1

x[n;]=w[n;]x[n]

(2)

where

n=1

X(!; )=

x[n;]e

(1)

j!n

X1

The Fouriertransform of the windowedspeech waveform(STFT) is given by

this time. :-)

veform under a sliding window - pretty muchlike we did in the previous lab, but in frequency domain

the short-time FourierTransform- STFT consists of separate FourierTransformson pieces of the wa-

whole speech signal cannot capture the time-varying frequency content of the waveform. In contrast,

eventsare so radically dierentboth in time and in frequency that a single Fouriertransform overthe

In the previous lab, you have seen that speech consists of a sequence of dierent events. These

1.1 Short Time FourierAnalysis - Spectrogram

using long analysis windows in time.

analysis includes the use of short analysis windowsin time, whereas narrowbandanalysis is performed

spectrogram. The spectrogram can be produced using wideband or narrowband analysis. Wideband

You will rst familiarize yourself with the time-frequency representation of speech, the so-called

1 Theoretical Background

age (adult or children) detection.

The nal goal of this lab is to implement a simple system for speaker gender (male, female) and

basic components on speech, such as the pitch.

of speech signals, with the help of Short-Time FourierAnalysis (spectrogram), and you will calculate

sonants, using the FastFourierTransform(FFT). Youwill learn about time-frequency representation

willexplorethefrequencydomainstructureofthemostbasicspeechelements,suchasvowelsandcon-

During this lab, you will have a rst contact on frequency domain analysis of speech signals. You

February25, 2021

Frequency-domainprocessing

A second hands-on lab on SpeechProcessing

Part2

Project 0:

Evaluation Warning : The document was created with Spire.PDF for java.

2

speech segment in time results in a frequency convolutionbetweenthe corresponding spectra. Hence,

the mainlobe (and thus, its bandwidth). Also, you should know that multiplying a window with a

already know that the length of the window aects its spectral characteristics, and mainly the size of

whereas a short window (i.e. a pitch period or even less) results in wideband analysis. You should

length that is used. A long window (i.e. up to 3 or 4 pitch periods) results in narrowband analysis,

Nowbasedonthatexpression,therearetwodierenttypesofSTFTanalysis,accordingtothewindow

k=1

P2

k

S(!; )=



H(!)G(!)W(!!

;)



(5)

1





2



X1



Thus, the spectrogram can be expressed as

x[n;]=w[n;](p[n]g[n]h[n])

(4)

ow overone cycle, g[n]:

k=1

of periodically placed impulses, p[n ]=

[nkP], with P being the pitchperiod, and a glottal

1

P

system with impulse response h[n] and with a glottal ow input given by the convolution of a series

For voiced speech, we can approximate the speech waveform as the output of a linear time-invariant

S(!;)=jX(!;)j

(3)

2

ristics and it can be described mathematically as

spectrogramis a graphical 2D display of the squared magnitude of the time-varyingspectral characte-

represents the windowed speech segment as a function of the center of the window, at time . The

Figure 1: Narrowband analysis of speech.

Evaluation Warning : The document was created with Spire.PDF for java.

3

spectrogram(s, 30\*10^(-3)\*fs, 20\*10^(-3)\*fs, 1024, fs, 'yaxis')

figure; subplot(211); plot(t, s); xlabel('Time (s)'); subplot(212);

% Window length of 30 msec and step of 10 msec

t = 0:1/fs:length(s)/fs - 1/fs;

[s, fs] = wavread('H.22.16k.wav');

A typical narrowband spectrogram is given in Figure 1. The code that generated it is given:

k=1

k

k

k

S(!;)

jG(!

)H(!

)j

jW(!!

;)j

(6)

2

2

X1

transform are negligible, we can approximately state that

of the shifted window Fourier transforms are non-overlapping, and that the sidelobes of the window

of duration of at least twopitch periods (more than 20 ms). Under the condition that the main lobes

As we said, the narrowband spectrogram, a "long" - in time - analysis window is used, typically

1.1.1 Narrowband analysis

Figure 2: Wideband analysis of speech.

analysis.

the underlying speech spectrum. Keeping this in mind, let us discuss the wideband and narrowband

simply speaking, the spectrum of the analysis window is "placed" around and on the harmonics of

Evaluation Warning : The document was created with Spire.PDF for java.

4

frame\_v = hamming(L1).\*frame1;

% Windowing it

L1 = length(frame1);

frame1 = s(3600:4400);

% Extracting a frame

[s,fs] = wavread('H.22.16k.wav');

% Loading the waveform

apply a FFT on it, what we haveis in Figure 3. The necessary MATLABcode is given:

If we select a voiced speech portion, long enough to resolve the harmonics in the spectrum, and

be used.

spectral contentof the source, but rather the the envelopeof speech. Thus,a narrowbandanalysis will

us nd out! :-) For our purpose, a wideband analysis is not convenient, since it does not reveal the

formants. However,aquestionshouldbe: howisthespectralcontentofdierentspeechelements? Let

dingtothetypeofanalysis,wegeteithertheharmonicstructureoranapproximationofthevocaltract

So, it is obvious that the STFTs are generated by "concatenated slices" of Fourierspectra. Accor-

1.2 FourierTransformand Spectral Content of Speech

Colours havethe same meaning as in narrowband spectrogram.

spectrogram(s, 5\*10^(-3)\*fs, 2\*10^(-3)\*fs, 1024, fs, 'yaxis');

figure; subplot(211); plot(t, s); xlabel('Time (s)'); subplot(212);

% Window length of 5 msec and step of 3 msec

waveform. A wideband spectrogram is depicted in Figure 2. The code is given below:

striations arise because the short window is sliding through uctuating energy regions of the speech

formants of the vocal tract in frequency, but also gives vertical striations over time. These vertical

where E[] is the energy of the waveformunder the sliding window. Thus, the spectrogram shows the

S(!;)jH(!)G(!)j

E[]

(7)

2

the wideband spectrogram can be very roughly approximated as

providedbywidebandanalysis,butgoodtimeresolutionisprovided. Forasteady-statevoicedsegment,

jH(!)G(!)j due to the vocal tract and glottal ow contributions. Thus, poor frequency resolution is

window transform and smear out the harmonic line structure, roughly revealing the spectral envelope

transform of the window,when "placed" on the harmonics, will overlapand add with its neighbouring

pitch period. By shortening the window length, its Fouriertransform "widens". This "wide" Fourier

For the wideband spectrogram, a "short" window is chosen with a duration of less than a single

1.1.2 Wideband analysis

color is for low magnitude areas (and thus, low energy regions).

intense red or black color corresponds to high magnitude values(high energy), whereas yellowor blue

revealneperiodicitychangesovertime. Itshouldbenotedthatcolorsinspectrogramhaveameaning:

poortimeresolution,becausethelonganalysiswindowcoversseveralpitchperiodsandthusisunableto

theharmonicsareeectivelyresolved(horizontalstriationsonthespectrogram). However,italsogives

thevocaltracttransferfunction. Thenarrowbandspectrogramgivesgoodfrequencyresolution because

- which are shaped by the magnitude of the product of the Fouriertransform of the glottal ow and

set of narrow "harmonic" lines - whose width is determined by the Fourier transform of the window

We can see that using a long window in time on a voiced segment gives a STFT that consists of a

Evaluation Warning : The document was created with Spire.PDF for java.

5

1Recent studies, however,haveshown that speech is (quasi-)harmonic up to 16 kHz!!

% Windowing it

L2 = length(frame2);

frame2 = s(4800:5500);

% Extracting a frame

[s,fs] = wavread('H.22.16k.wav');

% Loading the waveform

gure is given below:

in the spectrum, and apply a FFT on it, what we have is in Figure 3. The code that produces this

If weselect an unvoicedspeechportion, long enough to resolveanystructure (surely not harmonic)

and age detection system.

spectrum is mostly covered by noise

. The spectrum and its peaks are a means to build our gender

1

of the FFT spectra. It is also clear that the speech harmonics are up to 4 kHz, and the rest of the

It is clear that the horizontal striations that are seen in Figure 1 come from the harmonic peaks

ylabel('FFT Magnitude'); xlabel('Frequency (Hz)'); grid;

subplot(212); plot(freq, 20\*log10(X1(1:NFFT/2)));

subplot(211); plot(frame1); xlabel('Time (samples)'); grid;

% Plot

freq = [0:fs/NFFT:fs/2-1/fs];

% Make frequency bins into frequencies

X1 = abs(fft(frame\_v, NFFT));

NFFT = 1024;

% Apply FFT and then take the absolute value in 1024 points

Figure 3: FFT spectrum of voiced speech.

Frequency (Hz)

0

1000

2000

3000

4000

5000

6000

7000

8000

−60

−40

−20

0

20

40

Fourier Spectrum

Time (s)

0.22

0.23

0.24

0.25

0.26

0.27

0.28

0.29

−0.4

−0.2

0

0.2

0.4

Voiced speech

Evaluation Warning : The document was created with Spire.PDF for java.

6

is no spectral content-not even noise- above6 kHz

2In all speech sample waveformdepicted in these gures, the signal is lowpass-ltered at 6 kHz, and that is why there

mass of the vocal cords and the pressure of the forced expiration also called the sub-glottal pressure.

The pitch of speech is determined by four main factors. These include the length, tension, and

and fundamental frequency of speech are used interchangeablyin literature.

sensation of the fundamentalfrequency of the pulses of airowfrom the glottal folds. The terms pitch

speechsignals. Theinverseoftheperiodisthefundamental frequencyofspeech. Pitchistheperceived

The periodic opening and closing of the vocal folds results in the harmonic structure in voiced

1.3 Pitch

Figure 4: FFT spectrum of unvoicedspeech.

Frequency (Hz)

0

1000

2000

3000

4000

5000

6000

7000

8000

−100

−80

−60

−40

−20

0

Time (s)

0.3

0.305

0.31

0.315

0.32

0.325

0.33

0.335

−0.02

−0.01

0

0.01

0.02

Unvoiced speech

parts in Figure 1).

.

2

speech as white noise, and the spectrogram information that we get in unvoicedregions (see unvoiced

There is no harmonic structure. This representation is consistent with the approximation of unvoiced

We can see that the spectrum of unvoiced speech is almost at and covers the whole spectrum.

ylabel('FFT Magnitude'); xlabel('Frequency (Hz)'); grid;

subplot(212); plot(freq, 20\*log10(X2(1:NFFT/2)));

subplot(211); plot(frame2); xlabel('Time (samples)'); grid;

% Plot

freq = [0:fs/NFFT:fs/2-1/fs];

% Make frequency bins into frequencies

X2 = abs(fft(frame\_unv, NFFT));

NFFT = 1024;

% Apply FFT and then take the absolute value in 1024 points

frame\_unv = hamming(L2).\*frame2;

Evaluation Warning : The document was created with Spire.PDF for java.

7

of the highest peaks in the waveform,wecan see that it is D =0:24340:2376=0:0058, whichis the

0

0

0

isattimet

=0:005875sec,whichcorrespondstof

=1=t

=170:2128Hz. Ifwemeasurethedistance

ples above,weget the result of Figure 5. As youcan see, the rst peak of the autocorrelation function

Ifweapplytheautocorrelationfunctioninthevoicedspeechsegmentthatispresentedintheexam-

in the autocorrelation function.

0;P;2P;::: That is, the pitchperiod can be estimated bynding the location of the rst maximum

be easily shown that for periodic signals, the autocorrelation function attains a maximum at samples

i.e. the autocorrelation function of a periodic signal is also periodic with the same period. It can also

(k)=(k+P)

(9)

with period P samples, then it can be shown that

The autocorrelation is a measure of similarity between signals. Forexample, if the signal is periodic

m=1

(k )=

x[m]x[m+k]

(8)

X1

The autocorrelation function of a discrete-time deterministic signal is dened as

courses. Wewill remind you here the most basic notions of the autocorrelation theory.

The autocorrelation function is (or should be :-) ) known to you from Digital Signal Processing

2.1 Short-time autocorrelation method

gender and the age of the speaker.

ecient :-) ) methods for pitch estimation. Our pitch estimates can then give us an idea about the

For our purpose, we will implement and compare a pair of rather simple (and for that, not very

several algorithms in literature, the robust estimation of pitch is still a relatively open subject.

Pitch tracking is still a very hot topic of research in speech signal engineering. Although there are

2 Pitch TrackingTechniques

and therefore a simple gender+age detection system can be implemented.

asimpletime-domainandasimplefrequency-domainmethodforestimatingthepitchofvoicedspeech,

should implement some simple techniquesfor pitchtracking. Forthis, we will describe and implement

Hence, we can detect the gender and the age of a speaker by tracking his/her pitch. :-) Thus, we

havea high-pitched speech signal of 300400 Hz.

(b)Ageandstateofhealth. Pitchcanalsosignalage,weightandstateofhealth. Forexample,children

pitch is the main indicator of gender.

averagepitch for females is about 200 Hz whereas the averagepitch for males is bout 110 Hz. Hence,

(a) Gender is conveyedin part by the vocal tract characteristics and in part by the pitch value. The

Among others, the following information is contained in the pitch signal:

tory of pitch (and other harmonics) overtime.

speaking manner, emotion, and accent. Figure 1 illustrates an example of the variationsof the trajec-

The pitchvariationscarry most of the intonationsignals associated with prosody (rhythmsof speech),

Evaluation Warning : The document was created with Spire.PDF for java.

8

1. Load one of the provided waveforms that end in -pout.wav. These signals are purely voiced,

Foryour convenience,follow the next steps:

Youwillusethepitchtrackersdescribedaboveinordertodesignyourage+genderdetectionsystem.

3 Age+Gender Detection System Implementation

parameters. :-)

pitchis unknown,so wecannot validateour result, unless wecreate a syntheticsignal that has known

signal is not strictly periodic, or due to the resolution of the FFT (1024 points). Of course, the actual

0

f

=171:9 Hz, whichis veryclose to 170:2 Hz. However,the mismatchcan be due to the fact that the

selecttherstsignicantpeak, wewillseetheresultofFigure6. Therstpeakislocatedatfrequency

Forexample,letustakealookatthemagnitudespectrumoftheusualvoicedspeechspectrum,and

spectrum and revealnot only the pitchbut the whole harmonic structure a voicedspeechsegment! :-)

frequency(andthus,thepitch),wecandevelopanalgorithmthatcanperformpeak-pickingonanFFT

fundamental frequency. Since the rst signicant peak of the spectrum is related to the fundamental

ain: it is dominated by sharp peaks at frequency locations that are nearly harmonically related to the

As it is shown in the previous sections, voiced speech has a certain structure in frequency dom-

2.2 Peak picking

it was this function that generated the result in Figure 5.

Foryour convenince,MATLABhas its own function for correlation measurements. It is xcorr and

0

same result, and thus the pitch is f

=1=0:0058=170:2128 Hz. :-)

speech

Figure 5: Upper panel: Voiced speech waveform. Lower panel: Autocorrelation function of voiced

Time (s)

−0.05

−0.04

−0.03

−0.02

−0.01

0

0.01

0.02

0.03

0.04

0.05

−10

−5

0

5

Y: 7.304

10

X: 0.005875

Time (s)

0.22

0.23

0.24

0.25

0.26

0.27

0.28

0.29

−0.4

−0.2

0

0.2

Y: 0.2113

Y: 0.2128

X: 0.2376

X: 0.2434

0.4

Voiced speech

Evaluation Warning : The document was created with Spire.PDF for java.

9

function of MATLABwill become useful.

 Interpolateyourpitchestimatesusingsplinesinordertoobtainapitchcontour. Theinterp1

 Use an FFT resolution of 2048 points.

like.

Thenyoucansaveitasfunction\\_that\\_does\\_something.mleandcallitwheneveryou

% End of function

out2 = %CODE

out1 = %CODE

%CODE

%CODE

%CODE

% FUNCTION\_THAT\_DOES\_THAT takes in1, in2, in3 arguments and returns out1, out2

% Comments

function [out1, out2] = function\_that\_does\_something(in1, in2, in3)

like this:

dicult - a simple rst derivativecriterion is enough). A MATLABfunction can be written

on your speech segment! You also have to write your own peak picking algorithm (not so

MATLAB's built-in functions t and xcorr. Do not forget to apply a Hamming window

 Estimate the pitchfor eachframe using both algorithms - FFT peak-pickingand ACF.Use

msec.

 Do a frame-by-frame analysis, with an analysis window of 30 msec and a frame rate of 10

using an approach similar to the one used in VUS discriminator:

0

s

synthetic speech, with known f

and sampling frequency F

= 8 kHz. Perform pitch estimation

Figure 6: Upper panel: Voicedspeech waveform. Lowerpanel: Magnitude Spectrum and rst peak

Frequency (Hz)

0

1000

2000

3000

4000

5000

6000

7000

8000

−60

−40

−20

0

20

Y: 24.05

40

X: 171.9

Magnitude Spectrum

Time (s)

0.22

0.23

0.24

0.25

0.26

0.27

0.28

0.29

−0.4

−0.2

0

0.2

Y: 0.2113

Y: 0.2128

X: 0.2376

X: 0.2434

0.4

Voiced speech

Evaluation Warning : The document was created with Spire.PDF for java.

10

If you haveANY questions on this lab, please send an e-mail to : hy578-list@csd.uoc.gr

7. Delivery deadline: 11 March 2021

parts with your VUS discriminator and set the pitch to zero in these time intervals).

parts: sincetheACFandthepeaksdonotcorrespondtoanypitch,youcanpre-detectnon-voiced

build the pitch contour for a full speech waveform! :-) (Care should be taken for the non-voiced

6. Optional: UsetheVUSdiscriminatorofthepreviouslabandthepitchtrackerofyourchoice,and

a plot of the pitch contour, and a text string, 'adult male', 'adult female', 'child'.

5. Accordingtothepreviousnote,theoutputofyoursystemshouldbeaplotofthespeechwaveform,

and the pitch of a female adult lies in the range 160250 Hz.

whereas a child has a pitch range from 300 to 500 Hz. A male adult ranges from 70 to 150 Hz,

4. For gender+age detection, you are given that an adult has a pitch ranging from 70 to 250 Hz,

3. Which method performs better? Why?

2. Which contour is closer to the true frequency given in the name of the -pout.wav les?

the frequencygrid of the voiced speech waveform.

 Optional: perform peak pickingin ALL peaks of the spectrum and construct an estimate of

Evaluation Warning : The document was created with Spire.PDF for java.