Linköping University
Logopedprogrammet
8LOA55 / HT 2014
Robert Eklund
2014-10-23

Acoustics: Crib Sheet

This document very succinctly lists the commonest terms covered by the course, with a brief explanation of how this is used in acoustic–phonetic analysis. For more detailed information, turn to lecture slides and software user guides and manuals.

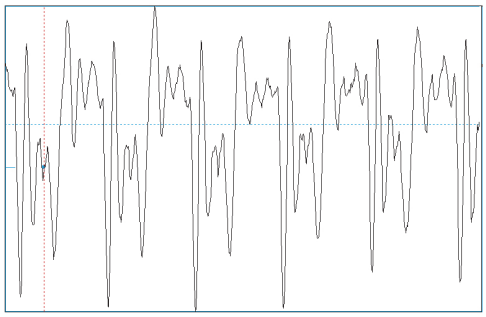
Sound

Sound consists of fluctuating pressure in a medium, e.g. air. Its velocity is dependent on the density of the medium (sound moves faster in iron than in air), but also other factors such as air humidity etc. Everything that fluctuates can be described as a wave, and in much literature sound is compared to ripples on water when you throw in a little rock. Since sound fluctuates between higher-than-normal and lower-than-normal air pressure the terms *peaks* (higher pressure) and *valleys* (lower pressure) are commonly used.

Waveform

A speech waveform is quasi-periodic (only sine waves are completely periodic) and indirectly tells you how regularly the vocal folds open and close. Note that only voiced speech sounds can be described as a wave. Waveform analysis gives you an idea as to how regularly the vocal folds move.

Note, however, that waveform analysis is not a *direct* measure of vocal folds oscillation, but a measure of the effect vocal folds oscillation has on air pressure once it has passed through the filter (see below) and is radiated through/by the lips.



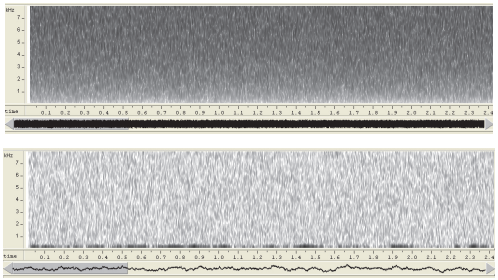
Typical speech waveform.
Peaks indicate higher-than-baseline air pressure.
Valleys indicate lower-than-baseline air pressure.

Noise

The definition of *noise* in data collections is "everything that you are not interested in". The opposite is *signal* (what you are interested in).

In speech noise refers to the complete lack of waveforms (see above). In-between noise and regular waveforms you have semi-regular sounds. like dog growls.

Some speech sounds consist of noise only, like the fricatives /f/ and /s/ and in pathological voices noise can be seen in several speech sounds.



Top: White noise. Bottom: Brown noise.
Complete lack of recurring periods.
White: Equal energy in all frequency bands.
Brown: Different energy in different frequency bands.

Frequency

Frequency ("frekvens") is a general term for how often something happens per time unit. Some examples are Christmas = 1/year; Olympic Games 0.25/year (i.e. once every four years). In speech analysis (and music) the term Hertz (Hz) is used which means times per second. The tone "standard a" in music is defined as 440 Hz. A typical male voice lies at around 100 Hz, while that of a woman lies at 200 Hz and that of a child at 300 Hz. Hz cannot be used to compare ranges between subjects given that Hz have different "values" at different frequency bands; please see Semitones.

Semitones

Semitones (ST; "halvtoner") correspond to the keys of a piano and divide the octave "double/half frequency" into twelve steps. While an octave in lower registers might be 50 Hz (the difference between 50 Hz and 100 Hz), the same perceived difference (an octave) is 1000 Hz in a higher register (the difference between 1000 Hz and 2000 Hz). In both these cases the difference is 12 STs. This difference sounds "the same" to the human ear so STs should be used instead of Hz since STs are always the same when the perceived sound "sounds the same" unlike Hz where values/figures will differ a lot dependent on the register/frequency band.

Pitch

Pitch ("tonhöjd") denotes *perceived* frequency. While frequency is an objective measure that can be accurately measured, pitch is what the ear/brain perceives. Although the two are related there is no one-to-one relationship between the two. If a tone above 2000 Hz is amplified it will be perceived as higher in pitch, while a tone under 1000 Hz which is amplified will be perceived as lower in pitch. In the register 1000–2000 Hz the pitch of a tone is more or less unaffected by amplitude changes. Several scales have been conceived to accurately describe human frequency-vs-pitch discrimination, but let's not get into that here and now.

Amplitude

Amplitude ("amplitud") is the term used to describe the objective measurement of energy in a signal. The most common unit used is Decibel (dB) which is a relative measure as well as a logarithmic measure. Although dB actually only compares two sounds and should not be used to denote absolute levels its most common use assumes that 0 dB is the human hearing threshold (and is left unmentioned). Note that although acoustic analysis software can provide you with dB values amplitude cannot be measured in the computer but is measured with a sound level meter on location at the time of recording and with the distance from the source carefully measured. Software does not provide you with the amplitude of the *sound* but of the sound *file* which is highly dependent on the microphone used, the sound card of the recording equipment, and is also subject to manipulation (in e.g. WaveSurfer, Audacity and Cool Edit).

Loudness

Tantamount to the distinction between frequency and pitch, the difference between amplitude and loudness ("ljudstyrka") is that the former is an objective measure and the latter is used to describe *perceived* amplitude (by humans). Consequently there are different dB scales used, and the one that represents human hearing is called dB(A) or dB(a). Three standard dB levels are:

3 dB = twice the energy/intensity

6 dB = half/double the distance from the source

10 dB = perceived halving/doubling of loudness

Source–Filter Theory

Theory presented by Gunnar Fant (1960) that describes speech production as a:

1: *sound source* (the lungs to provide air/energy) and the vocal folds (that act as an oscillator),

… and a:

2: *filter* (the vocal tract, tongue, cheeks, nasal cavity, velum, lips, teeth) that filters the signal and that wave creates resonances which in speech are called formants.

Since the source and the filter are considered linearly separable (they act independently), one can cancel out the filter in order to obtain source characteristics; see inverse filtering.

Partials

The tones that constitute the building blocks of voiced speech. When an oscillator (something that vibrates) creates regular fluctuations ("waves") this creates a tone. Perfect oscillation, i.e. fluctuation with perfect periodicity (the waves occur with perfect time and amplitude) creates a sine wave ("sinuston") which does not exist naturally.

Voiced speech (above all vowels, but also voiced consonants) consists of quasi-periodic fluctuations, although the term periodic is commonly used. If the fundamental tone ("grundton") is 100 Hz, the upper partials ("övertoner") will occur at 200, 300, 400 (etc) Hz, i.e. as even multiples ("jämna multiplar") of the fundamental. If the fundamental is 150 Hz, the upper partials will be 300, 450, 600 (etc) Hz.

In Swedish you can either talk about "deltoner" where the fundamental is "delton 1" followed by higher "deltoner" that begin with "delton 2" etc, or you can talk about "grundton" plus "övertoner" where the first "överton" is numbered "1" which then corresponds to "delton 2", i.e. the numbering is shifted by one.

Partials are then created by the oscillating vocal folds, i.e. the source, and serve as the "raw material" of speech, which are then shaped by the vocal tract, i.e. the filter, to produce all voiced speech sounds.

In normal speech partials taper off by 6–12 dB per octave, thus creating a slope in the spectrum which exhibits characteristic shapes depending on the voice type (laboration 3); see narrowband spectrogram graphic below.

Resonances and formants

In speech, the vocal tract acts as a filter that creates peaks or dips in the spectral slope. Increased amplitude is seen as a peak and is called a *resonance* (in general acoustics) or a *formant* (in human speech and sometimes for monkey vocalizations).

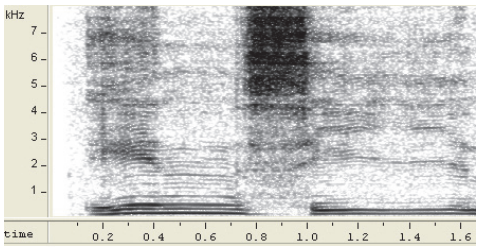
In a wideband spectrogram they appear as dark horizontal bands; see Spectrogram: wideband graphic on the next page.

Spectrogram: narrowband

Spectrograms display sound (e.g. human speech) in three dimensions: time (*x* axis), frequency (*y* axis) and amplitude (*z* axis, indicated by blackness).

A narrowband spectrogram (~45 Hz/0.03 seconds) displays the partials as narrow dark horizontal lines; see graphic to the right.

Useful for analysis of *voice* characteristics.

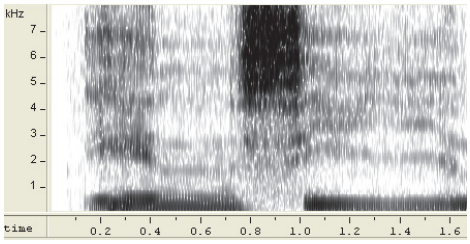


Narrowband spectrogram (in WaveSurfer).
Partials are shown as narrow horizontal bands

Spectrogram: wideband

A wideband spectrogram (~250 Hz/0.005 seconds) works like a narrowband spectrogram but displays the formants instead of partials; see graphic to the right.

Useful for analysis of *speech* characteristics.



Wideband spectrogram (in WaveSurfer).
Formants are shown as wide horizontal bands.

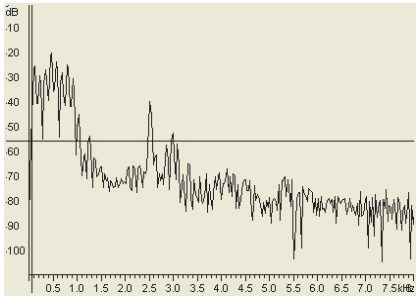
Note on bandwidth/time

In WaveSurfer bandwidth is specified by bandwidth (250 Hz).
In Praat bandwidth is specified by duration (0.005 seconds).
The formula is `1.2982804 / bandwidth`

Spectrum: FFT

A spectrum takes a "slice" out of a spectrogram and looks at it "head-on" with frequency on the *x* axis and amplitude on the *y* axis.

An FFT (for "Fast Fourier Transform") spectrum displays the partials (similar to a narrowband spectrogram) and is useful for analysis of *voice* characteristics. See graphic to the right.



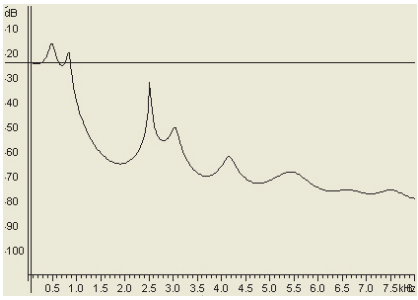
FFT Spectrum (in WaveSurfer).
Partials are shown as narrow peaks.

Spectrum: LPC

An LPC (for "Linear Predictive Coding") spectrum works like an FFT spectrum but instead displays the formants (similar to a wideband spectrogram) and is useful for analysis of *speech* characteristics. See graphic to the right.

Spectrogram and Spectra: software note

Since spectrogram and spectra are dynamically linked in WaveSurfer (and even be watches as little "movies"), and you can switch between all four of the above analyses in a second, WaveSurfer is recommended for these analyses.



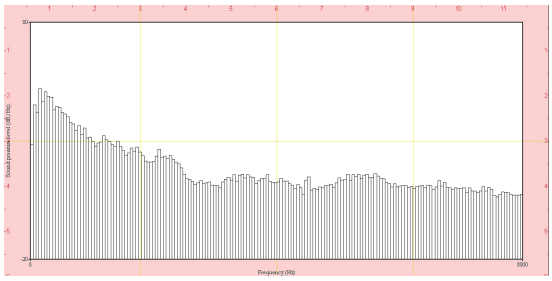
LPC Spectrum (in WaveSurfer).
Formants are shown as clearly defines peaks.
Relative and absolute distances between the peaks define what vowel sound is perceived.

Long-Term Average Spectrum (LTAS)

A long-term average spectrum is created from a large number of spectral slices (collected from ~20 seconds of speech) which are then averaged and turned into one "summarizing" spectrum

Unvoiced sounds should be avoided if possible. Since it is an analysis that is averaged over a fairly long stretch of speech, all speech characteristics are "blurred" out and the image can be seen as a kind of "voice print" of the particular voice.

LTAS analysis is readily available in WaveSurfer.



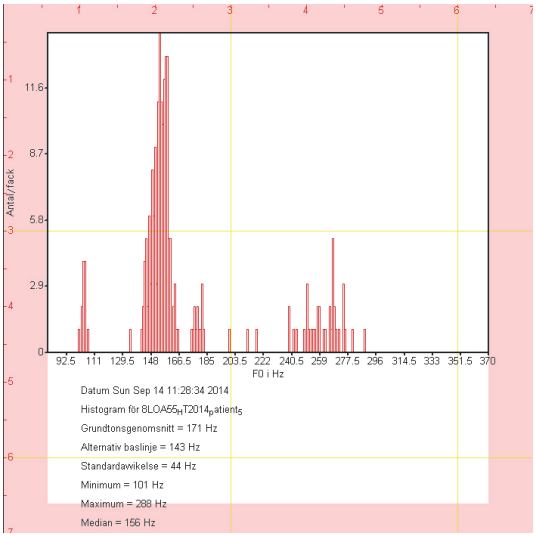
LTAS (in Praat).
Make sure that all data points appear within the image by adjusting the scales (axes values).

Histogram

A phonetic histogram provides general statistics of how often different frequency bands are used in a given voice. It can thus be likened to a kind of "voice print" top reveal voice characteristics. See graphic to the right.

Requires Voxalys plug-in.

Histograms cannot be created in WaveSurfer.



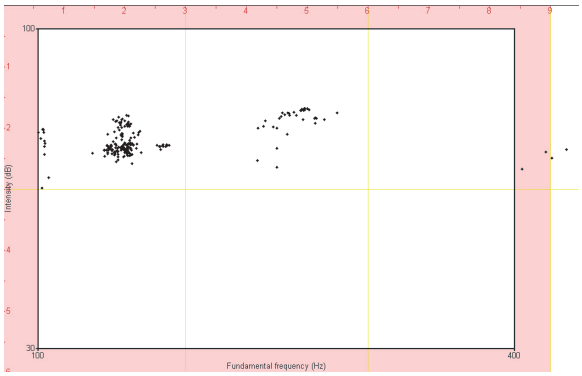
Histogram as shown (in Praat).

Phonetogram

A phonetogram is similar to a histogram but also includes amplitude. The data points thus provide information about the frequency *and* the amplitude of the data sample points, and that way also acts like a kind of voice print. See graphic to the right.

Phonetograms can be created in Praat "proper" or by using the Voxalys plug-in.

Phonetograms cannot be created in WaveSurfer.



Phonetogram (in Praat).
Make sure that all data points appear within the image by adjusting the scales (axes values).

Inverse analysis

An LPC analysis of the sound wave uses the air pressure differences in order to calculate what the filter configuration looks like (which gives you the formant values). Inverse analysis uses this but turns it "upside-down" in order to *cancel out* the effect of the filter (the so called "transfer function") in order to obtain a more "truthful" analysis of vocal fold oscillation.

The effect of inverse analysis can be verified by looking at LPC spectra before and after inverse analysis and should have the effect that the formants (i.e. what the filter creates) are cancelled out. Since neither LPC nor inverse analysis are "perfect", and since glottal oscillation is not perfect, either, the spectrum will still have "bumps", but it should be much smoother and lack obvious formant peaks. (There are other reasons for this, but again: let's not get into that here and now).

If you listen to a sound file that has been object to inverse filtering it should sound somewhat like a buzzing bee, which is not surprising since a buzzing bee constitutes a sound source (the wings – instead of vocal folds – act as oscillators) *without* a filter, and since inverse analysis "takes away" the filter the sound of an inverse audio file is what a human voice phonated where the vocal folds were exposed directly into the air (like bee wings), without a head (vocal tract) filtering the signal.

Jitter

While a healthy voice exhibits sound waves where the peaks and valleys both appear at regular distances (equidistant), i.e. stable frequency, and with similar height (i.e. stable amplitude), pathological voices exhibit irregularities in both these domains. Irregular frequency (distances between peaks in the waveform) is called jitter. This is a useful measure to assess the effect of treatment where "before–after" analysis can show you whether treatment has the desired effect.

Shimmer

See jitter above. Irregularities in peak and valley height (amplitude) are called shimmer. Note that pathological voices often exhibit both jitter and shimmer. This is a useful measure to assess the effect of treatment where "before–after" analysis can show you whether treatment has the desired effect.

Harmonics-to-Noise ratio (HNR)

Harmonics-to-Noise ratio measure the relative contribution of harmonic material (partials) in a voice compared to noise. As you noticed during the laboration, most of you had an HNR values of 20–26 dB (that is how much stronger your partials were compared to noise) while the pathological voices you analysed exhibited HNR values of around 3–6 dB. HNR is a useful measure to assess the effect of treatment where "before–after" analysis can show you whether treatment has the desired effect.

Final comments

Note that this document provides very condensed explanations and definitions which inevitably leads to a certain (over-)simplification. For more detailed descriptions please consult the lecture/seminar slides and the laboration instructions, as well as the T2 lecture slides and all other additional materials handed out during the acoustics course.