Basic Acoustics

When an object such as tuning fork is vibrating, we hear the sound because 'sound waves' are transmitted to our ear through the medium of air molecules. When the tuning fork is vibrating, the air surrounding the fork is also vibrating, meaning each air molecule surrounding the fork moves relatively little around a rest position. This vibrating movement of air molecules is due to the interplay of elasticity and inertia of air molecules. The pressure **wave** (the alternation of compression and rarefaction of air pressure) is transmitted away from the source as time proceeds (tuning fork, energy, displacement, rest position, cycle, sine waves).

Properties of sound waves

- The **amplitude** is the peak deviation of a pressure fluctuation from normal, atmospheric pressure
- The **frequency** is the rate at which the air molecules are vibrating. It is the number of cycles in a unit of time (second). So, often it has been called '**cps**' (cycles per second). 100cps = 100Hz (Hertz)
- A sound wave which has only one frequency is called a **pure tone**, e.g. sine waves. Most sound waves around us are often composed of more than one frequency. This type of sound is called a **complex tone** and has a complex waveform (they move in a complex manner).
- Complex tones can be either periodic or aperiodic
    a. **periodic**: pattern of vibration, however complex, repeats itself (e.g. vowels)
    b. **aperiodic** (noise): vibration is random and has no repeatable pattern (e.g. fricative)

Among many frequencies composing the **periodic complex waveform**, the lowest frequency or the basic frequency is called a **fundamental frequency (Fo)**. A speaker's fundamental frequency, which varies constantly during speech, determines the perceived **pitch** of his/her voice. Pitch variation is produced primarily by stretching the length of the vocal folds (function: intonation, tonal distinctions on vowels).

In addition to the fundamental frequency, additional frequencies forming the periodic complex tone are whole-number multiples of Fo and are called **harmonics**. E.g. if Fo is 100Hz, the 2nd harmonic is 200Hz, the 3rd harmonic is 300Hz, etc. Thus, if we know the value of the n^{th} harmonic, we can tell the value of Fo by dividing the n^{th} harmonic value by *n*.

There are two kinds of **aperiodic sound** (noise):

transient noise - producing a burst of noise of short duration
e.g. stop consonant, book dropping noise
continuous noise - turbulent air passing through a narrow constriction
e.g. hissing noise, fricatives

damping: the vibratory movement is reduced in amplitude (i.e. the amplitude of sound waves are getting weaker as waves progress in time, e.g. damping in piano)

A **waveform** is a graph showing the amplitude of an air molecule movement in a time course. 'Amplitude' in Y-axis, and 'Time' in X-axis.

The **spectrum** is an amplitude by frequency graph (power spectrum)

Source and Filters in the Speech Mechanism

In speech the larynx is the sound **source** and the vocal tract is a *system* of acoustic **filters**.

functions of larynx vibration in speech

- the glottal wave is a periodic complex wave (a pulse wave) composed of **fundamental frequency** (Fo) and a range of harmonics. It is a carrier wave of speech (vehicle) (**3,4**)
- variations in Fo (roughly between 60-500Hz) which are responsible for **intonation** are relations with respect to *direction* in which the fundamental frequency changes with time and are NOT dependent on some standard pitch scale.
- **voice switching** switching on and off of the vibratory activity of the vocal folds plays a linguistic function connected with phonemic differences and contrasts (voiced / voiceless).

modification of the glottal wave (larynx wave) in the vocal tract (resonant system)

- **Acoustic filters** keep out, or more accurately, reduce the amplitude of certain ranges of frequency while allowing other frequency bands to pass with very little reduction of amplitude.
- The speech mechanism makes extensive use of this filtering function of resonators. Thus, the sound waves radiated at the lips and the nostrils are a result of modifications imposed on the larynx wave by this resonating system (**6,7**).
- The response of the vocal tract (with the tongue in neutral position) is such that it imposes a pattern of certain natural *frequency regions* (e.g. 500Hz-1500Hz-2500Hz for a vocal tract 17cm in length) and reduces the amplitude of remaining harmonics of the glottal wave. This gives us a **schwa** vowel.
- The respective peaks of energy (**formants** F1 - F2 - F3) are retained regardless of variations in Fo of the larynx pulse wave (**7**). **The output of the resonance system always has the Fo of the glottal wave, and the formants F1, F2, F3 ... are imposed by the vocal tract!!!**
- Further modification of *formant structure* can be obtained by altering **tongue** and **lip** positions (**8**). So, there is a relation between articulation and formant structure (**9**). *Note a very low F2 for the English [o:] which is due to lip rounding, i.e. rounded and protruded lips lengthen the "horizontal" resonance chamber/cylinder and hence lowers its resonating properties.*

Acoustic Characteristics of Vowels

Source-filter theory : INPUT (source / glottal harmonics)
↓
FILTER (Vocal Tract / sound shaping / *tongue - lips - velum*)
↓
OUTPUT (speech signal)

Formants. If we know the length of the vocal tract, we can calculate its lowest natural frequency (F1), as well as the higher formant regions by multiplying the lowest frequency by 3, 5, 7... (odd multiples).

The length of the vocal tract is 1/4 of the length of the sound wave which is the lowest frequency compatible with that size of VT (vocal tract). The frequency can be calculated by dividing the speed of sound (35,000cm/s) by the wavelength (e.g. for a 17,5cm long vocal tract the lowest compatible wavelength is 17,5 x 4 = 70cm, i.e. for the wave to complete one cycle it must travel 70cm in space). 35,000 divided by 70 gives us a frequency of 500 cycles per second, i.e.500Hz. F1 = 500Hz.

Each vowel has a different pattern of output spectrum. Which is independent of the source.

- same source but different filter shape → gives different vowels in same pitch
- different source but same filter shape → gives same vowel in different pitch

*!!!*Even if the source is aperiodic noise produced in the glottis, the shape of the VT will produce formants. This is what is happening when we whisper (**1**).

definition: Vowels are defined by the physiological characteristic of their having no obstruction in VT, and by their function within a phonologically defined syllable (cf. semivowels [w],[j] unobstructed VT but no syllabic function).

acoustics to articulation relationship: the issue becomes crucial if we require such a one to one relationship for the purpose of linguistic description. In practice, one or the other system is used (cf. Jakobson, Fant and Halle 1952, KLV 1985 as opposed to Chomsky and Halle (SPE) and the post-SPE frameworks).

The importance of vowel height (F1). If a language has only two vowels it always uses the height distinction rather than *front / back*. For example: [ɨ - a] Margi (Chadic language), [ə - a] in Eastern Arrernte (Australian), Ubykkh and Abkhaz.

> *It seems that the use of front/back distinction (F2) is secondary to height in that it is present only in more complex systems. Likewise, lip rounding which affects F2 by its further lowering appears once the front/back space of the tongue movement has been exploited and further contrasts are required.*

Reasons for the prominence of F1 (height) in vowel systems

- **acoustic:** lower frequencies have typically higher amplitude (remember glottal wave?) (modes of vibration of a musical string) therefore it is not surprising that the lower peaks of energy (frequency region of F1) will be the first variable to pick in vowel systems. Only then, F2, and especially the relationship between F1-F2 begins to play a role.
- **auditory:** generally speaking, the perception of speech sounds by the human ear favours lower frequencies not necessarily because they are louder but because the human ear picks them up better (wait till we discuss basic audition). So, even though the differences in the frequency regions that F1 occupies in different vowels are not terribly distinct in linear terms, the auditory response to them is greater than to relatively greater distinctions at higher frequencies.

  The two arguments coupled together suggest that we should be very much surprised to find a primitive system which would employ front/back distinctions (F2 variables) before F1 is used.
- **functional:** The area in the mouth (more or less: **lips ⟶ soft palate**) is where *vowel articulation* coincides with most of the *consonant articulation*. Therefore, we expect that place of articulation of consonants will have influence on the shape of F2. This is not a bad thing *(we will soon learn that formant transitions of F2 constitute excellent acoustic cues for the place of articulation of both the preceding and the following consonant)* but if we are to choose between the **relative stability of F1** in coarticulation and the **relative instability of F2** in deciding on the basic parameters of our vowel system, the markedness criteria seem to be obvious (Marshalese, Irish?).

The Front / Back distinction (F2) occupies relatively little phonetic / acoustic space - up to three degrees FRONT, CENTRAL, BACK, where the introduction of the third distinction may involve **lip articulations** (NEUTRAL, COMPRESSED (raising of F2), PROTRUDED (lowering of F2)). In the great majority of world's languages the relationship between the phonetic Backness and Rounding dimensions, as well as Height and Roundness are predictable.

Other properties of vowels

- nasalisation (to be discussed: e.g. French)
- ATR tense / lax (e.g. Akan) Typically it is not only the tongue root gesture that is involved but in fact the whole pharyngeal cavity is normally enlarged (partly by tongue root movement and partly by lowering of the larynx) The additional articulatory effect is change in vowel height. Acoustically there is a considerable lowering of F1 and slight raising of F2.
- phonation types a) **creaky voice** (laryngealized, or *stiff voice*) characterised by concentration of energy in the first and second formants, more irregular vocal cord pulse rate (more jitter), b) **modal** (ordinary vowels), and c) **breathy voice** (more random energy, slack cords/voice, a larger noise component in the higher frequencies).

Acoustic characteristics of Consonants

(Voicing - Manner of Articulation - Place of Articulation for consonants in the spectrogram)

- **Voicing** (phonation types across languages to be discussed next time)
  voiced: Vertical *striations* corresponding to vocal fold vibration
  voiceless: Nothing during stop duration, but noise for aspiration or frication
- **Manner of Articulation**
  voiceless stops: Silent interval 70-140ms (=closure), long if unaspirated; strong release burst if aspirated, formants may be seen in noise, i.e. aspiration.

  voiced stops: Generally short closure (reason for having voiced geminates cross-linguistically?); voice bar during closure (not often for the whole duration and not always in English); weaker release burst; no aspiration

  voiceless: fricatives: Noise only; sibilant noise is much stronger than non-sibilant noise; voiceless fricatives are generally longer than voiced fricatives

  voiced fricatives: Weaker noise and voicing bar
  Non-sibilant ones may have no noise at all

  affricates: Silence for closure (not as long as a single stop) followed by a thin burst (not always) and then frication noise; voiced affricates have voicing bar at the bottom (low frequency)

  nasals: In general, formants in weak amplitude; resonance below 500Hz called *nasal murmur*, and anti-resonance (zero) between 500Hz and 2000Hz depending on the place of articulation. Often a discontinuity in amplitude between nasal and adjacent vowels. (Nasalised vowels or nasals show F1 with low amplitude and wide bandwidth)

  laterals: Vowel-like formant structure with weak amplitude. A discontinuity in formant amplitude from adjacent vowel. **/l/** has a low F1, low and weak F2 (a little higher than 1000Hz), and high and weak F3. Velarised /l/ has even lower F2

  approximants: Formants are like the corresponding vowels but lower amplitude and lower F1 (all consonants have lower F1 than vowels). In general, stronger amplitude than laterals and nasals; **/w/** has low F2 and weak lowish F3; **/j/** has high F2 and high F3; **/ɹ/** has a very low F3

  flap: very short closure duration (around 30ms). A clear discontinuity in amplitude from adjacent vowels. Often voiceless...

	F1	F2	F3
/ɹ/	low	low	**low**
/l/	low	lowish	**high** (higher than /w/)
/j/	low	**high**	high (higher than //l/)
/w/	low	low	lowish (close to /a/'s F3)

the order of F3: from the highest to the lowest:

j	>	l	>	w	>	ɹ
close to 3000Hz		close to 3000Hz		mid 2000Hz		below 2000Hz

- **Place of Articulation**
  - is cued by frequency of burst or frication noise. Burst/frication is formed from front cavity (shorter front cavity → higher frequency)
  - is also cued by frequency of aspiration and the location of F2 locus, which is predicted from adjacent vowel's F2

labial: 'diffuse falling' spectrum
Lack of front cavity filter. Weak and diffuse spectrum but lower frequencies stronger. Main peak in burst, if any, at F2 of the following vowel. Constriction at lips lowers all frequencies relatively. **Locus of F2 about 700~1200Hz**

alveolar: 'diffuse rising' spectrum
Stronger energy in high frequency: 4000Hz or higher. Strong release burst in apicals especially; laminals may be affricated. Frequency of frication noise depends on the size of front cavity: anterior is higher than non-anterior. Main peak in burst at or above F4 of the following vowel. **Locus of F2 about 1700~1800Hz.**

velar: 'compact' spectrum
F2 and F3 similar in frequency, so transitions converge (known as **velar pinch**). Higher frequency for front velars than back velars. Main peak in burst at F3 vs. F2 of the following vowel. **Locus of F2 about 3000Hz**

Acoustic/auditory cue robustness – contexts and markedness effects

(based on Wright 1996)

Cues to place: F2 transitions, stop release bursts, nasal pole-zero patterns, fricative noise

- *external*: strongest cues to place are found in the brief transitional period between a consonant and an adjacent segment, i.e. F2 transitions, burst.
  -why?: relatively better perception of transient sounds, cracks, bursts and rapid transitions
  -what adjacent segment?: vowel or consonant (glides and fricatives best)
- *internal*: cues to place "within" a consonant, e.g. noise in fricatives, formants in nasals liquids and glides. (fricative noise just as formants in sonorants (including vowels) can act as a "carrier" of external cues of preceding or following consonants because they have continuous noise).

> *A viable hypothesis:* the external cues (though better) will be more vulnerable in certain contexts than the internal cues. E.g. in the word-final context (C#), F2 transitions following the C will be absent, sometimes also the burst (in unreleased stops).

- **F2 transitions** are a) periodic with formant structure, b) transient and dynamic
  both the transitions **into** and **out of** the consonant constriction provide cues for place → C ←
- **fricative noise** is aperiodic with relatively long duration. Its spectrum is shaped primarily by the cavity in front of the noise source.
  -sibilants [s ʃ] easy to distinguish in terms of place (frequency region of the noise component)
  -for non-sibilants [f v θ ð] F2 necessary to distinguish the Place of Articulation. [v ð] least reliably distinguished: this is also a very rare contrast in the world's languages (Maddieson 1984)
- **stop release bursts** are aperiodic with a duration of approximately 5-10ms. Bursts play an important role in the perception of place, but if this cue conflicts with the F2 transition cue then the listeners rely on the latter more.
- **nasal cues** a) F2 transitions in the adjacent vowels, b) marked weakening in the upper formants due to the antiresonance (zero) and a low frequency resonance (pole) below 500Hz. F2 transitions are more reliable than the pole-zero patterns (**summary in (1)**).

Cues to manner: all oral constrictions will result in an attenuation of the signal. The relative degree of attenuation is a strong cue to the manner of a consonant.

- **abrupt attenuation:** of the signal in all frequencies (excepting the Fo frequency in voiced stops) is a cue to the presence of a stop. (*insertion of a period of silence in a signal, either between vowels or between a fricative and a vowel, results in the listener perceiving a stop (Bailey and Summerfield 1980)*)
- **noise:** a complete attenuation of the harmonic signal but with fricative noise provides the listener with cues to the presence of a fricative**.**
- **nasal murmur:** a less severe drop in amplitude accompanied by nasal murmur and a nasal pole and zero are cues to nasal manner. Nasalisation of the preceding vowel provides look-ahead cues to the nasal manner.
- **relative gradualness of F2 transitions:** as opposed to rapid spectral changes during the formant transitions suggests glides. (*lengthening the duration of synthesised formant transitions changes the listener's perception of manner from stop to glide!!!*) (**summary in (2)**)

Cues to voicing contrasts: periodicity in the signal is an obvious cue to voicing, and also: VOT, the presence and the amplitude of aspiration noise, durational cues.

- **periodicity** (formants and the voice bar in the spectrogram). This cue can be absent in English obstruents, especially stops in final position (C#). The listener must rely on other cues then.
- **VOT:** for syllable initial stops, the primary cue appears to be VOT lag (*the time between the release burst and the onset of voicing*). The relationship between VOT and voicing is language dependent, but generally:
  - -a negative VOT is a cue to voicing
  - -a short VOT is a cue to voicelessness
  - -a long VOT is a cue to aspiration

- **relative amplitude of the release burst (English):**
  - -a low amplitude burst cues voiced stops
  - -a high amplitude burst cues voiceless stops
- **duration of the preceding vowel:** voiceless is perceived if the vowel is short (**summary in (3)**)

CONCLUSION: There are many cues with differing acoustic characteristics distributed throughout the signal that point to a single contrast. There are also a number of acoustic aspects of the signal that are cues to more than one contrast (e.g. *the consonant release burst contains place, manner and voicing cues*)

Preferred contexts for cue preservation and markedness of syllabic and segmental types.

- CV > VC (*onset maximisation, *onsetless syllables, transitions are just as important for the perception of the vowel, hence: preferred CV and VC combinations will be those with more prominent F2 transitions: iwi > uwu*)
- #C... > ...C# (**no final coda, CVCV > VCV > CVC > VC, markedness of possible "codas": sonorants > fricatives > neutral stops > stops with VOT, neutralisation and markedness:* $p > p^h/b$*, why final devoicing and not voicing? why devoicing at all if the cue (lead) is internal to the segment?*)
- ...C_1C_2... **overlap** (see **4** on the other handout) *Generally:* C_1 *is a bad place for a stop*
  a) # C_1C_2V vs. b) V C_1C_2V vs. c) V C_1C_2#
  - best context for C_1 is **b** & **c;** best context for C_2 is **a** & **b**
  - best segment for C_1 is one with strong internal cues if C_2 is a stop

why is [a**rt**a] less marked than
[a**tr**a]? i.e. ↓ Rhyme-Onset
Branching Onset

Theories of organisation of speech

- Overlap principle (Mattingly 1981): a driving force in the organisation of speech is the need for both a *maximal speed of signal transmission* and a *robust encoding of the information* in the signal. This is achieved through **gestural overlap** in production that increases the transmission rate and robustly encodes information about the articulations. **How?**: The gestural overlap results in compression of information and a decreased transmission time. It also leads to *redundancy* of information in the signal. e.g. F2 transitions in CV provide cues for C and V (**aperture based!**). Redundancy makes the information in the signal more robust: if a portion of the signal is lost the segmental makeup of the string is still recoverable. ***Too great a degree of overlap can result in information loss.***

- Modulation principle (and increased salience) (Ohala 1992): change (modulation) along an acoustic dimension, such as *frequency* or *amplitude*, will result in increased salience of the cues in the portion of the signal where the change occurs. The grater the modulation and more dimensions that are involved, the greater the salience. The more modulation there is in the acoustic signal, the better the segmental organisation. **Reason?**: Rapid changes in frequency and amplitude result in a dramatic increase in activity in the auditory nerve fibres. (e.g. *this model predicts that the transition from a labial constriction with relatively low F2 into a front vowel with a relatively high F2 will be more salient than the transition into a back vowel with a relatively low F2, hence: certain co-occurrence restrictions!*)

☺ CVCV a) overlap and robust encoding, b) optimal signal modulation

☹ geminates a) little perceptual benefit from overlap, b) little signal modulation

😐 stop+glide+vowel+glide+stop are nearly as good as CVCV (but depend on sufficient formant modulation)

“coda” (VC#, VCCV) consonants with internal cues to place, manner and voicing will suffer less:

- **nasals** will do OK for manner and voicing, but they are only a little bit better than stops in terms of place because these cues are very weak in nasals. Thus: we expect a loss of place contrasts in coda nasals.
- **fricatives** are also better in preconsonantal and final position than stops (strong cues to place in their friction noise (particularly sibilants)

Word initially (#CCV) fricative+stop and nasal+stop are predicted to be much more common than stop+stop, and even stop+fricative, because they are not as reliant on vowel transitions for perceptual cues.

- **nasal** will be even more susceptible to loss of place [nd.., mb..]
- **fricatives** will be limited to the sibilants which have strong internal cues [sC.., ʃC..]

Characteristics of dispreferred sequences

- *small redundancy because of too much overlap and cue loss* (stop+stop) must obey the *audible-cue timing principle:* A language will permit a sequence of consonants that violate the principles mentioned above only in so far as cues to the segmental makeup can be preserved e.g. [**kt**o **tk**ać **czcz**y **dżdż**ysty **ss**ak] in Polish.
- *little modulation* e.g. [*ji, wu vs. ju, vi]

Basic Audition

Hearing: Auditory system transforms physical vibration of air into electrical signal that the brain can interpret. Thus, a sound input reached in our hearing system is not like a spectrogram. It is an ‘auditory spectrogram’ or ‘cochleagram’ (distance in F1-F2 is wider than in normal acoustic spectrogram) which reflects the sensitivity of frequency and amplitude in human ears.

Human auditory system is not a high-fidelity system.

- amplitude is compressed
- frequency is warped and smeared
- adjacent sounds may be smeared together

Auditory system in brief (*stages in the translation of the soundwave into neural activity*)

Sound waves impinge upon the **outer ear**, and travel down the **ear canal** to the **eardrum** in the **middle ear**. The eardrum is a thin membrane of skin which moves in response to air pressure fluctuations (*conversion of sound pressure into vibrations*). These movements are conducted by a chain of three tiny bones in the middle ear, through the **oval window**, to the fluid-filled **inner ear**. There is a membrane (the **basilar membrane**) that runs down the middle of the conch-shaped inner ear (the **cochlea**). The cochlear fluid transmits vibrations to the membrane. This membrane is thicker at one end than the other. The thin end responds to the high-frequency components in the acoustic signal, while the thick end responds to low-frequency components. Each **auditory nerve** fibre innervates a particular section of the basilar membrane, and thus carries information about a specific frequency component in the acoustic signal *(transform mechanical vibrations into electric impulses, hence, we are talking about* ***firing*** *of auditory nerves).* In this way the inner ear performs a kind of Fourier analysis of the acoustic signal, braking it down into separate frequency components.

What happens to the acoustic signal at different stages of reception?

- **ear canal** boosts frequencies 3500-4000HZ plus a large bandwidth around that. Why? It is a 2,5cm tube closed at one end so it acts like a quarter wave resonator (remember source and filter lecture and schwa production?).
- **ossicular chain** (three tiny bones in middle ear) **a)** attenuates particularly intense sounds (85dB and above, by muscular stiffening of the chain), **b)** amplifies the signal by ~5dB (to help overcome the greater impedance of the fluid-filled inner ear).
- **oval window** 18 times smaller area than that of the ear drum. This results in a 25 dB boost
- **basilar membrane** responds to different frequencies but quickly damps out higher frequencies, particularly above 4000-5000Hz. So it acts like a band of filters. The Bark scale is proportional to the distance along the basilar membrane. Roughly 60% of the length of the membrane responds to signals below 4000Hz

some non-lenearities

- non-linear scaling (frequency: **Bark-** greater sensitivity to changes in lower frequenncies, loudness: **Sones** – greater sensitivity to changes in lower intensities)
- onset of stimulus is more noticeable than offset (hV vs. Vh, F2 in _C)
- onset of sound sounds louder after silence
- upward masking – signal of lower frequency tends to suppress the response to and adjacent signal of higher frequency (ia, ua vs. ai, au?)
- forward masking – one signal drives down the response to a following signal with a lower intensity and similar frequency components (hV vs. Vh)
- after 50ms of steady duration response decayes