Information in Spoken Language

A quantitative approach

Rob van Son

Chair of Phonetic Sciences
ACLC
University of Amsterdam

LOT winterschool 2006

AMSTERDAM CENTER FOR LANGUAGE AND COMMUNICATION ACLC

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Introduction

Spoken language is mostly sound: Speech

- To understand speech we must understand speaking and hearing
- Hearing limits the information we can distill from sound
- Articulation limits what we can produce within these limits
- What information can be extracted from speech sounds?
- How can we study this?
- How can we study human speech processing?

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

- The speech chain is not symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around phonemic segments and words
- However, speech does contain neither, just sound

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

- The speech chain is *not* symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around phonemic segments and words
- However, speech does contain neither, just sound

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

- The speech chain is *not* symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around phonemic segments and words
- However, speech does contain neither, just sound

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

- The speech chain is *not* symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around phonemic segments and words
- However, speech does contain neither, just sound

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

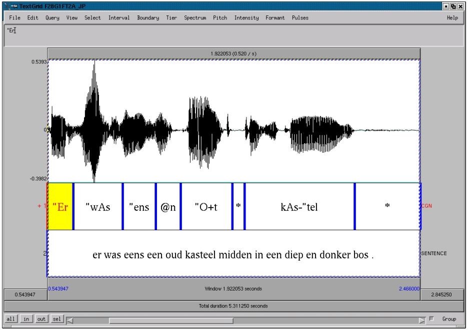
- The speech chain is *not* symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around *phonemic segments* and *words*
- However, speech does contain neither, just sound

Acoustics of speech

The speech chain constist of articulation, sound, and hearing

- The speech chain is *not* symbolic
- Articulation is preceeded by formulation
- Hearing is followed by recognition
- Both formulation and recognition center around *phonemic segments* and *words*
- However, speech does contain *neither*, just sound

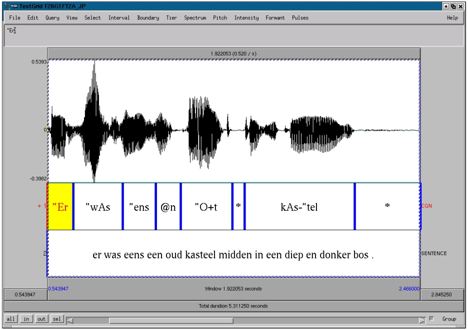
Acoustics of speech: Oscillogram



"Er was eens een oud kasteel"

- Display of presure versus time
- Words are aligned with sound
- Using computer readable (SAMPA) phoneme symbols

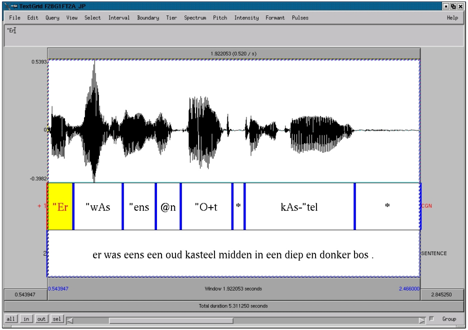
Acoustics of speech: Oscillogram



"Er was eens een oud kasteel"

- Display of presure versus time
- Words are aligned with sound
- Using computer readable (SAMPA) phoneme symbols

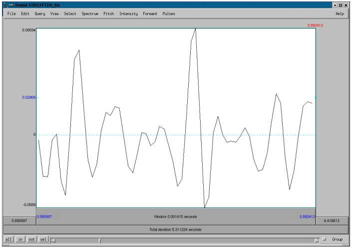
Acoustics of speech: Oscillogram



"Er was eens een oud kasteel"

- Display of presure versus time
- Words are aligned with sound
- Using computer readable (SAMPA) phoneme symbols

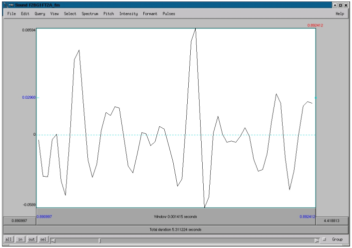
Acoustics of speech: Digital sound and band-width



1.5 ms of an /s/ sound from "was"

- Samples taken at 44.1 kHz (CD audio)
- Quantisize at 16 bit ($\approx$ 65000 amplitude levels)
- Maximum audio frequency 22.05 kHz (Nyquist frequency) but generally *much* less
- Dynamic Range $\approx 96db$ ($\approx 6db/bit$)

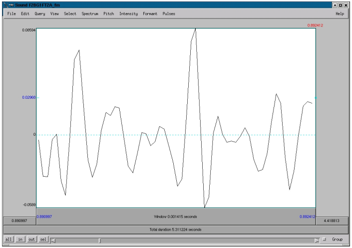
Acoustics of speech: Digital sound and band-width



1.5 ms of an /s/ sound from "was"

- Samples taken at 44.1 kHz (CD audio)
- Quantisize at 16 bit ($\approx$ 65000 amplitude levels)
- Maximum audio frequency 22.05 kHz (Nyquist frequency) but generally *much* less
- Dynamic Range $\approx 96db$ ($\approx 6db/bit$)

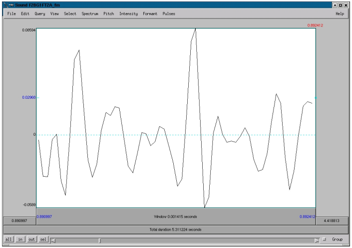
Acoustics of speech: Digital sound and band-width



1.5 ms of an /s/ sound from "was"

- Samples taken at 44.1 kHz (CD audio)
- Quantisize at 16 bit ($\approx$ 65000 amplitude levels)
- Maximum audio frequency 22.05 kHz (Nyquist frequency) but generally *much* less
- Dynamic Range $\approx 96db$ ($\approx 6db/bit$)

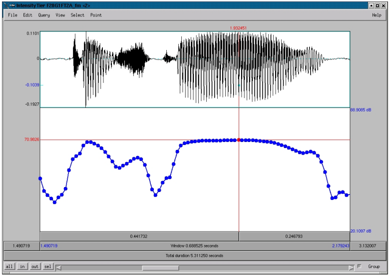
Acoustics of speech: Digital sound and band-width



1.5 ms of an /s/ sound from "was"

- Samples taken at 44.1 kHz (CD audio)
- Quantisize at 16 bit ($\approx$ 65000 amplitude levels)
- Maximum audio frequency 22.05 kHz (Nyquist frequency) but generally *much* less
- Dynamic Range $\approx 96db$ ($\approx 6db/bit$)

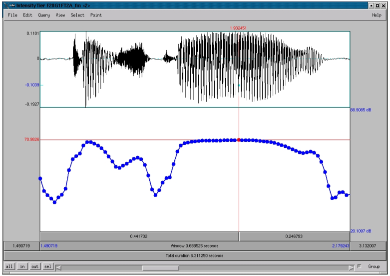
Acoustics of speech: Amplitude and sound level



Intensity contour of "Kasteel"

- Intensity versus amplitude
- Intensity in db ($10 \cdot log_{10}(SoundEnergy)$)
- Intensity you hear is not the intensity you measure $\Rightarrow$ correct for human perception (dBA)

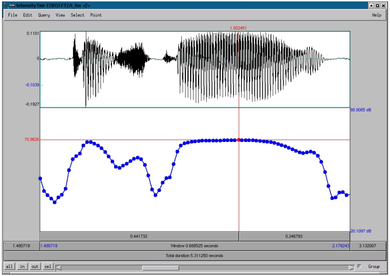
Acoustics of speech: Amplitude and sound level



Intensity contour of "Kasteel"

- Intensity versus amplitude
- Intensity in db ($10 \cdot log_{10}(SoundEnergy)$)
- Intensity you hear is not the intensity you measure ⇒ correct for human perception (*dBA*)

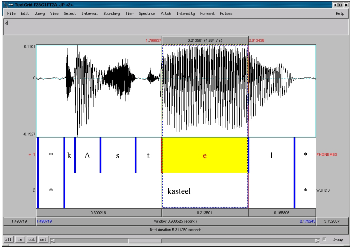
Acoustics of speech: Amplitude and sound level



Intensity contour of "Kasteel"

- Intensity versus amplitude
- Intensity in db $(10 \cdot log_{10}(SoundEnergy))$
- Intensity you hear is not the intensity you measure ⇒ correct for human perception (dBA)

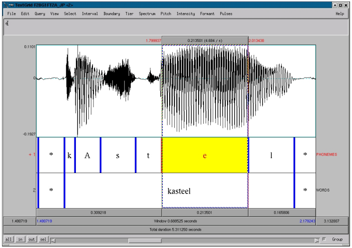
Acoustics of speech: Durations



Phoneme segmentation of "Kasteel"

- Determine the boundaries of words, syllables and phonemes
- Use waveform, ear, and spectrum
- Segmentation is ambiguous and laborious
- Start with automatic segmentation (for speed)

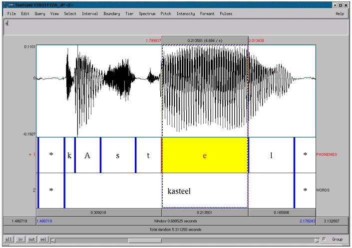
Acoustics of speech: Durations



Phoneme segmentation of "Kasteel"

- Determine the boundaries of words, syllables and phonemes
- Use waveform, ear, and spectrum
- Segmentation is ambiguous and laborious
- Start with automatic segmentation (for speed)

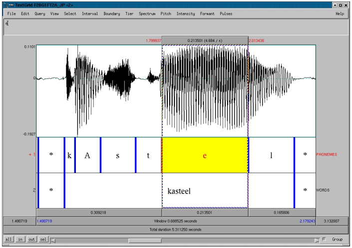
Acoustics of speech: Durations



Phoneme segmentation of "Kasteel"

- Determine the boundaries of words, syllables and phonemes
- Use waveform, ear, and spectrum
- Segmentation is ambiguous and laborious
- Start with automatic segmentation (for speed)

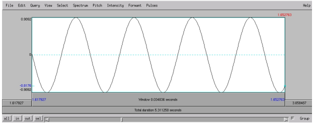
Acoustics of speech: Durations

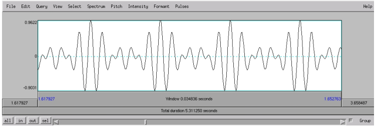


Phoneme segmentation of "Kasteel"

- Determine the boundaries of words, syllables and phonemes
- Use waveform, ear, and spectrum
- Segmentation is ambiguous and laborious
- Start with automatic segmentation (for speed)

Acoustics of speech: The perception of tones: F_0

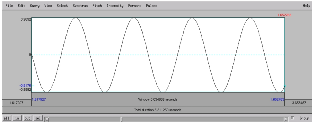


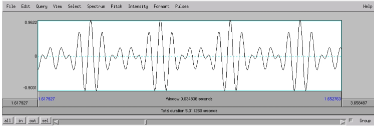


Pitch or F_0 is the *perception* of a harmonic sequence.
Generally, perceived *pitch* is the:

- frequency of a pure tone (top, 125 Hz)
- distance between the components in a mixture of harmonic tones (eg, 125 Hz)
- closest harmonic fit in complex sounds (bells)

Acoustics of speech: The perception of tones: F_0

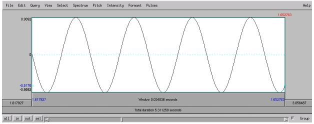


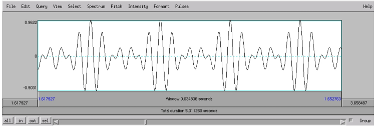


Pitch or F_0 is the *perception* of a harmonic sequence. Generally, perceived *pitch* is the:

- frequency of a pure tone (top, 125 Hz)
- distance between the components in a mixture of harmonic tones (eg, 125 Hz)
- closest harmonic fit in complex sounds (bells)

Acoustics of speech: The perception of tones: F_0

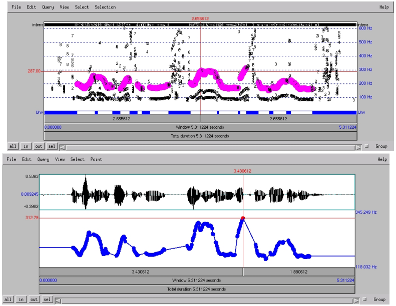




Pitch or F_0 is the *perception* of a harmonic sequence.
Generally, perceived *pitch* is the:

- frequency of a pure tone (top, 125 Hz)
- distance between the components in a mixture of harmonic tones (eg, 125 Hz)
- closest harmonic fit in complex sounds (bells)

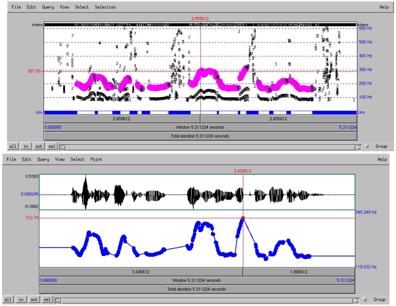
Acoustics of speech: Measuring F_0



The best F_0 candidates are determined

- from the possible repeat frequencies using an autocorrelation function
- from the best fitting harmonics using a *harmonic sieve*

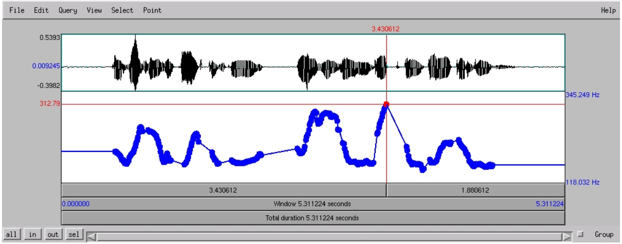
Acoustics of speech: Measuring F_0



The best F_0 candidates are determined

- from the possible repeat frequencies using an autocorrelation function
- from the best fitting harmonics using a *harmonic sieve*

Acoustics of speech: Pitch contours

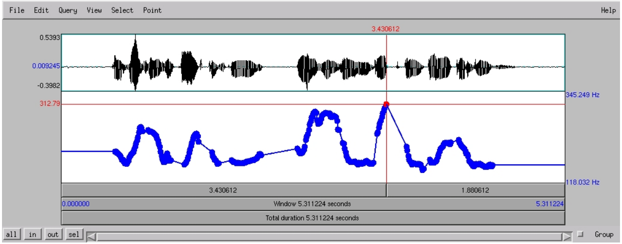


Hummed sound

F_0 makes the melody, or intonation, of an utterance

- There is a general decrease of F_0 over an utterance: The *declination*
- F_0 movements indicate emphasized words: pitch *accents*
- F_0 movements and *declination resets* indicate boundaries

Acoustics of speech: Pitch contours

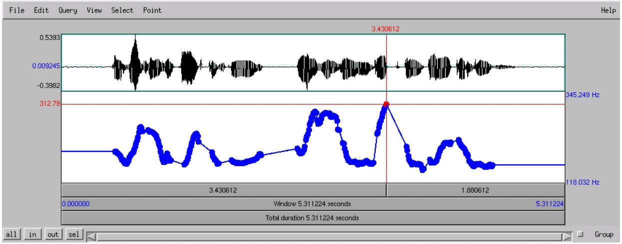


Hummed sound

F_0 makes the melody, or intonation, of an utterance

- There is a general decrease of F_0 over an utterance: The *declination*
- F_0 movements indicate emphasized words: pitch *accents*
- F_0 movements and *declination resets* indicate boundaries

Acoustics of speech: Pitch contours

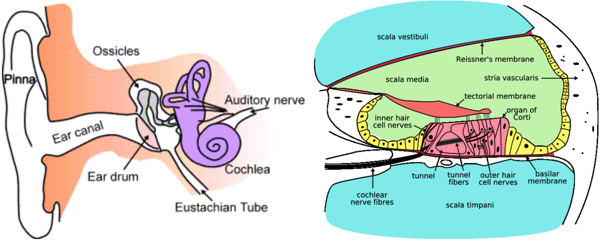


Hummed sound

F_0 makes the melody, or intonation, of an utterance

- There is a general decrease of F_0 over an utterance: The *declination*
- F_0 movements indicate emphasized words: pitch *accents*
- F_0 movements and *declination resets* indicate boundaries

Hearing: The human ear

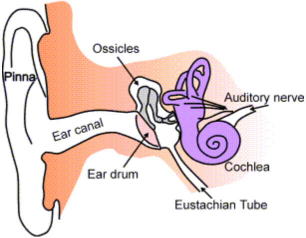


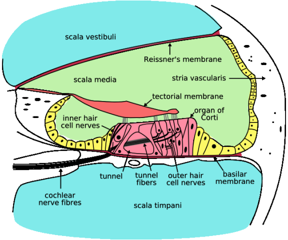
The anatomy of the human ear (l) and cochlea (r)

- Sound enters the ear canal (left)
- The ear drum transfers the sound over the ossicles to the oval window into the cochlea
- The cochlea (right) converts sound into nerve exitations

http://en.wikipedia.org/wiki/Human_ear

Hearing: The human ear



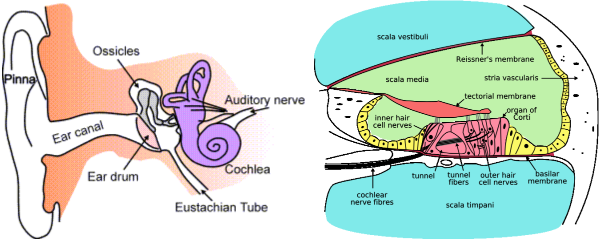


The anatomy of the human ear (l) and cochlea (r)

- Sound enters the ear canal (left)
- The ear drum transfers the sound over the ossicles to the oval window into the cochlea
- The cochlea (right) converts sound into nerve exitations

http://en.wikipedia.org/wiki/Human_ear

Hearing: The human ear

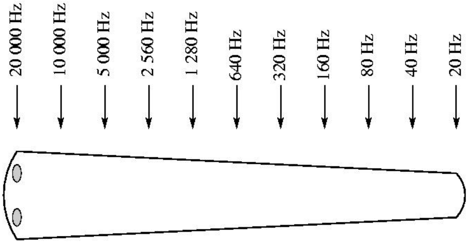


The anatomy of the human ear (l) and cochlea (r)

- Sound enters the ear canal (left)
- The ear drum transfers the sound over the ossicles to the oval window into the cochlea
- The cochlea (right) converts sound into nerve exitations

http://en.wikipedia.org/wiki/Human_ear

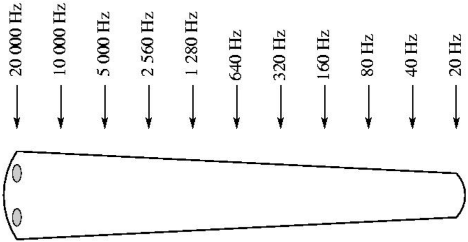
Hearing: The basilar membrane



Frequency map of the basilar membrane from [Moore(2003)]

- The ear analyses sounds roughly into *Log* (*Power* (*Frequency*)) vs. *Log* (*Frequency*)
- Speech is analyzed in the same way
- Use power spectra of sounds

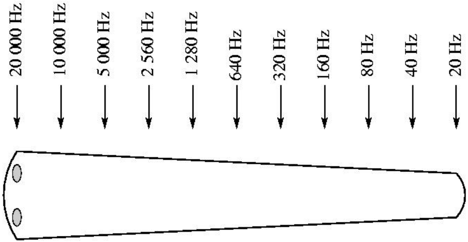
Hearing: The basilar membrane



Frequency map of the basilar membrane from [Moore(2003)]

- The ear analyses sounds roughly into *Log* (*Power* (*Frequency*)) vs. *Log* (*Frequency*)
- Speech is analyzed in the same way
- Use power spectra of sounds

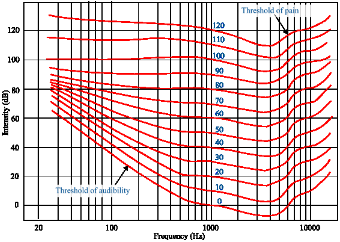
Hearing: The basilar membrane



Frequency map of the basilar membrane from [Moore(2003)]

- The ear analyses sounds roughly into *Log* (*Power* (*Frequency*)) vs. *Log* (*Frequency*)
- Speech is analyzed in the same way
- Use power spectra of sounds

Hearing: Loudness

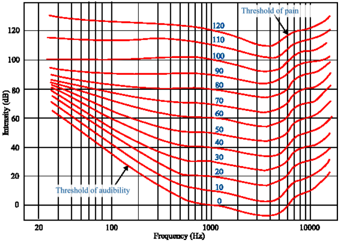


Relation between perceived loudness and frequency

- The absolute threshold varies with frequency
- Best hearing between 800Hz and 6000Hz
- Human speech needs 400Hz - 3400Hz (telephone speech)
- Loudness JND is around 1 dB (soft sounds)

http://en.wikipedia.org/wiki/Threshold_of_hearing

Hearing: Loudness

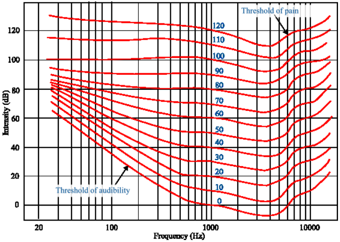


Relation between perceived loudness and frequency

- The absolute threshold varies with frequency
- Best hearing between 800Hz and 6000Hz
- Human speech needs 400Hz - 3400Hz (telephone speech)
- Loudness JND is around 1 dB (soft sounds)

http://en.wikipedia.org/wiki/Threshold_of_hearing

Hearing: Loudness

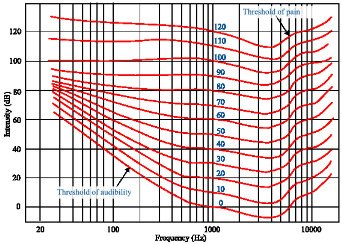


Relation between perceived loudness and frequency

- The absolute threshold varies with frequency
- Best hearing between 800Hz and 6000Hz
- Human speech needs 400Hz - 3400Hz (telephone speech)
- Loudness JND is around 1 dB (soft sounds)

http://en.wikipedia.org/wiki/Threshold_of_hearing

Hearing: Loudness

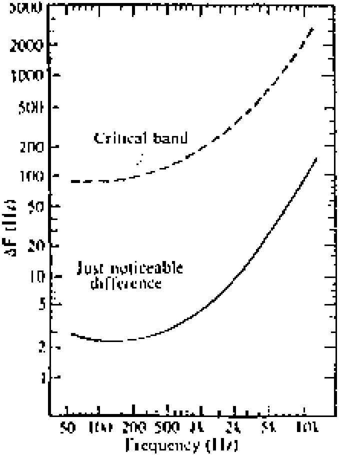


Relation between perceived loudness and frequency

- The absolute threshold varies with frequency
- Best hearing between 800Hz and 6000Hz
- Human speech needs 400Hz - 3400Hz (telephone speech)
- Loudness JND is around 1 dB (soft sounds)

http://en.wikipedia.org/wiki/Threshold_of_hearing

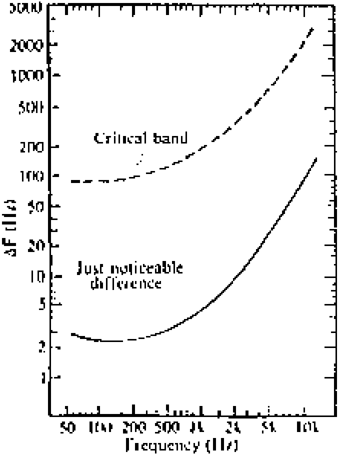
Hearing: Critical bands



http://www.music.gla.ac.uk/~george/audio/psy/psy.html

- Sounds grouped in *critical bands* (CB)
- CB is around 30 times JND
- One signal per critical band
- 3 CB per Octave
- Time resolution $\approx 25ms$

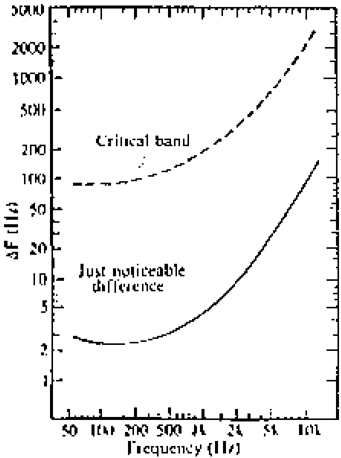
Hearing: Critical bands



http://www.music.gla.ac.uk/~george/audio/psy/psy.html

- Sounds grouped in *critical bands* (CB)
- CB is around 30 times JND
- One signal per critical band
- 3 CB per Octave
- Time resolution $\approx 25ms$

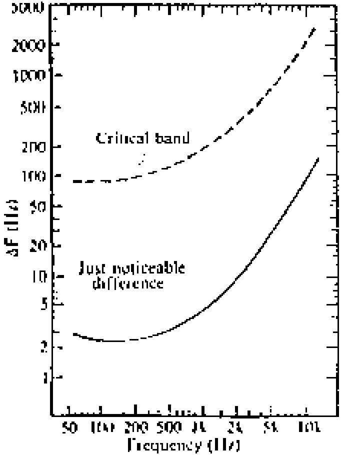
Hearing: Critical bands



http://www.music.gla.ac.uk/~george/audio/psy/psy.html

- Sounds grouped in *critical bands* (CB)
- CB is around 30 times JND
- One signal per critical band
- 3 CB per Octave
- Time resolution $\approx 25ms$

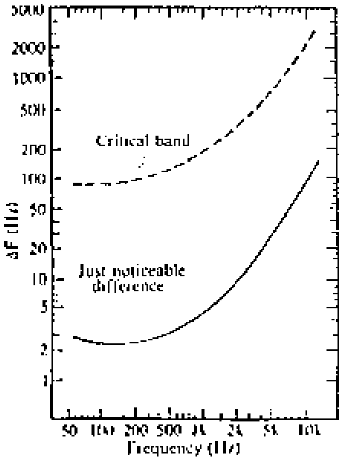
Hearing: Critical bands



http://www.music.gla.ac.uk/~george/audio/psy/psy.html

- Sounds grouped in *critical bands* (CB)
- CB is around 30 times JND
- One signal per critical band
- 3 CB per Octave
- Time resolution $\approx 25ms$

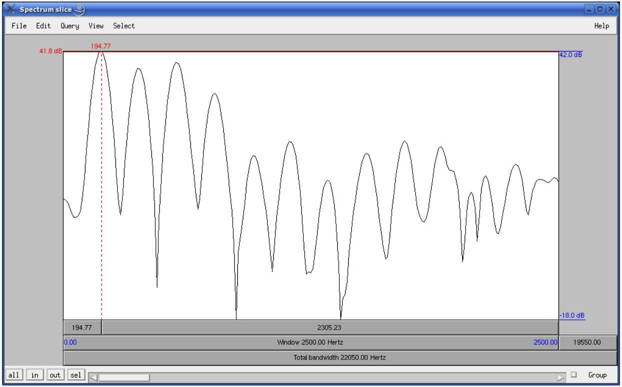
Hearing: Critical bands



http://www.music.gla.ac.uk/~george/audio/psy/psy.html

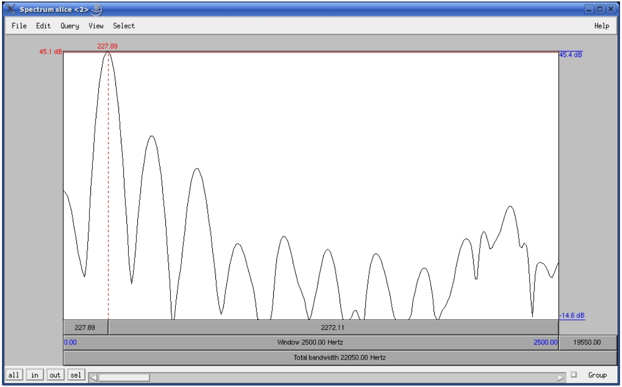
- Sounds grouped in *critical bands* (CB)
- CB is around 30 times JND
- One signal per critical band
- 3 CB per Octave
- Time resolution $\approx 25ms$

Hearing: Example of /ɛ/



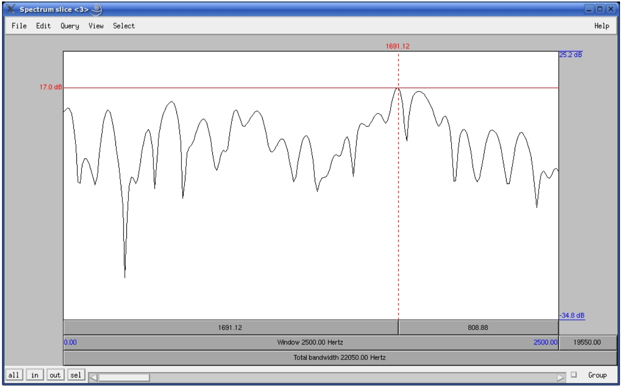
Note the harmonic structure and the "bumps"

Hearing: Example of /n/



Note the harmonic structure and the low level of high frequencies

Hearing: Example of /s/



Note the noisy structure and the broad bandwidth

Hearing

How much information can the ear extract at a minimum from speech?

- Frequencies: 400Hz - 3400Hz $\rightarrow$ 3 octaves $\approx$ 9 CBs
- Dynamic range $\approx 30dB$ per CB, $\rightarrow$ 5 bit / CB
- Time resolution $\approx 25ms$
- Information available: $\frac{9CB \cdot 5bit}{0.025sec} \approx 1800$ bit/second
- Just an order of magnitude estimation

Hearing

How much information can the ear extract at a minimum from speech?

- Frequencies: 400Hz - 3400Hz $\rightarrow$ 3 octaves $\approx$ 9 CBs
- Dynamic range $\approx 30dB$ per CB, $\rightarrow$ 5 bit / CB
- Time resolution $\approx 25ms$
- Information available: $\frac{9CB \cdot 5bit}{0.025sec} \approx 1800$ bit/second
- Just an order of magnitude estimation

Hearing

How much information can the ear extract at a minimum from speech?

- Frequencies: 400Hz - 3400Hz $\rightarrow$ 3 octaves $\approx$ 9 CBs
- Dynamic range $\approx 30dB$ per CB, $\rightarrow$ 5 bit / CB
- Time resolution $\approx 25ms$
- Information available: $\frac{9CB \cdot 5bit}{0.025sec} \approx 1800$ bit/second
- Just an order of magnitude estimation

Hearing

How much information can the ear extract at a minimum from speech?

- Frequencies: 400Hz - 3400Hz $\rightarrow$ 3 octaves $\approx$ 9 CBs
- Dynamic range $\approx 30dB$ per CB, $\rightarrow$ 5 bit / CB
- Time resolution $\approx 25ms$
- Information available: $\frac{9CB \cdot 5bit}{0.025sec} \approx 1800$ bit/second
- Just an order of magnitude estimation

Hearing

How much information can the ear extract at a minimum from speech?

- Frequencies: 400Hz - 3400Hz $\rightarrow$ 3 octaves $\approx$ 9 CBs
- Dynamic range $\approx 30dB$ per CB, $\rightarrow$ 5 bit / CB
- Time resolution $\approx 25ms$
- Information available: $\frac{9CB \cdot 5bit}{0.025sec} \approx$ 1800 bit/second
- Just an order of magnitude estimation

Hearing

How much information is actually needed?

- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are ≈ 1600 diphones ($40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate ≤ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

Hearing

How much information is actually needed?

- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are ≈ 1600 diphones $(40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate ≤ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

Hearing

How much information is actually needed?

- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are $\approx$ 1600 diphones ($40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate $\leq$ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

Hearing

How much information is actually needed?

- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are ≈ 1600 diphones ($40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate ≤ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

Hearing

How much information is actually needed?

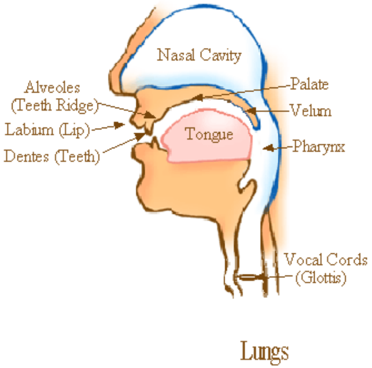
- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are ≈ 1600 diphones ($40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate ≤ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

Hearing

How much information is actually needed?

- Concatenated phoneme transitions (diphones) are perfectly intelligble
- There are ≈ 1600 diphones $(40 \cdot 40$ allophones)
- 1600 diphones fit easily in 11 bit
- Articulation rate ≤ 15 phones/second
- Information needed for recognition: $11 \cdot 15 = 165$ bits/second
- Compare to 1800 bits available, ample redundancy

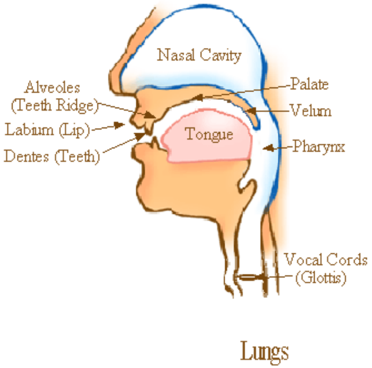
Speaking: Articulators



Vocal trackt and primary articulators

- The lungs pump air through the glottal folds
- The glottal folds can vibrate (or not)
- The oral cavity can be constricted at many points
- Alternatively, air can flow through the nasal cavity

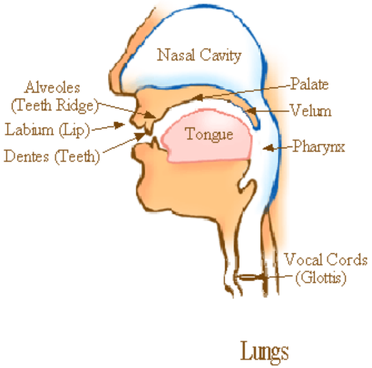
Speaking: Articulators



Vocal trackt and primary articulators

- The lungs pump air through the glottal folds
- The glottal folds can vibrate (or not)
- The oral cavity can be constricted at many points
- Alternatively, air can flow through the nasal cavity

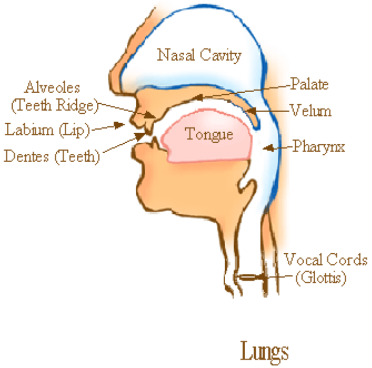
Speaking: Articulators



Vocal trackt and primary articulators

- The lungs pump air through the glottal folds
- The glottal folds can vibrate (or not)
- The oral cavity can be constricted at many points
- Alternatively, air can flow through the nasal cavity

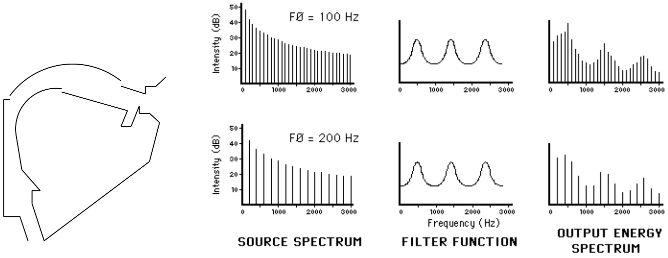
Speaking: Articulators



Vocal trackt and primary articulators

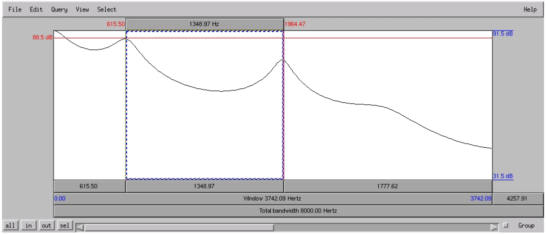
- The lungs pump air through the glottal folds
- The glottal folds can vibrate (or not)
- The oral cavity can be constricted at many points
- Alternatively, air can flow through the nasal cavity

Speaking: Source Filter model of speech



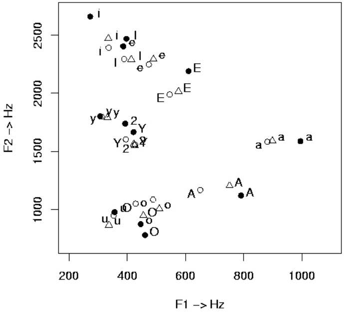
Sound enters the oral cavity (vocal tract) from below and is filtered by the resonances of the cavity

Speaking: Resonances and formants



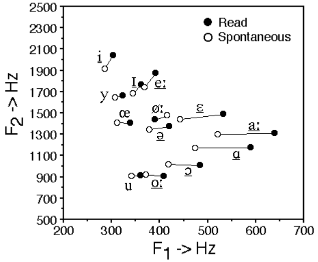
Oral cavity filter function of /ɛ/ (LPC model). Peaks are formants F_1 and $F2$.
The resonances of the vocal tract are called Formants, and numbered from below, i.e., F_1, F_2, F_3, $\cdots$. Normally, the first three are sufficient to describe (voiced) speech.

Speaking: Vowel Formant space



Vowel formant space of Dutch.
Only two formant values, F_1 and F_2, suffice to identify a vowel (in the ideal case). However, in normal speech, there is so much overlap and variation that it remains almost impossible.

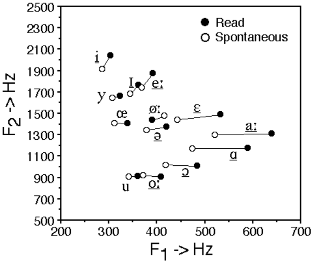
Speaking: Reduction



Articulation takes effort and humans are "lazy"

- Unimportant, unstressed, items are articulated "less well"
- Leads to less distinctive sound segments
- Visible as a smaller vowel triangle (see above)
- Also found in consonants

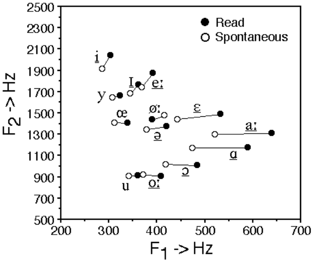
Speaking: Reduction



Articulation takes effort and humans are "lazy"

- Unimportant, unstressed, items are articulated "less well"
- Leads to less distinctive sound segments
- Visible as a smaller vowel triangle (see above)
- Also found in consonants

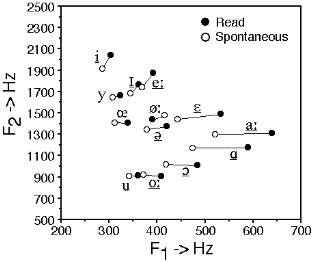
Speaking: Reduction



Articulation takes effort and humans are "lazy"

- Unimportant, unstressed, items are articulated "less well"
- Leads to *less distinctive* sound segments
- Visible as a smaller vowel triangle (see above)
- Also found in consonants

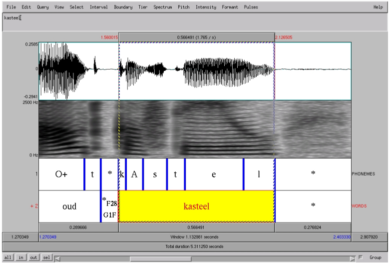
Speaking: Reduction



Articulation takes effort and humans are “lazy”

- Unimportant, unstressed, items are articulated “less well”
- Leads to *less distinctive* sound segments
- Visible as a smaller vowel triangle (see above)
- Also found in consonants

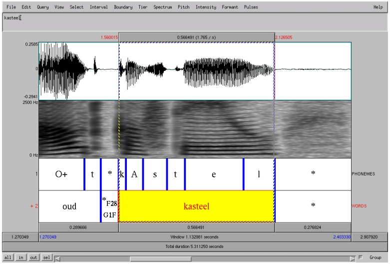
Spectrograms



A spectrogram shows the development of the spectrum in time (darker is more power)

- A spectrogram shows the harmonics
- Vowels, fricatives, and plosives are visible

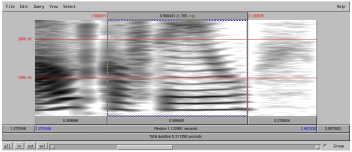
Spectrograms

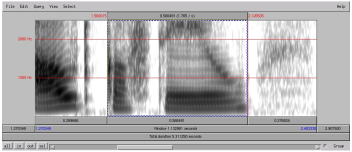


A spectrogram shows the development of the spectrum in time (darker is more power)

- A spectrogram shows the harmonics
- Vowels, fricatives, and plosives are visible

Spectrograms: Narrow versus Wide band

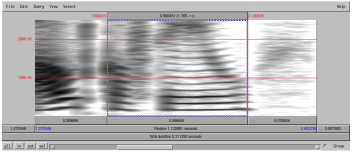


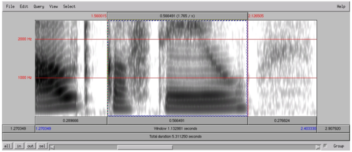


Two views on spectrograms

- Narrow-band (top): High frequency resolution, low time resolution
- Wide-band (bottom): Low frequency resolution, high time resolution

Spectrograms: Narrow versus Wide band

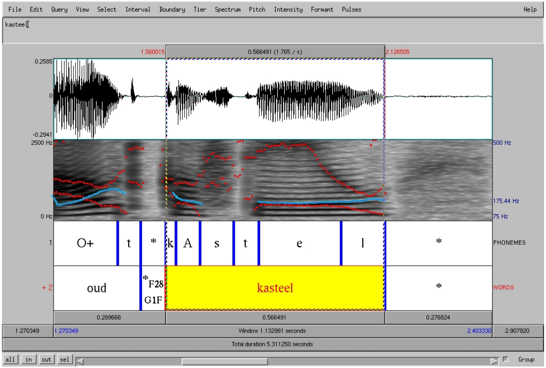




Two views on spectrograms

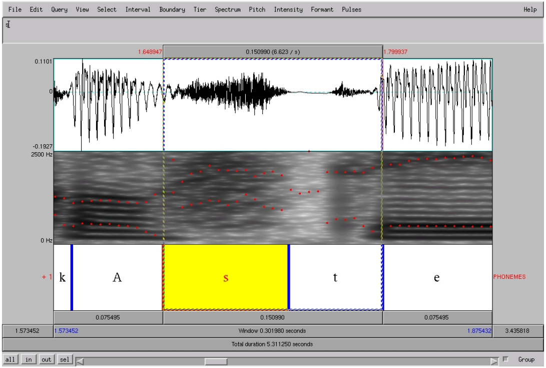
- Narrow-band (top): High frequency resolution, low time resolution
- Wide-band (bottom): Low frequency resolution, high time resolution

Spectrograms: Formant and Pitch tracking



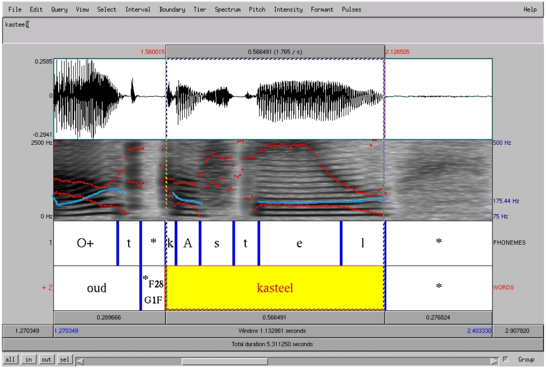
Formants (red dots) and Pitch (blue line) can be automatically determined and plotted into a spectrogram.

Spectrograms: Noise and bursts



Fricatives are visible as gray noise patches. Plosives as a silent part followed by a noisy burst.

Spectrograms: Spectrogram reading



It is actually possible, after a few weeks training, to read spectrograms. All the information needed to "understand" the speech is in the spectrogram [Lander and Carmell(1997)].

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transliteration

Before anything can be done with speech, it has to be written down and transcribed

- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic Part-of-Speech tags and Syntax

Corpora: Transcription

Human annotator transcriptions: Difficult and expensive

- Accents, stress, and boundaries (always ambiguous)
- Handcorrected word-boundaries
- Handcorrected phoneme-boundaries (always ambiguous)
- Check Part-of-Speech tags
- Check Syntax

Corpora: Transcription

Human annotator transcriptions: Difficult and expensive

- Accents, stress, and boundaries (always ambiguous)
- Handcorrected word-boundaries
- Handcorrected phoneme-boundaries (always ambiguous)
- Check Part-of-Speech tags
- Check Syntax

Corpora: Transcription

Human annotator transcriptions: Difficult and expensive

- Accents, stress, and boundaries (always ambiguous)
- Handcorrected word-boundaries
- Handcorrected phoneme-boundaries (always ambiguous)
- Check Part-of-Speech tags
- Check Syntax

Corpora: Transcription

Human annotator transcriptions: Difficult and expensive

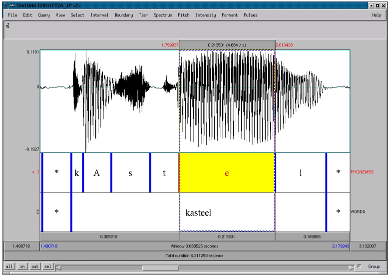
- Accents, stress, and boundaries (always ambiguous)
- Handcorrected word-boundaries
- Handcorrected phoneme-boundaries (always ambiguous)
- Check Part-of-Speech tags
- Check Syntax

Corpora: Transcription

Human annotator transcriptions: Difficult and expensive

- Accents, stress, and boundaries (always ambiguous)
- Handcorrected word-boundaries
- Handcorrected phoneme-boundaries (always ambiguous)
- Check Part-of-Speech tags
- Check Syntax

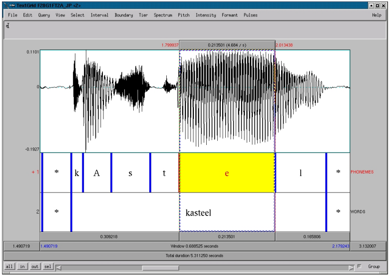
Corpora:
Identifying and annotating phonemes



Phonemes are not pearls on a string

- Phonemes always overlap and are extremely variable
- A phoneme you hear can appear absent in the waveform
- It is often unclear what phonemes were uttered
- Sometimes, even the order is unclear

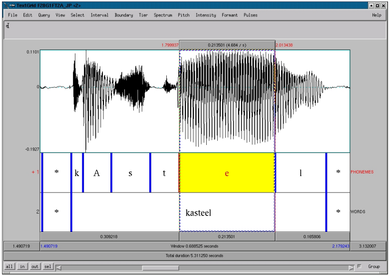
Corpora:
Identifying and annotating phonemes



Phonemes are not pearls on a string

- Phonemes always overlap and are extremely variable
- A phoneme you hear can appear absent in the waveform
- It is often unclear what phonemes were uttered
- Sometimes, even the order is unclear

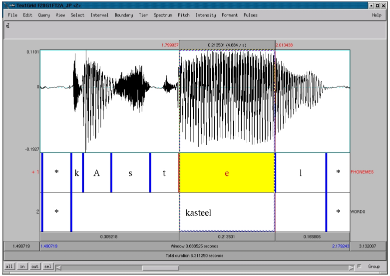
Corpora: Identifying and annotating phonemes



Phonemes are not pearls on a string

- Phonemes always overlap and are extremely variable
- A phoneme you hear can appear absent in the waveform
- It is often unclear what phonemes were uttered
- Sometimes, even the order is unclear

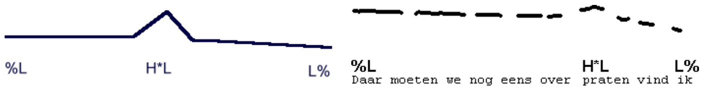
Corpora: Identifying and annotating phonemes



Phonemes are not pearls on a string

- Phonemes always overlap and are extremely variable
- A phoneme you hear can appear absent in the waveform
- It is often unclear what phonemes were uttered
- Sometimes, even the order is unclear

Corpora:
ToBI like systems for intonation transcription



ToDI symbols (IP: Intonational Phrase)

[Gussenhoven et al.(2003)Gussenhoven, Rietveld, Kerkhoff, and Terken]

High	Low	description
H*	L*	high/low accent
H	L	upward/downward movement after L*/H*
H%	L%	rising/low ending of IP
%H	%L	high/low beginning of IP
%HL		Initial falling pitch not marking accent
%		half-completed fall/rise at end of IP
!H*		downstepped H*

Corpora

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application/research is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Corpora

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application/research is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Corpora

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application/research is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Corpora

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application/research is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Corpora

There is no data like more data

- Speech and Language are extremely complex
- Large amounts of data are necessary to model them
- "The best application/research is the one with the largest corpus"
- 10-1000 hours of speech recordings needed
- 10^8 - 10^9 word text corpus needed

Corpora: for Speech and Language research

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed): Information on the items
- Normalization (fixed): All items must adhere to certain guidelines
- Data (fixed): Immutable text or speech records
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
- Storage, distribution, access, and software (volatile): Practical use

Corpora: for Speech and Language research

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed): Information on the items
- Normalization (fixed): All items must adhere to certain guidelines
- Data (fixed): Immutable text or speech records
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
- Storage, distribution, access, and software (volatile): Practical use

Corpora: for Speech and Language research

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed): Information on the items
- Normalization (fixed): All items must adhere to certain guidelines
- Data (fixed): Immutable text or speech records
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
- Storage, distribution, access, and software (volatile): Practical use

Corpora: for Speech and Language research

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed): Information on the items
- Normalization (fixed): All items must adhere to certain guidelines
- Data (fixed): Immutable text or speech records
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
- Storage, distribution, access, and software (volatile): Practical use

Corpora: for Speech and Language research

A language corpus is a documented collection of coherent text, speech, video, and transcriptions and annotations of these

Requirements

- Meta data (fixed): Information on the items
- Normalization (fixed): All items must adhere to certain guidelines
- Data (fixed): Immutable text or speech records
- Transcriptions and annotations (cumulative): Added value of interpretations and analysis
- Storage, distribution, access, and software (volatile): Practical use

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: Spoken Dutch Corpus (CGN) [NTU(2004)]

Contents ($\frac{2}{3}$ Dutch, $\frac{1}{3}$ Flemish)

- 500 hours (5,650,000 words) recorded in The Netherlands
- 300 hours (3,250,000 words) in Flanders
- 9 million words
- 4250 speakers
- 15 Styles/genres (formal reading to informal multilogues)
- Field recordings with Sony Minidisk
- 16/16 and 8/8 kHz/bit encoding

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Corpora: CGN Annotations

Annotations and transcriptions

- Orthographic transcription (the full 8,900,000 words)
- Manually verified POS tagging and lemmatization (all)
- Lexicon and identification of multi word units (all)
- Automatic time alignment and phonetic transcription at the word level (all)
- Manually verified broad phonetic transcription (1,000,000 words)
- Manually verified time alignment at the word level (1,000,000 words)
- Syntactic annotation (1,000,000 words)
- Two independent prosodic annotations (250,000 words)

Bibliography I

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

Carlos Gussenhoven, Toni Rietveld, Joop Kerkhoff, and Jacques Terken.
ToDI: Transcription of Dutch Intonation.
Web, 2003.
URL http://todi.let.kun.nl/ToDI/home.htm.
Courseware.

Terri Lander and Tim Carmell.
Structure of Spoken Language: Spectrogram Reading.
Web, 15 March 1997.
URL http://speech.bme.ogi.edu/tutordemos/SpectrogramReading/cse551html/cse551/cse551.html.

Guy Moore.
Physics 224: the Physics of Music.
Web, 2003.
URL http://www.physics.mcgill.ca/~guymoore/ph224/notes/lecture6.html.
Lecture 6.

NTU.
Spoken Dutch Corpus (CGN).
Corpus, 2004.
URL http://www.tst.inl.nl/cgn.htm.
Metadata (MPI) - http://www.mpi.nl/world/ISLE/overview/Overview_CGN.html Contents - http://www.elis.ugent.be/cgn/ Descriptions and references - http://lands.let.kun.nl/cgn/ehome.htm.

Bibliography II

A.C.M. Rietveld and F.J. Koopmans-van Beinum.
Vowel reduction and stress.
Speech Communication, 6:217–229, 1987.

Bibliography

Copyright ©2005,2006 R.J.J.H. van Son, GNU General Public License [FSF(1991)]

Bibliography II

Bibliography III

3. If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.
Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.
In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

   1. Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

Bibliography V

2. Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
3. Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.
If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

Bibliography VI

Bibliography VIII

11 Because the program is licensed free of charge, there is no warranty for the program, to the extent permitted by applicable law. Except when otherwise stated in writing the copyright holders and/or other parties provide the program "as is" without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the program is with you. Should the program prove defective, you assume the cost of all necessary servicing, repair or correction.

12 In no event unless required by applicable law or agreed to in writing will any copyright holder, or any other party who may modify and/or redistribute the program as permitted above, be liable to you for damages, including any general, special, incidental or consequential damages arising out of the use or inability to use the program (including but not limited to loss of data or data being rendered inaccurate or losses sustained by you or third parties or a failure of the program to operate with any other programs), even if such holder or other party has been advised of the possibility of such damages.

End of Terms and Conditions

Appendix: How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.
To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

> *one line to give the program's name and a brief idea of what it does.*
> *Copyright (C) yyyy name of author*
> *This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.*
>
> *This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.*
>
> *You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA.*

Also add information on how to contact you by electronic and paper mail.
If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

> *Gnomovision version 69, Copyright (C) yyyy name of author*
> *Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type 'show w'.*
> *This is free software, and you are welcome to redistribute it under certain conditions; type 'show c' for details.*

Bibliography X

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than `show w` and `show c`; they could even be mouse-clicks or menu items—whatever suits your program.
You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

> *Yoyodyne, Inc., hereby disclaims all copyright interest in the program*
> *'Gnomovision' (which makes passes at compilers) written by James Hacker.*
> *signature of Ty Coon, 1 April 1989*
> *Ty Coon, President of Vice*

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.