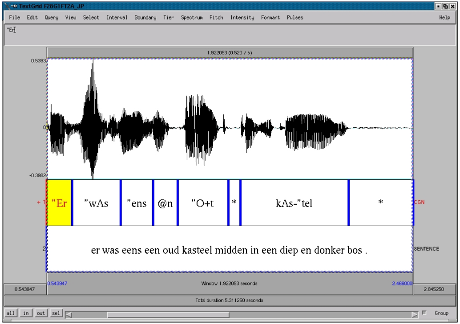
Speech recognition and synthesis

Introduction

All technology starts with quantitative modelling

- Speech technology is about speech sounds
- Only limited knowledge of human speech production and perception is necessary for modelling speech sounds
- In practice, knowledge about human speech is only used implicitely

Waveforms: Oscillogram



"Er was eens een oud kasteel"

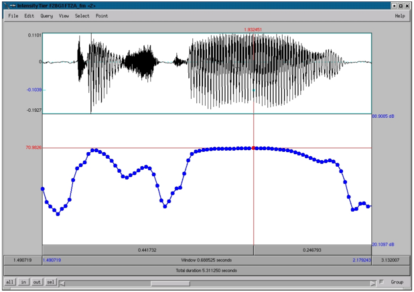
- Display of presure versus time
- Words are aligned with sound
- Using computer readable (SAMPA) phoneme symbols

Waveforms: 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 $\approx 96dB$ ($\approx 6dB/bit$)

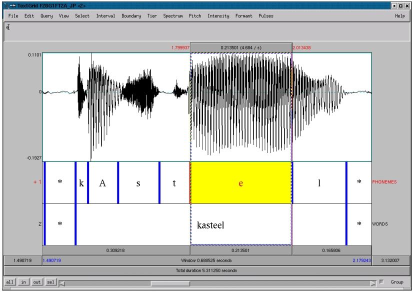
Waveforms: Amplitude and sound level



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*)

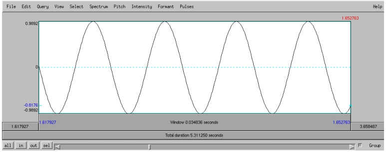
Waveforms: Durations

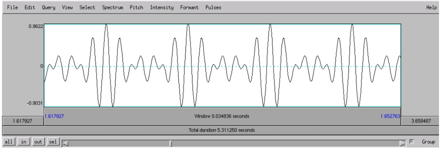


Phoneme segmentation of "Kasteel"

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

Pitch and F0: The perception of tones: F_0

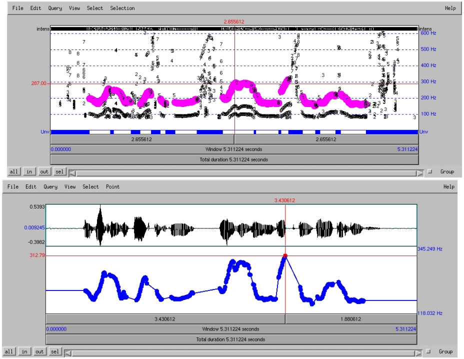




Pitch or F_0 is the *perception* of a harmonic sequence. Generally, perceived *pitch* is the:

- 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)

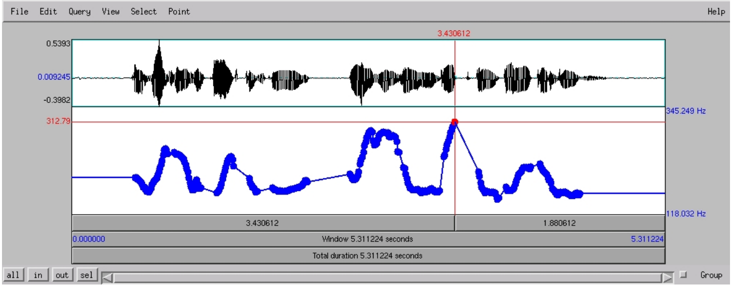
Pitch and F0: Measuring F_0



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*

Pitch and F0: Pitch contours

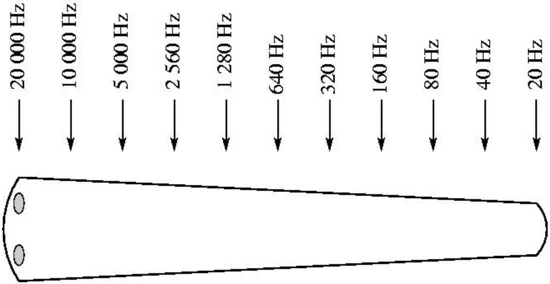


Hummed sound

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

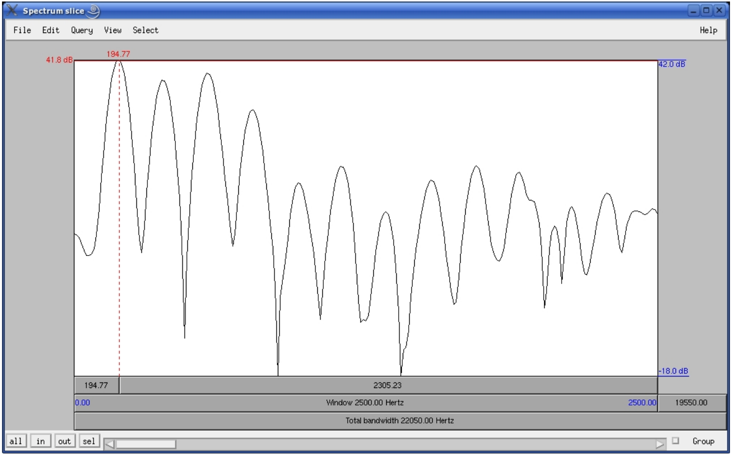
Spectrum: The Ear (again)



Frequency map of the cochlea from [Moore(2003)]

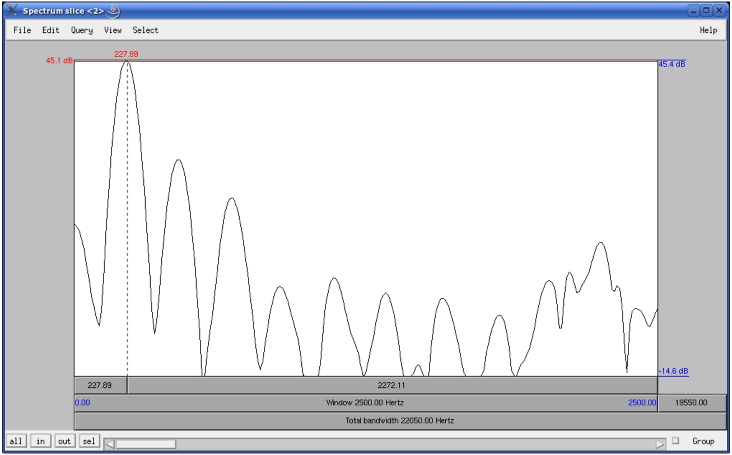
- The ear analysis sounds roughly into *Log* (*Power* (*Frequency*)) vs. *Log* (*Frequency*)
- Speech is analyzed in the same way
- Use power spectra of sounds

Spectrum: Example of /ɛ/



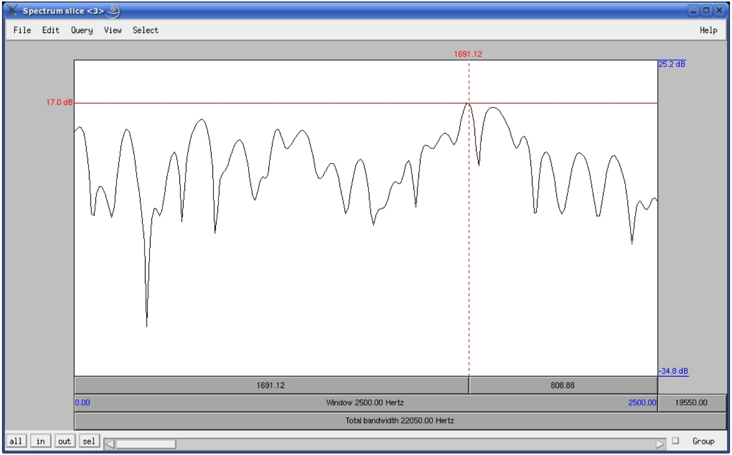
Note the harmonic structure and the "bumps"

Spectrum: Example of /n/



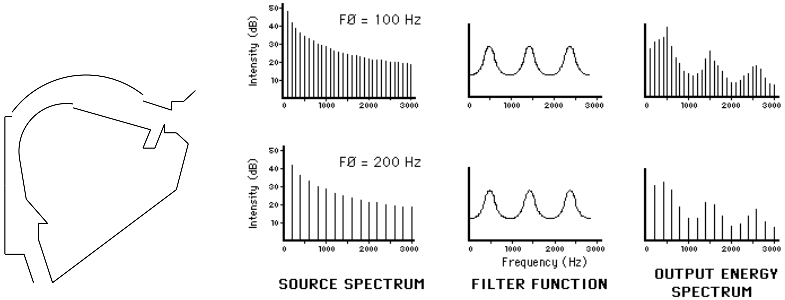
Note the harmonic structure and the low level of high frequencies

Spectrum: Example of /s/



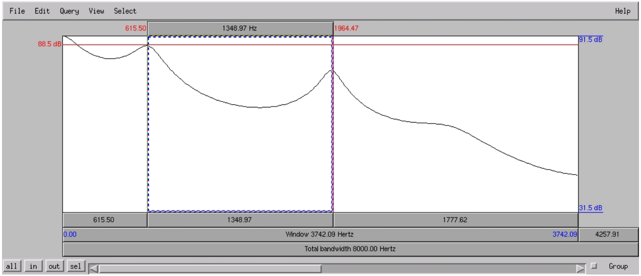
Note the noisy structure and the broad bandwidth

Spectrum: Source Filter model of speech



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

Spectrum: Resonances and formants

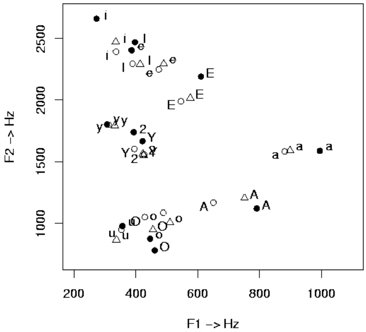


Oral cavity filter function of /ɛ/ (LPC model).

Peaks are formants F_1 and $F2$.

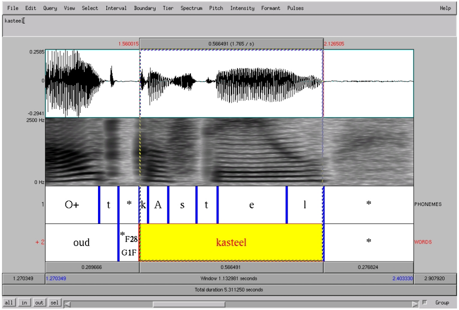
The resonances of the vocal tract are called Formants, and numbered from below, i.e., F_1, F_2, F_3, $\cdots$. Normally, the first three are sufficient to describe (voiced) speech.

Spectrum: 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.

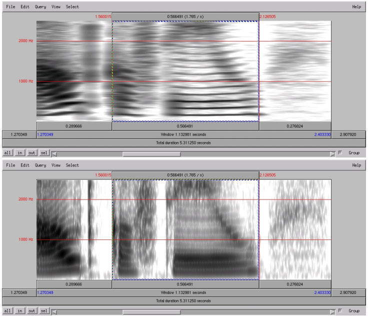
Spectrograms



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

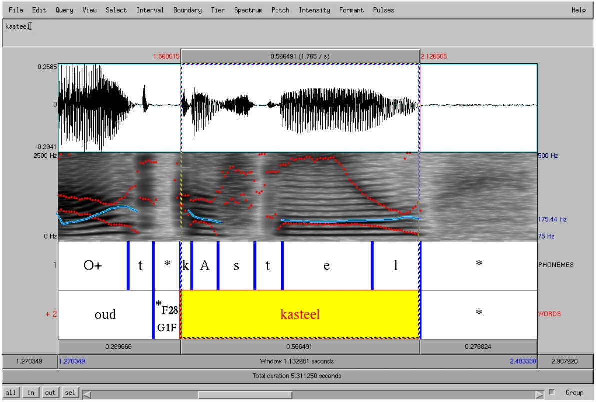
Spectrograms: Narrow versus Wide band



Two views on spectrograms

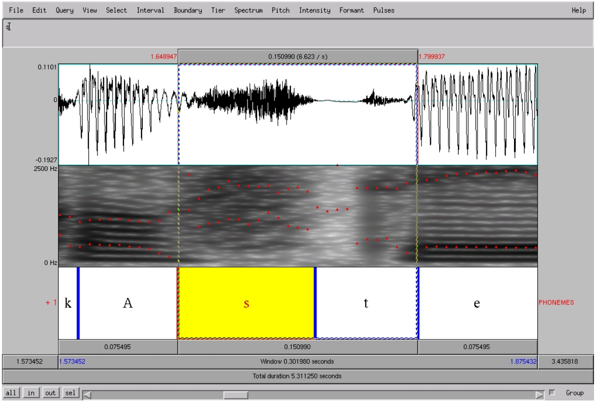
- Narrow-band (top): High frequency resolution, low time resolution
- Wide-band (bottom): Low frequency resolution, high time resolution

Spectrograms: Formant and Pitch tracking



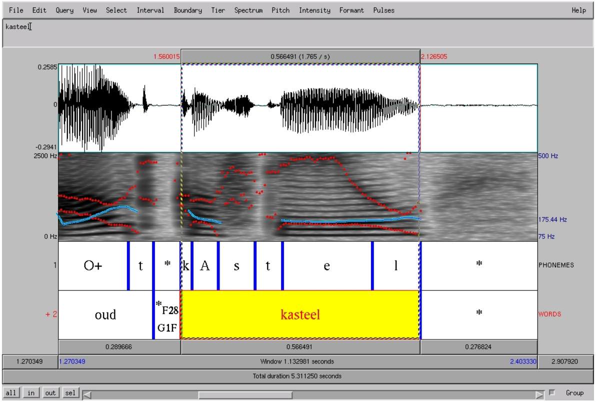
Formants (red dots) and Pitch (blue line) can be automatically determined and plotted into a spectrogram.

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)].

Transcription: Transliteration

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

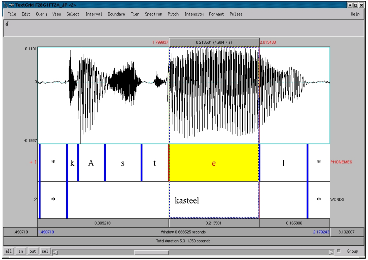
- Write out orthographically what was said (and check it)
- Align chunks of text roughly with the stretches of speech
- Transcribe the text automatically into phonemes using a lexicon
- Split the orthographic/phonemic text into words
- Align the words/phonemes automatically with the speech
- Add automatic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

Transcription: 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

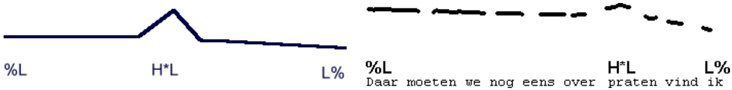
Transcription: Identifying and annotating phonemes



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

Transcription: ToBI systems for intonation transcription



ToDI symbols (IP: Intonational Phrase)

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*

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

Assignment- Week 2 Spectrogram and spectrum

See BlackBoard for full description

- Use a recorded sentence (assignment 1)
- Determine pitch maxima and minima. On which words do you find the maxima? Where inside the word?
- Determine the spectrum of a (monophthong) vowel and draw it
- Compare the formant frequencies of this vowel with the spectrum
- Determine the formants at the center of all (at most 10) monophthong vowels
- Using the formant values, where do these vowels fit into the vowel triangle? (F_1/F_2 space)
- Hand in your report as a PDF

Further Reading I

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

Marcus Filipsson.
Speech Analysis Tutorial.
Web, 1995.
URL http://www.ling.lu.se/research/speechtutorial/tutorial.html.
Courseware.

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.ru.nl/ToDI/home.htm.
Courseware.

Further Reading II

Peter Ladefoged.
Vowels and Consonants.
Wiley-Blackwell, Malden, 2005.
URL http://linguistlist.org/pubs/books/get-book.cfm?BookID=16055.

Peter Ladefoged and Ian Maddieson.
The Sounds of the World's Languages.
Wiley-Blackwell, Malden, 1995.
URL http://linguistlist.org/pubs/books/get-book.cfm?BookID=3034.

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.

W.J.M. Levelt.
The Skill of Speaking, volume 1 of *International perspectives on psychological science*, pages 89–103.
Lawrence Erlbaum Associates, 1994.
URL http://hdl.handle.net/2066/15531.

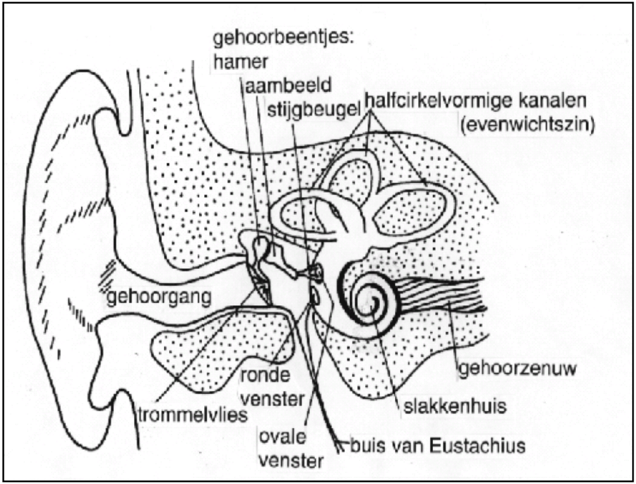
W.J.M. Levelt.
Waar komen gesproken woorden vandaan?
De Psycholoog, 31:434–437, 1996.
URL http://hdl.handle.net/2066/15548.

Further Reading III

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

Appendix A

The inner ear



Copyright License