# Older People's Korean Speech Data Processing with Voice Conversion

Eun Ju Woo*, Byeong Rae Lee**
*Korea National Open University
* pawoo2645@knou.ac.kr, **brlee@knou.ac.kr

**ABSTRACT**

Although Korea got to an aging society, older people's speech data is not familiar research theme. This paper is the result of voice conversion by using the older people's Korean speech data made by National Institute of Korean Language. The main system for the experiments for this paper is 'sprocket', open-source software for voice conversion, and with setting target data with parallel speech data from young people, we discovered that voice conversion can help enhancing the quality of speech data.

## 1. Introduction

For the human machine interaction in all industrial fields, speech is convenient, natural and intuitive method for a long time. Especially. speech interface is essential for the older and patients who are weak and cannot have normal life like young and healthy people.

Korea is an 'aged' society from a few years ago [1]. Compared to other countries, speech processing research according to speaker's age is not performed widely today in Korea. All the people experience this situation: As a person gets older, his voice changed a lot compared to their young age. Compared to under 40's, There are some older voice characteristics.

At senescence, their tongue's movement range, duration and thickness has diminished. Due to this point, older people's speech speed is low, silent sections increase, and their speech is hard to be recognized automatically [2].

For improving old people's speech quality, in this paper, we apply voice conversion technology to enhance for accurate pronunciation, loudness and timbre for listening clearly.

## 2. Experiments

Voice conversion (VC) is an area of speech processing that deals with the conversion of the perceived speaker identity [3]. For this technique, there needs two speakers: the speech signal uttered by a first speaker, the source speaker, is modified to sound as if it was spoken by a second speaker, referred to as the target speaker. In this paper, sprocket [4], open-source voice conversion software developed by Nagoya University's researchers is the main tool for this voice conversion experiments. This system was developed by the concept of statistical voice conversion and the result through sprocket is speech data whose voice such as target speaker. Table 1. shows the Korean corpus used in the experiments.

### 2.1. Experimental Data Set

For this paper, we used Korean Seoul dialect Speech Corpus produced by National Institute of Korean Language in 2002 and unveiled in 2005 [5]. This corpus is parallel corpus, and the main contents are Korean children's stories and novels. We selected this corpus because this was arranged well by speaker's age and parallel corpus. For using sprocket, we had to use parallel corpus. Quality of converted speech is not great by using non-parallel corpus in the level of current technology.

We selected each 2 persons (male 2, female 2) per gender they are the 1st and 2nd oldest among the male and female speakers. They are the source speakers. As we listened to all the 20's and 30's speakers' speech, with considering their voice and pronunciation, we selected each one target speaker as Table 1. shows.

Table 1. Speech Data for Voice Conversion [6]

| 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 |

We selected 133 utterances for each speaker. 100 utterances (70%) were used for training, and 33 utterances (30%) were used for evaluation. The fact that men's and women's voice quite different, we performed cross gender voice conversion one time for each gender.

### 2.2. Parameters for experimental settings

To implement each experiment exactly, we must set F0 (Fundamental Frequency) range and proper threshold

according to each speech data's property. We set min and max F0 and proper threshold(dB) for each speaker following the feature extraction graph made by sprocket's functions. Table 2. shows each experiment parameter settings.

Table 2. Parameter Settings

| Speaker | Minimum F0 (Hz) | Maximum F0 (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 |

As you can see, women's voice has wide $F_0$ range and their speech data has higher frequencies compared to the men's.

### 3. Experimental Result and Analysis

Voice conversion system, sprocket was developed focused on timbre change of speech, so strengthening pitch, loudness and pronunciation correction is not perfect regarding older people's voice. By using this system, we could enhance speech data's pronunciation, loudness, clarity and so on according to the experimental result. Figure 1. is the original speech wave and spectrogram of the speaker, 'MZ09'. This waveform and spectrogram can be seen by the software, PRATT. (script: 나무꾼은 가슴을 치며 후회하고, 눈물을 흘렸지만)

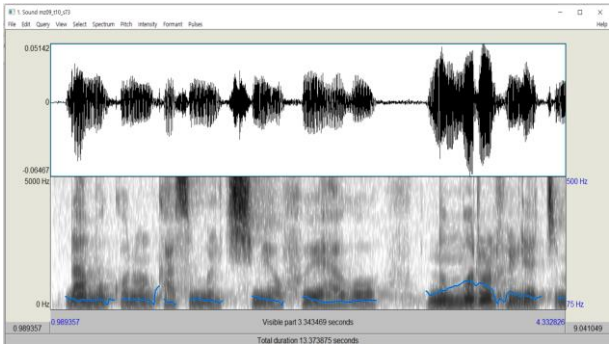

Figure 1. Original waveform and spectrogram of 'MV09'

Figure 2. shows the converted speech data corresponding to Figure 1. Compared to Figure 1, spectrogram gets clearer and the sound of speech is evident compared to original sound. Waveform pitch was changed after voice conversion.

After 10 times experiments of older people's voice conversion, we got to know that although this technique was developed general timbre change, it can also have an impact on changing old people's speech clearly.

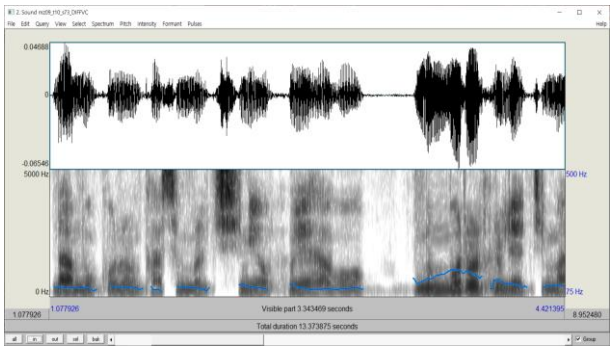

Figure 2. converted waveform and spectrogram of 'MZ09'

Each voice conversion results are normal VC and DIFFVC. Overall, the quality of speech converted by DIFFVC (Vocoder-Free voice conversion) is better than normal voice conversion in sprocket.

### 4. Further Study

In sprocket, there is no function for dealing with non-parallel corpus, and there is no training step with state-of-the-art deep learning speech processing algorithms. The voice conversion system for non-parallel speech data will be used widely if it will be developed, so our next research is for that system. Furthermore, objective standards for evaluating voice conversion result needs to be developed for the better evaluation of voice conversion system performance.

### References

[1] Korea is now an 'aged' society http://koreajoongangdaily.joins.com/news/article/article.aspx?aid=3052445

[2] Seoungjun Lee, Soonil Kwon. (2014) Elderly Speech Analysis for Improving Elderly Speech Recognition, *Communications of the Korean Institute of Information Scientists and Engineers 32(11)*, 16-20.

[3] Nurminen, J., Silén, H., Popa, V., Helander, E., & Gabbouj, M. (2012). 0 Voice Conversion.

[4] K. Kobayashi & T. Toda. (2018). sprocket: open-source voice conversion software. *Proc. Odyssey 2018*, 203-210.

[5] Yoon, Tae-Jin. & Kang, Yoonjung. (2014). Monophthong Analysis on a Large-scale Speech Corpus of Read-Style Korean. *Phonetics and speech sciences v.6 no.3,* 139 – 14

[6]https://ithub.korean.go.kr/user/total/referenceManager.do