Deeply Read Korean Speech Corpus

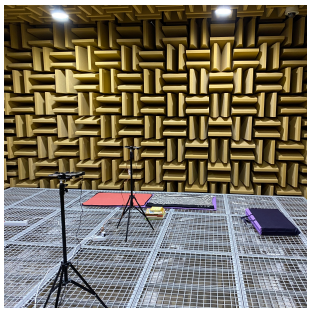
Summary

Pairs of Korean speakers reading a script with ***3 distinct text sentiments (negative, neutral, positive)***, with ***3 distinct voice sentiments (negative, neutral, positive)***, are recorded. The recordings took place in 3 different types of places, which are an ***anechoic chamber***, ***studio apartment***, and ***dance studio***, of which the level of reverberation differs. And in order to examine the effect of the distance of mic from the source and device, every experiment is recorded at ***3 distinct distances*** with ***2 types of smartphone***, iPhone X, and Galaxy S7.

Recording contents	A pair of adults reading scripts containing 3 distinct text sentiments(negative, neutral, positive) with 3 distinct voice sentiments(negative, neutral, positive). (Script: movie reviews(positive, negative), everyday conversation(neutral))
Recording environment	Anechoic Chamber (no reverb), Studio apartment (moderate reverb), Dance studio (high reverb)
Device	iPhone X (iOS), Samsung Galaxy S7 (Android)
Distance from the source	0.4m, 2.0m, 4.0m
Volume	~ 290 hours, ~ 190,000 utterances, ~ 107 GB
Format	wav(44100Hz, 16-bit, mono), or h5(16000Hz, 16-bit, mono)
Language	Korean
Demographics	34 Korean adults, with 26% males and 74% females, and 47% are in 20s, 20.5% in 30s, 17.5% in 40s, 6% in 50s, 9% in 60s.

<Fig1. Studio apartment>

<Fig 2. Dance studio>



<Fig 3. Anechoic chamber>

What's inside the Deeply Korean Read Speech Corpus?

The Read Speech dataset consists of ***289.9 hours*** of audio clips of reading the scripts with ***3* text sentiments** with ***3 voice sentiments*** recorded at ***3 distinct places*** using ***2 different smartphones*** running under different operating systems. The participants are encouraged to record repetitively in all ***3 types of place (anechoic chamber, studio apartment, dance studio)***, and every recording is conducted systematically at ***3 ordinal distances(0.4m, 2.0m, 4.0m)*** with ***2 types of device(iPhone X and Galaxy S7)***.

'Negative text sentiment', ***'neutral text sentiment'*** , ***'positive text sentiment'*** indicate that the contents being vocalized are negative, neutral, and positive respectively. Specifically, for the negative and positive text sentiments, negative/positive movie reviews, containing degradations, criticisms or compliments, were used. And, for the neutral text sentiment, everyday conversations without typical emotions were used.

'Negative voice sentiment' indicates that the speaker vocalized the script with a negative tone of voice, for the sake of consistency, we instructed the speakers to vocalize as if they were angry. ***'Neutral voice sentiment'*** indicates that the speaker vocalized the script with a neutral tone of voice, with any emotions involved. Finally, ***'positive voice sentiment'***

indicates that the speakers vocalized the script with a positive tone of voice, especially as if they were happy. Each type of voice sentiment was vocalized regardless of the content of the script (text sentiment), for example, the speakers were also asked to vocalize the script positively even though the content was negative.

The dataset also includes metadata such as a script(speech-to-text aligned), speaker, age, sex, noise, type of place, distance, and device. The type of text sentiments and voice sentiments is categorized as follow:

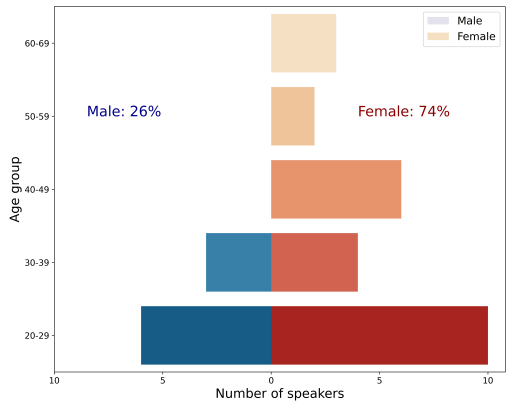
Information & Statistics

Figures 4 demonstrates the demographic information of the speaker. The participants comprise 26% of males and 74% of females, and 67%of males and 40% of females are in their 20s.

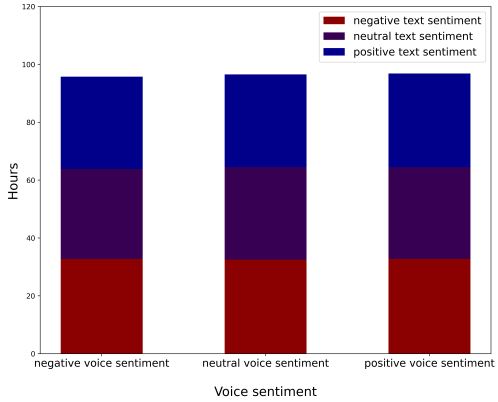
Figure 5 illustrates the total length of the utterances by text and voice sentiment. The bars displayed horizontally show that the total length of the utterances with distinct voice sentiment types is almost equal to one another. Also, by looking at the bars stacked vertically, you could observe that the length of the utterances with distinct text sentiment types within the same voice sentiment is almost identical to one another.

Figure 6 illustrates the average length per utterance by text and voice sentiment. Since the scripts of the negative and positive contents were the reviews of movies, they were relatively longer than those of the neutral text scripts.

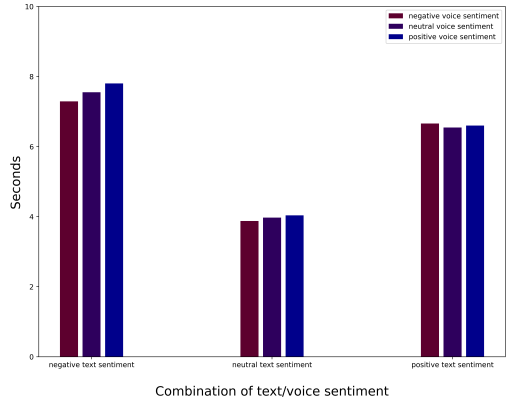
Lastly, Figure 7 illustrates the distribution of the length of the utterances by text and voice sentiment. It shows that, while reading the scripts with the same text sentiment, it doesn't seem to change the distribution of the length of the utterances even though you change your voice sentiment. And that every distribution is positively skewed.



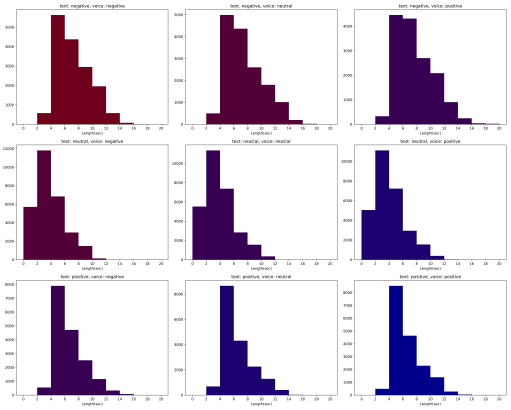
<Fig 4. Age distribution of the speakers by sex>



<Fig 5. Total length(hours) by text, voice sentiment>



<Fig 6. Average length(seconds) by text, voice sentiment



<Fig 7. Length distribution by text, voice sentiment>

Filename convention
Recorded wav files are named under following format:
{subject ID}_{yyyy}_{mm}_{dd}_{sex_a}{sex_b}_{age_a}{age_b}_{location}
_{distance}_{device}_{voice_sentiment}.wav
Example: sub2001_2020_11_29_00_22_0_0_0_-1.wav

Subject ID is a unique 4-digit alphanumeric code representing speaker pair, ***{yyyy}_{mm}_{dd}*** is a date of recording, ***sex_a*** is a digitized code indicating the sex of speaker_a(parent), ***sex_b*** is a digitized code indicating the sex of speaker_b(child), ***age_a*** is indicating the first digit of the age group of speaker_a(parent), ***age_b*** is indicating the age of speaker_b(child), ***location*** is a digitized code indicating where the recording took place, ***distance*** indicates the distance at which the recording was taken place from the source, and ***device*** is a digitized code indicating the device which was used to record.

How to decode?

Class
Text sentiment
-1: 'negative'
0: 'neutral'
1: 'positive'

Voice sentiment
-1: 'negative'
0: 'neutral'
1: 'positive'

Speaker
a: speaker a
b: speaker b

Age
First digit of the age (real age in h5 attributes, metadata.json)

Sex
{0: 'Female', 1: 'Male'}

Location
{0: 'Studio apartment', 1: 'Dance studio', 2: 'Anechoic Chamber'}

Distance
{0: 0.4 m, 1: 2.0 m, 2: 4.0m}

Device
{0: iPhone, 1: Samsung Galaxy S7}

Noise
{0: 'Noiseless', 1: 'Indoor noise', 2: 'Outdoor noise', 3: 'Both indoor and outdoor noise'}

License

Contact & Purchase

contact@deeplyinc.com