Exercise #1:
Manual speech transcription, analysis & synthesis, basic signal representations

Exercise summary

Every speech technologist should know their data. The first part of the exercise is about getting hands-on experience with speech signals, including how to transcribe speech into linguistic constituents and how to conduct basic acoustic analyses with existing tools, namely Praat. In the second part of the exercise, basic signal reading and time and frequency plotting are rehearsed in MATLAB.

Data and existing resources: An utterance from a speech corpus (LibriSpeech): `251-136532-0016.flac` available in `data/`. Pre-defined MATLAB script frame `E1_main.m` a. A pre-provided script at `aux_scripts/readTextGrid.m` for reading your annotation file.

Software: Praat. Available from https://www.fon.hum.uva.nl/praat/ for all standard operating systems. MATLAB, available on university computers and with a student license (see https://intra.tuni.fi/content/software/10)

Tasks:

Praat:

1) Manual segmentation and labelling of phones
2) Measurement of segment durations
3) Measurement of segment formant frequencies
4) Measurement of utterance min, max, and mean F0
5) Synthesis by segment concatenation

MATLAB:

6) Waveform reading and resampling
7) Windowing, waveform and spectrum plotting

Deliverables: ***A written report*** with a table of acoustic measurements and answers to questions + ***synthesized .wav*** + ***annotation TextGrid-file***, packaged into a .zip file `E1_firstname_surname_studentID.zip`.

Learning goals: hands-on experience on speech transcription and analysis, basic understanding of phones, words, formants, F0, and coarticulation.

Note 1: All exercises of the course, including this one, will consist of a written report and additional files (e.g., code, sounds, annotation files) that are to be created and submitted for evaluation. In the report, always add your name and student ID to the beginning of the report. Use complete English sentences or paragraphs of text to answer the exercise questions, and number each response according to the question numbers in the exercise instructions. Some of the questions have strictly correct answers while others can have multiple valid responses. All requested exercise files are to be placed inside the main folder inside the .zip file without any subfolders.

Note 2: The provided MATLAB code template will save key variables at several points of the script. These variables will be used to evaluate correctness of the provided solutions. Do not comment out these lines of code, remember to execute them when producing your results, and do not change the naming of the key variables defined in the template.

Task 1: Manual segmentation and labeling of words and phones

The first task is about transcribing word- and phone-level units of speech in the given audio clip. Annotation consists of 1) marking unit boundaries in time, and 2) writing down the unit types ("identities") for each segment. Figure 1 shows an example of word and phone annotation for a speech clip.

Task 1.1: Create a Praat annotation tier named "words" and annotate words "*the*", "*Terran*", "*public*", "*wanted*", "*to*", "*hear*", "*about*", "*Martians*" as their own segments, and so that each word segment begins at where the previous one ends.

Task 1.2: Create a Praat annotation tier named "phones" and annotate phones of the first four words ("*the Terran public wanted*"). For simplification, you can assume the following phone string:

[DH] [AH] [T] [EH] [R] [AH] [N] [P] [AH] [B] [L] [IH] [K] [W] [A] [N] [IH] [D]

(see Appendix A for conversion of symbols to IPA phones).

Task 1.3: Save your annotation as a `251-136532-0016.TextGrid` file when you are ready. Remember to also save your intermediate progress regularly.

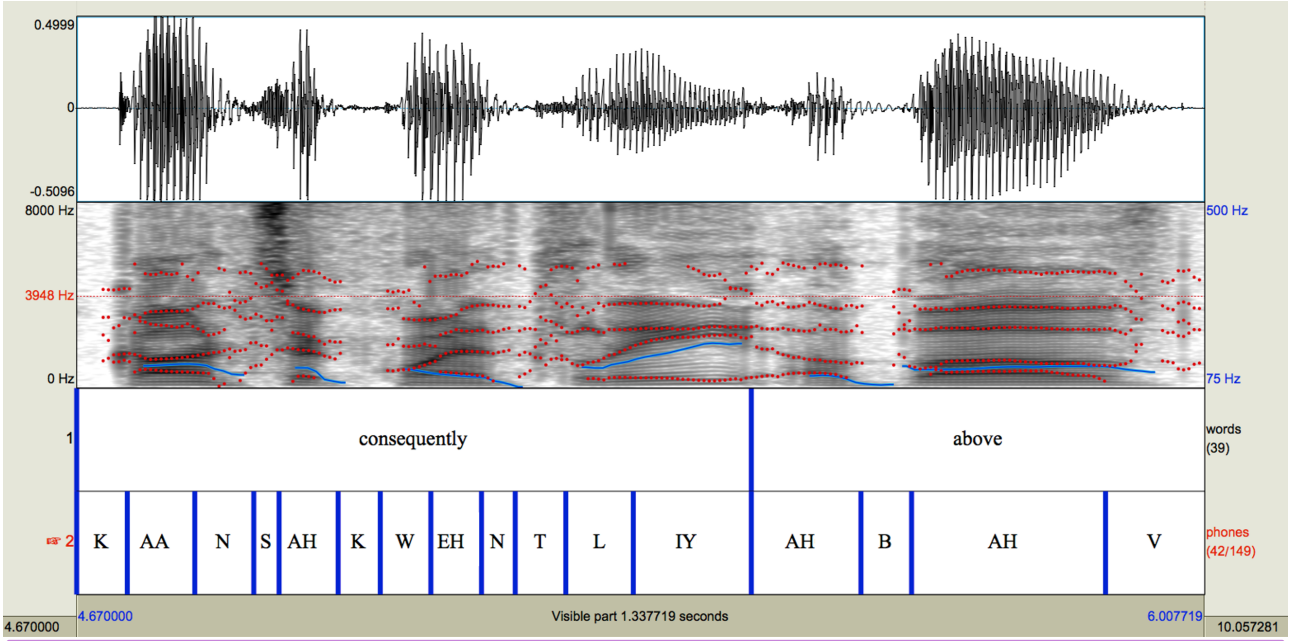


Figure 1: Example annotation from another speech sample. Waveform and spectrum with formants (red dots) and F0 (blue line) are shown in the two top panels. Word and phone annotation tiers are shown at the two bottom panels.

Guidance:

1) Load 251-136532-0016.flac in Praat, either by "Open with…" functionality of your OS, or by starting Praat and then from Praat objects list **[Read from file]".**
2) Create annotation tiers for your signal from Praat objects list by selecting "Sound 251-136532-0016" and clicking **[Annotate] → [to TextGrid]**, and then defining two tiers in the prompt: "words phones". Leave "point tier" field empty.

3) Go to annotation mode by selecting both the audio and the "TextGrid 251-136532-0016" items from the Praat objects list and clicking **[View & Edit]**.
4) Start annotation by clicking on the correct tier, then paint the chosen segment from waveform or spectrum with a mouse. To add a boundary, click **[Boundary]** → **[Add on selected tier]** (or see the menu for hotkeys). By having the segment selected, you can type in name of the segment (word or phone identity).
5) You can start the next segment from the previous one by clicking on the boundary and then defining the new endpoint while keeping shift-key pressed down.
6) Remember to save your .TextGrid file. To save, select the TextGrid annotation object, click "Save" at Praat object menu, and "Save as text file".

You can also find guidance to Praat annotation tool from videos, such as https://www.youtube.com/watch?v=w5CJE-KpEqI .

Hint #1: you can move the boundaries with mouse after you have defined them. Click on the bars at the bottom of the view to listen to the segments and selections.
Hint #2: you can zoom-in and out of the waveform and spectrogram from the **[View]** menu or using the keyboard shortcuts listed there.
Hint #3: Plosives, such as [*t,p,k,d,b,g*] consist of a closure ("silence") and burst phases that both belong to the consonant segment. However, identification of closure duration at the beginning of speech may not be possible, as the beginning of the closure can be inaudible and hence indistinguishable from the leading silence.

Question 1.1: Which ones were more difficult to segment in time: words or phones?

Question 1.2: Can you clearly delineate all changepoints from a phone to another? Are some sound combinations more challenging? If so, why?

Task 2: Duration estimation

Task 2.1: Measure durations of phone segments in words "The Terran" ([DH] [AH] [T] [EH] [R] [AH] [N]) and record the values to the second column of Table 1. Use milliseconds as the unit of measurement.

Guidance: Segment duration is shown at the bottom of Praat window when you highlight the segment of interest.

Question 2.1: How would you characterize the measured phone durations? Are they uniform or varying? Assuming that the present sample is representative, how many phones per second would this type of read speech approximately have?

Task 3: Formant estimation

Formants are resonances of the vocal tract during voiced speech.

Task 3.1: Visually estimate values of the first two formants (F1 and F2) of the voiced phone segments in words "The Terran" ([DH] [AH] [T] [EH] [R] [AH] [N]) and record the values to the

third and fourth columns of Table 1. If the formants are changing throughout the segments, visually estimate the average values per segment.

Guidance: you can get visualization of formant estimates by clicking **[Formant]→ [Show formants]** from the top menu. Formants will show up as red dots during voiced speech. Clicking a position on the spectrogram will show the corresponding frequency on the left edge of the window. You can also view amplitude spectrum of a painted selection by clicking **[Spectrum] → [View spectral slice]**.

Question 3.1: What are the typical frequency ranges for F1 and F2 that you observe in this speech sample? Do they align with typical formant frequencies of American English reported by Hillenbradt et al. (1995) and shown in Appendix B?

Task 4: Fundamental frequency estimation

Fundamental frequency (F0) is the vibration frequency of the vocal folds, and therefore the longest periodic component in speech.

Task 4.1 Estimate average fundamental frequency of voiced phone segments in words "The Terran" ([DH] [AH] [T] [EH] [R] [AH] [N]), and record the values to the last column of Table 1.

Guidance: you can get visualization of the fundamental frequency by clicking **[Pitch] → [Show pitch]** from the top menu. F0 contour will show up as a blue continuous line during voiced speech. Clicking on the spectrogram will show the corresponding frequency on F0 scale at the right edge of the window.

Question 4.1: How would you characterize the behavior of the fundamental frequency (F0) contour across the whole clip? Do levels or changes of F0 align with phone and/or word segments?

Task 5: Manual concatenative synthesis

One approach to create synthetic speech is to concatenate suitable pre-recorded speech segments to form new sentences.

Task 5.1: Create a synthetic utterance "*Erin's red moon*" by concatenating suitable speech segments from the audio sample. Save the created audio as E1_synthesis.wav

Guidance:

1) Start by highlighting the segment of choice from `251-136532-0016.flac` and click "**[File] → [Extract selected sound (time from 0)]**" to create a new audio clip in Praat list.
2) You can add more segments to the new signal using copy & paste from the old one (**[copy selection to sound clipboard]** + **[paste after the selection]**). Note that copying of selected audio segments works only in audio signal [View & Edit] mode, not in annotation mode (= [View & Edit] with both the signal and TextGrid file selected). If you want to see the annotation while selecting segments, you can open the audio signal

twice from Praat objects list: once where you only open the audio (for which copy and cut works), and using another window where you open both the annotation and signal at the same time. Any segment selection/zooming operations conducted in the annotation view will automatically update the selections in the signal-only view as well.

3) You can save the signal as .wav from the Praat objects list in Praat main view.

Hint: you can try lengthening the sounds by pasting some (shorter) segments several times. Pay attention to waveform continuation at the segment edges to avoid distortion artefacts.

Question 5.1: How would you characterize the quality of the resulting speech?

Question 5.2: Consider and discuss possible ways of making the concatenative synthesis sound more natural.

Task 6 (MATLAB):

Waveform reading and resampling

Task 6.1: Start building your scripts based on the template frame `E1_main.m`. Read the utterance file `251-136532-0016.flac` located in `data/` folder into variable '`x`', and sampling rate to variable '`fs`'. Resample the signal to 16,000 Hz.

Guidance: You can use `resample()`-function to implement sampling. Remember to redefine '`fs`' based on the resampling process.

Question 6.1: Why is 16-kHz sampling rate suitable for typical speech analysis tasks?

Task 7: Windowing, waveform and spectrum plotting

Task 7.1: Extract a 20-ms Hamming-windowed waveform segment from middle part of sound 'IH' (/i/) of word "*public*" you annotated earlier.

Task 7.2: Calculate logarithmic magnitude spectrum of the windowed signal (i.e., on dB scale).

Task 7.3: Create a plot with two panels, where the top panel shows the windowed signal waveform and the bottom panel shows the corresponding log-magnitude spectrum. Remember to define x- and y-axes of the plots correctly and informatively. See Fig. 2 for an example.

Guidance:
The provided code template reads your .TextGrid annotation file from Task 1 and finds the onset and offset timestamps for the [IH] segment.

Hamming window function is built-in to MATLAB as `hamming()`, fast Fourier transform is `fft()`, absolute value is `abs()`, and 10-base logarithm is `log10()`. For plotting: use

`figure()`, `plot()`, and `subplot()`. To set the axes and their labels correctly, you can use `xlim()`, `ylim()`, `xlabel()`, `ylabel()`, `grid()`.

Complex-valued Fourier spectrum X(k) of a signal **x** = {x_0, x_1, …, x_{N-1}} is calculated with discrete Fourier transform (DFT) as

$$\mathrm{X}(k) = DFT(\mathbf{x}, k) = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \quad (2.1)$$

where $k = \{0, 1, \ldots, \frac{N}{2}\}$ can be interpreted as signal frequencies $f = fs \cdot k/N$. In practice, DFT is calculated with fast Fourier transform (FFT). Magnitude spectrum is obtained from X(k) with

$$\mathrm{X}_{\mathrm{magn}}(k) = |\mathrm{X}(k)| \quad (2.2)$$

and logarithmic spectrum as

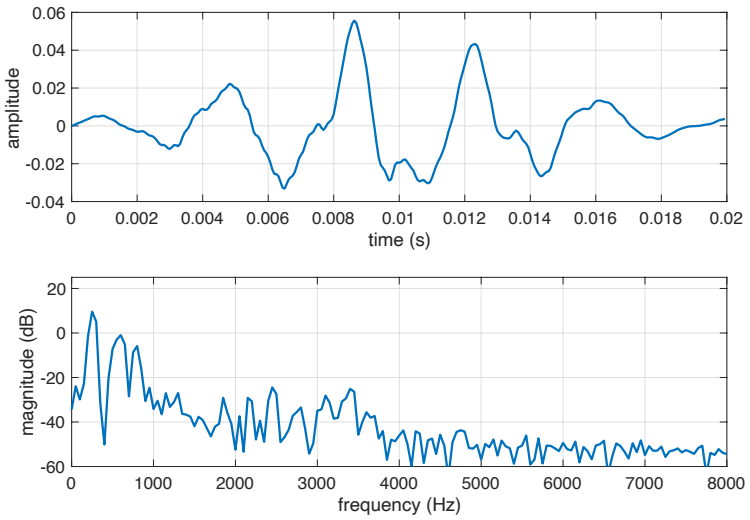
$$\mathrm{X}_{\mathrm{log-magn}}(k) = 20 \cdot \log_{10}\{|\mathrm{X}(k)|\} \quad (2.3)$$

or, in case of power-spectrum |Y(k)|^2, as

$$\mathrm{X}_{\mathrm{log-magn}}(k) = 10 \cdot \log_{10}\{|\mathrm{X}(k)|^2\} \quad (2.4)$$

Hamming window is defined as:

$$h(n) = 0.54 + 0.46 \cos\left\{\frac{2\pi n}{N}\right\} \quad (2.5)$$



F0 (Hz)

Appendix A: IPA to ARPAbet conversion table

		IPA Symbol	ARPAbet (SV)	ARPAbet (UV)	Examples
Vowels	Front	i	i	IY	beat
		ɪ	I	IH	bit
		e	e	EY	bait
		ɛ	E	EH	bet
		æ	@	AE	bat
	Back	ɑ	a	AA	Bob
		ɔ	c	AO	bought
		o	o	OW	boat
		ʊ	U	UH	book
		u	u	UW	boot
	Mid	ɝ	R	ER	bird
		ə	x	AX	ago
		ʌ	A	AH	but
Diphthongs		ɑɪ	Y	AY	buy
		ɑʊ	W	AW	down
		ɔɪ	O	OY	boy
		ɨ	X	IX	roses
Stop Consonants	Voiced	b	b	B	bat
		d	d	D	deep
		g	g	G	go
	Unvoiced	p	p	P	pea
		t	t	T	tea
		k	k	K	kick
Fricatives	Voiced	v	v	V	vice
		ð	D	DH	then
		z	z	Z	zebra
		ʒ	Z	ZH	measure
	Unvoiced	f	f	F	five
		θ	T	TH	thing
		s	s	S	so
		ʃ	S	SH	show
Semivowels	Liquids	l	l	L	love
		l̩	L	EL	cattle
		r	r	R	race
	Glides	w	w	W	want
		ʍ	H	WH	when
		j	y	Y	yard
Nasal	Non vocalic	m	m	M	mom
		n	n	N	noon
		ŋ	G	NX	sing
	Vocalic	m̩	M	EM	some
		n̩	N	EN	son
Affricates		tʃ	C	CH	church
		dʒ	J	JH	just
Others	Whisper	h	h	HH	help
	Vocalic	ɾ	F	DX	batter
	Glottal stop	ʔ	Q	Q	

Appendix B: Table of average American English formant frequencies

Table reproduced from Hillenbrandt et al. (1995) for pedagogical purposes.

TABLE V. Average durations, fundamental frequencies, and formant frequencies of vowels produced by 45 men, 48 women, and 46 children. Averages are based on a subset of the tokens that were well identified by listeners (see text for details). The duration measurements are in ms; all others are in Hz.

		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Dur	M	243	192	267	189	278	267	283	265	192	237	188	263
	W	306	237	320	254	332	323	353	326	249	303	226	321
	C	297	248	314	235	322	311	319	310	247	278	234	307
*F*0	M	138	135	129	127	123	123	121	129	133	143	133	130
	W	227	224	219	214	215	215	210	217	230	235	218	217
	C	246	241	237	230	228	229	225	236	243	249	236	237
*F*1	M	342	427	476	580	588	768	652	497	469	378	623	474
	W	437	483	536	731	669	936	781	555	519	459	753	523
	C	452	511	564	749	717	1002	803	597	568	494	749	586
*F*2	M	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
	W	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588
	C	3081	2552	2656	2267	2501	1688	1210	1137	1490	1345	1546	1719
*F*3	M	3000	2684	2691	2605	2601	2522	2538	2459	2434	2343	2550	1710
	W	3372	3053	3047	2979	2972	2815	2824	2828	2827	2735	2933	1929
	C	3702	3403	3323	3310	3289	2950	2982	2987	3072	2988	3145	2143
*F*4	M	3657	3618	3649	3677	3624	3687	3486	3384	3400	3357	3557	3334
	W	4352	4334	4319	4294	4290	4299	3923	3927	4052	4115	4092	3914
	C	4572	4575	4422	4671	4409	4307	3919	4167	4328	4276	4320	3788