

INTERSPEECH 2020


© 2020 The University of Sheffield

What everyone working on spoken language processing needs to know about spoken language

SPEECH 101

Prof. Roger K. Moore
Chair of Spoken Language Processing
Dept. Computer Science, University of Sheffield
(Visiting Prof., Language Sciences, University College London)

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 1



1

INTERSPEECH 2020


© 2020 The University of Sheffield

~~THEORIES OF HOW HUMAN'S PROCESS SPOKEN LANGUAGE~~

versus

PROPERTIES OF SPOKEN LANGUAGE

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 2



2

INTERSPEECH 2020

© 2020 The University of Sheffield

Speech is not just Audible Text

Speech is ...

- variable
- ambiguous
- effortful
- contrastive
- prosodic
- adaptive
- context-dependent
- meaningful
- referential
- indexical
- rhetorical
- personalised
- affective
- multimodal
- contaminated



bbc.co.uk home



http://www.bbc.co.uk/voices/

communicative

embodied

ostensive

pragmatic

situated

grounded



The University Of Sheffield.

INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 3



3

INTERSPEECH 2020

© 2020 The University of Sheffield

Topics to be Covered

- Sound
- Speaking
- Hearing
- Phonetics
- Phonology
- Prosody
- Behaviour





The University Of Sheffield.

INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 4



4

© 2020 The University of Sheffield

INTER_SPEECH 2020


SOUND

Speech 101

The physics of speech

The University Of Sheffield.

INTER_SPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 5

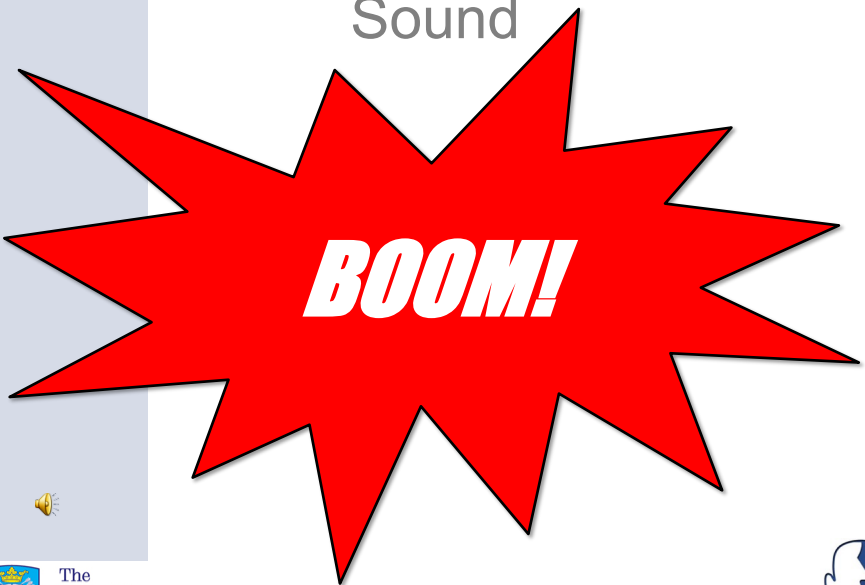


5

© 2020 The University of Sheffield


INTER_SPEECH 2020

Sound



The University Of Sheffield.

INTER_SPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 6



6

INTERSPEECH 2020

© 2020 The University of Sheffield

Sound is ...

http://commons.wikimedia.org/wiki/File:Ondes_compression_2d_20.gif

- acoustic energy
- a mechanical wave
- vibrations that travel through the air (*or other medium*)
- variation in (*air*) density

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 7

7

INTERSPEECH 2020

© 2020 The University of Sheffield

Measuring Sound

- Waveforms
(*time vs. amplitude*)
- Spectra
(*amplitude vs. frequency*)
- Spectrograms
(*amplitude vs. frequency vs. time*)

WARNING
It is incorrect to refer to "the frequencies present in a signal" – it should be "the energy at different frequencies"

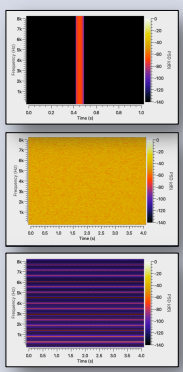
The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 8

8

© 2020 The University of Sheffield

Sound Characteristics

- The generation of sound requires energy
- Sound sources can be ...
 - 'impulsive'** (e.g. an explosion or a handclap): energy at all frequencies, but only for a short period of time
 - 'noisy'** (e.g. the wind or the surf on a beach): randomly varying energy at all frequencies
 - 'repetitive'** (e.g. a buzzing insect or a vibrating doorbell): energy at the 'fundamental' frequency of vibration and its harmonics
- These reflect the different ways in which sound pressure can vary over time



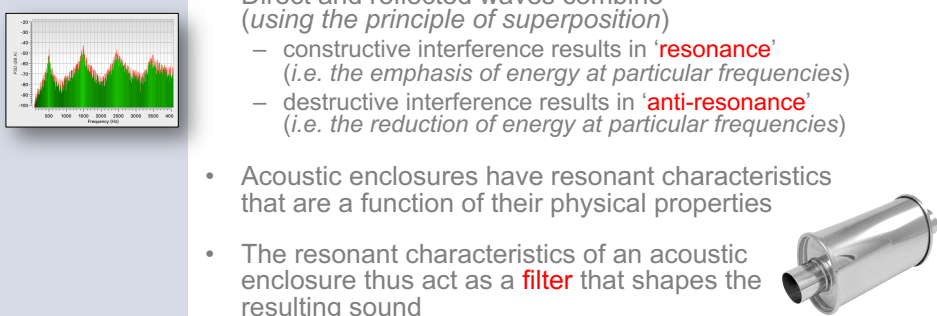
The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 9 SPANDI

9

© 2020 The University of Sheffield

Shaping Sound

- If a sound wave encounters an obstacle, a proportion of its energy will be reflected
- Direct and reflected waves combine (using the principle of superposition)
 - constructive interference results in **'resonance'** (i.e. the emphasis of energy at particular frequencies)
 - destructive interference results in **'anti-resonance'** (i.e. the reduction of energy at particular frequencies)
- Acoustic enclosures have resonant characteristics that are a function of their physical properties
- The resonant characteristics of an acoustic enclosure thus act as a **filter** that shapes the resulting sound
- A simple acoustic structure (such as a tube) has straightforward resonant characteristics



The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 10 SPANDI

10

INTERSPEECH 2020 © 2020 The University of Sheffield

Resonant Structures

$\lambda = 4L$

The resonant frequency f of a cylinder closed at one end and open at the other is given by ...

$$f = \frac{(2n-1)v}{4L}$$

where ...

- v is the speed of sound
- L is the length of the tube
- n is a positive integer (1, 2, 3, ...)

$n = 1$ gives the frequency of the 'fundamental' resonance
 $n = 2, 3, \dots$ gives the frequencies of the 'harmonics'

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 11 SPANDI

11

INTERSPEECH 2020 © 2020 The University of Sheffield

Resonant Structures

$\lambda = 4L$

The resonant frequency f of a cylinder closed at one end and open at the other is given by ...

$$f = \frac{(2n-1)v}{4L}$$

where ...

- v is the speed of sound
- L is the length of the tube
- n is a positive integer (1, 2, 3, ...)

$n = 1$ gives the frequency of the 'fundamental' resonance
 $n = 2, 3, \dots$ gives the frequencies of the 'harmonics'

input

output

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 12 SPANDI

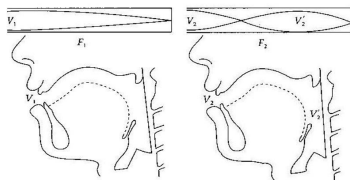
12

© 2020 The University of Sheffield


The Human Vocal Tract

- In the **neutral position** ("uh"), the human vocal tract is equivalent to a tube that is open at one end (*the mouth*) and closed at the other (*the larynx*)

WARNING
It is easy to confuse the fundamental frequency of a repetitive sound source with the fundamental frequency of a resonant structure – these are two different systems

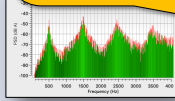



$$f = \frac{(2n-1)v}{4L}$$




So, the resonances for a 17.5 cm (*adult male*) vocal tract are ...

- fundamental = 490 Hz
- 1st harmonic = 1470 Hz
- 2nd harmonic = 2450 Hz
- 3rd harmonic = 3430 Hz
- etc.

The University Of Sheffield.

INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 13



13

© 2020 The University of Sheffield

SPEAKING

Speech 101

The speech production system



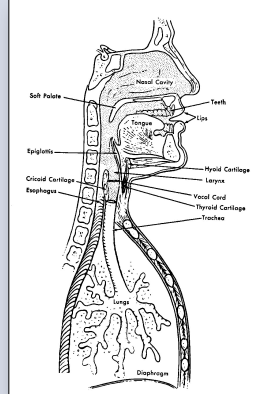
The University Of Sheffield.

INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 14



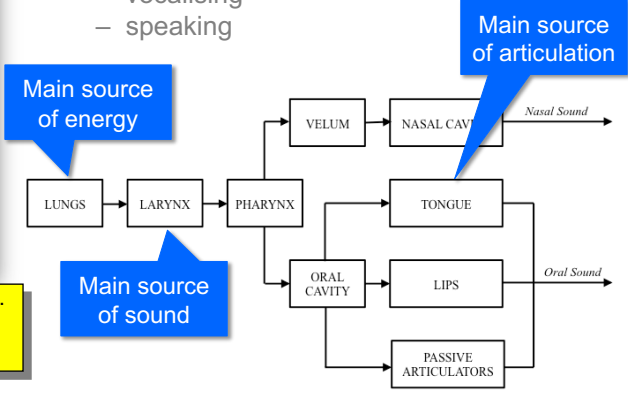
14

The Human Vocal Organs



The vocal apparatus has evolved for ...

- breathing
- eating
- vocalising
- speaking



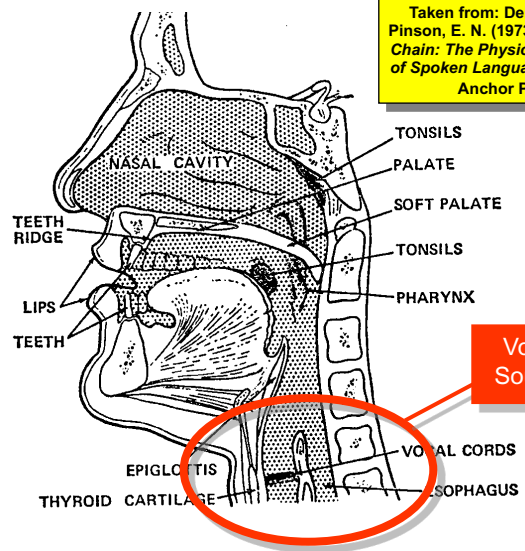
Taken from: Denes, P. B., & Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*. New York: Anchor Press.

The Human Vocal Tract



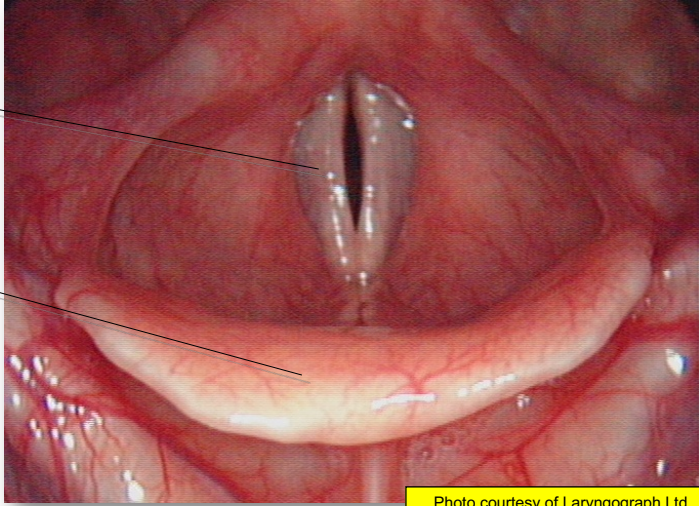
<http://sail.usc.edu/span>

Taken from: Denes, P. B., & Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*. New York: Anchor Press.



INTERSPEECH 2020 © 2020 The University of Sheffield

The Human Larynx



Vocal Fold

Epiglottis

Photo courtesy of Laryngograph Ltd.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 17 SPANDI

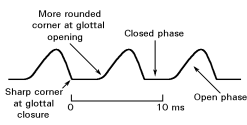
17

INTERSPEECH 2020 © 2020 The University of Sheffield

Voice 'Source'

WARNING
The vocal folds are also known as vocal cords, but they are not "vocal chords"!

- Air pressure from the lungs builds up behind closed **'vocal folds'**
- The vocal folds are repeatedly forced apart and pulled together again, producing a series of small pulses of air
- This modulation of the airstream is known as **'phonation'**
- The tension in the muscles attached to the vocal folds determines their rate of vibration and hence the **'fundamental frequency'** (and *perceived pitch*) of the voice
- Because the vibration is not a pure tone, there is energy at the fundamental frequency and its harmonics



WARNING
Referring to the fundamental frequency as "F0" is misleading (*it is not the zeroth resonance*) - "Fx" is preferable.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 18 SPANDI

18

INTERSPEECH 2020 © 2020 The University of Sheffield

The Human Vocal Tract

Taken from: Denes, P. B., & Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*: New York: Anchor Press.

USC SPAN
<http://sail.usc.edu/span>

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 19 SPANDH

19

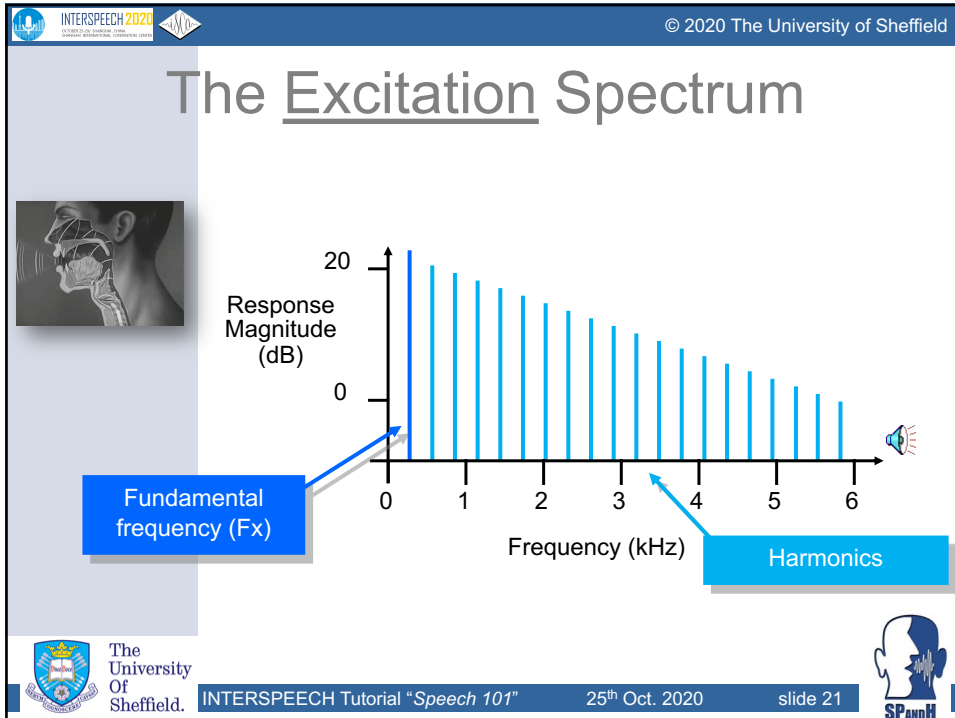
INTERSPEECH 2020 © 2020 The University of Sheffield

Voice 'Filter'

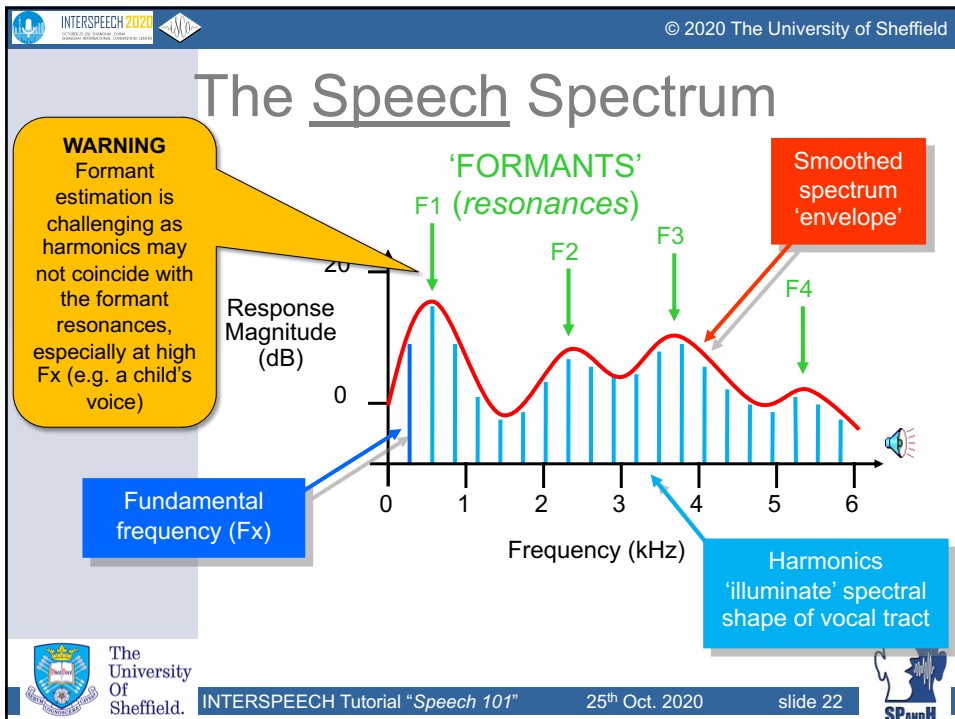
- The vocal tract forms a **resonator** with a complex shape
- Resonances are known as **formants**
- Speech is produced by using the **articulators** to change the shape of the vocal tract, hence modifying its resonant characteristics
- Different configurations of the vocal tract enhance some of the harmonics of the pitch, and suppress (*damp*) others
- The principal articulator is the **tongue**, but the jaw, lips, soft palate and teeth are also involved

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 20 SPANDH

20



21



22

© 2020 The University of Sheffield

Vocal Tract Shape and Spectra

96 ELEMENTS OF ACOUSTIC PHONETICS

THE PRODUCTION OF SPEECH

Peaks = resonances = 'formants'

F1, F2

Taken from: Ladefoged, P. (1962). *Elements of Acoustic Phonetics*: London: University of Chicago Press.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 23 SPANDH

23

© 2020 The University of Sheffield

Simulating Vocal Tract Shapes

Example of an electrolytic

Cancer Research UK

Glottis Lips

3D printed vocal tracts courtesy of Takayuki Arai

Adapted from: Vary, P. & Martin, R. (2006), *Digital Speech Transmission*, Wiley.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 24 SPANDH

24

© 2020 The University of Sheffield

Simulating the Vocal Tract

<https://dood.ai/pinktrampoline/>

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 27 SPANDI

27

© 2020 The University of Sheffield

HEARING

Speech 101

The auditory system

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 28 SPANDI


28

© 2020 The University of Sheffield


The Human Ear

- The auditory system has evolved for ...
 - acoustic sensing
 - sound localisation
 - communication
- Its primary function is **frequency analysis**
- The main percepts are ...
 - 'pitch' (for a repetitive sound)
 - 'loudness' (logarithmically related to the sound pressure level)
 - 'timbre' (related to the energy at different frequencies)

WARNING
Perceived pitch may not correspond to the fundamental frequency of a complex sound!



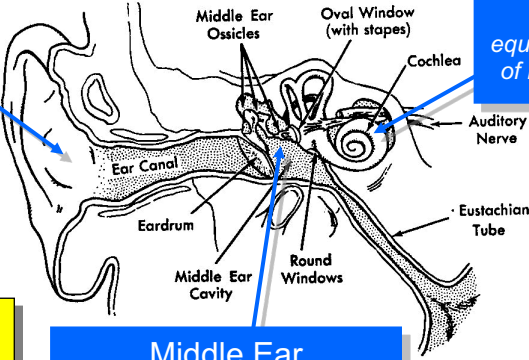
The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 29



29

© 2020 The University of Sheffield

The Human Ear




Outer Ear
directionally-sensitive

Inner Ear
place-dependent
liquid-to-neural
transduction
equivalent to a bank
of bandpass filters

Middle Ear
air-to-liquid transduction with
mechanical amplification and
overload protection

Taken from: Denes, P. B., & Pinson, E. N. (1973). *The Speech Chain: The Physics and Biology of Spoken Language*. New York: Anchor Press.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 30



30

INTERSPEECH 2020 © 2020 The University of Sheffield

Action of the Cochlea

<https://youtu.be/dyenMluFaUw>

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 31 SPANDI

31

INTERSPEECH 2020 © 2020 The University of Sheffield

Wideband Speech Spectrogram

Taken from: Holmes, J. N., & Holmes, W. (2002). *Speech Synthesis and Recognition*: Taylor & Francis.

Vertical striations correspond to the sudden closure of the vocal folds

THE NEW BRICKS FOR THE OVEN

ð ə n j u z b r ɪ k s f ɔ r ð ə u v ə

Frequency (kHz)

Time (s)

Short time analysis window provides ...

- good time resolution
- poor frequency selectivity

Horizontal bands correspond to formant resonances

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 32 SPANDI

32

© 2020 The University of Sheffield

Narrowband Speech Spectrogram

Taken from: Holmes, J. N., & Holmes, W. (2002). *Speech Synthesis and Recognition*: Taylor & Francis.

Horizontal striations correspond to harmonics of Fx

Frequency (kHz)

Time (s)

Pitch rising

In this region the harmonics are moving so fast that they are not seen at their expected spacing

Pitch falling

irregular very low pitch

Section point (a)

Long time analysis window provides ...

- good frequency selectivity
- poor time resolution

Harmonics have higher energy at formant resonances

WARNING
The term 'spectrogram' is underspecified unless the window size is given

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 33

33

© 2020 The University of Sheffield

Spectrogram vs. Cochleagram

"They enjoy it when I audition"

Frequency [Hz]

Time [sec]

SPECTROGRAM

Frequency [Hz]

Time [sec]

COCHLEAGRAM

Note difference in frequency scales: spectrogram is linear, cochleagram is approximately logarithmic

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 34

34

© 2020 The University of Sheffield

Demo: Real-Time Speech Analysis

Frequency (Hz)

Time (s)

PSD (dB)

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 35

35

© 2020 The University of Sheffield

Time for a BREAK

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 36

36

© 2020 The University of Sheffield

INTER_SPEECH 2020


SOUNDS & SYMBOLS

Speech 101

'Visible Speech'

The University Of Sheffield.

INTER_SPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 37



37

© 2020 The University of Sheffield

INTER_SPEECH 2020

Writing Systems

Writing
writing
Writing


WARNING
Some languages (e.g. Korean) have a closer relationship between spelling and pronunciation than others (e.g. English)

- Writing is an ancient technology that has been invented by human beings as a **'communication aid'**
- Writing allows information to be:
 - transmitted over space
 - stored over time
- The basic unit in most languages is the **'word'** (*because it carries meaning*)
- Meaningful **'utterances'** are made up from sequences of words
- Languages depict words using alphabets, syllabaries, logographs or ideographs
- Writing systems that use alphabets and syllabaries are **'phonetic'**, i.e. based on how the words sound

WARNING
Not all languages have a writing system

The University Of Sheffield.

INTER_SPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 38



38


© 2020 The University of Sheffield

Sound-Based Writing Systems

C	С	С	С
Е	Э	Е	Э
Q	Q	Q	Q
В	В	В	В
Г	Г	Г	Г
Д	Д	Д	Д
Е	Е	Е	Е
Ж	Ж	Ж	Ж
З	З	З	З
И	И	И	И
К	К	К	К
Л	Л	Л	Л
М	М	М	М
Н	Н	Н	Н
О	О	О	О
П	П	П	П
Р	Р	Р	Р
С	С	С	С
Т	Т	Т	Т
У	У	У	У
Ф	Ф	Ф	Ф
Х	Х	Х	Х
Ц	Ц	Ц	Ц
Ч	Ч	Ч	Ч
Ш	Ш	Ш	Ш
Щ	Щ	Щ	Щ
Ъ	Ъ	Ъ	Ъ
Ы	Ы	Ы	Ы
Ь	Ь	Ь	Ь
Э	Э	Э	Э
Ю	Ю	Ю	Ю
Я	Я	Я	Я

- Sound-based writing can be thought of as a kind of visible speech
- **'Visible Speech'** was the name given to a sound-based writing system invented in 1867 by Alexander Melville Bell (*father of Alexander Graham Bell*)
- Visible Speech was intended to help hearing-impaired people learn spoken language
- The symbols were representations of the positions of the vocal organs, hence it was independent of a particular language or dialect

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 39



39

© 2020 The University of Sheffield

Phonetic Transcription

WARNING
Not "phonemes"!


I
P
A

Founded
in 1886

- Languages select from 200-300 basic speech sounds known as **'phones'**
- The International Phonetic Association (IPA) has agreed a set of standard symbols for representing any speech sound
- A trained **'phonetician'** can use the IPA symbols to transcribe spoken utterances from any language (*even a newly discovered one*) or any speaker
- The symbols in an IPA phonetic transcription are written between square brackets, e.g. ...
"hello" → [heləʊ]
"speech processing" → [spi:tʃ pɹəʊsesɪŋ]

<https://www.internationalphoneticassociation.org>

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 40



40

© 2020 The University of Sheffield

The International Phonetic Alphabet

Consonants

Vowels

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

Phonetic	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		ɳ̠	ɲ	ŋ	ɴ		
Trill		ʙ		ʀ							
Tap or Flap				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			

<https://www.internationalphoneticassociation.org>

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 41

41

© 2020 The University of Sheffield

IPA Consonants

CONSONANTS (PULMONIC)											
	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		ɳ̠	ɲ	ŋ	ɴ		
Trill		ʙ		ʀ					ʀ		
Tap or Flap				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			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

English Consonants

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 42

42

IPA Consonants

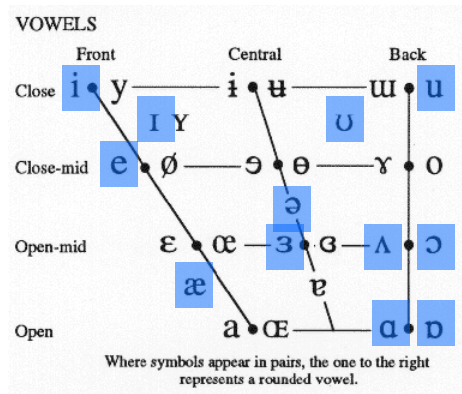
CONSONANTS (PULMONIC)

	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			ʀ					ʀ		
Tap or Flap				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			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Cantonese

IPA Vowels



English Vowels

© 2020 The University of Sheffield

IPA Vowels

Where symbols appear in pairs, the one to the right represents a rounded vowel.

Arabic Vowels

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 45

45

© 2020 The University of Sheffield

SOUND SYSTEMS

'physiophonic'
(language-independent)

'psychophonic'
(language-dependent)

PHONETICS

- auditory
- acoustic
- articulatory

PHONOLOGY

- contrast
- distribution
- function

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 46

46

INTERSPEECH 2020


© 2020 The University of Sheffield

ARTICULATORY PHONETICS

Speech 101

The description of speech sounds in terms of the physical actions performed in their production

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 47

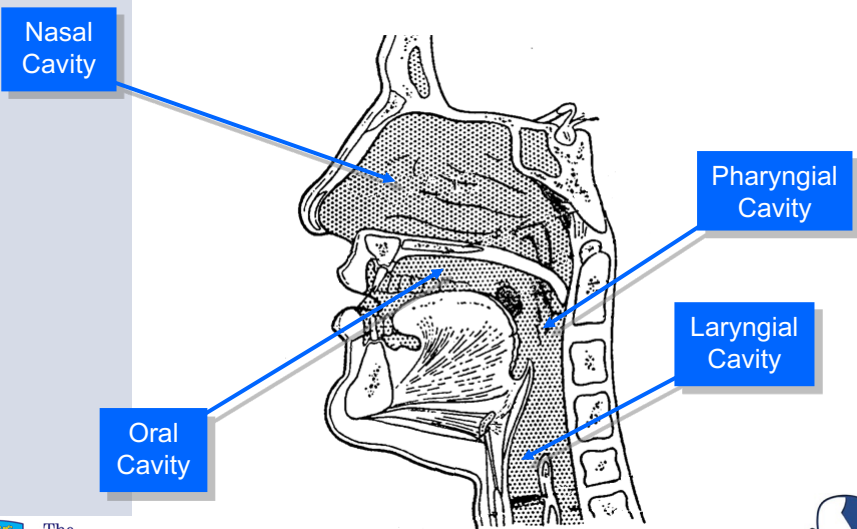


47

INTERSPEECH 2020

© 2020 The University of Sheffield

The Resonant Cavities




Nasal Cavity

Oral Cavity

Pharyngeal Cavity

Laryngeal Cavity

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 48



48

© 2020 The University of Sheffield

The 'Articulators'

Alveolar Ridge

Upper Teeth

Upper Lip

Lower Lip

Tongue

Hard Palate

Soft Palate (Velum)

Uvula

Vocal Cords

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 49 SPANDI

49

© 2020 The University of Sheffield

The Tongue

Tongue Front

Tongue Blade

Tongue Tip

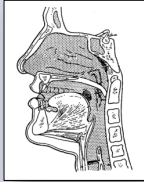
Tongue Back

Tongue Root

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 50 SPANDI

50

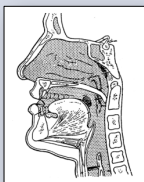
The Syllable



- The 'syllable' is the shortest stretch of speech
- Syllables consist of ...
 - 'vowels': sound segments produced using an unobstructed configuration of the vocal tract
 - 'consonants': sound segments in which the airflow is at least partly obstructed
- A simple 'CVC' (*c*onsonant-*v*owel-*c*onsonant) syllable corresponds to the opening and closing of the mouth
- Words can be ...
 - monosyllabic (*having one syllable*)
 - polysyllabic (*having two or more syllables*)



Speech Sounds

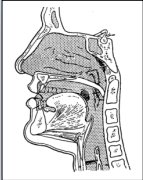


- In articulatory phonetics speech sounds are classified according to ...
- where the air stream comes from
 - whether air is going in or out
 - whether the vocal cords are vibrating
 - the location of any constriction
 - how the sound is made
 - the position of the tongue
 - the shape of the lips



INTERSPEECH 2020 © 2020 The University of Sheffield

Consonants



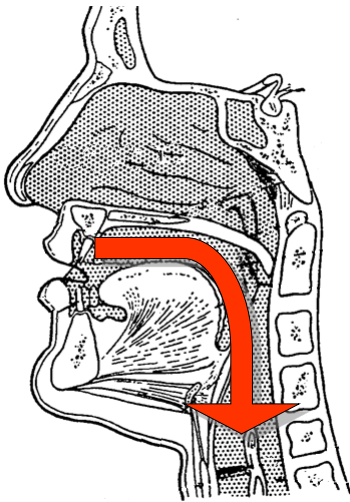
- **‘Voice’**
 - degree of voicing (*voiced-voiceless*)
 - voice quality (e.g. *modal, creaky, falsetto, breathy, etc.*)
- **‘Place’**
 - the physical location of the constriction in the vocal tract
- **‘Manner’**
 - the way in which the airstream is modified by the primary and secondary articulators
 - the degree of stricture (*closure, narrowing or approximant*)

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 53 SPANDI

53

INTERSPEECH 2020 © 2020 The University of Sheffield

‘Place’ of Articulation



- Bilabial
- Labiodental
- Dental
- Alveolar
- Postalveolar
- Retroflex
- Palatal
- Velar
- Uvular
- Pharyngeal
- Glottal

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 54 SPANDI

54

© 2020 The University of Sheffield

‘Place’ of Articulation

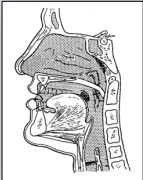
Place of Articulation	Active Articulator	Passive Articulator	Example Sounds
Bilabial	upper and lower lips	<i>none</i>	[p b m]
Labiodental	lower lip	upper front teeth	[f v]
Dental	tongue tip	upper front teeth	[θ ð]
Alveolar	tongue tip or blade	alveolar ridge	[t d n l s z]
Postalveolar	tongue tip or blade	rear of alveolar ridge	[ʃ ʒ]
Retroflex	tongue tip	hard palate	[ʈ ɖ ɳ]
Palatal	tongue front	hard palate	[j ɲ]
Velar	tongue back	soft palate	[k g ŋ]
Uvular	tongue back	uvula	[q ɢ]
Pharyngeal	tongue root	rear wall of pharynx	[ħ]
Glottal	vocal folds	<i>none</i>	[h ʔ]

Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 55

55

© 2020 The University of Sheffield

‘Manner’ of Articulation



- Stops
 - complete blockage of the airstream
 - can be produced at many different places of articulation
 - e.g. ‘**plosives**’ [p t k b d g], ‘**affricates**’ [tʃ dʒ], ‘**nasals**’ [n m]
- Fricatives
 - sound is produced with a very narrow opening
 - the resulting airflow is turbulent and hence noisy
 - e.g. [f s h]
- Approximants
 - stricture is not narrow enough to cause turbulence
 - e.g. ‘**lateral**’ [l], ‘**glides**’ [w v j]
- Taps & Trills
 - a single closure much shorter than for a plosive, e.g. [ɾ]
 - a series of rapid tap-like closures, e.g. [r]

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 56

56

INTERSPEECH 2020 © 2020 The University of Sheffield

Voice, Place, Manner

I
P
A

Manner

Place

CONSONANTS (PULMONIC)											
	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				ʀ					ʁ		
Tap or Flap				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			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Voice

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 57

57

INTERSPEECH 2020 © 2020 The University of Sheffield

Vowels

- Vowels are articulated by ...
 - raising the front or the back of the tongue towards the roof of the oral cavity
 - shaping the lips
 - ... thereby changing the 'vowel quality'
- Vowel quality is governed by ...
 - 'vowel height': *high / low*
 - 'vowel location': *front / back*
 - 'lip position': *rounded / unrounded*
- The mid-central vowel [ə] is called "schwa"

The University Of Sheffield.

INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 58

58

© 2020 The University of Sheffield

Vowels

Location

Front Central Back

Close i • y i • ʉ ʊ • u

Unrounded/rounded

Open-mid ɛ • œ ə • ɘ ɤ • ɞ

Open æ • ɶ ɶ • ɶ ɒ • ɔ

Height

“schwa”

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 59

59

© 2020 The University of Sheffield

ACOUSTIC PHONETICS

Speech 101

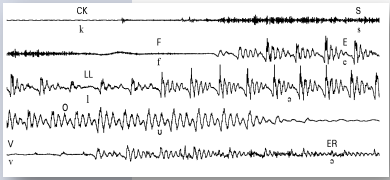
The description of speech sounds in terms of the acoustic consequences of their production

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 60

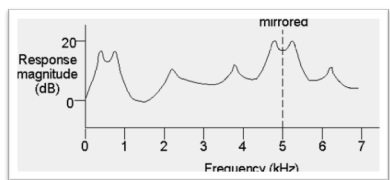
60

© 2020 The University of Sheffield

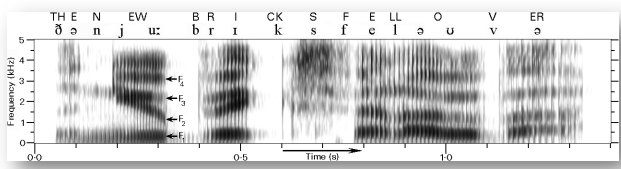
Acoustic Consequences



Waveforms



Spectra



Spectrograms

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 61

61

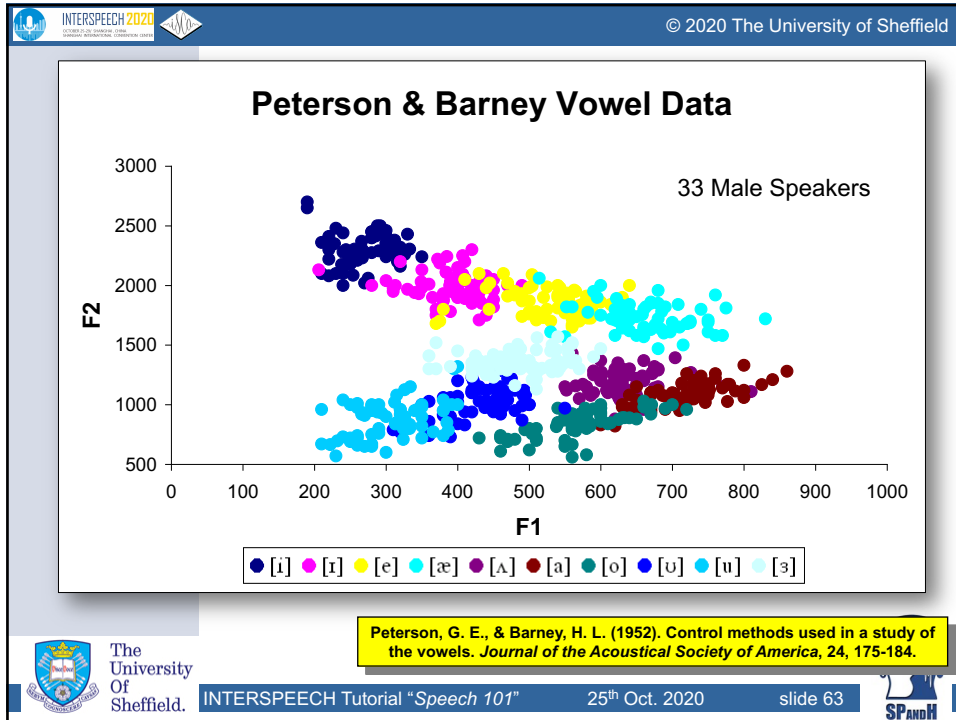
© 2020 The University of Sheffield

Stationary Sounds

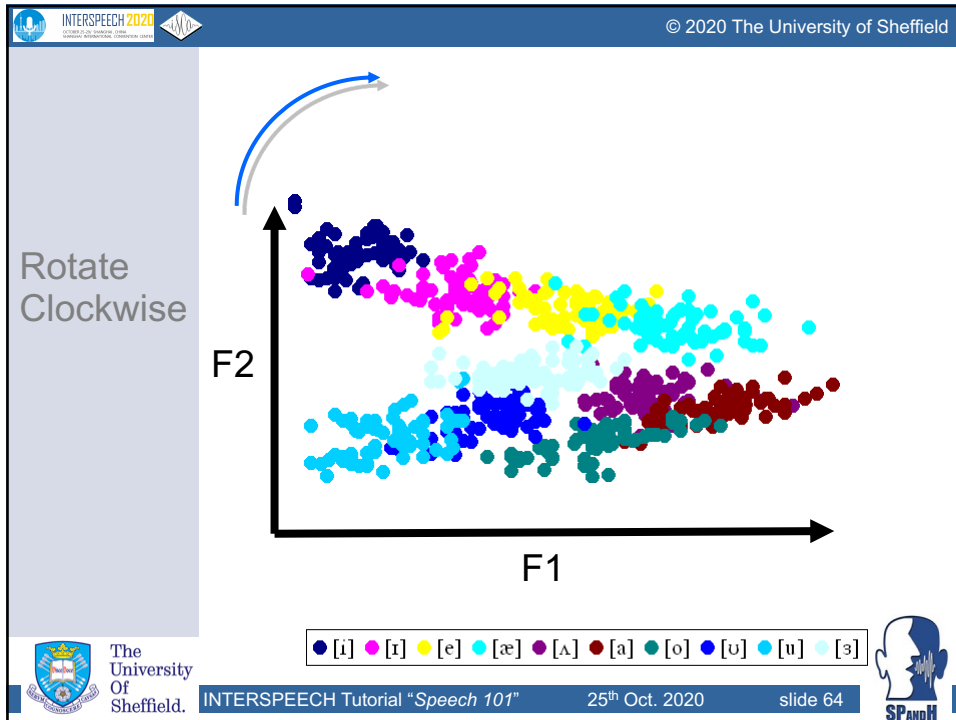
- Some speech sounds ('**continuants**') can be sustained over time and retain their phonetic quality
- For example ...
 - monophthong vowels, e.g. [i: æ u:]
 - all fricatives, e.g. [f θ v ð s ʃ z ʒ]
 - some approximants, e.g. [ɹ l]
 - all nasals, e.g. [n m ŋ]

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 62

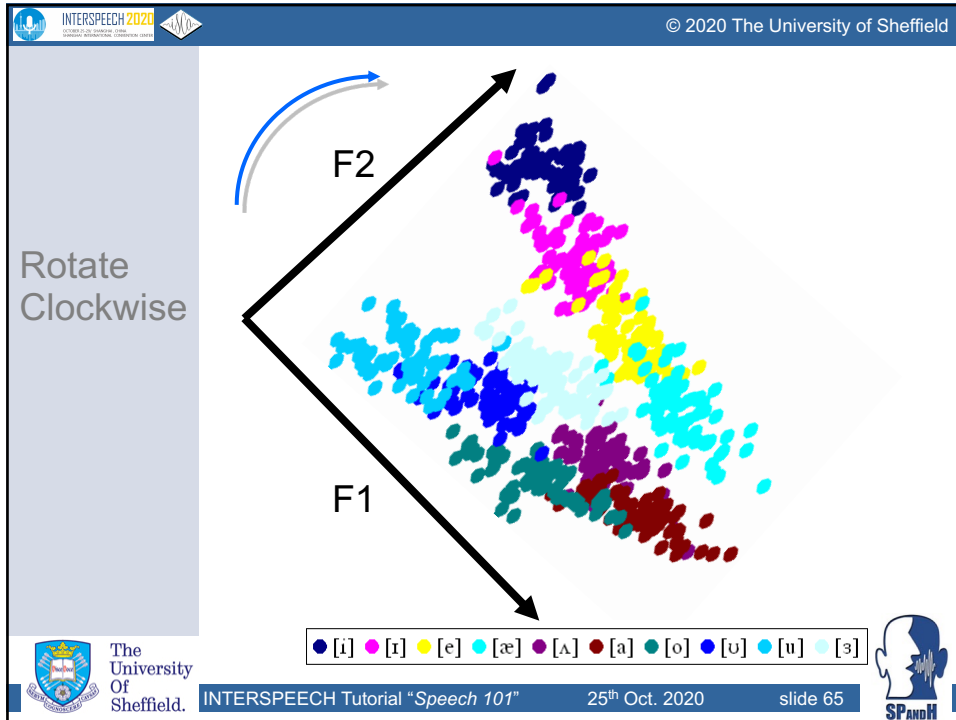
62



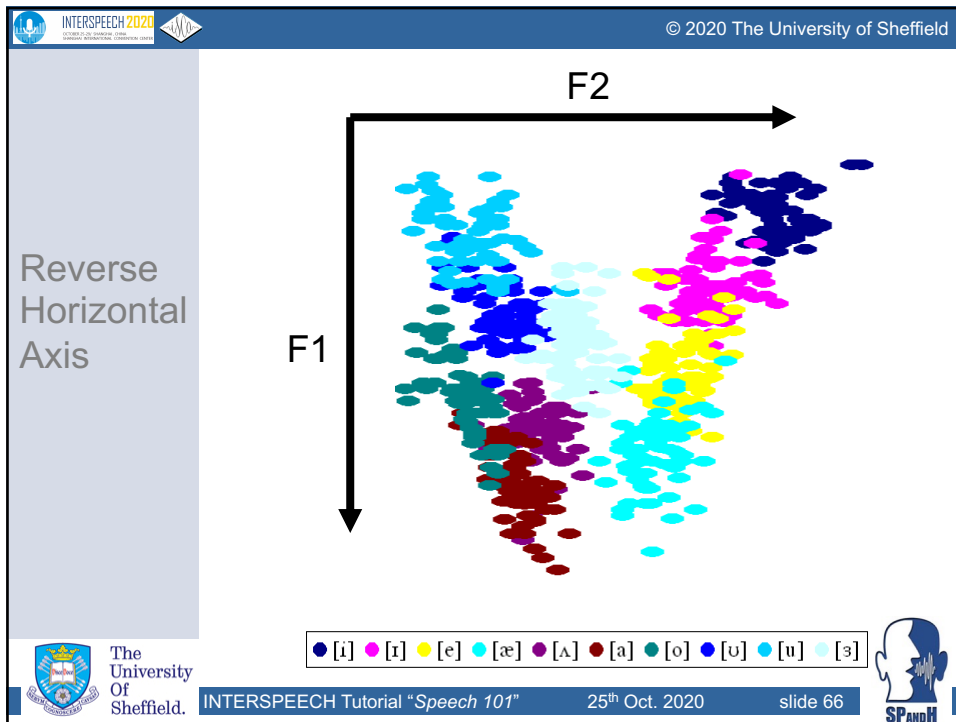
63



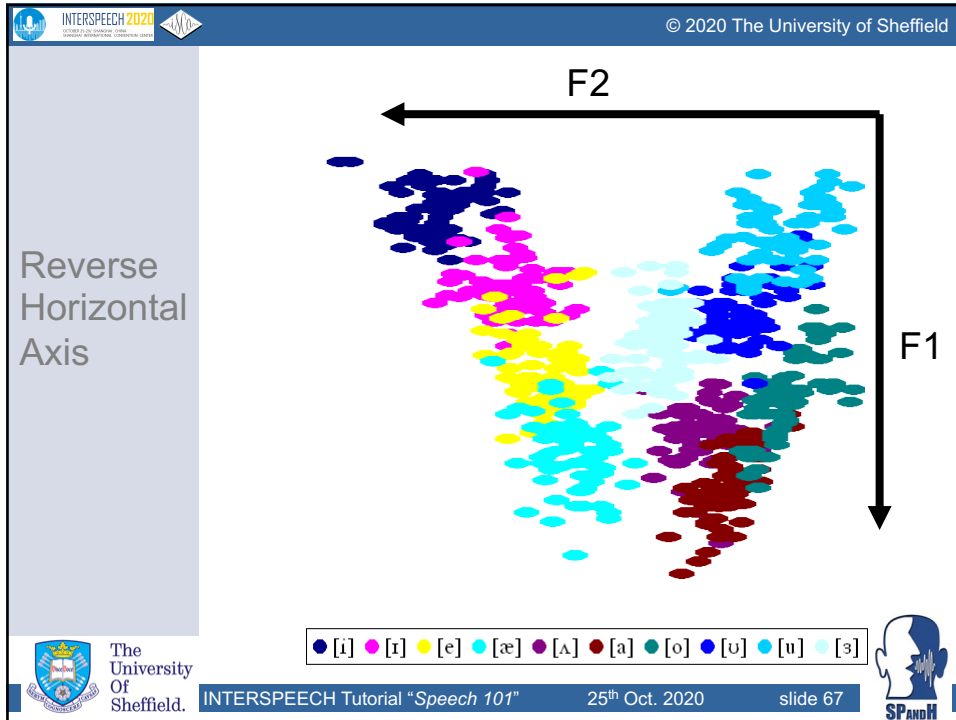
64



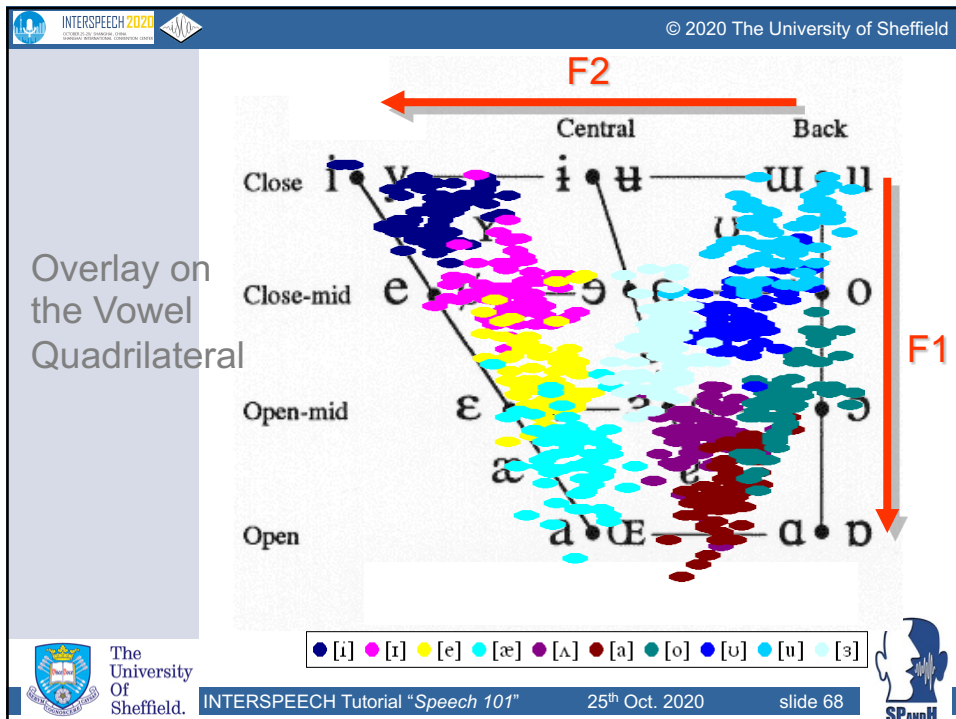
65



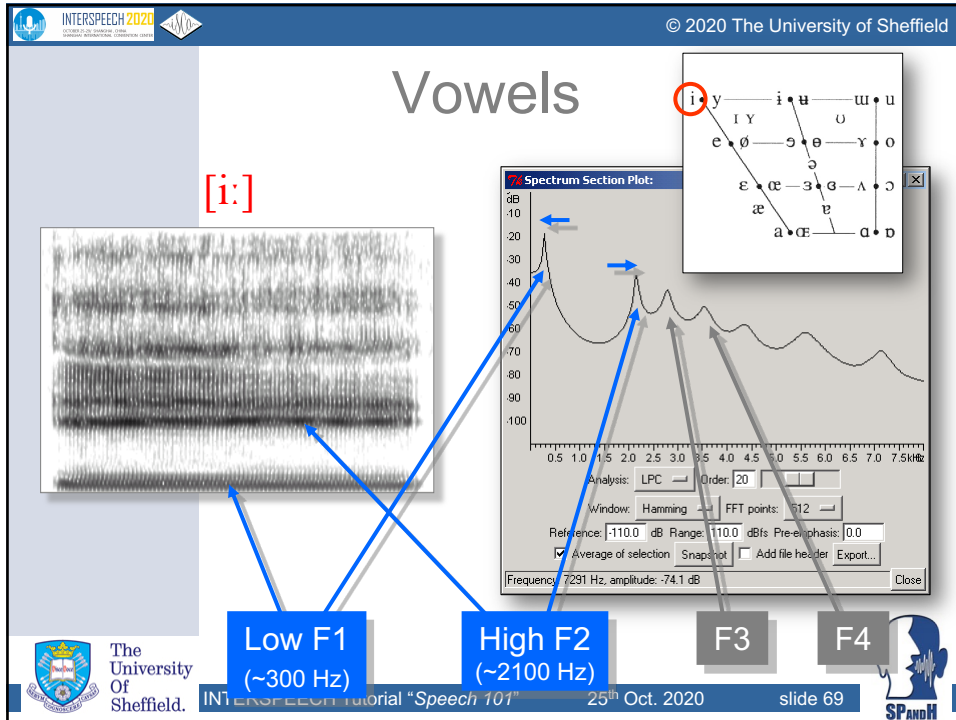
66



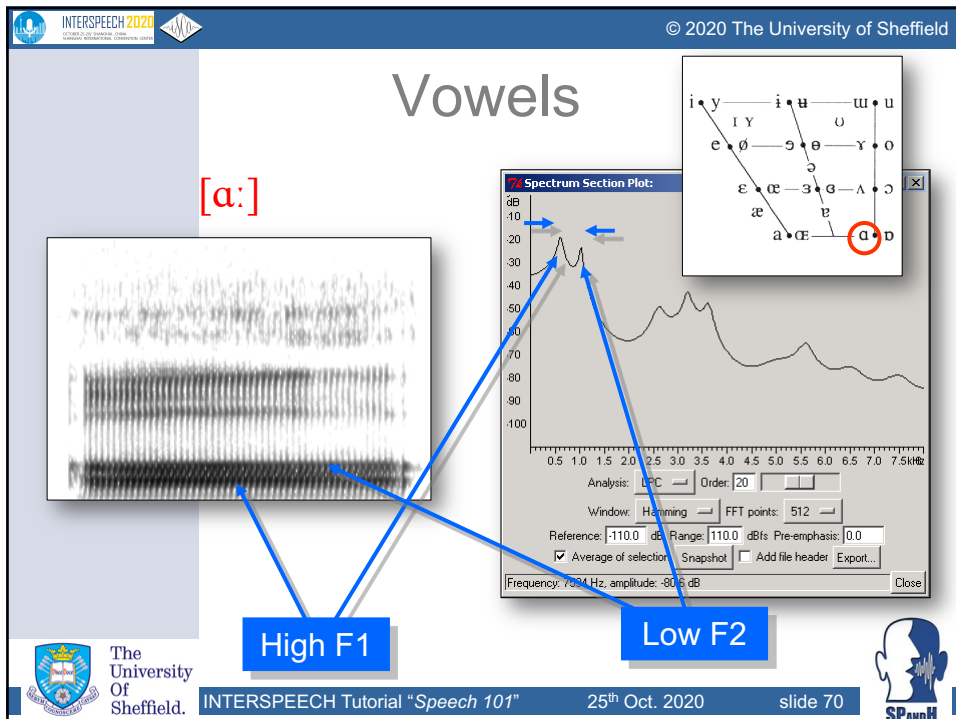
67



68



69



70

© 2020 The University of Sheffield

Two Vowels

[i:] [ɑ:]

[i:ɑ:]

F2 → → F2
F1 → → F1

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 71

71

© 2020 The University of Sheffield

Fricatives

[f] [θ] [s] [ʃ]

[v] [ð] [z] [ʒ]

No appreciable energy at low frequencies

Voicing energy at low frequencies

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 72

72

© 2020 The University of Sheffield

Nasals

[m] [n] [ŋ]

Nasal Formant

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 73

73

© 2020 The University of Sheffield

Non-Stationary Sounds

- The essence of some speech sounds is that they change over time
- For example ...
 - diphthongs, e.g. [eɪ əʊ]
 - plosives, e.g. [p b t d k g]
 - affricates, e.g. [tʃ dʒ]
 - glides, e.g. [w j]

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 74

74

© 2020 The University of Sheffield

Plosives

[a:pa:]

[a:ta:]

[a:ka:]

[a:ba:]

[a:da:]

[a:ga:]

Voice plosives have shorter 'voice onset times' (VOTs)

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 75

75

© 2020 The University of Sheffield

Question: *What's This?*

Answer: *whispered speech*

no voicing

Clue

"spectrogram"

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 76

76

© 2020 The University of Sheffield


Coarticulation

- ‘**Coarticulation**’ refers to the influence of one sound on another
- To a speech technologist/engineer, this is a form of ‘**context-dependency**’
- To a phonetician, it is a consequence of efficient **motor planning**
- Coarticulations can be both short and long range (e.g. *the influence of nasalisation or lip rounding can span several syllables*)

[[p]:[s]_[i:]]

WARNING
Coarticulation is not limited to adjacent phones

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 77




77

© 2020 The University of Sheffield

Coarticulation


“*Why are you early you owl*”



WARNING
This shows that the ‘boundaries’ between phones are a perceptual not acoustic phenomenon


[w aɪ ɑ: j u: ɜ: l i: j u: aʊ l]

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 78



78

© 2020 The University of Sheffield



Time for
a BREAK

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 79 SPANDI

79

© 2020 The University of Sheffield

PHONOLOGY

Speech 101

The study of the distribution and function of speech sounds in a given language

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 80 SPANDI

80

INTERSPEECH 2020


© 2020 The University of Sheffield

Phonemic Contrast

- The phonetic description of speech is language-independent
- Earlier, it was implied that only a certain number of phones occur in any given language ... this is not strictly correct
- In principle, all phones can occur in all languages
- However, acoustically distinct speech sounds are only perceived as different by native listeners if they distinguish between one word and another in that language
- Such sounds are described as **‘contrastive’**

WARNING
This is the ‘*psychophonic*’ phenomenon observed by early phoneticians

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 81







81

INTERSPEECH 2020


© 2020 The University of Sheffield

Phonemic Contrast

- The contrastive phones in a given language are called **‘phonemes’**
- The phonemic inventory of a language is found by exploring all of the possible **‘minimal pairs’**
- For example, ...
 -  English: *“fussy”* [fʌsɪ] versus *“fuzzy”* [fʌzɪ]
 -  French: *“choux”* [ʃy] versus *“joue”* [ʒy]
 -  Italian: *“fate”* [fato] versus *“done”* [fat.o]
 -  Spanish: *“dog”* [pero] versus *“but”* [pero]

WARNING
Phonemes are – *by definition* – language-specific

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 82



82

© 2020 The University of Sheffield


The 'Phoneme'

WARNING
This is a potential cause of confusion

- Phonemes are written using IPA symbols
- IPA symbols between **square brackets** signify a 'phonetic transcription', e.g. "hello" → [heləʊ]
- IPA symbols between **slashes** signify a 'phonemic transcription', e.g. "hello" → /heləʊ/
- A phonemic transcription is an idealised representation of an utterance, whereas a phonetic transcription represents the actual sounds used
e.g. "law and order" → /lə: ænd ɔ:dɜ:/ → [lə:ɪæno:də]
- The term 'phoneme' is often misunderstood and therefore misused

WARNING
The terms 'phone' and 'phoneme' are not synonymous


The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 83



83

© 2020 The University of Sheffield

Misuse of the Term 'Phoneme'



"We have 252 *phonemes*, of which there are 213 Mandarin and 39 English."

"There are 144 traditional *phoneme* states in a mono phone system."

"... below the minimum duration of a *phoneme* (30 ms) are considered as spurious regions."

"... treating filled pause as a special '*phoneme*'."

"... isolated *phonemes* extracted from CS [continuous speech] sentences."

"... trained with context-independent *phoneme* states as targets."


"Even though they all use the same *phoneme* symbols, each language and accent imposes its own coloring or 'twang'."

"We propose a language-independent *phoneme* segmentation method."

"Diphthongs and triphthongs [sic] are split into their constituent phones to reduce the number, and enforce sharing, of *phonemes*."

Moore, R. K., & Skidmore, L. (2019). On the use/misuse of the term 'phoneme'. In INTERSPEECH-2019. Graz, Austria.


The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 84



84

© 2020 The University of Sheffield

The 'Phoneme'



Daniel Jones (1881-1967)

Allophones can be thought of as 'equivalence classes'

- Daniel Jones defined the 'phoneme' as

"... a family of uttered sounds (segmental elements of speech) in a particular language which count for practical purposes as if they were one and the same."

Jones, D. (1973). The history and meaning of the term "phoneme". In E. C. Fudge (Ed.), *Phonology: Selected Readings* (pp. 17–34). Harmondsworth, UK: Penguin Books.
- The set of phones that make up a given phoneme are known as 'allophones', e.g. ...
 - Japanese: [ɺ] and [ɺ̚] are allophones of /ɺ/
 - English: [t̚] and [t̪] are allophones of /t/
 - Spanish: [d̪] and [d̪̞] are allophones of /d/
 - US English: [tʰ] and [t̚] are allophones of /t/

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 85

85

© 2020 The University of Sheffield

Phonological Processes

Phonemic 'intention' → Phonetic 'realisation'

- 'Assimilation' (feature spreading)
 - "can be"

/kæ̃n bi:/ → [kæ̃mbi:]
- 'Elision' (deletion)
 - "I don't know"

/aɪ dəʊ̃nt nəʊ/ → [adnəʊ]
- 'Epenthesis' (insertion)
 - "vanilla ice cream"

/vənɪlə aɪskri:m/ → [vənɪlə.ɪaɪskri:m]
- 'Reduction' (neutralised vowel quality)
 - "and"

/æ̃nd/ → [ɛ̃nd]

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 86

86

© 2020 The University of Sheffield

Phonological Processes

“Do you actually know any solicitors?”


Phonemic transcription: /du: ju: æktʃʊəli nəʊ enɪ səlɪsɪtə:z/

Orthographic transcription: Do you actually know any solicitors?

Phonetic transcription: [dʒæʃlinəʊwenɪslɪstəz]

Processes shown: elision, assimilation, epenthesis, reduction

The University of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 87



87

© 2020 The University of Sheffield


The ‘Phoneme Restoration Effect’

- Another consequence of the ‘psychophonic’ nature of phonemes is that they may be perceived even if they are not physically present
- E.g. [ɑdnəʊ] → /aɪ dəʊnt nəʊ/ → “I don’t know”
- If a short section of speech is cut out and replaced by another sound (such as a cough), listeners cannot detect that anything is missing ...

WARNING
Confirmation that phonemes are a perceptual phenomenon

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917), 392–393.

The University of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 88



88

© 2020 The University of Sheffield

PROSODY

Speech 101

The suprasegmental properties of speech

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 89 SPANDI

89

© 2020 The University of Sheffield

Prosody

'Prosody' is one way in which speakers are able to signal the important information in what they are saying

It allows different words to be *emphasised* and is critical to communicating unambiguous meanings

It's a bit like the 'punctuation' in text,
but prosody carries much more information!

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 90 SPANDI

90

INTERSPEECH 2020


© 2020 The University of Sheffield

Prosody

- Prosodic features span ...
 - several speech segments
 - several syllables
 - whole utterances
- Such **'suprasegmental'** behaviour includes ...
 - lexical stress
 - lexical tone
 - rhythmic stress
 - intonation

All carried by pitch, timing and loudness

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 91



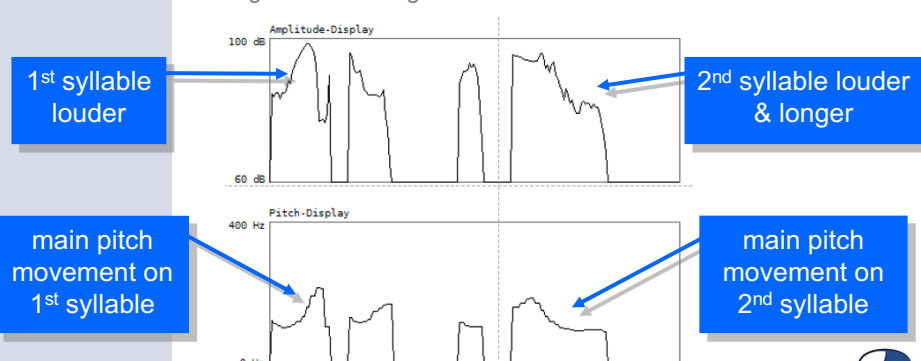
91

INTERSPEECH 2020

© 2020 The University of Sheffield

Lexical Stress

- Lexical stress refers to the **'prominence'** of syllables in words
- A stressed syllable is usually *louder* and/or *higher pitch* and/or *longer* than its neighbours ...



Amplitude-Display

Pitch-Display

1st syllable louder


2nd syllable louder & longer

main pitch movement on 1st syllable

main pitch movement on 2nd syllable

“written” → [ˈɪtən] “return” → [ɪˈtʃ:ɪn]

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 92



92

INTERSPEECH 2020 © 2020 The University of Sheffield

Lexical Tone

- About half of the world's languages use the pitch pattern to distinguish between one word and another
- Modern Standard Chinese has four lexical tones
 - high level, high rising, low falling-rising, high falling

施氏食獅史 **Lion-Eating Poet in the Stone Den**

石室詩士施氏，嗜獅，誓食十獅。 *In a stone den was a poet Shi, who was a lion addict, and had resolved to eat ten.*
 氏時時適市視獅。 *He often went to the market to look for lions.*
 十時，適十獅適市。 *At ten o'clock, ten lions had just arrived at the market.*
 是時，適施氏適市。 *At that time, Shi had just arrived at the market.*

氏視是十獅，持矢勢，使是十獅逝世。 *He saw those ten lions, and using his trusty arrows, caused the ten lions to die.*
 氏拾是十獅屍，適石室。 *He brought the corpses of the ten lions to the stone den.*
 石室濕，氏使侍拭石室。 *The stone den was damp. He asked his servants to wipe it.*
 石室拭，氏始試食是十獅。 *After the stone den was wiped, he tried to eat those ten lions.*
 食時，始識是十獅，實十石獅屍。 *When he ate, he realized that these ten lions were in fact ten stone lion corpses.*
 試釋是事。 *Try to explain this matter.*

http://en.wikipedia.org/wiki/Lion-Eating_Poet_in_the_Stone_Den

The University of Sheffield. INTERSPEECH tutorial "Speech 101" 25th Oct. 2020 slide 93 SPANDI

93

INTERSPEECH 2020 © 2020 The University of Sheffield

Intonation

- Intonation refers to pitch variation that doesn't affect the meaning of the words, but does affect the meaning of an utterance ...

Falling pitch contour **Rising pitch contour**

Amplitude-Display Amplitude-Display

Pitch-Display Pitch-Display

“It’s cold!” “It’s cold?”

The University of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 94 SPANDI

94

© 2020 The University of Sheffield

Intonational Accent

“This is my laptop”

“This is my laptop”

“This is my laptop”

“This is my laptop”

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 95 SPANDI

95

© 2020 The University of Sheffield

Intonational Accent

The primary intonational accent or ‘**nucleus**’ indicates a contrast in meaning ...

- “This is **my** laptop” (not yours)
- “**Jack** likes fish” (but George doesn’t)
- “Jack **likes** fish” (he doesn’t hate them)
- “Jack likes **fish**” (but not meat)

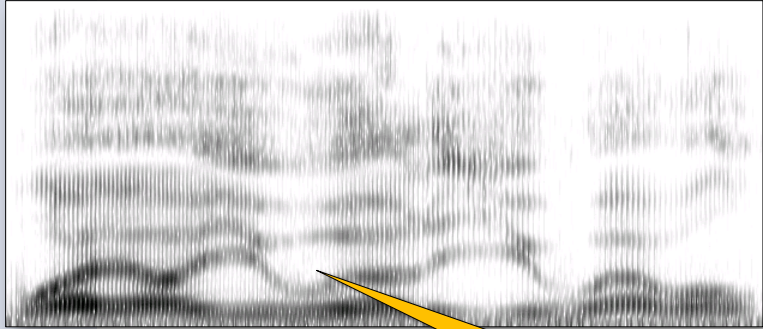
The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 96 SPANDI

96

© 2020 The University of Sheffield

Speech is ... *continuous*

“Why are you early you owl?”



WARNING
There are usually no gaps between words

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 99 SPANDI

99

© 2020 The University of Sheffield

Speech is ... *continuous*

“We were away a year ago.”



WARNING
Gaps can occur within a word

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 100 SPANDI

100

© 2020 The University of Sheffield

Speech is ... *variable*

العربي български català 中国话 hrvatski česky
 english ελληνικά עברית हिंदी italiano 日本語
 한국어 românește русский српски

There are 3000-8000 languages:

- 6500 'living' languages
- 83% are limited to single countries
- 52% are spoken by less than 10,000 people
- 28% are spoken by less than 1000 people

WARNING
 There are ~8,000,000,000 people in the world ... and they all speak differently

WARNING
 Half of the world's languages will be out of use by 2100




The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 101

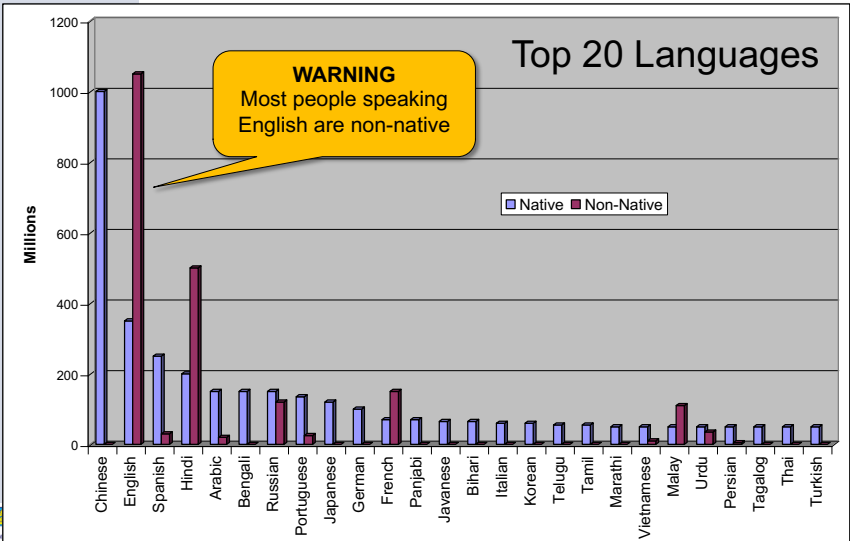


101

© 2020 The University of Sheffield


Speech is ... *variable*

Top 20 Languages




WARNING
 Most people speaking English are non-native

Language	Native (Millions)	Non-Native (Millions)
Chinese	1000	0
English	350	1050
Spanish	250	0
Hindi	200	500
Arabic	150	0
Bengali	150	0
Russian	150	0
Portuguese	150	0
Japanese	120	0
German	100	0
French	100	0
Punjabi	100	0
Javanese	100	0
Bihar	100	0
Italian	100	0
Korean	100	0
Telugu	100	0
Tamil	100	0
Marathi	100	0
Vietnamese	100	0
Malay	100	0
Urdu	100	0
Persian	100	0
Tagalog	100	0
Thai	100	0
Turkish	100	0



The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 102



102

Speech is ... *variable*



WARNING
This is not random, it is adaptation to the communicative context

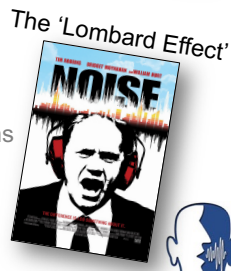
- Within any given language, groups of individuals use local dialects and accents
 - a '**dialect**' is when different words are employed
 - an '**accent**' is when different sounds are employed
- Accents and dialects reflect regional and/or social differences
- Even within a given language/dialect/accnt group, different individuals sound different
 - '**inter-speaker variation**' arising from age, gender, physical characteristics, social habits etc.
 - '**intra-speaker variation**' arising from physiological, psychological and external factors
- Also, people can speak loudly or softly, quickly or slowly, clearly or mumbled, formally or casually ... and this can change in the course of a single utterance

Speech is ... *adaptive*



WARNING
The 'Lombard Effect' is more than just speaking louder, it also involves speaking more clearly ('hyper-articulating')

- to the listener
 - a child ('parentese')
 - a non-native person
 - a hearing-impaired individual
 - an animal
 - a machine (!)
- to the task
 - casual conversation
 - reading out loud
 - public speaking
- to the cognitive load
 - interaction with other tasks
 - stressful/emotional situations
- to the environment
 - noise
 - reverberation



© 2020 The University of Sheffield

Speech is ... *effortful*

“Where’s the newspaper?”

[əəəə]

“I DO NOT KNOW!”
 “I do not know”
 “I don’t know”
 “I dunno”
 “dunno”

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373-405.

Designed by pch.vector / Freepik

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 105

105

© 2020 The University of Sheffield

Speech is ... *efficient*

“fish and chips” → “fish’n’chips”
 “temporary” → “tem’pry”
 “can be” → “cam be”
 “bread and butter” → “bre’m butter”

WARNING
 This is not ‘sloppy’ speaking! This is a normal part of communicative behaviour that balances the demands of speaking against the demands of listening

The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 106

106

© 2020 The University of Sheffield


Speech is ... *contrastive*

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Kluwer Academic Publishers.

- Signalling involves physical/mental effort
- Large effort creates clear signals but uses more energy (*and vice versa*)
- The ‘target’ is a perception not a signal
- So optimisation is over competing perceptions not competing signals
- The intention is sufficient **contrast** at the pragmatic level (*leading to suitable compensations at the semantic, syntactic, lexical, phonemic, phonetic and acoustic levels*)
- The obstacles are ...
 - alternative interpretations (*internal*)
 - competing signals (*external*)

Referred to as ‘pragmatics-first’

WARNING
This has huge implications as it shows that speech is not an absolute signalling system




The University Of Sheffield.

INTERSPEECH Tutorial “Speech 101”

25th Oct. 2020

slide 107



107

© 2020 The University of Sheffield

Speech is ... *ambiguous*



- Variability means that signals we’d like to be the same are actually quite different ...

speech speech speech speech speech



- Ambiguity arises when signals we’d like to be different end up being the same ...

“to” vs. “two” vs. “too”
“hear” vs. “here”

‘homophones’



The University Of Sheffield.

INTERSPEECH Tutorial “Speech 101”

25th Oct. 2020

slide 108



108

INTERSPEECH 2020 © 2020 The University of Sheffield

Speech is ... *ambiguous*






“fork handles” vs. *“four candles”*



The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 109

109

INTERSPEECH 2020 © 2020 The University of Sheffield

Speech is ... *ambiguous*



“four candles” vs. *“fork handles”*

“great ape” vs. *“grey tape”*

“law and order” vs. *“Laura Norder”*

“Joe is office head” vs. *“Joe is off his head”*

“This nudist play will wreck a nice beach”
vs.
“This new display will recognise speech”



The University Of Sheffield. INTERSPEECH Tutorial “Speech 101” 25th Oct. 2020 slide 110

110

© 2020 The University of Sheffield

Speech is ... *ambiguous*

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 111 SPANDI

111

© 2020 The University of Sheffield

Speech is ... *multimodal*

- Speech is not just an acoustic signal, it is also visual
- Listeners can derive a large benefit from seeing a speaker's lips
 - ~12 dB signal-to-noise gain
- The visual information can even *override* the acoustic information
 - the 'McGurk Effect'


Enough to render unintelligible speech intelligible

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 112 SPANDI

112

INTERSPEECH 2020 © 2020 The University of Sheffield

Speech is ... *multimodal*



The 'McGurk Effect'

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 113 SPANDI

113

INTERSPEECH 2020 © 2020 The University of Sheffield

Speech is ... *d'd'dis-disfluent*

- Spoken utterances are not as 'well-formed' as written sentences
- Normal spontaneous speech contains:
 - false-starts 🗣️
 - repeats 🗣️
 - filled pauses ("uhms" and "errs") 🗣️
 - overlaps
- So-called 'disfluencies' actually make speech easier to understand (*by revealing aspects of the planning process*)

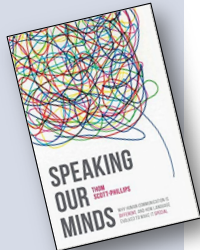
WARNING
This is true for human listeners

The University of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 114 SPANDI

114

Speech is ... *spoken language*

- Utterances are ...
 - meaningful
 - referential
 - indexical
 - structured
 - intentional
 - pragmatic
- But speech is also ...
 - real-time
 - continuous
 - interactive
 - coupled
 - incremental
 - synchronous



Communicative behaviour is founded on “ostensive inferential recursive mind-reading”

Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. London, New York: Palgrave MacMillan.



University Of Sheffield.


INTERSPEECH Tutorial “Speech 101”

25th Oct. 2020

slide 115

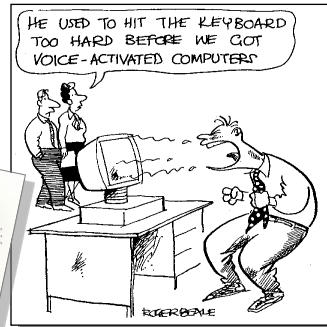


Speech is ... *more than language*

- Rich in ‘extra-linguistic’ information
 - breathing noises
 - lip-smacks
- Rich in ‘para-linguistic’ information 
 - individuality
 - personality
 - attitude
 - emotion



WARNING
There is a difference between ‘felt’ emotions and ‘expressed’ emotions



University Of Sheffield.

INTERSPEECH Tutorial “Speech 101”

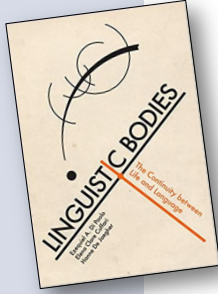
25th Oct. 2020

slide 116



© 2020 The University of Sheffield

Speech is ... *situated & embodied*



- Utterances are tied to ...
 - time
 - place
 - individuals (physically and psychologically)
- Dialogues emerge to ...
 - enact participatory sense-making
 - co-regulate social encounters

WARNING
Breaking the visual, vocal and behavioural coherence can lead to the 'uncanny valley'

"Without the possibility ... of conflict and misunderstanding, there is no participatory sense-making ... no communication ... no reason for communication" (p117)

Di Paolo, E. A., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic Bodies: The Continuity between Life and Language*. MIT Press.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 117 SPANDI

117

© 2020 The University of Sheffield

Speech is ... *learnt from small data*




- Babies only have to hear a word once for it to become part of their vocabulary
- A two year-old has heard ~1000 hours of speech
- A 10 year-old has heard ~10,000 hours

WARNING
This implies that speech exploits huge priors, e.g. 'theory-of-mind' (ToM). Hence, unrestricted spoken language communication may only be possible in human-human interaction.

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *EUROSPEECH03* (pp. 2581–2584). Geneva.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 118 SPANDI

118

© 2020 The University of Sheffield

Finally, speech is ...




“... the most sophisticated behaviour of the most complex organism in the known universe!”

Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. *Speech Communication*, 49, 418-435.



University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 119

119

© 2020 The University of Sheffield

Thank You

Speech 101

University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 120

120

INTERSPEECH 2020 © 2020 The University of Sheffield

Where to Find Out More

Speech 101

<https://isca-speech.org/iscaweb/index.php/scot>

  SCOOT: Speech Communication Online Training


The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 121 

121

INTERSPEECH 2020 © 2020 The University of Sheffield

Summary

- In recent years, the field of spoken language processing has been moving at a very fast pace.
- The impact of deep learning coupled with access to vast data resources has given rise to unprecedented improvements in the performance of speech processing algorithms and systems.
- However, the availability of such pre-recorded datasets and open-source machine-learning toolkits means that practitioners – especially students – are in real danger of becoming detached from the nature and behaviour of actual speech signals.
- This tutorial is aimed at providing an appreciation of the fundamental properties of spoken language, from low-level phonetic detail to high-level communicative behaviour, with a special emphasis on aspects that may have significance for current and future research.

The University Of Sheffield. INTERSPEECH Tutorial "Speech 101" 25th Oct. 2020 slide 122 

122