustic theory must consider the effects of the following:

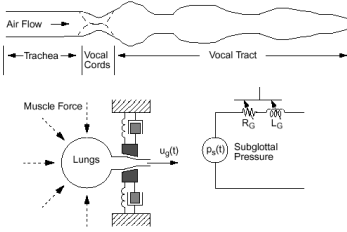
- Time variation of the vocal tract shape
- Losses due to heat conduction and viscous friction at the vocal tract walls
- Softness of the vocal tract walls
- Radiation of sound at the lips
- Nasal coupling
- Excitation of sound in the vocal tract

Let us begin by considering a simple case of a lossless tube:

$p = p(x, t)$



How do we couple energy into the vocal tract?



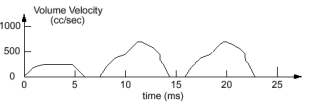
The glottal impedance can be approximated by:

$$Z_G = R_G + j\omega L_G$$

The boundary condition for the volume velocity is:

$$U(0, \omega) = U_G(\omega) - P(0, \omega) / Z_G(\omega)$$

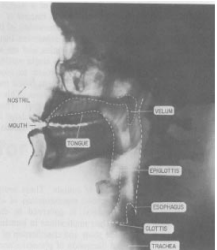
For voiced sounds, the glottal volume velocity looks something like this:



LML Audio Processing and Indexing

17

Models for Speech Production

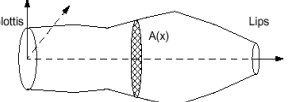
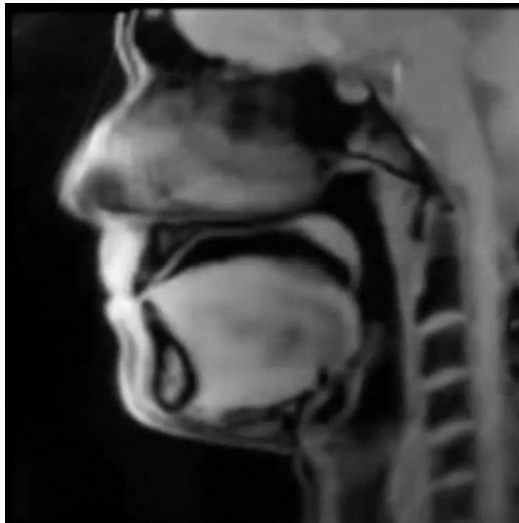


A detailed acoustic theory must consider the effects of the following:

- Time variation of the vocal tract shape
- Losses due to heat conduction and viscous friction at the vocal tract walls
- Softness of the vocal tract walls
- Radiation of sound at the lips
- Nasal coupling
- Excitation of sound in the vocal tract

Let us begin by considering a simple case of a lossless tube:

$p = p(x, t)$

LML Audio Processing and Indexing

18

Generation of diphthong [ai]

using the adaptive grid



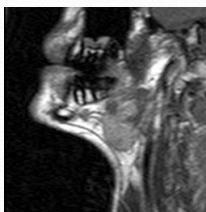
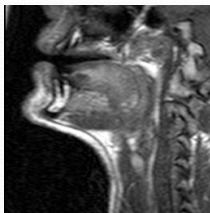
Modeling the VocalTract: Finite Element Methods: generating a diphthong.
https://www.youtube.com/watch?v=2jMqrd_pA8w

LML Audio Processing and Indexing

19

Vocal Tract Modeling

MRIs of the vocal tract shape.



1.

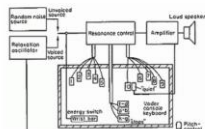


S. Fujita, K. Honda, An experimental study of
Acoustic characteristics of hypopharyngeal cavities
Using vocal tract solid models. Acoust. Sci. & Tech,
26, 4 (2005)



Vocal Tract Models

Block diagram and image of VODER at World Fair in 1939.
Courtesy of <http://www2.ling.su.se/staff/hartmut/kempine.htm>



Joseph Faber's speech machine called Euphonia. Image courtesy of <http://www2.ling.su.se/staff/hartmut/kempine.htm>

1846



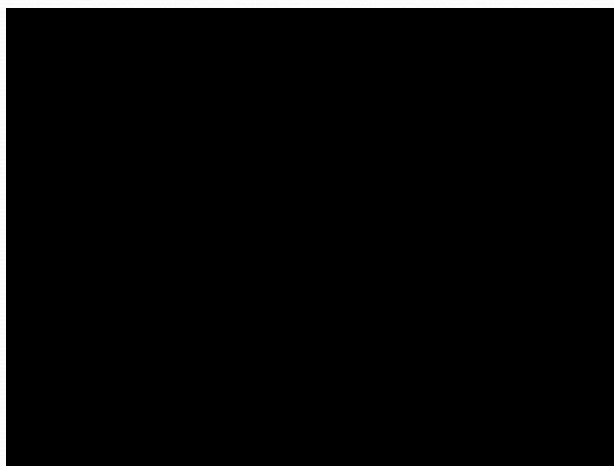
A modern model of Darwin's speech machine from 1771
courtesy of
BirminghamStories.co.uk

From: <https://muse.union.edu/griffiths-capstone/different-types-of-models/>

Arai, T., "Vocal-tract models and their applications in education for intuitive understanding of speech production," *Acoust. Sci. & Tech.*, 37(4), 148-156, 2016

LML Audio Processing and Indexing

21

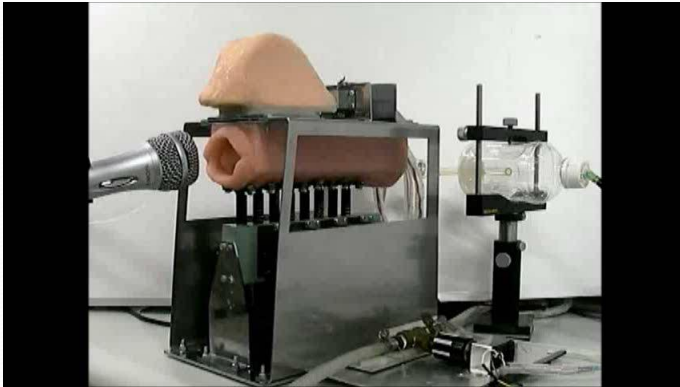


From: Arai T. 2007: http://www.splab.net/Vocal_Tract_Model/index.htm

LML Audio Processing and Indexing

22

Physical Models of Vocal Tract



<http://www.eng.kagawa-u.ac.jp/~sawada/index.html>

LML Audio Processing and Indexing

23

D. M. Howard, The Vocal Tract Organ: a new musical instrument using 3-D printed vocal tracts* <http://dx.doi.org/10.1016/j.jvoice.2017.09.014>
University of London, UK



LML Audio Processing and Indexing

24

Ian S. Howard ROBOTIC ACTUATION OF A 2D MECHANICAL VOCAL TRACT
Konferenz Elektronische Sprachsignalverarbeitung 2017, Saarbrücken

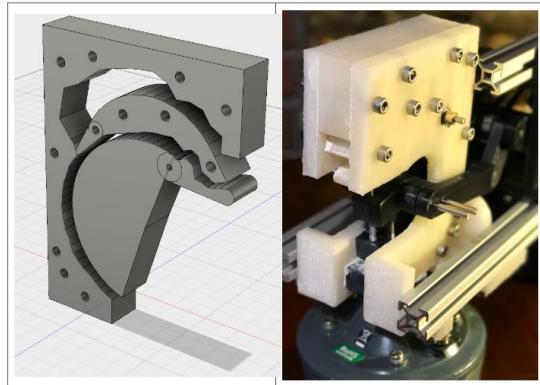


Figure - 2 LHS: AutoCAD Fusion design of the central section of vocal apparatus. Here the sides are not shown so that the tongue mechanism vocal and nasal cavities and the velar flap can be seen. **RHS:** Oblique front view of the 3D printed vocal apparatus. The black movable tongue body and white tongue tip can be seen, as can the white side plates of the apparatus.

LML Audio Processing and Indexing

25

Vocal Tract Workshop

Articulatory Speech Synthesis using Vocal Tract Lab
(P. Birkholz, 2013).

Note: Also API for Matlab, C++ and Python available.

Reading material:

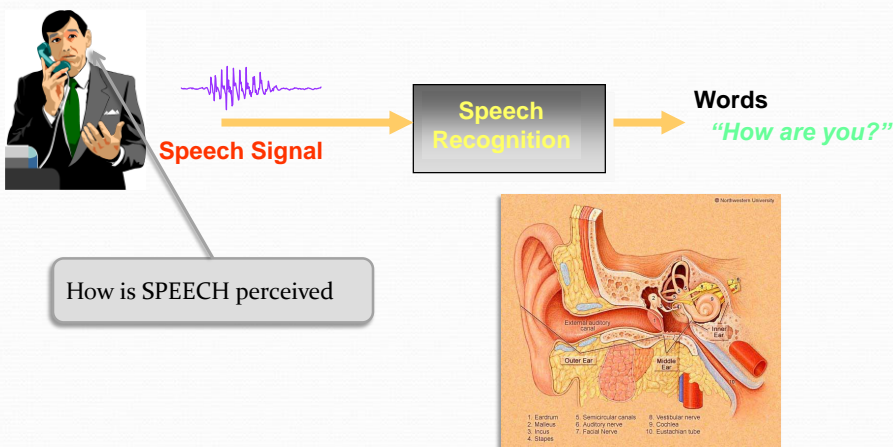
P. Birkholz, D. Jackel, A Three-Dimensional Model of the Vocal Tract for Speech Synthesis. In Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2597-2600, Barcelona, Spain, 2003.

See: <http://www.vocaltractlab.de/>

LML Audio Processing and Indexing

26

The Perception of Speech



LML Audio Processing and Indexing

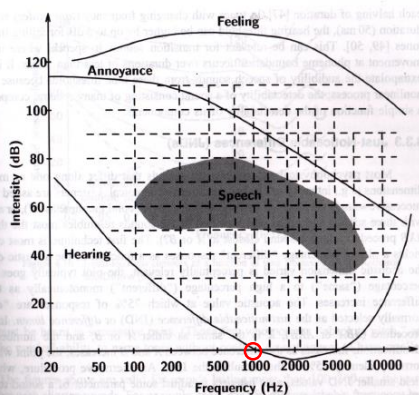
27

The Perception of Speech Sound Pressure

- The ear is the most sensitive human organ. Vibrations on the order of angstroms are used to transduce sound. **It has the largest dynamic range (~140 dB) of any organ in the human body.**
- The lower portion of the curve is an audiogram - hearing sensitivity. It can vary up to 20 dB across listeners.
- Above 120 dB corresponds to a nice pop-concert (or standing nearby a Boeing 747 when it takes off).
- Typical ambient office noise is about 55 dB.

$$x \text{ dB} = 10 \log_{10}(x/x_0)$$

x_0 = 1kHz signal with intensity that is just hearable.



LML Audio Processing and Indexing

28

dB (SPL)	Source (with distance)
194	Theoretical limit for a sound wave at 1 atmosphere environmental pressure; pressure waves with a greater intensity behave as shock waves.
188	Space Shuttle liftoff as heard from launch tower (less than 100 feet) (source: acoustical studies [1] [2]).
180	Krakatoa volcano explosion at 1 mile (1.6 km) in air [3]
160	M1 Garand being fired at 1 meter (3 ft); Space Shuttle liftoff as heard from launch pad perimeter (approx. 1500 feet) (source: acoustical studies [4] [5]).
150	Jet engine at 30 m (100 ft)
140	Low Calibre Rifle being fired at 1m (3 ft); the engine of a Formula One car at 1 meter (3 ft)
130	Threshold of pain; civil defense siren at 100 ft (30 m)
120	Space Shuttle from three mile mark, closest one can view launch. (Source: acoustical studies) [6] [7]. [Train horn]] at 1 m (3 ft). Many foghorns produce around this volume.
110	Football stadium during kickoff at 50 yard line; chainsaw at 1 m (3 ft)
100	Jackhammer at 2 m (7 ft); inside discothèque
90	Loud factory, heavy truck at 1 m (3 ft), kitchen blender
80	Vacuum cleaner at 1 m (3 ft), curbside of busy street, PLV of city
70	Busy traffic at 5 m (16 ft)
60	Office or restaurant inside
50	Quiet restaurant inside
40	Residential area at night
30	Theatre, no talking
20	Whispering
10	Human breathing at 3 m (10 ft)
0	Threshold of human hearing (with healthy ears); sound of a mosquito flying 3 m (10 ft) away

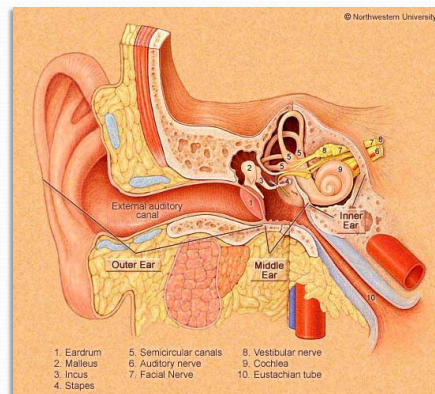
LML Audio Processing and Indexing

29

The Perception of Speech: The Ear

Outer and middle ear

- The **outer** and **middle** ears reproduce the analog signal (impedance matching).
- The **outer** ear consists of the external visible part and the auditory canal. The tube is about 2.5 cm long. (ex: sweep 3000 – 3500 Hz)
- The **middle** ear consists of the eardrum and three bones (malleus, incus, and stapes). It converts the sound pressure wave to displacement of the oval window (entrance to the inner ear).



Inner ear

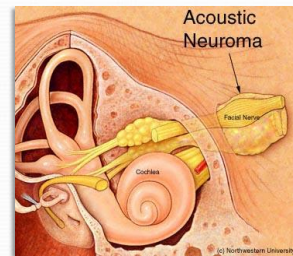
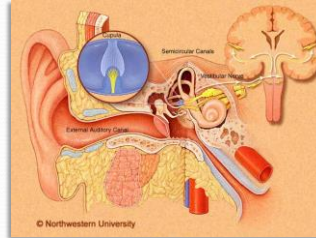
- the **inner** ear transduces the pressure wave into an electrical signal.

LML Audio Processing and Indexing

30

The Perception of Speech: The Ear

- The inner ear primarily consists of a fluid-filled tube (cochlea) which contains the basilar membrane. Fluid movement along the basilar membrane displaces hair cells, which generate electrical signals.
- There are a discrete number of hair cells (30,000). Each hair cell is tuned to a different frequency.
- Place vs. Temporal Theory: firings of hair cells are processed by two types of neurons:
 - onset chopper units for temporal features
 - transient chopper units for spectral features.



LML Audio Processing and Indexing

31

Perception Psychoacoustics

- **Psychoacoustics**: a branch of science dealing with hearing, the sensations produced by sounds.
- **Perceptual attributes of a sound** vs **measurable physical quantities**:
- Many physical quantities are perceived on a logarithmic scale (e.g. loudness).
- Perception is often a nonlinear function of the absolute value of the physical quantity being measured (e.g. equal loudness).
- Timbre can be used to describe why musical instruments sound different.
- What factors contribute to speaker identity?

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental Frequency	Pitch
Spectral Shape	Timbre
Onset/Offset Time	Timing
Phase Difference (Binaural Hearing)	Location

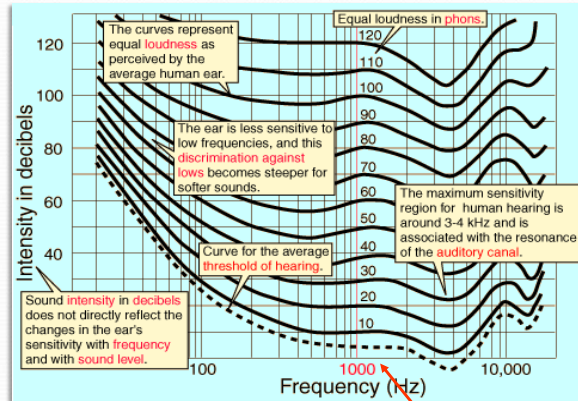
LML Audio Processing and Indexing

32

Perception

Equal Loudness

- Just Noticeable Difference (JND): The acoustic value at which 75% of responses judge stimuli to be different (limen)
- The perceptual loudness of a sound is specified via its relative intensity above the threshold. A sound's loudness is often defined in terms of how intense a reference 1 kHz tone must be heard to sound as loud.



Sweep: 0 – 3kHz



0 dB

LML Audio Processing and Indexing

33

Perception, Non-Linear Frequency Warping: Bark and Mel Scale

- **Critical Bandwidths:** correspond to ~ 1.5 mm width 'bands' along the basilar membrane,
=> 24 bandpass filters.
- **Critical Band:** can be related to a bandpass filter whose frequency response corresponds to the tuning curves of auditory neurons. A frequency range over which two sounds will sound like they are fusing into one.

- **Bark Scale:**
$$Bark = 13 \operatorname{atan}\left(\frac{0.76f}{1000}\right) + 3.5 \operatorname{atan}\left(\frac{f^2}{(7500)^2}\right)$$

- **Mel Scale:**
$$mel \text{ frequency} = 2595 \log_{10}(1 + f/700.0)$$

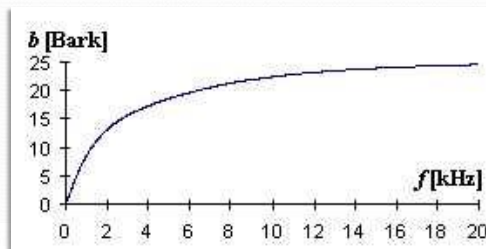
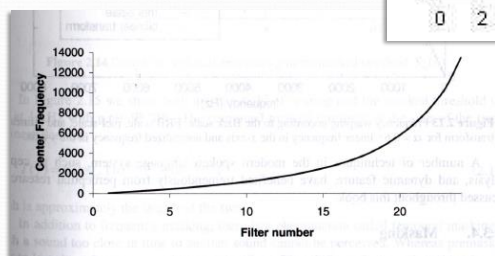
LML Audio Processing and Indexing

34

Perception

Bark and Mel Scale

- The Bark scale implies a nonlinear frequency mapping



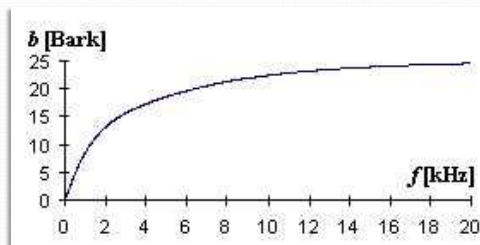
LML Audio Processing and Indexing

35

Perception:

Bark Scale and Mel Scale

- Filter Banks used in ASR:
- The Bark scale implies a nonlinear frequency mapping



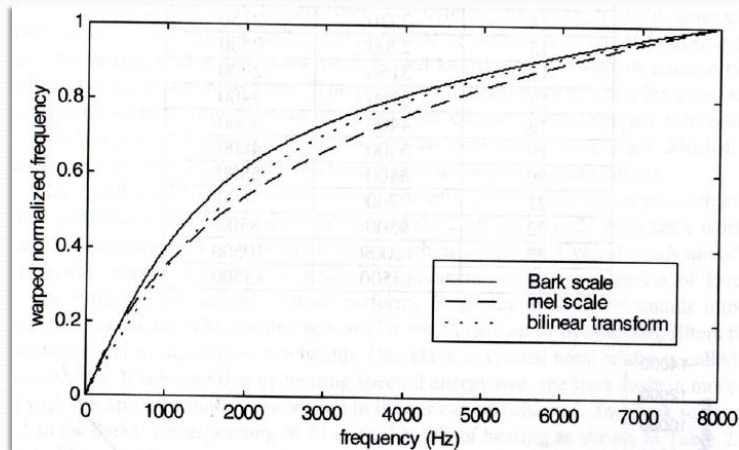
(BW = Bandwidth)

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

LML Audio Processing and Indexing

36

Comparison of Bark and Mel Space Scales



LML Audio Processing and Indexing

37

Perception: Frequency Masking

Frequency masking:

One sound cannot be perceived if another sound close in frequency has a high enough level.

- Thresholds are frequency and energy dependent.
- Thresholds depend on the nature of the sound as well.

LML Audio Processing and Indexing

38

Perception: Tone-Masking Noise



Tone-masking noise:

Noise with energy EN (dB) at Bark frequency f_{noise} masks a tone at Bark frequency f_{tone} if the tone's energy is below the threshold:

$$TT(b) = EN - 6.025 - 0.275g + Sm(f_{noise} - f_{tone}) \quad (\text{dB SPL})$$

where the *spread-of-masking* function $Sm(f)$ is given by:

$$Sm(f) = 15.81 + 7.5(f + 0.474) - 17.5 * \sqrt{1 + (f + 0.474)^2} \quad (\text{dB})$$

$SPL = \text{Sound Pressure Level}$

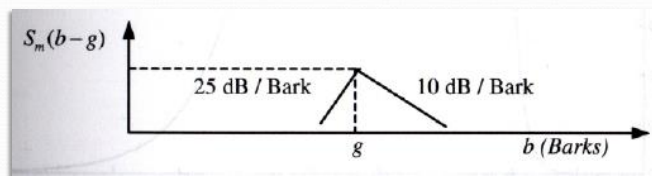
Perception: Noise-Masking Tone

- **Noise-masking tone:** a tone at Bark frequency f_{tone} with energy E_{tone} (dB) masks noise at Bark frequency f_{noise} if the noise energy is below the threshold:

$$E_{\text{threshold}}(f_{noise}) = E_{\text{tone}} - 2.025 - 0.17g + Sm(f_{noise} - f_{tone}) \quad (\text{dB SPL})$$

$SPL = \text{Sound Pressure Level}$

- Masking thresholds are commonly referred to as Bark scale functions of *just noticeable differences* (JND).
- Thresholds are not symmetric.
- Thresholds depend on the nature of the noise and the sound.



Perception: Temporal Masking

Temporal Masking:

Onsets of sounds are masked in the time domain through a similar masking process.

Perception: Echo and Delay

- Humans are used to hearing their voice while they speak - real-time feedback (side tone).
- When this side-tone is delayed, it interrupts our cognitive processes, and degrades our speech.
- This begins at delays of approximately 250 ms.
- When we place headphones over our ears, which dampens this feedback, we tend to speak louder.
- **Lombard Effect:** Humans speak louder in the presence of ambient noise.
- Modern telephony systems have been designed to maintain delays lower than this value.
- Digital speech processing systems can introduce large amounts of delay due to non-real-time processing.

Perception: Adaptation (1/2)

- **Adaptation** refers to changing sensitivity in response to a continued stimulus, and is likely a feature of the mechano-electrical transformation in the cochlea.
- Neurons tuned to a frequency where energy is present do not change their firing rate drastically for the next sound.
- Additive broadband noise does not significantly change the firing rate for a neuron in the region of a formant.

LML Audio Processing and Indexing

43

Perception: Adaptation (2/2)

J. Medina, Brain Rules.

Visual Adaptation

- **The McGurk Effect** is an auditory illusion which results from combining a face pronouncing a certain syllable with the sound of a different syllable. The illusion is stronger for some combinations than for others.
- For example, an auditory 'ba' combined with a visual 'ga' is perceived by some percentage of people as 'da'. A larger proportion will perceive an auditory 'ma' with a visual 'ka' as 'na'. Some researchers have measured evoked electrical signals matching the "perceived" sound.

LML Audio Processing and Indexing

44

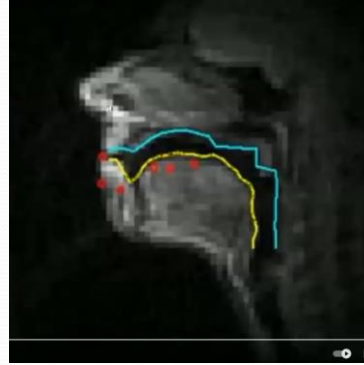
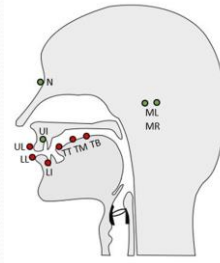
Perception: Timing (1/2)

- Temporal resolution of the ear is crucial.
- Two clicks are perceived mono-aurally as one unless they are separated by at least 2 ms.
- 17 ms of separation is required before we can reliably determine the order of the clicks. (~58bps or ~353obpm)
- Sounds with onsets faster than 20 ms are perceived as "plucks" rather than "bows".

Perception: Timing (2/2)

- Short sounds near the threshold of hearing must exceed a certain intensity-time product to be perceived.
- Humans do not perceive individual "phonemes" in fluent speech - they are simply too short. We somehow integrate the effect over intervals of approximately 100 ms.
- Humans are very sensitive to long-term periodicity (ultra low frequency) - this has implications for random noise generation.

Articulography



Movement sensors.

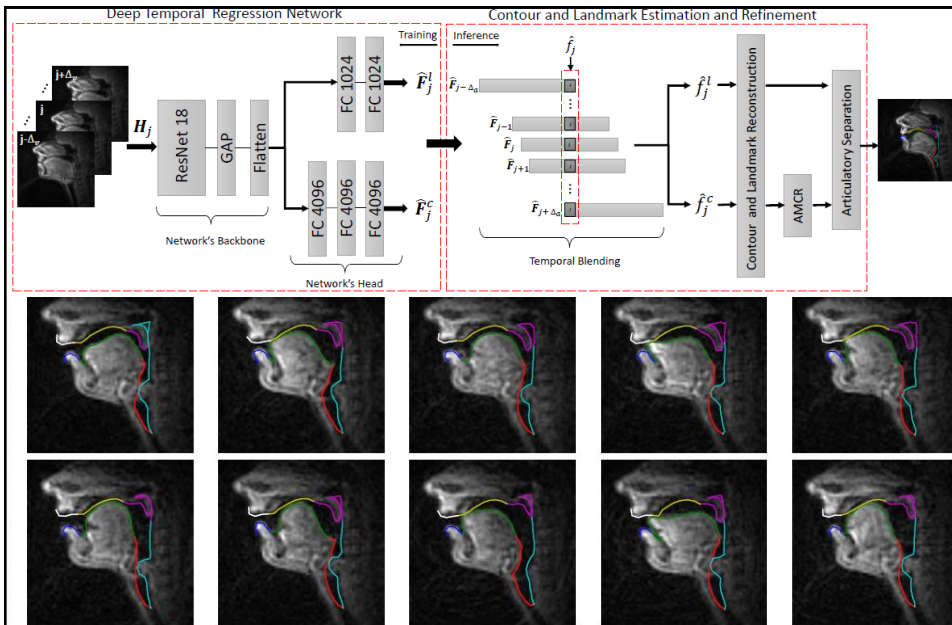
Green: reference sensors.
Red movement sensors.

See: <https://www.youtube.com/watch?v=NaaChEnX6IM>
https://sail.usc.edu/old/software/Registration_EMA_rtmr/

Rebernik, T. & Jacobi, J. & Jonkers, R. & Noiray, A. & Wieling, M., (2021) "A review of data collection practices using electromagnetic articulography", Laboratory Phonology 12(1), p.6. doi: <https://doi.org/10.5334/labphon.237>

LML Audio Processing and Indexing

47



In [1] S. Asadiabadi et al. used a DTRN+ACMR model on real time MRI videos to estimate the depicted contours of the vocal tract while the subject pronounced different words.

LML Audio Processing and Indexing

48

Human-Computer Interaction
UNIVERSITY OF CALIFORNIA
MERCED



Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI

Laxmi Pandey, Ahmed Sabbir Arif
Human-Computer Interaction Group, UC Merced

Play (h)

0:02 / 1:58

CC

LML Audio Processing and Indexing49



SINGING IN THE MRI

Play (h)

0:00 / 4:14

CC

<https://www.youtube.com/watch?v=icKWZZV-gog>
<https://www.tyleyrossvoice.com/videos>

LML Audio Processing and Indexing50

Vocal Tract Workshop

See:

<http://liacs.leidenuniv.nl/~bakkerem2/api/>

or

(the same but shorter)

www.liacs.nl/~erwin/api

Dynamic MRI at the University of Illinois at Urbana-Champaign

Maojing Fu, Aaron Johnson, Zhi-Pei Liang, Brad Sutton

In collaboration with the
Biomedical Imaging Center

BECKMAN INSTITUTE
FOR ADVANCED SCIENCE AND TECHNOLOGY

Play (k)

0:00 / 1:19

CC BY-NC-SA

<https://beckman.illinois.edu/news/2015/04/new-super-fast-mri-technique>

References

- Some of the slides in these lectures are adapted from the presentation: “Can Advances in Speech Recognition make Spoken Language as Convenient and as Accessible as Online Text?”, an excellent presentation by: Dr. Patti Price, Speech Technology Consulting Menlo Park, California 94025, and Dr. Joseph Picone Institute for Signal and Information Processing Dept. of Elect. and Comp. Eng. Mississippi State University
- Several anatomic graphics are from Wikipedia
- **Fundamentals of Speech Recognition** by Lawrence Rabiner, and Biing-Hwang Juang (Hardcover, 507 pages; Publisher: Pearson Education POD; ISBN: 0130151572; 1st edition, April 12, 1993)
- NIH: <https://training.seer.cancer.gov/anatomy/respiratory/passages/larynx.html>

References

- [1] Sasan Asadiabadi, Student Member, IEEE, Engin Erzin, Vocal Tract Contour Tracking in rtMRI Using Deep Temporal Regression Network, IEEE/ACM Trans. On Audio, Speech, and Language Processing, Vol. 28, pp 3053 – 3064, 2020
- [2] https://linguistics.berkeley.edu/acip/appendix/vocal_tracts/usc/
- [3] J. Esling, <https://www.youtube.com/watch?v=rOqAZJfiZkk>
- [4] Marc Arnela, Saeed Dabbaghchian, Oriol Guasch and Olov Engwall, “MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs”, IEEE/ACM Transactions on Audio, Speech and Language Processin, 2019
- [5] Vocal Tract rtMRI Playlist: <https://www.youtube.com/playlist?list=PLG-dco7bxh-GNo3HjbLgKyXofXHcSvhl->