TIME-DOMAIN PITCH ESTIMATION USING AMDF

Instructor: Dr. Ninh Khanh Duy

Nguyen Phi Truong
Student ID: 102210082
Class: 21TCLC_DT1

Table of content

1

Problem

Input: Random audio of voice contains voiced, unvoiced, silent signal

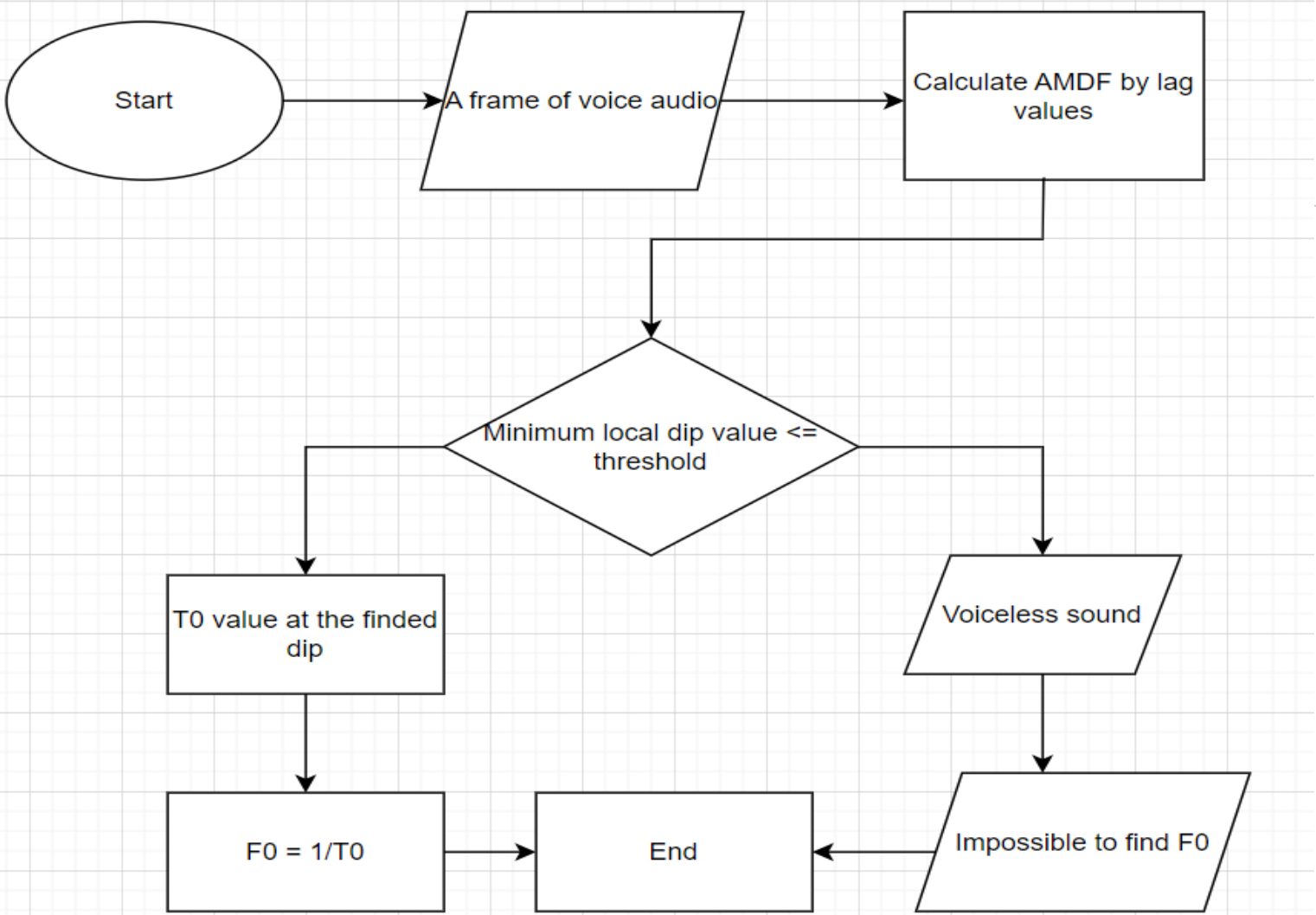
Output: Estimate fundamental pitch
F0 mean and F0 std, plot the F0 contour and AMDF of voiced, unvoiced frame

Constraint: The variability of people’s fundamental frequency is 70 – 400 Hz

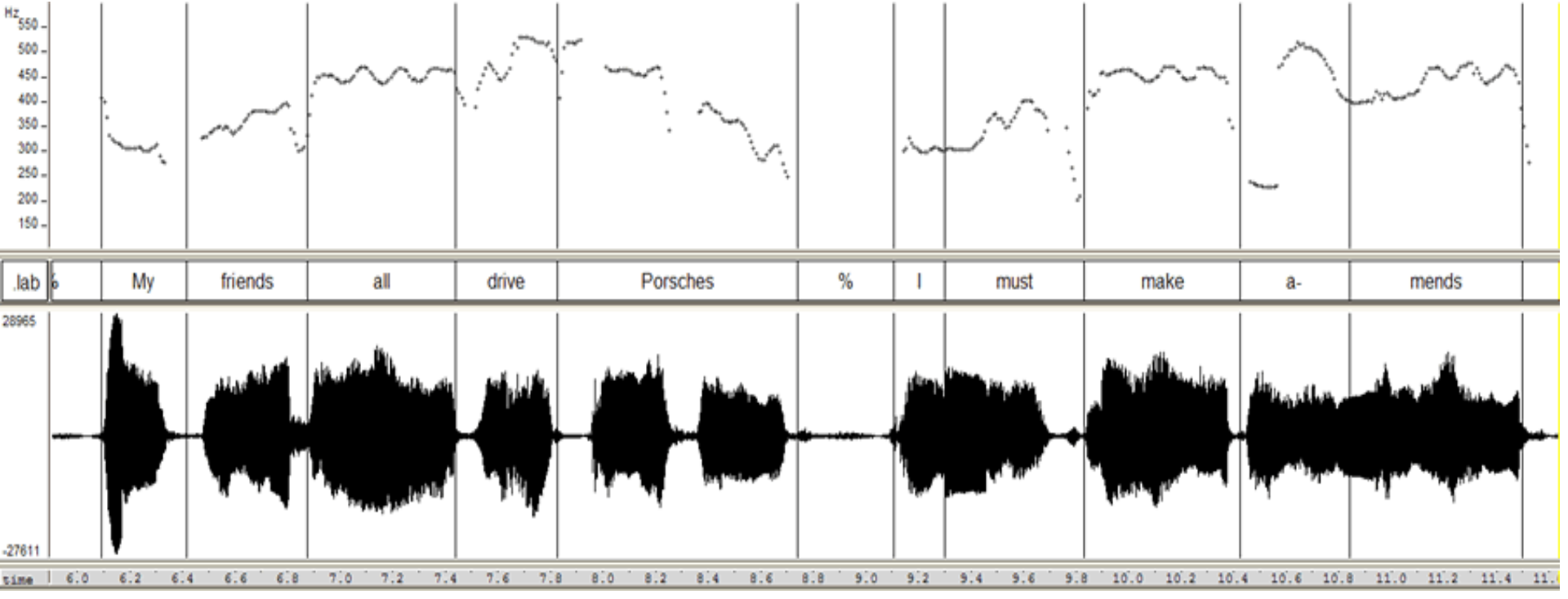
2

Algorithm

- Finding F0: using short-term average magnitude difference function (**AMDF**)
- Smoothing F0 contour: using median filter
- Finding local minimum dips
- Finding VU discrimination threshold


Start
A frame of voice audio
Calculate AMDF by lag values
Minimum local dip value <= threshold
T0 value at the finded dip
Voiceless sound
F0 = 1/T0
End
Impossible to find F0

Median filter



Median filter

```
[135.59322034 135.59322034 137.93103448 136.75213675 137.93103448
 134.2991546  132.19700305 129.77637055 126.84887477 124.48712
 122.77533475 123.01598759 124.29678659 126.79700912 129.90870568
 133.46383676 136.15262466 138.352084   139.88403599 142.03189952
 143.52904311 144.82038796 146.69453971 148.77705158 149.66635314
 146.62027469 143.03882035 138.13329318 132.93199914 126.70113354
 123.78685096 121.42359649 119.33765276 118.39048965 118.22815885
 118.33475646 117.8694591  117.51404536 115.73678829 114.10800963
 114.23898114 114.51196106 114.99358035 115.72600943 116.4868363
 115.60583785 114.67135245 113.94655084 113.75962795 113.98627023
 115.04740243 116.83265478 116.97215666 116.80504993 116.14209401
 114.59593143 112.20830066 111.03396314 109.73764993 111.82176987
 109.84688472 108.49530203 107.9280038  107.36099761 103.3860069
 103.80454465 104.50382606 100.          99.37888199 104.5751634
 108.84353741 100.62893082]
```

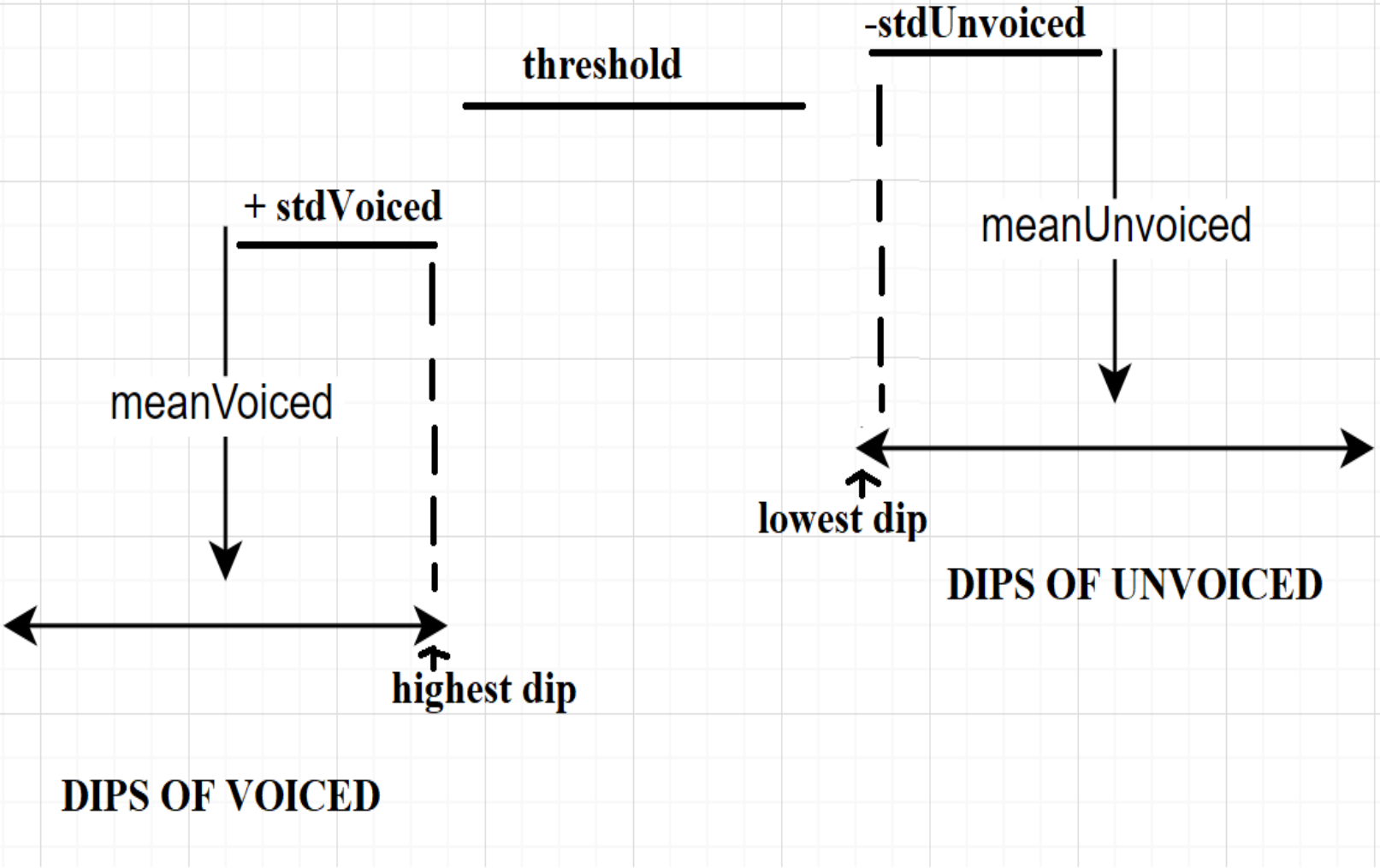
Observation: The neighborhood pitch won’t shift too much from each other

→

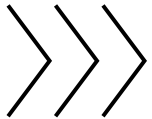
The pitch which has too much shift from its neighborhood pitch is virtual pitch

Finding VU discrimination threshold

File	MeanVoiced	StdVoiced	MeanUnvoiced	StdUnvoiced
phone_M2	0.263	0.152	0.616	0.097
phone_F2	0.272	0.150	0.554	0.114
studio_M2	0.245	0.165	0.553	0.090
studio_F2	0.217	0.166	0.558	0.150


-stdUnvoiced
threshold
+ stdVoiced
meanUnvoiced
meanVoiced
lowest dip
DIPS OF UNVOICED
highest dip
DIPS OF VOICED

Threshold must be between highest dip of voiced and lowest dip of unvoiced



Threshold = avg(highest dip of voiced, lowest dip of unvoiced

Taking average of all audio training input threshold, we have the final threshold to discriminate VU

File	MeanVoiced	StdVoiced	MeanUnvoiced	StdUnvoiced	(meanVoice + stdVoice + meanSilent - stdSilent) / 2
Phone_M2	0.263	0.152	0.616	0.697	0.467
Phone_F2	0.272	0.150	0.554	0.114	0.431
Studio_M2	0.245	0.165	0.553	0.090	0.436
Studio_F2	0.217	0.166	0.558	0.150	0.395
Threshold = AVG(COLUMN(6))					0.4325

3

Experimental results

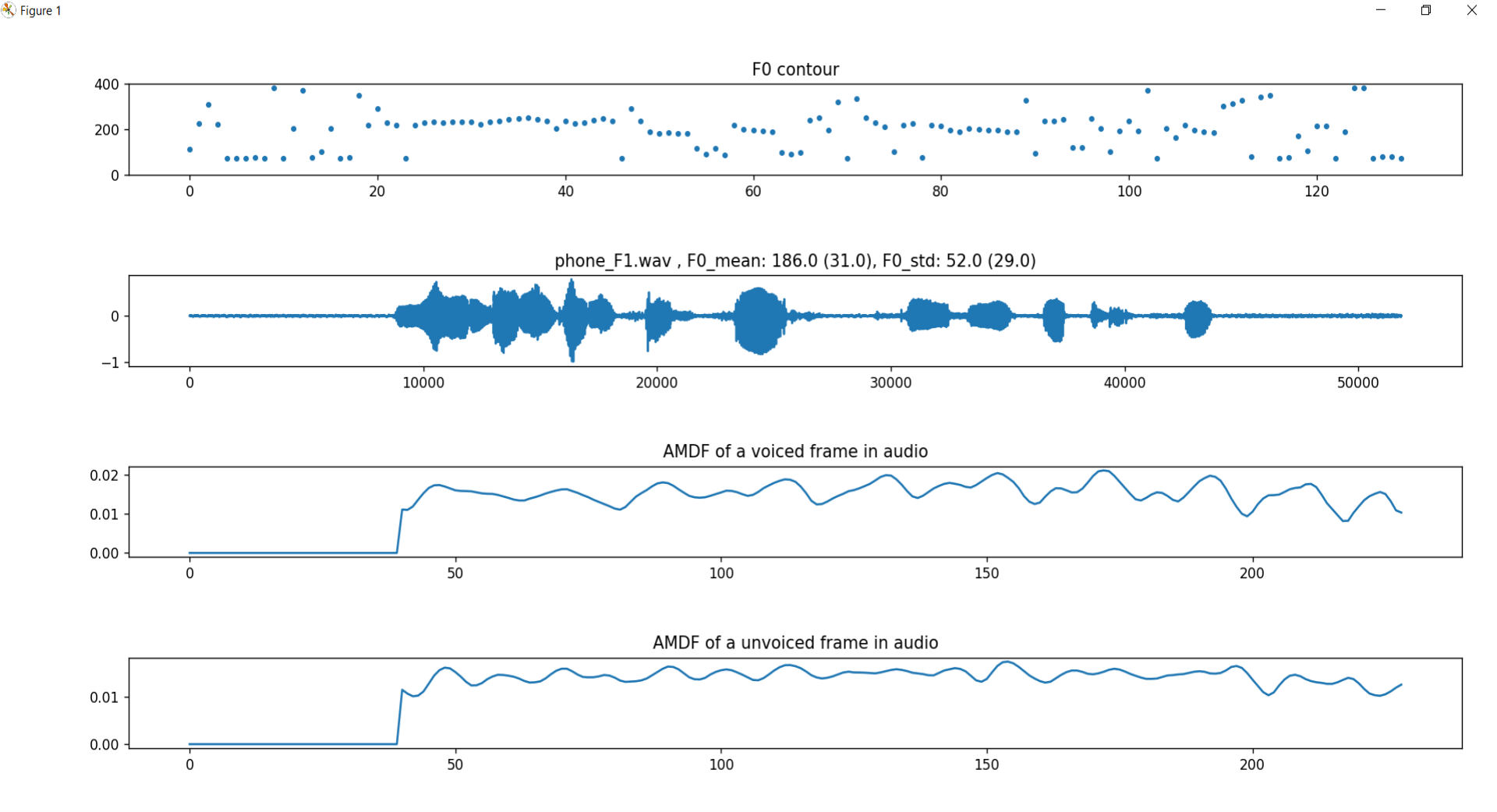


Figure 1
F0 contour
400
200
0
0
20
40
60
80
100
120
phone_F1.wav , F0_mean: 186.0 (31.0), F0_std: 52.0 (29.0)
0
−1
0
10000
20000
30000
40000
50000
AMDF of a voiced frame in audio
0.02
0.01
0.00
0
50
100
150
200
AMDF of a unvoiced frame in audio
0.01
0.00
0
50
100
150
200

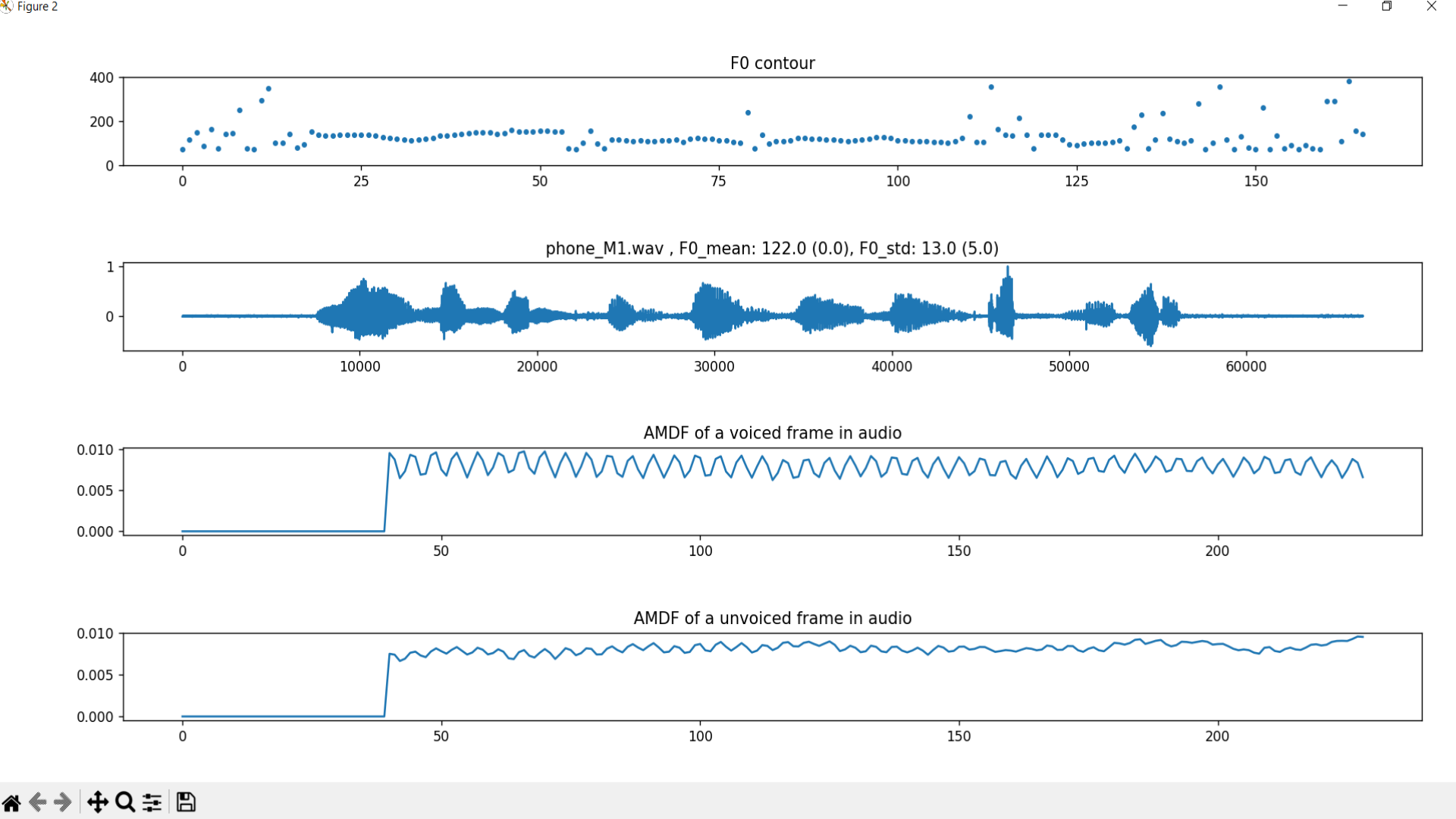


Figure 2
F0 contour
400
200
0
0
25
50
75
100
125
150
phone_M1.wav , F0_mean: 122.0 (0.0), F0_std: 13.0 (5.0)
1
0
0
10000
20000
30000
40000
50000
60000
AMDF of a voiced frame in audio
0.010
0.005
0.000
0
50
100
150
200
AMDF of a unvoiced frame in audio
0.010
0.005
0.000
0
50
100
150
200

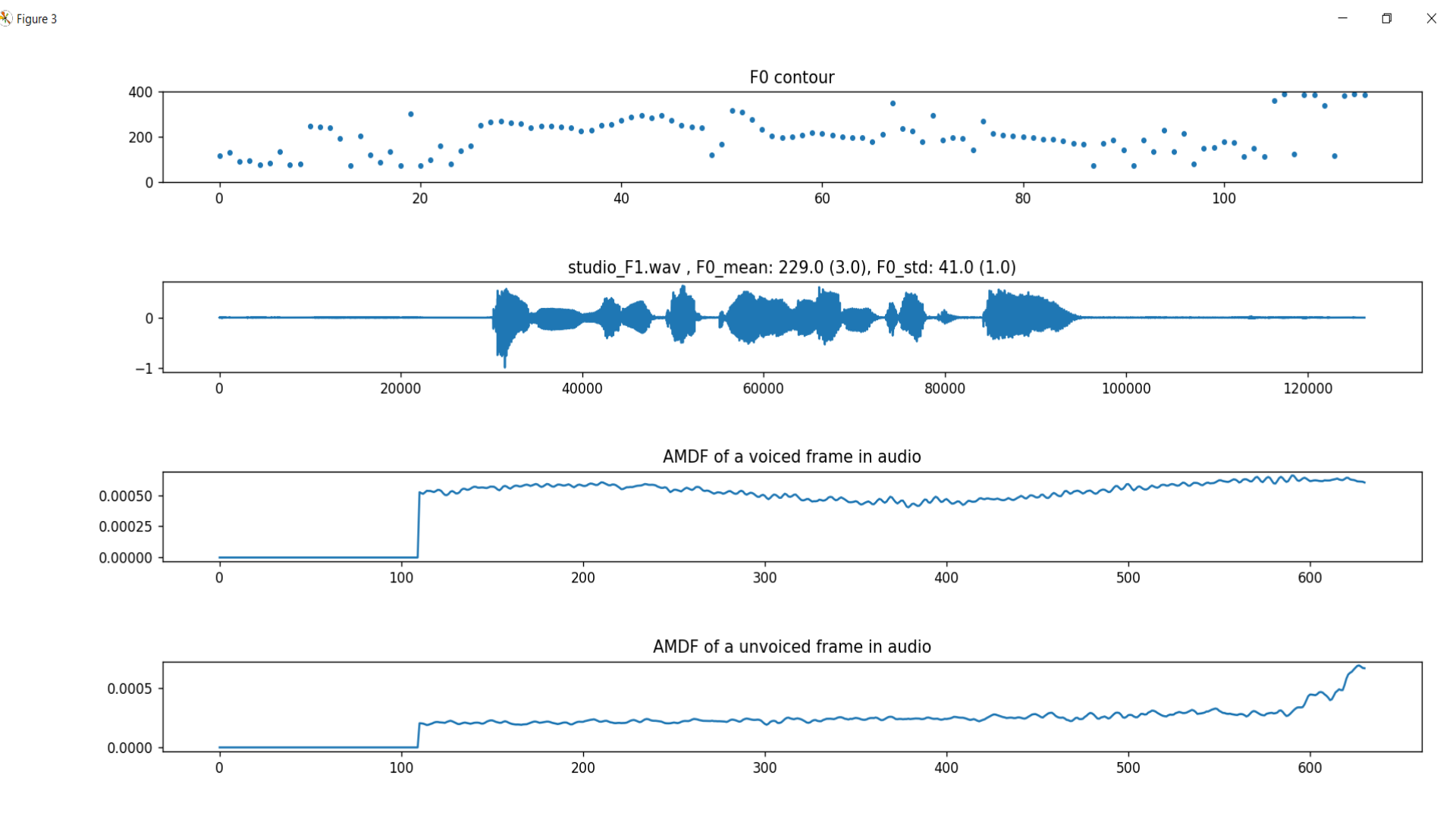


Figure 3
F0 contour
400
200
0
0
20
40
60
80
100
studio_F1.wav , F0_mean: 229.0 (3.0), F0_std: 41.0 (1.0)
0
−1
0
20000
40000
60000
80000
100000
120000
AMDF of a voiced frame in audio
0.00050
0.00025
0.00000
0
100
200
300
400
500
600
AMDF of a unvoiced frame in audio
0.0005
0.0000
0
100
200
300
400
500
600

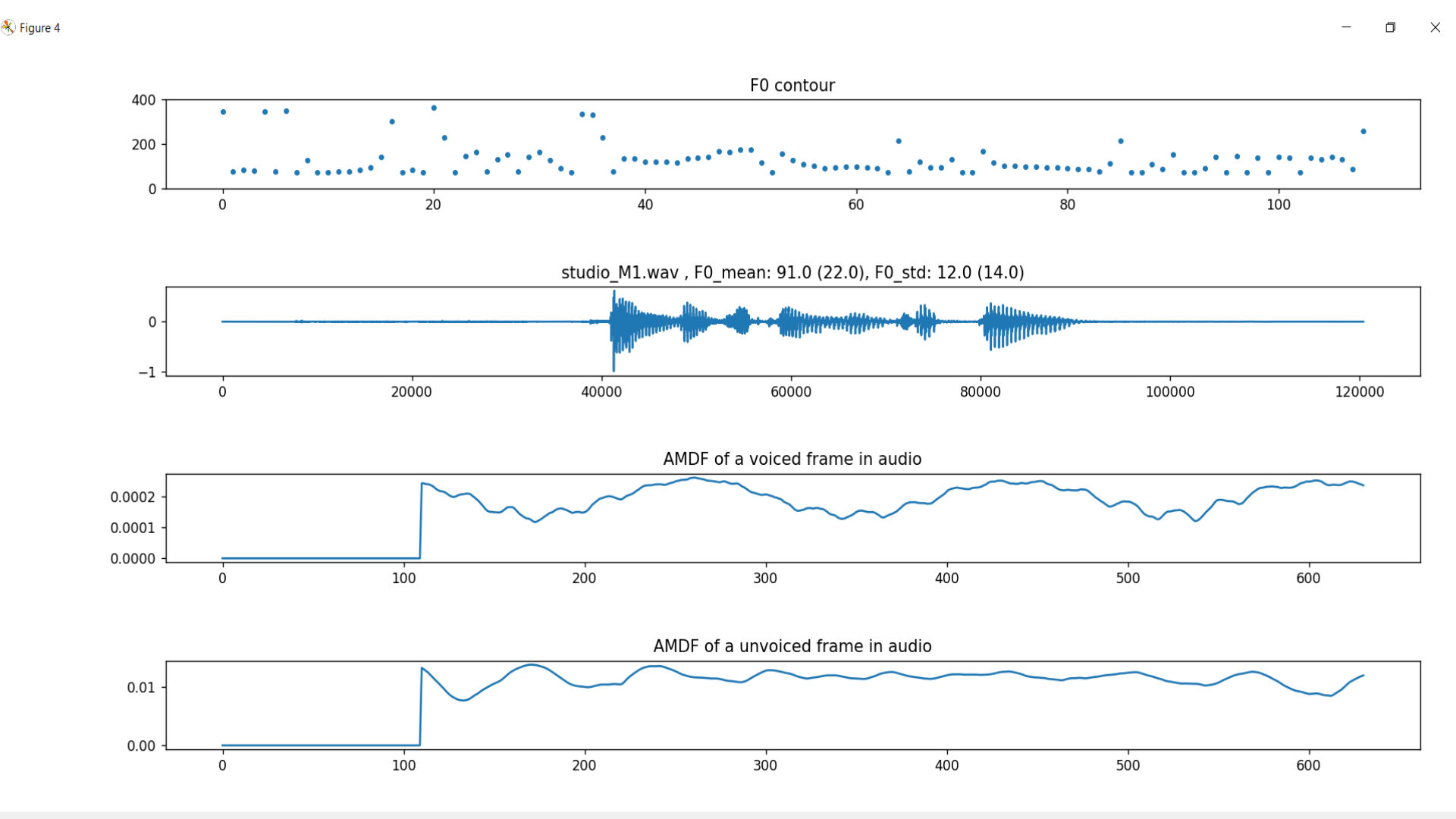


Figure 4
F0 contour
400
200
0
0
20
40
60
80
100
studio_M1.wav , F0_mean: 91.0 (22.0), F0_std: 12.0 (14.0)
0
−1
0
20000
40000
60000
80000
100000
120000
AMDF of a voiced frame in audio
0.0002
0.0001
0.0000
0
100
200
300
400
500
600
AMDF of a unvoiced frame in audio
0.01
0.00
0
100
200
300
400
500
600

File	Type	AMDF (Hz)	Lab (Hz)	Relative error (%)
Phone_F1	F0 mean	186	217	14.29
	F0 std	52	23	126.09
Phone_M1	F0 mean	122	122	0
	F0 std	13	18	27.78
Studio_F1	F0 mean	229	232	1.29
	F0 std	41	40	2.5
Studio_M1	F0 mean	91	113	19.47
	F0 std	12	26	53.85

Lowest F0 mean relative error: Phone_M1 (0%)
Lowest F0 std relative error: Studio_F1 (2.5%)
Highest F0 mean relative error: Studio_M1 (19.47%)
Highest F0 std relative error: Phone_F1 (126.09%)

Studio_F1 has the best result: F0 mean relative error (1.29%) and F0 std relative error (2.5%)

Phone_F1 has the worst result: F0 mean relative error (14.29%) and F0 std relative error (126.09%)

3

Conclusions

- The median filter is only efficient when filling out the isolated pitch, when there are multiple neighboring pitchs the function became less efficient thus it can not filter out all virtual pitchs

=> Find a way to handle multiple neighboring pitchs

- The discrimination threshold is not accurate with all audio input

=> Find a way to automate the finding threshold process (maybe using statistic distribution …)

Thanks!