

Older People's Korean Speech Data Processing with Voice Conversion

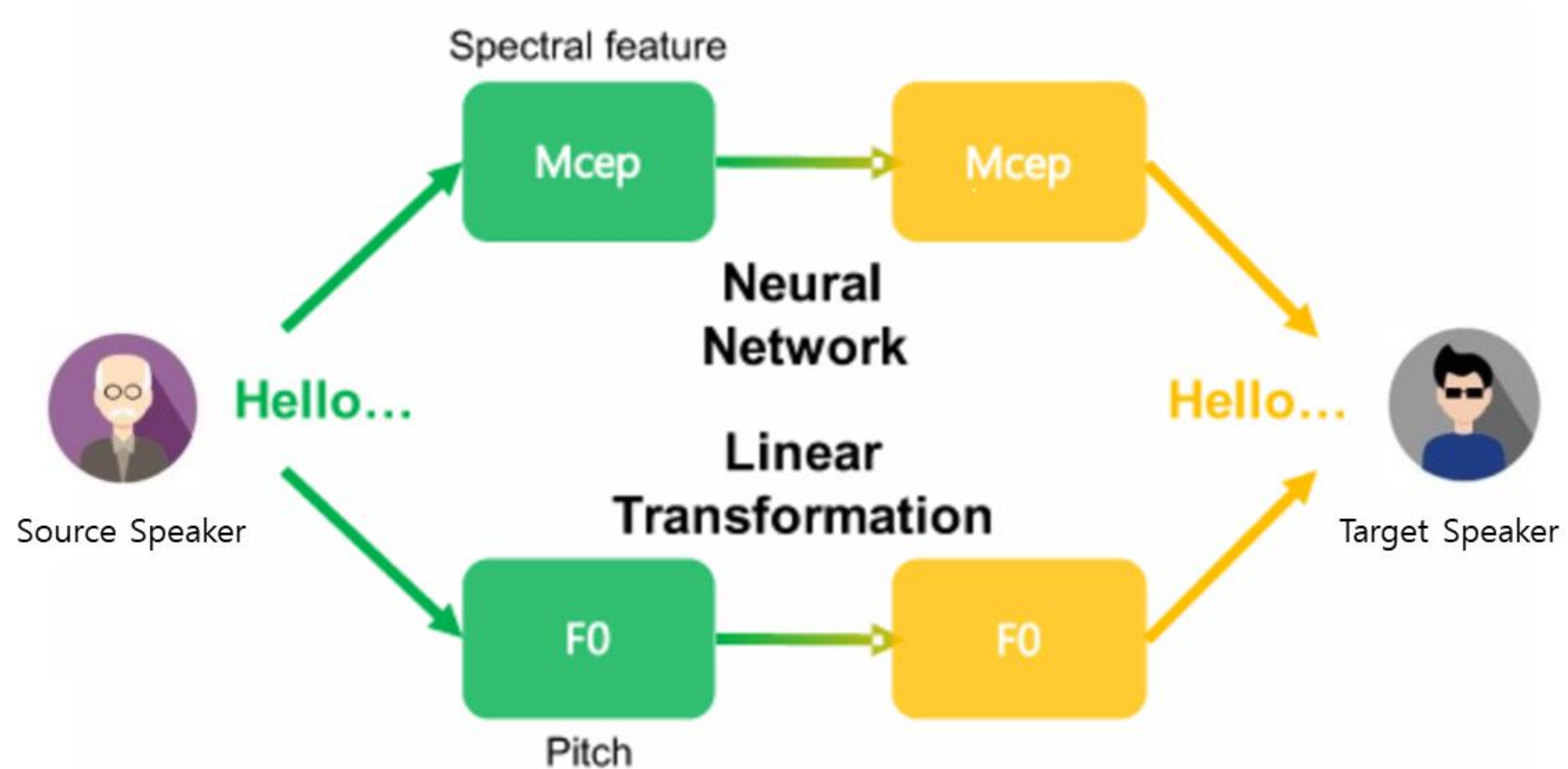
2019-11-16

Eun-Ju, Woo

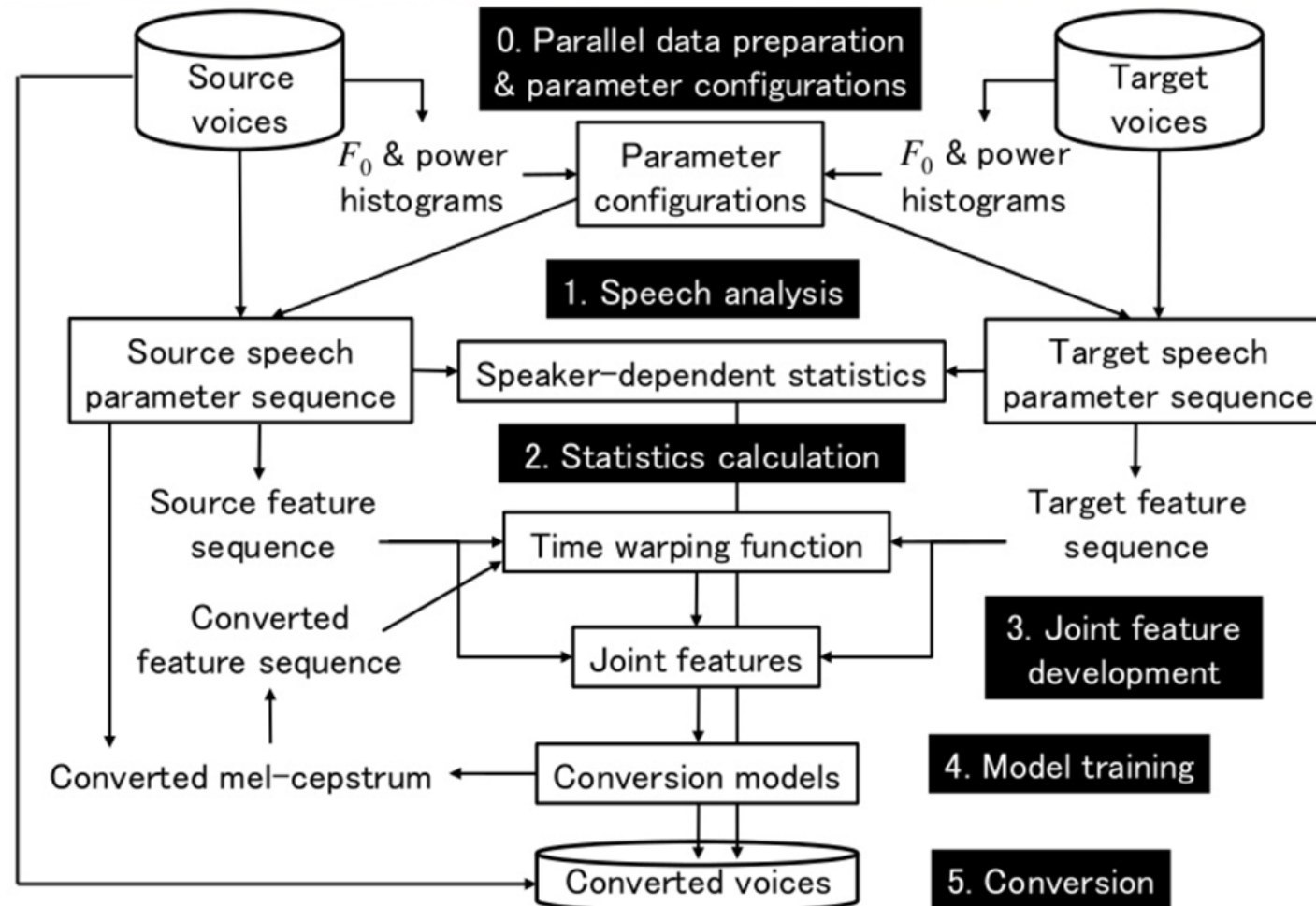
Objective of this study

- Korea = 'aged' society
- Almost no study regarding older people's speech data
- ASR(Automatic Speech Recognition) is not suitable for older people
- Speech data preprocessing study is valuable

Voice Conversion

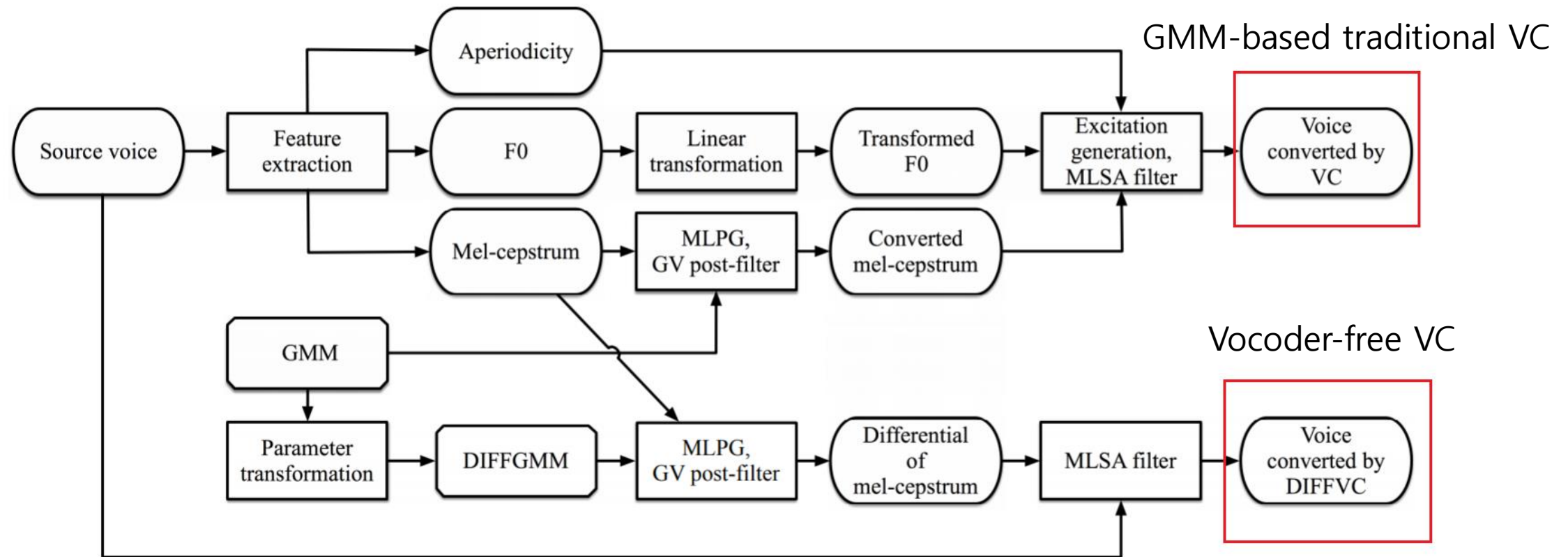


SPRocket 📌 Voice conversion system made by K. Kobayashi et al.



SPRocket

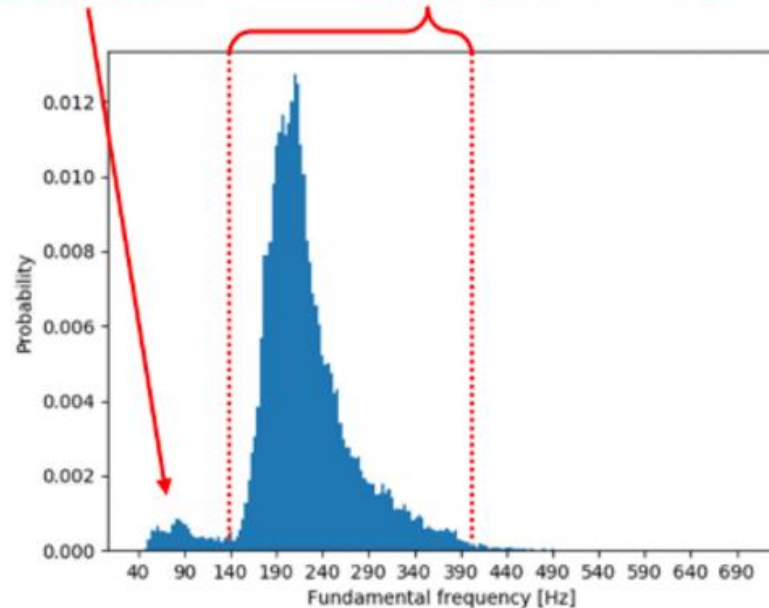
👉 Voice conversion system made by K. Kobayashi et al.



Parameter Settings $\leftarrow F_0$ Histogram

Supposed to be
half F_0 error

Proper F_0 search range
might be 140 – 400 Hz

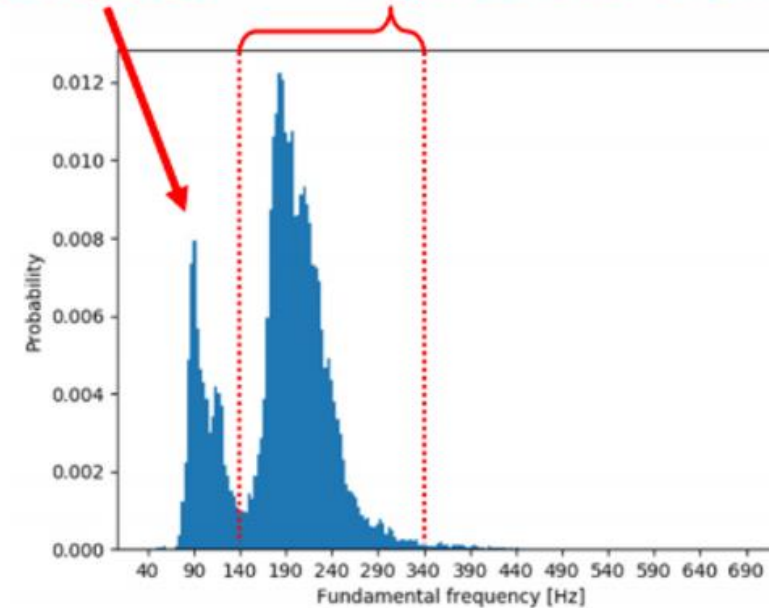


Source speaker: SF1

conf/figure/SF1_f0histogram.png

Supposed to be
half F_0 error

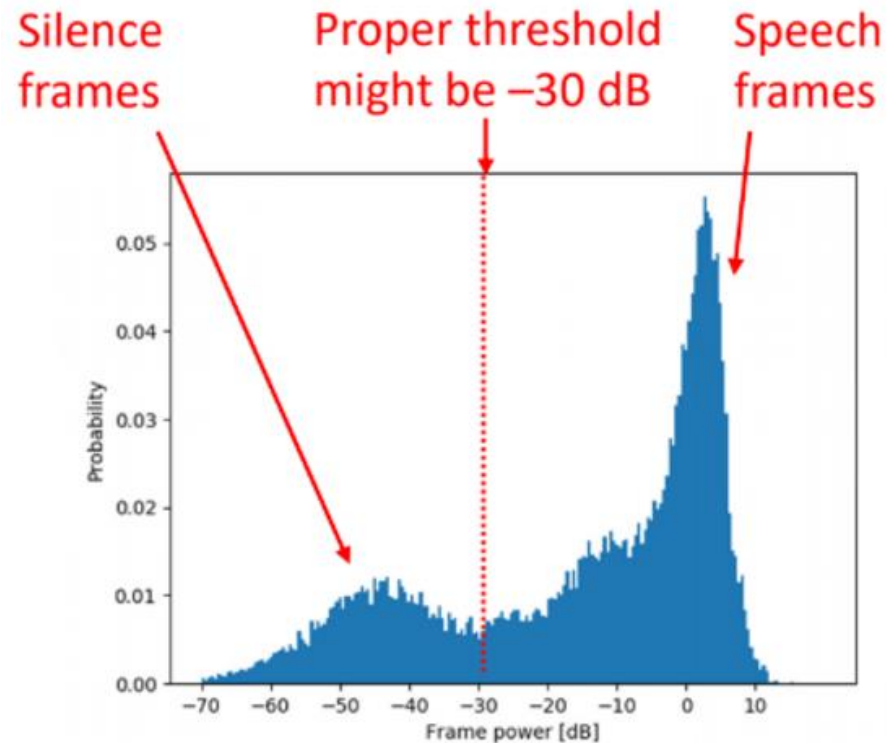
Proper F_0 search range
might be 140 – 340 Hz



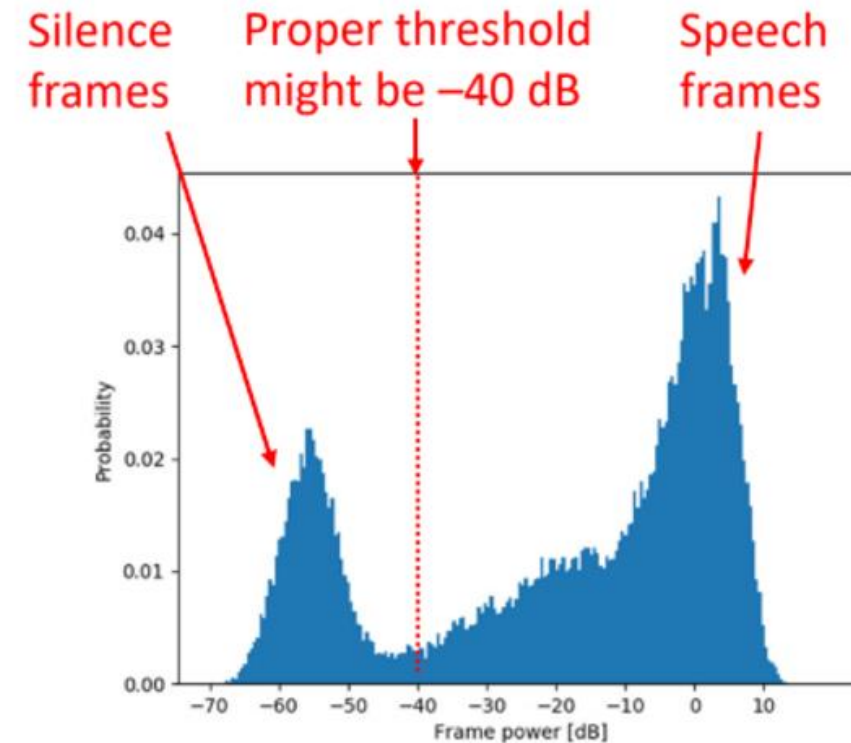
Target speaker: TF1

conf/figure/TF1_f0histogram.png

Parameter Settings ← Normalized Power Histogram



Source speaker: SF1



Target speaker: TF1

Modification of yaml files

```
YML IN_M_SS_80.yml x
1 wav:
2   fs: 16000
3   bit: 16
4   fftl: 1024
5   shiftms: 5
6   f0:
7     minf0: 140
8     maxf0: 480
9   mcep:
10    dim: 24
11    alpha: 0.410
12    power:
13      threshold: -38
14    analyzer: world
```

Source Speaker

```
YML UK_M_TS_80_wav.yml x
1 wav:
2   fs: 16000
3   bit: 16
4   fftl: 1024
5   shiftms: 5
6   f0:
7     minf0: 80
8     maxf0: 250
9   mcep:
10    dim: 24
11    alpha: 0.410
12    power:
13      threshold: -3
14    analyzer: world
15
```

Target Speaker

Experiments for strengthening elders' speech

Source Speaker			Target Speaker			
Sex	Speaker ID	Age	Sex	Speaker ID	Age	
M	MZ05	68	M	MV13	25	
				MW01	30	
	MZ09	71		MV13	25	
				MW01	30	
		F	FV01	23		
F	FZ06	65	F	FV01	23	
				FV13	29	
	FZ05	68		FV01	23	
				FV13	29	
		M	MW01	30		

Setting Parameters

Speaker	Minimum F_0 (Hz)	Maximum F_0 (Hz)	Power (dB)
MV13	80	190	-25
MW01	80	190	-20
MZ05	45	190	-20
MZ09	45	190	-25
FV01	140	340	-30
FV13	120	340	-15
FZ05	90	290	-25
FZ06	90	240	-20

Voice Conversion Result (Male)

Script: 아, 내 옷! 날개옷이 없어졌네.

Source Speaker

Target Speaker

Result

MZ09_71

MV13_25

MZ09-MV13_VC



MZ09-MV13_DIFF_VC



Voice Conversion Result (Female)

Script: 아, 내 옷! 날개옷이 없어졌네.

Source Speaker

Target Speaker

Result

FZ06_65



FV01_23



FZ06-FV01_VC



FZ06-FV01_DIFF_VC



No.	대본
41	아 내 옷 날개옷이 없어 졌네
42	내게 오셨으면 하늘로 올라갈 수 있는데 이를 어쩌나
43	선녀는 발을 동동 굴렀지만 날개옷을 찾을 수가 없었어요
44	애들 아 시간이 다 되었구나
45	어서 가자
46	올라 갈 시간이 되자 날개옷을 잃어 버린 손녀를 남겨두고 소녀들은 한 사람씩 하늘로 날아가 버려쎄요
47	내 곁을 찾거든 빨리 올라오너라
48	혼자 남은 선수는 너무 슬퍼서 흐느껴 울고 있었어요
49	여보세요 손님
50	옷을 잃어 버렸군요
51	추 우실 텐데 우선 이 옷이라도 입으세요
52	나무꾼은 미리 준비해 놓을 선수에게 주었어요
53	선녀는 나무꾼이 건네준(주는) 옷을 입고 하는 수 없이 나무꾼을 따라가서 살게 되었지요
54	그 (후) 예쁜 선녀는 나무꾼의 아내가 되었고 예쁜 소녀를 아내로 맞은 나무꾼은 더욱 열심히 (일)해 쎄요
55	선녀도 하늘 나라 일들은(을) 모두 잊어 버리고 나무꾼의 사랑을 듬뿍 받으며 행복하게 살았어요
56	한해 두해 세월이 지나는 사이에 선녀는 두 아이를 낳아 기르게 되었지요
57	어느 날 나무꾼은 선녀와 이런 이야기 저런 이야기를 하다가 그만 날개옷을 감춘 이야기까지 하고 말았어요
58	선녀는 날개옷 입고 싶어 건딜 수가 없었어요
59	여보 당신이 내 날개를 감 쳤다(추었다)고요
60	아 내 날 교사 한번만 입어 보았으면
61	나무꾼은 날개옷을 보고 싶은 선녀 가여워 감춰 두었던 날개옷을 꺼내 주었어요
62	아 내 나요
63	선녀는 날개옷을보다 너무나 기뻐서 날개옷을 입어 보았어요
64	그런데 이게 웬일이죠
65	눈 깜짝할 사이에서는 두 아이를 한판 판명 시간 꼭 하늘로 훨훨 날아가버리는 거였어요
66	여보여 보어딜 가는 거요
67	나무꾼이 울부짖으며 따라갔지만 아무 소용이 없었지요
68	선녀와 두 아이는 하늘 높이 까마득하게 날아가서 00 보이지 않았어요
69	나무꾼은 땅바닥에 털썩 주저앉자 한숨을 쉬며 사슴이 부탁한 말을 생각해 쎄요
70	아 사람이 한 말을 꼭 지켜야 하는 것을
71	사슴 이야기를 셋 낳을 때까지 날개옷을 내주지 말라고 한 뜯은 새 아이를 두 팔로는 알 수가 없었기 때문이었구나
72	아 이를 어쩌나
73	나무꾼은 가슴을 치며 후회하고 눈물을 흘렸지만 섬유와 두 아이들은 영영 나무꾼이 사는 나라로 돌아오지 않았다는 슬픈 옛날 이야기랍니다

WER (Word Error Rate)

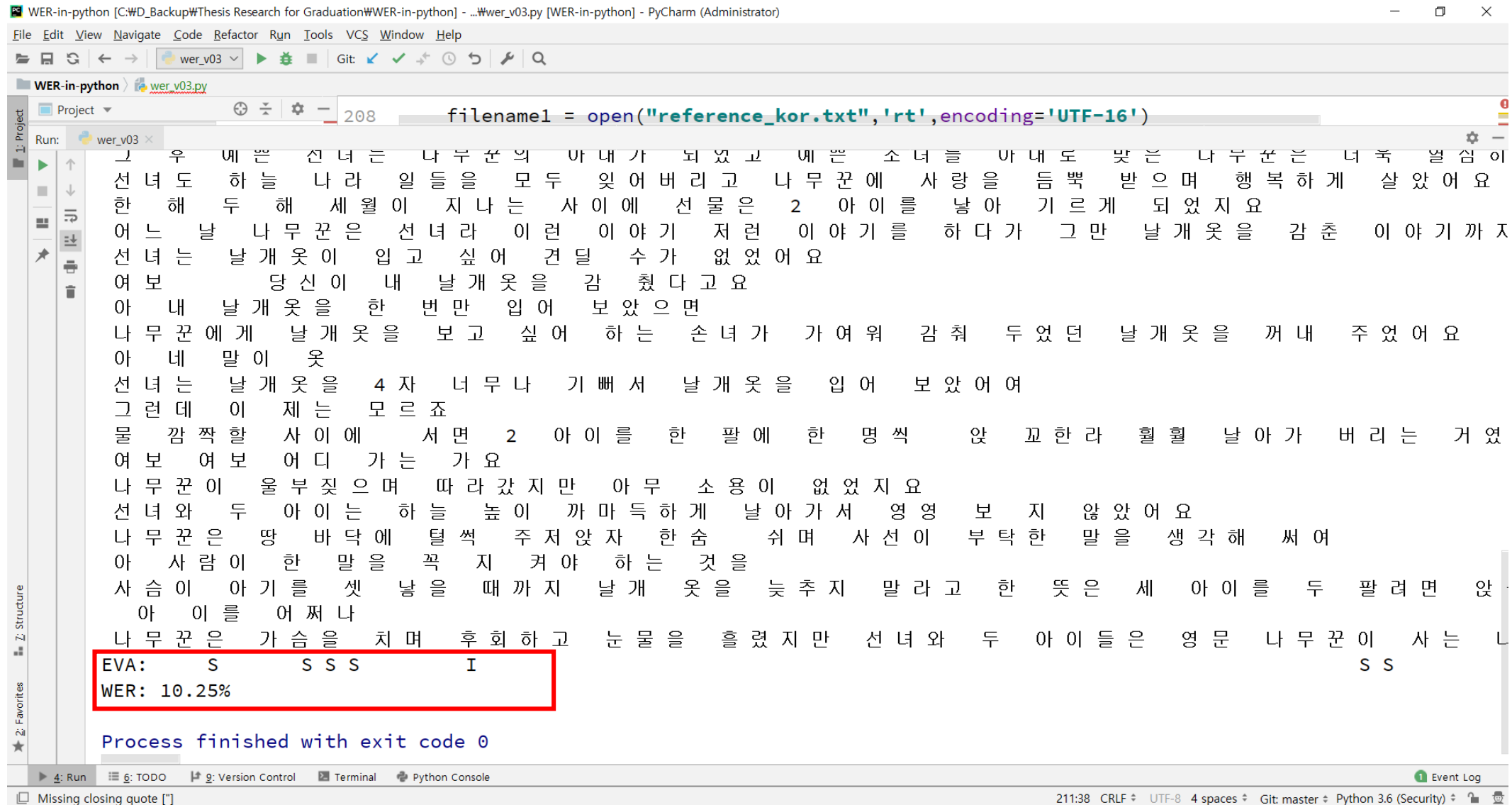
- ETRI_API (http://aiopen.etri.re.kr/guide_recognition.php)
- Speech Attribute: 16,000Hz, mono
- WER Equation

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference ($N=S+D+C$)

WER (Word Error Rate) – Python Programming



```
WER-in-python [C:\WD_Backup\Thesis Research for Graduation\WER-in-python] - ...wer_v03.py [WER-in-python] - PyCharm (Administrator)
File Edit View Navigate Code Refactor Run Tools VCS Window Help
wer_v03
WER-in-python wer_v03.py
Project 208 filename1 = open("reference_kor.txt",'rt',encoding='UTF-16')
Run: wer_v03 x
그 우 메 쓴 신너는 나무꾼의 아내가 되었고 메 쓴 소녀들 아내도 맞은 나무꾼은 너 죽 얼심 이
선 너 도 하늘 나라 일들을 모두 잊어 버리고 나무꾼 에 사랑을 듬뿍 받으며 행복 하게 살았 어요
한 해 두 해 세월이 지나 는 사이에 선물은 2 아이를 낳아 기르게 되었 지요
어느 날 나무꾼은 선너라 이런 이야기 저런 이야기를 하다가 그만 날개옷을 감춘 이야기 까지
선너는 날개옷이 입고 싶어 견딜 수가 없었 어요
여보 당신이 내 날개옷을 감춰 달라고요
아 내 날개옷을 한 번만 입어 보았 으면
나무꾼에게 날개옷을 보고 싶어 하는 손녀가 가여워 감춰 두었던 날개옷을 꺼내 주었 어요
아 네 말이 옷
선너는 날개옷을 4자 너무나 기빠서 날개옷을 입어 보았 어여
그런데 이 제는 모르 죠
물 감 짝 할 사이에 서면 2 아이를 한 팔에 한 명씩 앉 교한라 훨훨 날아가 버리는 거였
여보 여보 어디 가는 가요
나무꾼이 울부짖 으며 따라 갔 지만 아무 소용이 없었 지요
선너와 두 아이는 하늘 높이 까마득 하게 날아가서 영영 보 지 않았 어요
나무꾼은 땅 바닥에 털썩 주저앉 자 한숨 쉬며 사선이 부탁한 말을 생각해 써여
아 사람이 한 말을 꼭 지켜야 하는 것을
사슴이 아기를 셋 낳을 때 까지 날개옷을 훑추지 말라고 한 뜻은 세 아이를 두 팔려면 앓
아 이를 어찌 나
나무꾼은 가슴을 치며 후회 하고 눈물을 흘렸 지만 선너와 두 아이들은 영문 나무꾼이 사는 L
EVA: S S S S I S S
WER: 10.25%
Process finished with exit code 0
4: Run 6: TODO 9: Version Control Terminal Python Console 1 Event Log
Missing closing quote ["] 211:38 CRLF UTF-8 4 spaces Git: master Python 3.6 (Security)
```

WER (Word Error Rate) Results

Sex	Speaker_ID (Source Target)	RAW WER (%)	VC WER (%)	DIFF_VC WER (%)	Lower WER
Female	FZ05_68 FV01_23	7.47	9.1	11.3	VC
	FZ05_68 FV13_29	7.47	9.48	9.67	VC
	FZ05_68 MW01_30	7.47	8.91	8.24	DIFF_VC
	FZ06_65 FV01_23	10.06	14.94	16.48	VC
	FZ06_65 FV13_29	10.06	15.33	13.31	DIFF_VC
Male	MZ05_68 MV13_25	9.29	12.26	11.97	DIFF_VC
	MZ05_68 MW01_30	9.29	11.21	11.11	DIFF_VC
	MZ09_71 FV01_23	11.69	16.67	14.27	DIFF_VC
	MZ09_71 MV13_25	11.69	13.12	12.74	DIFF_VC
	MZ09_71 MW01_30	11.69	16.57	15.52	DIFF_VC

Text Analysis

MZ05_68_RAW.txt	MZ05_68-MV13_25_DIFF_VC.txt
안에 옷 날개옷이 없어 졌네	아 내 옷 날개옷이 없어 졌네
날개옷이 없으면 하늘로 올라갈 수 없는데 이를 어쩌나	날개옷이었으면 하늘로 올라갈 수 없는데 이를 어쩌나
소년은 발을 동동 굴렀지만 날개옷을 찾을 수가 없었어요	선녀는 다른 동동 그렇지만 날개옷을 찾을 수가 없었어요
얘들 아 시간이 다 되었구나	얘들 아 시간이 다 되었구나
어서 가자	어서 가자
올라 갈 시간이 되자 날개옷을 잃어 버린 손녀를 남겨두고 손녀들은 한 사람씩 하늘로 날아가 버려쎄요	올라 갈 시간이 되자 날개옷을 잃어 버린 선녀를 남겨두고 선녀들은 한 사람씩 하늘로 날아가 버려쎄요
날개옷을 찾거든 빨리 올라오너라	날이었을 찾거든 빨리 올라오너라
혼자 남은 소년은 너무 슬퍼서 흐느껴 울고 있었어요	혼자 남은 3년은 너무 슬퍼서 흐미 즐고 있었어요
여보세요 소년님	여보세요 선녀님
무엇을 잃어 버렸군요	무엇을 잃어 버렸군요

ETRI speech recognition API's limitation

- Word Error Rate is almost 10% although speaker's age is 20's
- Result is not the same whenever speech recognition experiment performed
- WER cannot be the absolute standard value for evaluating voice conversion

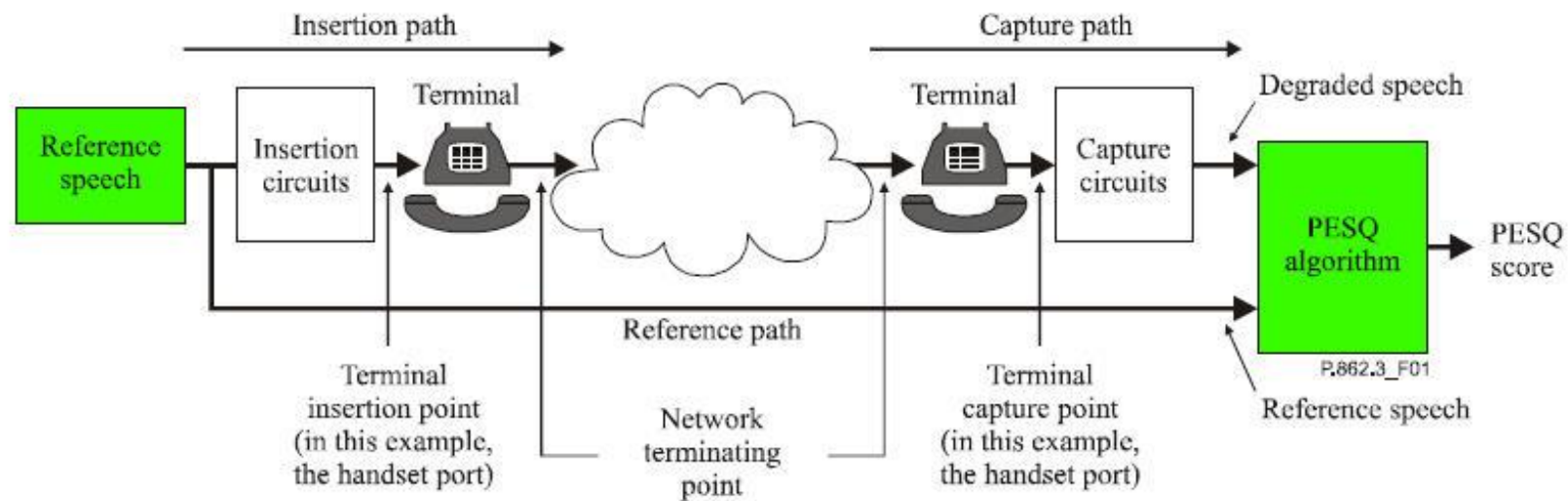
👉 Just reference data

Speaker_ID	WER
mv_13_25	9.87%
fv_01_23	10.54%

Better Choice for Reducing WER

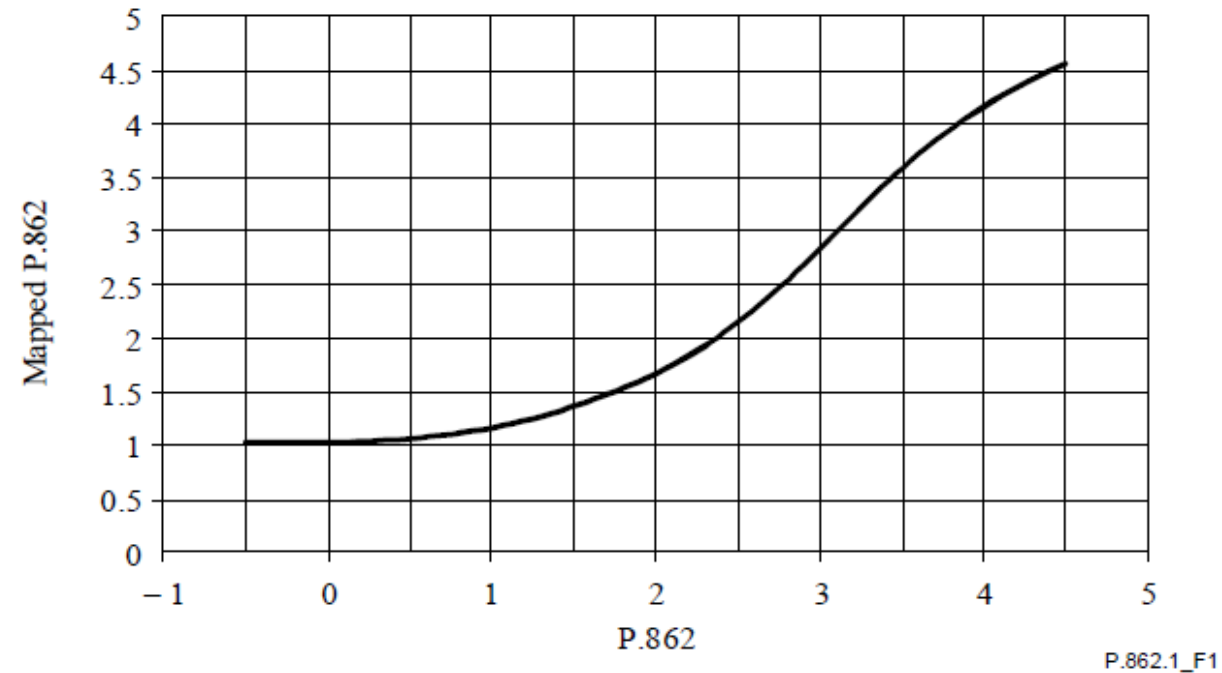
- For Male: DIFF_VC
- For Female: Overall VC (VC: 3, DIFF_VC: 2)
- Intra-Gender Voice Conversion: DIFF_VC

PESQ (ITU-T P.862)



- Evaluation speech quality with comparing between original signals and reducing signals through communication systems
- For using this value, reference data is essential

PESQ (ITU-T P.862) & MOS $1.02 \leq \text{MOS-LQO} \leq 4.56$



- PESQ cannot be used simultaneously with MOS because of PESQ's range
- Mapping function from ITU-T P.862.1's PESQ to MOS-LQO (Listening Quality Objective)

PESQ & MOS Result

Sex	Speaker_ID (Ref. Eval.)	Mean PESQ	Mean_MOS	Better VC
Female	FZ05_68 FV01_23_DIFF_VC	1.385	1.281	DIFF_VC
	FZ05_68 FV01_23_VC	1.045	1.172	
	FZ05_68 FV13_29_DIFF_VC	1.727	1.455	DIFF_VC
	FZ05_68 FV13_29_VC	1.059	1.175	
	FZ05_68 MW01_30_DIFF_VC	1.389	1.282	DIFF_VC
	FZ05_68 MW01_30_VC	1.036	1.169	
	FZ06_65 FV01_23_DIFF_VC	1.538	1.350	DIFF_VC
	FZ06_65 FV01_23_VC	1.051	1.173	
	FZ06_65 FV13_29_DIFF_VC	1.752	1.472	DIFF_VC
	FZ06_65 FV13_29_VC	1.051	1.173	

PESQ & MOS Result

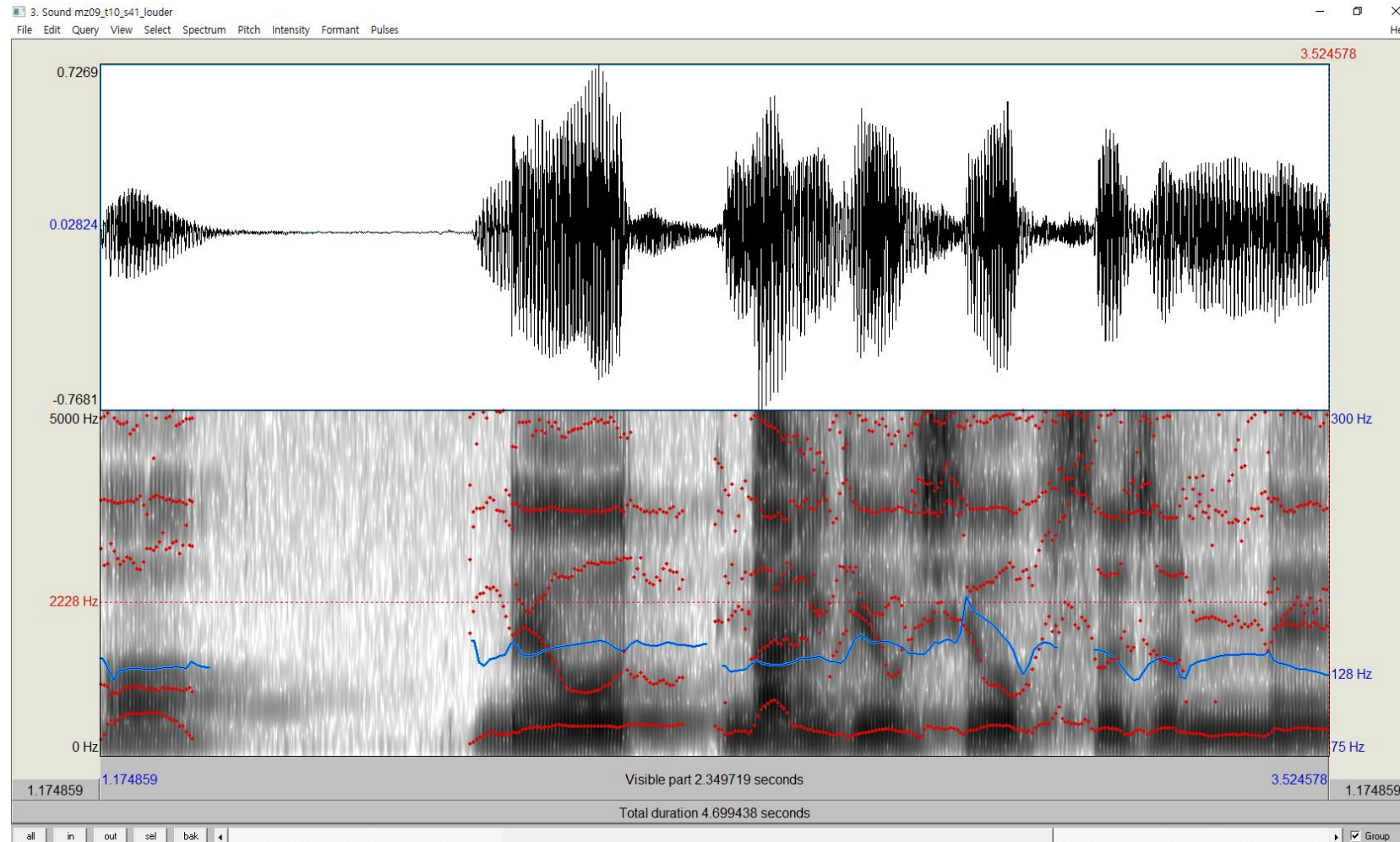
Sex	Speaker_ID (Ref. Eval.)	Mean PESQ	Mean_MOS	Better VC
Male	MZ05_68 MV13_25_DIFF_VC	1.355	1.270	DIFF_VC
	MZ05_68 MV13_25_VC	1.239	1.227	
	MZ05_68 MW01_30_DIFF_VC	1.705	1.441	DIFF_VC
	MZ05_68 MW01_30_VC	1.282	1.242	
	MZ09_71 FV01_23_DIFF_VC	1.238	1.227	DIFF_VC
	MZ09_71 FV01_23_VC	1.047	1.172	
	MZ09_71 MV13_25_DIFF_VC	1.423	1.296	DIFF_VC
	MZ09_71 MV13_25_VC	1.157	1.202	
	MZ09_71 MW01_30_DIFF_VC	1.457	1.312	DIFF_VC
	MZ09_71 MW01_30_VC	1.085	1.182	

PESQ & MOS Result Analysis

- For Male & Female: DIFF_VC has better results
- Vocoder-free voice conversion's Mean PESQ and mean MOS-LQO value is higher than conventional voice conversion

Waveform in PRAAT (MZ09_71_RAW)

— : pitch
● : formant



Waveform in PRAAT (MZ09_71_RAW)

Pitch: Draw

Time range (s): 0.0 0.0 (= all)

Frequency range (Hz): 75.0 300.0

☒ Garnish

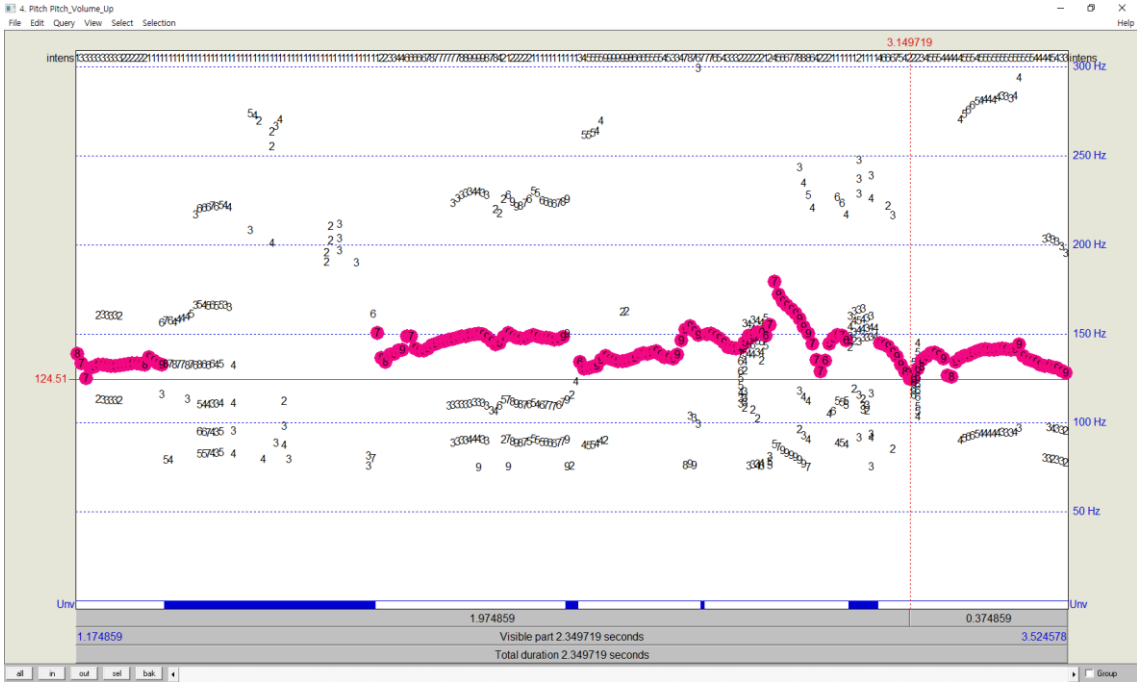
Help

Standards

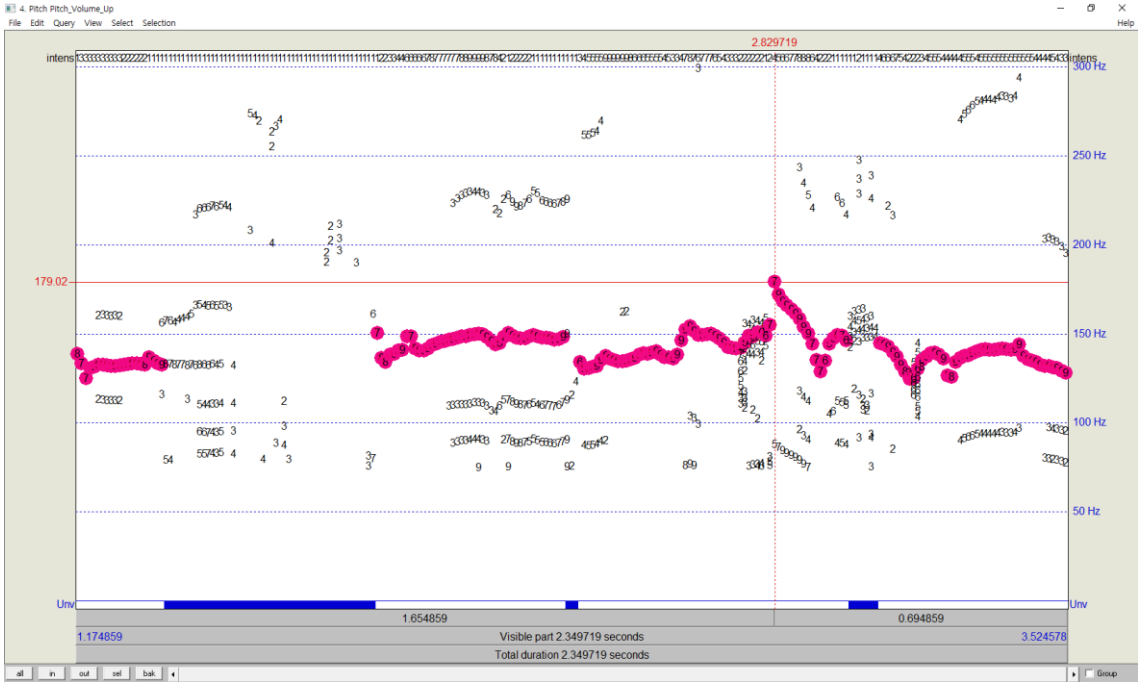
Cancel

Apply

OK

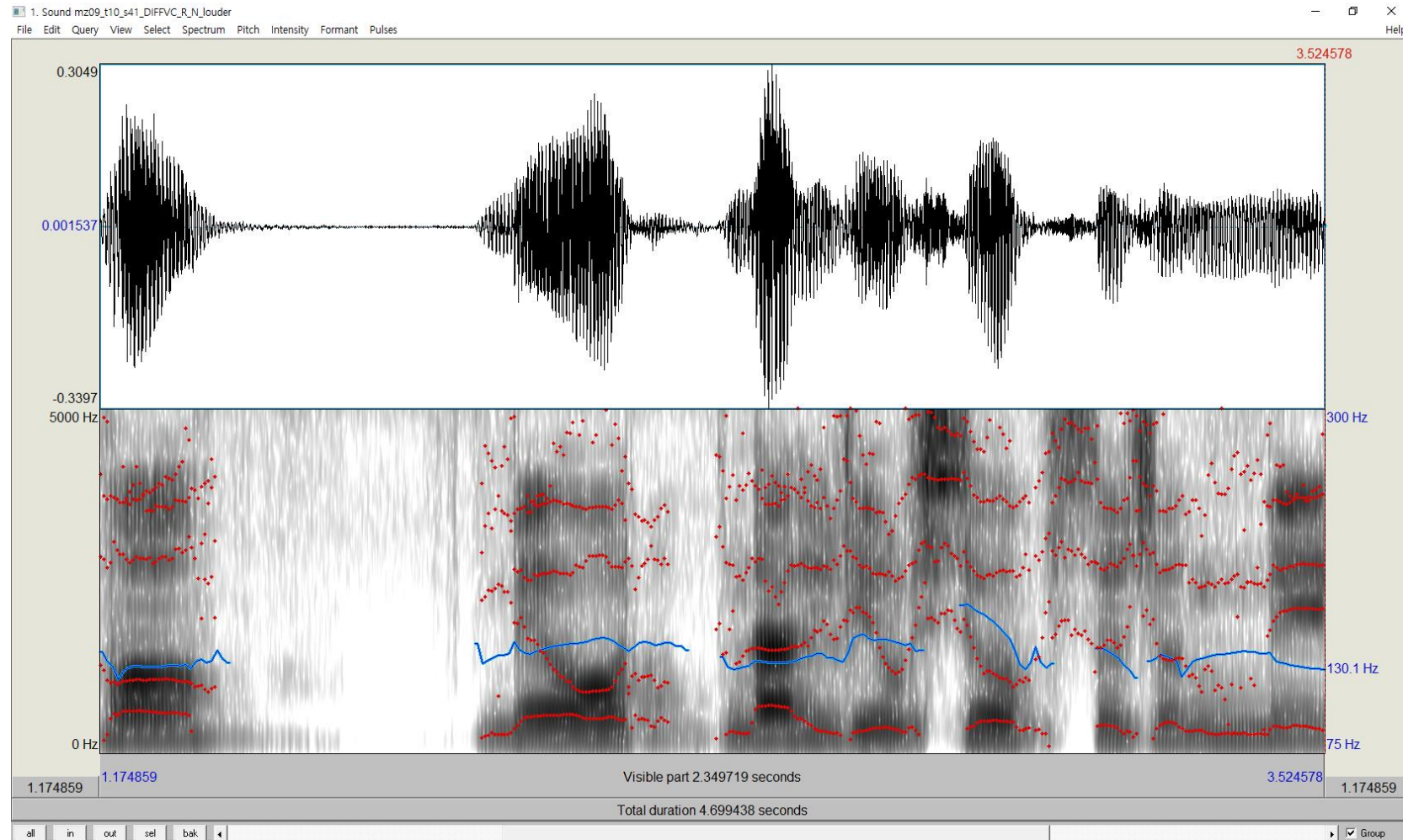


Min F₀: 124.14 Hz



Max F₀: 180.18 Hz

Waveform in PRAAT (MZ09_71-MV13_25)



— : pitch
● : formant

Waveform in PRAAT (MZ09_71-MV13_25)

Pitch: Draw

Time range (s):

0.0

0.0 (= all)

Frequency range (Hz):

75.0

300.0

☒ Garnish

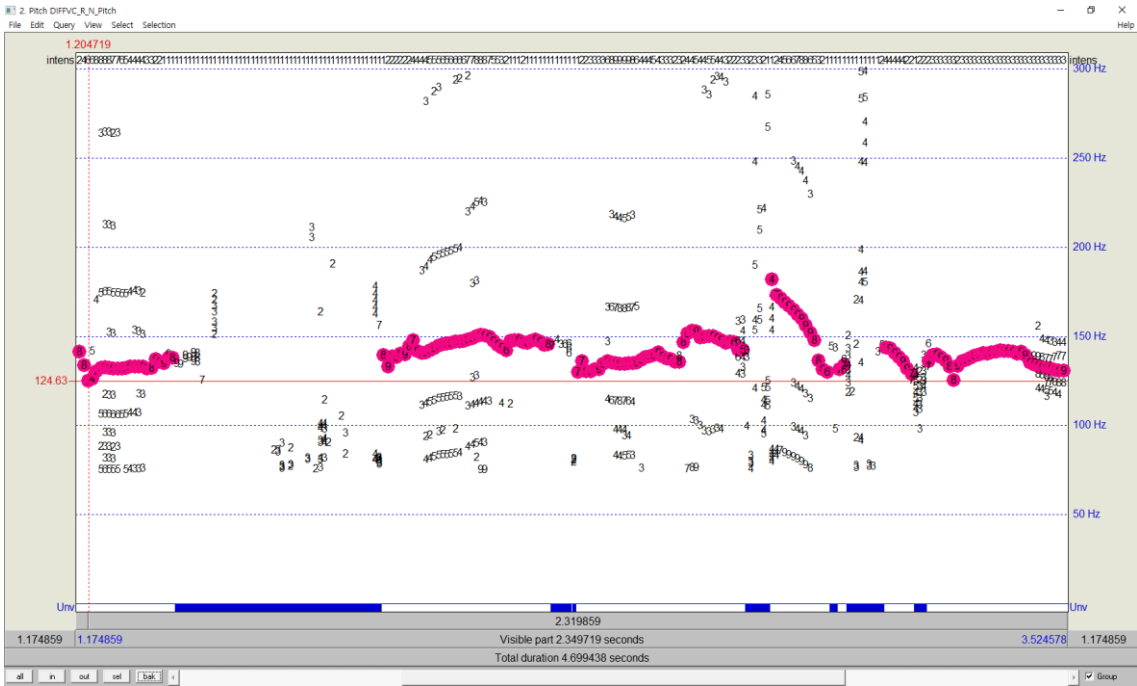
Help

Standards

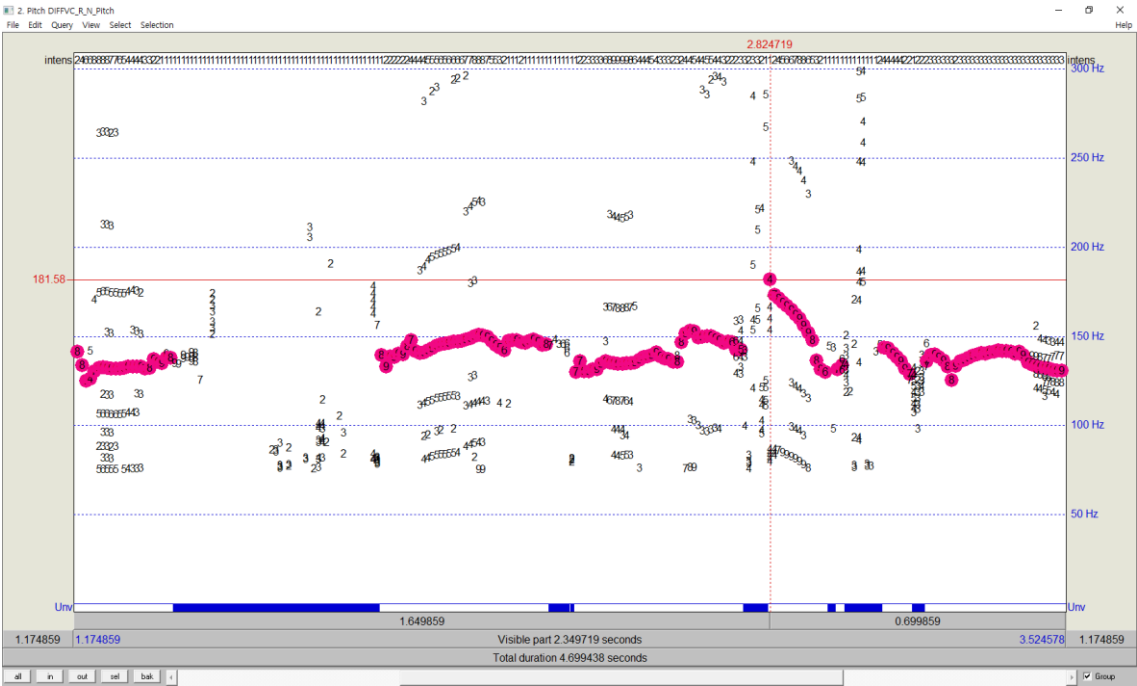
Cancel

Apply

OK



Min F₀: 123.51 Hz



Max F₀: 172.57 Hz

Conclusions

- Normalization, volume up and high / low pass filter cannot be the perfect solution for older people's speech data processing
- Not only mimicking target speaker's style results, Voice conversion can be the mean of data preprocessing
- Vocoder-free voice conversion's data is better according to the WER , PESQ, MOS-LQO calculation



Thank You!