Phonetically Balanced Text Corpus Design Using a Similarity Measure for a Stereo Super-Wideband Speech Database

Yoo Rhee OH†*, Yong Guk KIM†**, Mina KIM†***, Nonmembers, Hong Kook KIM†(†), Member, Mi Suk LEE††, and Hyun Joo BAE†††, Nonmembers

SUMMARY In this paper, we propose a text corpus design method for a Korean stereo super-wideband speech database. Since a small-sized text corpus for speech coding is generally required for speech coding, the corpus should be designed to comply with the pronunciation behavior of natural conversation in order to ensure efficient speech quality tests. To this end, the proposed design method utilizes a similarity measure between the phoneme distribution occurring from natural conversation and that from the designed text corpus. In order to achieve this goal, we first collect and refine text data from textbooks and websites. Next, a corpus is designed from the refined text data based on the similarity measure to compare phoneme distributions. We then construct a Korean stereo super-wideband speech (K-SW) database using the designed text corpus, where the recording environment is set to meet the conditions defined by ITU-T. Finally, the subjective quality of the K-SW database is evaluated using an ITU-T super-wideband codec in order to demonstrate that the K-SW database is useful for developing and evaluating super-wideband codecs.

**Key words:** text corpus design, similarity measure, phonetically balanced text, super-wideband speech database

1. Introduction

A text corpus is a bundle of text data that can be used as a script for recording speech utterances in speech-related research such as that for automatic speech recognition (ASR), text-to-speech (TTS), and speech coding. To date, there have been several text corpus design methods reported for speech recognition or speech synthesis [1]–[3]. In general, the pronunciation behavior of natural conversation features such as a phoneme structure and distribution can be obtained from a significantly large amount of randomly collected text data [3]. However, an optimal-sized or a minimal-sized text corpus is desired in order to reduce the recording or transcribing efforts for such speech utterances. As a typical scenario for constructing a text corpus, a large amount of text data is first collected from various sources including websites, newspapers, and books. Next, the text corpus is optimized or minimized from the collected texts by using a greedy algorithm based on a predefined criterion [1]–[3], which results in either a phonetically balanced or phonetically rich corpus. In other words, a phonetically balanced text corpus is constructed by selecting text data whose phoneme distribution is similar to that occurring from natural conversation, in an attempt to closely reflect the pronunciation behavior. On the other hand, a phonetically rich text corpus has sentences in which all the phonemes are distributed uniformly. These corpora are then used to ensure that all acoustic models of an ASR or TTS system are well trained.

Unlike the development of a text corpus for speech recognition or speech synthesis, the design methods described above are not appropriate to a database for speech coding. That is, the size of the text corpus for speech coding is relatively small in order to efficiently perform subjective quality tests, compared to the database for speech recognition or speech synthesis. Moreover, the text corpus for speech coding should fully reflect common pronunciation behaviors while working with only a small amount of text data. Although there have been few works on text corpus design for speech coding reported, the NTT Advanced Technology Corporation (NTT-AT) database has been popularly used for developing or evaluating narrow-band speech codecs [4]. Notably, it contains Korean speech utterances; however, the Korean text corpus lacks some pronunciation behaviors as it misses some Korean phonemes. This suggests that there should be a more efficient method of selecting text data that represent the phoneme distribution of a language. As another speech corpus for speech coding, a super-wideband speech database distributed by NTT-AT exists [5], which is referred to here as NTT-SW. Unfortunately, this database does not support Korean.

Therefore, in this paper, we propose a text corpus design method for speech coding. The proposed method is based on a similarity measure that calculates how close the phoneme distribution occurring from natural conversation is to that of the designed text corpus. To this end, text data containing predefined phonetically balanced words are first collected from various materials including websites, newspapers, and books. From the refined text data, a text corpus is
then designed by applying the similarity measure to generate a text corpus. This method is subsequently applied to design a Korean text corpus for a Korean stereo super-wideband speech (K-SW) database, where the recording environment is arranged to meet the conditions defined by ITU-T Recommendation P.800 [6]. Next, the designed corpus is compared with the Korean text corpus of the NTT-AT database [4]. Finally, the subjective quality of the K-SW database is evaluated using an ITU-T super-wideband speech codec, ITU-T Recommendation G.722.1.C [7], in order to ensure the appropriateness of the proposed corpus design method.

The remainder of this paper is as follows. Following this introduction, we propose a phonetically balanced text corpus design method for speech coding in Sect. 2. In Sect. 3, we describe the requirements for a text corpus for the K-SW database and also review an existing super-wideband speech database, the NTT-SW database [5]. In addition, we describe how to construct the K-SW database by applying the proposed text corpus design method. In Sect. 4, we evaluate the performance of subjective quality tests on the designed Korean text corpus and the K-SW database. Finally, we conclude this paper in Sect. 5.

2. Proposed Text Corpus Design for Speech Coding

In this section, we propose a text corpus design method based on a similarity measure for speech coding. The similarity measure compares the phoneme distribution occurring from natural conversations with that from the designed text corpus.

Figure 1 shows the overall procedure of the proposed text corpus design method. As a first step of the proposed method, we collect text data containing predefined phonetically balanced words (PBWs) in order to ensure that all phonemes of the language are included. Next, the collected text data are refined so that they become formal and meaningless texts. After that, each refined text is transcribed into a phoneme sequence, and then the transcribed phoneme sequences are randomly grouped into N sets, where each set consists of M sentences. If the total number of sentences is T, then the i-th set, $S(i)$ ($i = 0, \ldots, N - 1$), is defined as

$$S(i) = \{\text{sent}(M - L) \cdot i + k, k = 0, \ldots, M - 1\}$$

where $\text{sent}(l)$ is the l-th sentence of all the sentences. In Eq. (1), L is the number of sentences overlapped between adjacent sets and indicates an optimal interpolation interval among randomly collected sentences. That is, $[T/(M - L)] = N$, where $\lfloor x \rfloor$ is the largest integer not greater than $x$.

The next step is to apply a similarity measure to determine the most appropriate set among the N sets. In other words, the distance between the phoneme distribution of the i-th set and the phoneme distribution of natural conversations, $D(i)$, is computed using a similarity measure defined as

$$D(i) = \sum_{j=1}^{N_p} \frac{|P(j) - P(i, j)|}{P(j)}$$

where $N_p$ is the total number of phonemes and $P(j)$ is the occurrence frequency of the j-th phoneme of the natural conversation. In addition, $P(i, j)$ is the occurrence frequency of the j-th phoneme in the i-th set. It can then be said that a candidate text set is more similar to natural conversation if the distance defined in Eq. (2) is smaller. Finally, a text corpus having the smallest distance is selected. That is, the designed text corpus is obtained as the $i^\ast$-th set such that $i^\ast = \arg \min_{0 \leq i \leq N-1} D(i)$.

3. Korean Stereo Super-Wideband Speech Database

In this section, we first review the NTT-SW database and then describe the requirements for a Korean stereo super-wideband speech (K-SW) database. After that, we explain how to construct the K-SW database by applying the proposed text corpus design method described in Sect. 2.

3.1 NTT-SW Database

Since we aim to design the K-SW database in order to evaluate the speech quality of super-wideband speech codecs for Korean speech, the K-SW database should be comparable to the NTT-SW database in terms of text corpus and speech material. The text corpus of the NTT-SW database is first investigated in terms of the languages supported and the number of utterances per speaker, which are summarized in Table 1. As shown in the table, the NTT-SW database is composed of six languages such as Japanese, North American English, British English, French, German, and Chinese (Mandarin). Note here that it does not include Korean, as mentioned in Sect. 1. The text corpus for North American English consists of 120 sentences, whereas 100 sentences are used as the text corpus for the other languages. Each sentence is constructed so that it can be uttered in around 4 seconds.

To further highlight the characteristics of the text data, we select the North American English text corpus as an ex-
Table 1  Characteristics of the text corpus of the NTT-SW database.

| Item          | Description                  |
|---------------|------------------------------|
| Language      | Japanese, North American English, British English, French, German, Chinese (Mandarin) |
| No. of sentences per speaker | North American English: 120 Others: 100 |

Table 2  Characteristics of the North American English text corpus of the NTT-SW database.

| Item              | Description |
|-------------------|-------------|
| No. of sentences  | 250         |
| No. of vocabulary | 924         |
| Total no. of words| 7,449       |
| No. of consonants | 14,963      |
| No. of vowels     | 8,433       |
| Average no. of syllables per sentence | 8.75 |
| Missing phonemes  | -           |

Table 3  Characteristics of the speech material of the NTT-SW database.

| Item          | Description |
|---------------|-------------|
| Guideline     | ITU-T Recommendation P. 800[6] |
| No. of Speakers | North American English: 4 males, 4 females Others: 3 males, 3 females |
| Average duration of utterances | Around 4 seconds |
| TDOA between L & R microphones | 18–21 samples |
| Active speech level | $-26\text{ dB}_{\text{rel}}$ |

Table 4  Active speech levels of left/right channels and TDOA between two channels in the NTT-SW database.

| Speech utterance | Active speech level ($-26\text{ dB}_{\text{rel}}$) | TDOA (samples) |
|------------------|-----------------------------------------------|----------------|
| WA01F002         | -26.10 to -26.45                              | 18             |
| WA01F004         | -26.01 to -26.71                              | 18             |
| WA01M004         | -26.13 to -26.50                              | 19             |
| WA01M0017        | -26.07 to -26.54                              | 20             |
| WE01F001         | -27.47 to -25.26                              | 21             |
| WE01F003         | -27.48 to -25.17                              | 21             |
| WE01M001         | -27.41 to -25.30                              | 20             |
| WE01M002         | -27.19 to -25.41                              | 20             |
| WF01F001         | -27.87 to -25.04                              | 20             |
| WF01F002         | -27.75 to -25.19                              | 20             |
| WF01M001         | -27.12 to -25.63                              | 21             |
| WF01M002         | -27.17 to -25.63                              | 21             |
| Average          | -26.98 to -25.74                              | 19.92          |

Example. The text corpus is a subset of the Harvard sentence list and consists of 250 sentences, where the vocabulary size is 924 and each speaker utters 120 sentences. Table 2 shows the total number of words, consonants, and vowels. Actually, they are measured to be 7,449, 14,963, and 8,433, respectively. Note here that the average number of syllables is measured as 8.75 with no missing phonemes. Based on this analysis of the text corpus, it is suggested that the K-SW database should contain all the Korean phonemes and the numbers of words and phonemes are similar to those of the analyzed text corpus.

Next, we examine the speech material for the NTT-SW database in terms of the recording guidelines, the number of speakers, the average duration of speech utterances, the time delay of arrival (TDOA) between microphones L and R, and the active speech level, which are shown in Table 3. In this database, the speech data are recorded in accordance with the recording environment and procedure defined by ITU-T Recommendation P. 800 [6]. In other words, each utterance in the database is recorded at a sampling rate of 48 kHz using a pair of high-quality studio condenser microphones so as to obtain a frequency spectrum of up to 20 kHz. In addition, four male and four female speakers participate in developing the materials for North American English, whereas three male and three female speakers do for other languages. Each speech utterance is recorded at 16 bits/sample and stored in Windows WAVE file format [13] with an average utterance duration of around 4 seconds.

In order to further examine the detailed characteristics of the speech data, we randomly select 12 speech utterances; two male and two female utterances from each of the North American English, British English, and French speech data. For each stereo speech utterance, we measure the active speech levels using ITU-T Recommendation P.56 [14] and the TDOA between the two microphones by calculating the maximum lag of cross-correlation between the left and right channels. Table 4 shows the active speech levels and TDOA for each stereo speech utterance, and their average values over the 12 stereo speech utterances. Consequently, the active speech level of each speech utterance is measured as $-26\text{ dB}_{\text{rel}}$ and the TDOA between the two microphones is as around 18–21 samples, as summarized in the last two rows of Table 3.

3.2 Requirements of the Korean Stereo Super-Wideband Speech Database

As mentioned above, the K-SW database is going to be used to evaluate super-wideband speech codecs for Korean utterances. Thus, we design the text corpus and speech material for the K-SW database to have a size comparable to or larger than the size of the NTT-SW database. Table 5 summarizes the characteristics of the text corpus required for constructing the K-SW database. As shown in the table, the text corpus is comprised of 320 short sentences and eight dialogues. For the short sentences, we assign the number of syllables per sentence to be between 14 to 18 syllables, such that the duration of each sentence is around 4 seconds. This number is set by counting the syllables in the Korean text corpus of the NTT-AT database. On the other hand, each dialogue is prepared to evaluate the conversational speech quality of a super-wideband speech codec having a duration of around 1 minute.

Table 6 shows the characteristics of the speech material for the K-SW database. As shown in the table, the speech material is prepared by recording 40 short sentences from
each speaker, where four male and four female speakers are involved in constructing the 320 short sentences in total. Moreover, each pair of one male and one female speakers utters one dialogue. After that, the two short sentence utterances are merged into one speech sentence, resulting in 160 speech sentences for the K-SW database. Finally, similar to the NTT-SW database described in Sect. 3.1, the TDOA and the active speech levels are set to 20 samples and −26 dBovl, respectively.

### 3.3 Construction of the Korean Stereo Super-Wideband Speech Database

This section explains how to construct the K-SW database by applying the proposed text corpus design method described in Sect. 3.2.

#### 3.3.1 Korean Text Corpus Design

In order to design the text corpus for the K-SW database described in Sect. 3.2, we first collect authentic Korean text materials from a wide range of materials including textbooks and websites, corresponding to the first step of Fig. 1. In fact, we obtain texts that contain at least one word of 452 PBWs in [8], which ensures that the texts contain all of the Korean phonemes. As a result, we obtain about 450,000 sentences. Next, among the collected sentences, we delete ungrammatical sentences as well as sentences that contain non-Korean characters such as Chinese, English, or special characters and numbers. After that, we screen the sentences to eliminate meaningful words such as proper nouns or political/religious/regional terms. In addition, we revise awkward parts of speech (POS) and informal terms to correct POS and formal terms, respectively. From these refinement processes, we construct about 40,000 refined sentences, which are then further reduced based on the number of syllables in each sentence. In this paper, we select short sentences whose number of syllables ranged from 14 to 18.

Next, each sentence is transcribed into its corresponding phoneme sequence based on Korean pronunciation rules [9]; Korean pronunciations are shown in Table 7. Note that for clarity each Korean pronunciation is denoted in its romanized form instead of using Korean phonetic symbols. In order to select the text corpus whose phoneme distribution best matches that from natural conversation, the transcribed sequences of the selected short sentences are randomly grouped into 5,000 sets of 320 short sentences. After that, the similarity measure defined in Eq. (2) is applied for each text set, where \( L, M, N, \) and \( N_p \) are 312, 320, 5,000, and 40, respectively. \( M, N, \) and \( N_p \) are design parameters and set by the requirements of the K-SW database as described in Sect. 3.2. On the other hand, in order to set a proper number to \( L, \) an experiment regarding the text corpus selection using Eqs. (1) and (2) is performed by changing \( L \) from 1 to 319 with a step of 1. Consequently, \( L (=312) \) is selected which provides the smallest distance defined in Eq. (2). In addition, Table 8 shows the Korean phoneme distribution, \( P(j) \), obtained from [10], [11].

Finally, a set of 320 Korean short sentences is selected, which has the smallest distance among the 5,000 sets, and is subsequently used as the text corpus for the K-SW database. In this work, the measured distance for the set is \( D(i^*) = 7.32 \), where the maximum, mean, and standard derivation of distances are measured as 39.28, 29.98, and 22.31, respectively. On the contrary, the distance for the NTT-AT Korean subset is measured as 36.78, which will be discussed in Sect. 4.1.

The text corpus for eight dialogues is also generated in a similar manner as for the 320 short sentences, except for the inclusion of a preference test. That is, we first collect 200 text data from textbooks related to Korean fluency tests. Among them, we refine the data by removing dialogues that do not match Korean sentiments or those associated with political, religious, or regional issues, resulting in 50 dialogue text data. The selected dialogues are further refined by removing any awkward meanings and proper nouns and by modifying the text in dialogues if the POS is incorrectly organized. Here, we select dialogues with durations of about 1 minute by informally reading each dialogue out loud. Similar to the selection process of the short sentences, we then

### Table 5 Characteristics of the text corpus for the K-SW database.

| Item            | Short sentences | Dialogues |
|-----------------|-----------------|-----------|
| Language        | Korean          |           |
| No. of texts    | 320 sentences   | 8 dialogues |
| No. of syllables per sentence | 14–18 | - |

### Table 6 Characteristics of the speech material for the K-SW database.

| Item            | Description                                                                 |
|-----------------|-----------------------------------------------------------------------------|
| Guideline       | ITU-T Recommendation P.800[6]                                              |
| No. of speakers | 4 males, 4 females                                                          |
| Avg. utterance duration | Short sentences: around 4 seconds                                             |
| TDOA between L & R microphones | 20 samples                                                             |
| Active speech level | −26 dBovl                                                              |

### Table 7 List of Korean vowels and consonants.

| Classification (Number) | Prosodic position | Phonemes                                                                 |
|-------------------------|-------------------|--------------------------------------------------------------------------|
| Vowel (21)              | Nucleus           | † (a), † (Ja), † (v)                                                   |
|                         |                   | † (jv), † (o), † (jo), † (u), † (iu), † (e), † (je), † (iE), † (we) |
|                         |                   | † (wO), † (wi), † (xi)                                                 |
| Consonant (19)          | Onset (11)        | τ (G), τ (D), w (B), x (S), x (z), x (Z), x (c), x (k), x (t), x (p), x (h) |
|                         |                   | ο (N)                                                                   |
|                         |                   | τ (g), τ (n), τ (d)                                                     |
|                         |                   | ο (l), ο (m), ο (b), ο (n)                                              |

**List of Korean vowels and consonants.**
apply the similarity measure to the transcribed phoneme sequences for the dialogues. As a result, we obtain a set of 16 dialogues having the smallest distance for possible use in the text corpus. In order to construct dialogues for the K-SW database, a preference test is performed on the selected 16 dialogues, based on the opinions of six males and three females. During the preference tests, each examiner is requested to mark plain dialogue candidates, and the eight most preferred dialogues are finally selected for the corpus.

3.3.2 Construction of the Stereo Super-Wideband Speech Material

In accordance with ITU-T Recommendation P.800 guidelines, the recording room has a volume of 5.94 m³, a reflection time of 8.137 ms, and a noise level of less than 30 dBA. During recording, the microphone is located 140–200 mm from the speaker’s lips. To obtain a flat spectral response of up to 20 kHz, for a sampling rate of 48 kHz, we use high-quality studio condenser microphones. Moreover, super-wideband signals are recorded using four microphones under the configuration shown in Fig. 2. In this figure, there is an upright center microphone located 200 mm from the speaker’s lips. In addition, two microphones, L and R, are used to capture the stereo super-wideband signals, each located 100 mm from the center microphone. The fourth microphone, located at F, is 1,000 mm from the center microphone. Note that the pair of microphones at C and F are designed to record super-wideband speech signals by simulating the environment of a conference room, where each microphone is placed according to the location of speakers in the room.

As in the NTT-SW database, the K-SW database developed by the proposed text corpus design method is used to evaluate the speech quality of super-wideband speech/audio codecs for Korean speech. Therefore, the TDOA between microphones L and R is adjusted to around 20 samples, at a sampling rate of 48 kHz, which is similar to the TDOA for the NTT-SW database. Moreover, the two pairs of stereo speech utterances recorded from the four microphones, (L, R) and (C, F), are post-processed in order to adjust the active speech level to $-26$ dBovl via the ITU-T Recommendation P.56 [14].

Figure 3 shows a procedure for post-processing the recorded super-wideband speech utterances. In the figure, for each utterance from microphones at L and R, we measure and adjust the TDOA to 20 samples. After that, the active speech level of each speech utterance is normalized to $-26$ dBovl by using ITU-T Recommendation P.56 [14]. Similarly, the active speech levels X and Y of two utterances from microphones C and F are measured using ITU-T Recommendation P.56. In this case, however, the active speech level of the mono speech utterance at C is normalized to $-26$ dBovl, whereas the level of the speech utterance at F is normalized by adding Y and the difference ($\Delta$) between $-26$ dBovl and X.

### Table 8

| Phoneme | a | ja | v | jv | o | jo |
|---------|---|----|---|----|---|----|
| Relative Freq. | 10.24 | 0.45 | 5.04 | 1.69 | 4.00 | 0.58 |

| Phoneme | u | ju | U | i | wa | wv |
|---------|---|----|---|---|----|----|
| Relative Freq. | 2.62 | 0.10 | 5.81 | 7.28 | 0.47 | 0.23 |

| Phoneme | wi | jiE | O | xi | E | n |
|---------|----|-----|---|----|---|---|
| Relative Freq. | 0.01 | 0.10 | 0.25 | 0.02 | 1.74 | 2.28 |

| Phoneme | je | wi | jE | g | G | n |
|---------|----|----|----|---|---|---|
| Relative Freq. | 0.13 | 0.17 | 0.03 | 9.24 | 1.53 | 11.23 |

| Phoneme | d | D | i | m | b | B |
|---------|---|---|---|---|---|---|
| Relative Freq. | 4.68 | 1.12 | 6.70 | 3.96 | 2.16 | 0.25 |

| Phoneme | s | S | z | z’ | c | k |
|---------|---|---|---|----|---|---|
| Relative Freq. | 3.07 | 1.11 | 3.34 | 0.66 | 0.94 | 0.53 |

| Phoneme | t | p | h | N |
|---------|---|---|---|---|
| Relative Freq. | 0.72 | 0.45 | 2.89 | 2.19 |

![Fig. 2](image)  
Illustration of the microphone arrangement for recording speech utterances in the K-SW database.

![Fig. 3](image)  
Post-processing procedure for making two pairs of super-wideband stereo speech utterances from four microphones placed at L, R, C, and F.

### 4. Experiments

#### 4.1 Text Corpus Comparison

As mentioned in Sect. 1, one of the key issues pertaining to speech database design for speech coding is that the quality of the text corpus must reflect the pronunciation behavior of natural conversation. Since the NTT-AT database has been popularly used for testing speech coders, we compared the phoneme distribution occurring from the developed Korean text corpus with that from the Korean subset in [4].

Table 9 shows the number of sentences, number of consonants and vowels, list of missing phonemes, average number of syllables per sentence, and distance defined in Eq. (2) for each database.

\[ \text{Distance} = \sqrt{\frac{\sum_{i=1}^{n} (|X_i - Y_i|)^2}{n}} \]

From now on, the subset of the NTT database based on Korean speech utterances is referred to as the NTT Korean subset.
In this comparison, the NTT-AT Korean subset was composed of 192 short sentences, whereas the Korean text corpus in the K-SW database was composed of 320 short sentences and eight dialogues. For instance, each short sentence in both databases had a similar average number of syllables. However, the phoneme ‘jE’ was not included in the NTT-AT Korean subset, which implied that the quality of speech coders could not be tested for words including ‘jE.’ In contrast, the K-SW database did include this phoneme, thereby overcoming this problem. In the case of Korean dialogues, the average number of syllables per sentence was 18.56, indicating that the dialogue sentences were longer than the short sentences. Finally, the distances defined in Eq. (2) were measured as 36.78, 7.32, and 8.44 for the NTT-AT Korean subset, the short sentences, and the dialogues in the K-SW database, respectively. This implied that the Korean text corpus developed by the proposed method was a better match for the phoneme distribution of natural conversation in the Korean language than the NTT-AT Korean subset. Thus, it could be concluded that the proposed text corpus design method was appropriate for designing a text corpus for speech coding.

### 4.2 Speech Material Comparison

The speech utterances obtained by the procedure described in Sect. 3.3.2 were then compared to those in the NTT-SW database, except for the speech utterances obtained from the (C, F) pair of microphones. Table 10 shows the average active speech levels and root mean square (RMS) values of the speech utterances from the four microphones, L, R, C, and F, respectively, in the K-SW database. As shown in the table, the average active speech levels of the speech utterances at L, R, and C were measured as -26.01 dB_{stoi}, though the level at F was -37.68 dB_{stoi}. In addition, the average RMS value was measured as about -75 dB, and the TDOA between the microphones L and R was measured as 20.32 samples. The numbers described above indicated that the characteristics of the K-SW database were similar to those of the NTT-SW database.

### 4.3 Application of the K-SW Database: Subjective Quality Test

In order to make sure that the developed K-SW database is useful for developing and evaluating the performance of super-wideband codecs, we would like to evaluate the quality of the developed K-SW database by using an already existing super-wideband codec. In other words, we performed a multiple stimuli with hidden reference and anchor (MUSHRA) test [15,16], which is a commonly used subjective quality test method for assessing wideband and super-wideband speech. Here, ITU-T Recommendation G.722.1 Annex C, simply G.722.1.C, was selected as the reference codec since it was a super-wideband speech codec recently standardized by ITU-T. Note that though G.722.1.C had the ability to operate at bitrates of 24, 32, and 48 kbit/s, the 24 kbit/s mode of G.722.1.C was selected in this work because speech quality difference was better observed at a lower bitrate than a higher bitrate.

The procedure for the MUSHRA test was as follows. First, we randomly selected two female and two male speech utterances from the K-SW database. We then down-sampled each speech utterance from 48 kHz to 32 kHz, and split the down-sampled stereo speech utterance into two mono speech utterances. It was due to the fact that G.722.1.C only supports encoding/decoding for 32 kHz mono speech utterances. Next, each 32 kHz speech utterance was encoded into a bitstream via the G.722.1.C encoder at a bitrate of 24 kbit/s. Thus, the encoded bitstream was subsequently decoded into a speech utterance by the G.722.1.C decoder. Finally, we up-sampled each decoded speech utterance from 32 kHz to 48 kHz and merged them into a stereo speech utterance.

We also needed to prepare a hidden anchor, a 7 kHz anchor, and a 10 kHz anchor for the MUSHRA test. The 7 kHz and 10 kHz anchors were low-pass-filtered with cutoff frequencies of 7 kHz and 10 kHz, respectively, from an original speech utterance. Here, the hidden anchor was an original speech utterance that was not informed to the participants for the MUSHRA test. In this work, the test was performed by six males and two females for the evaluation of the K-SW database and by three males and two females for the evaluation of the English subset of NTT-SW database. Note here that each database was evaluated by the natives. That is, Koreans and Americans participated in the

---

**Table 9** Comparison of the NTT-AT Korean subset with the text corpus in the K-SW database.

| Measurement     | L  | R  | C  | F |
|-----------------|----|----|----|----|
| No. of sentences| 192| 320| 126|    |
| No. of consonants| 3,418| 6,087| 2,730| |
| No. of vowels    | 2,883| 3,063| 2,339| |
| Missing phoneme  | jE |    |    |    |
| Average no. of syllables per sentence | 15.02 | 15.8 | 18.56 |
| Distance (Eq. (2)) | 36.78 | 7.32 | 8.44 | |

---

According to the conditions and requirements for a MUSHRA test [16], we prepared one hidden reference and two hidden anchors, where the hidden anchors were obtained by filtering the hidden reference with low-pass filters having 7 kHz and 10 kHz bandwidths. Actually, the hidden reference was used to ensure the top of the scale, i.e., 100, while the 7 kHz anchor was used for the lower end of the scale and the 10 kHz anchor indicated somewhere in the middle.
Table 11  Comparison of MUSHRA scores for the K-SW database and the English subset of the NTT-SW database, where each database was evaluated by the natives corresponding to the language.

| Database     | Hidden 7kHz | 10 kHz | G.722.1.C |
|--------------|------------|--------|-----------|
| Korean (K-SW)| 100        | 67     | 90        | 89        |
| English (NTT-SW)| 100       | 67     | 83        | 82        |

MUSHRA test for the K-SW database and for the English subset of the NTT-SW database, respectively.

Table 11 shows the MUSHRA test results for the Korean speech utterances of the K-SW database and the English speech utterances of the NTT-SW database. It was shown from the first row of the table that for the Korean speech utterances of the K-SW database the MUSHRA score of G.722.1.C was similar to that of the 10 kHz anchor. Similarly, for the English speech utterances of the NTT-SW database, the MUSHRA score of G.722.1.C was also similar to that of the 10 kHz anchor, as shown in the second row of the table. Notice that a little difference in the MUSHRA scores between the Korean speech utterances of the K-SW database and the English speech utterances of the NTT-SW database came from the different nationality of the listeners but it was negligible [16]. This result implies that the speech codec, G.722.1.C, preserves a language independency among English and Korean. Actually, since the proposed text corpus method aims to have fully reflected the characteristics of a language, the developed K-SW database can provide a high reliability of a MUSHRA subjective quality test.

5. Conclusion

In this paper, we proposed a text corpus design method for a Korean stereo super-wideband speech (K-SW) database. We first defined a similarity measure to compare the phoneme distribution occurring from natural conversation with that obtained from a previously designed text corpus. Then, based on this similarity measure, we could select the most appropriate text corpus, resulting in a phonetically balanced text corpus having a small amount of text data. As an application of the proposed text corpus design method, 320 short sentences and eight dialogues were generated for the K-SW database. By comparing text corpora, it was shown that the designed Korean text corpus better matched the phoneme distribution occurring from natural conversation than the NTT-AT Korean subset. Next, the texts for the K-SW database were recorded by following the recording guidelines defined by ITU-T Recommendation P.800. It was also shown from a comparison of the speech materials that the speech utterances in the K-SW database had similar characteristics to those in the NTT-SW database. Finally, a subjective quality test of the super-wideband speech codec was performed and it was found that the proposed method could provide comparable speech quality for speech utterances in the K-SW database to those in the NTT-SW database.

Acknowledgments

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0023888) and by the Ministry of Knowledge and Economy, Korea, under the ITRC support program supervised by the National IT Industry Promotion Agency (NIPA-2010-C1090-1021-0007).

References

[1] A.A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, “Design and development of phonetically rich Urdu speech corpus,” Proc. Oriental COCOSDA, pp.38–43, Urumqi, China, Aug. 2009.
[2] L. Villasenor-Pineda, M. Montes-y-Gomez, D. Vaufreydaz, and J.-F. Serignat, “Experiments on the construction of a phonetically balanced corpus from the web,” Lect. Notes Comput. Sci., vol.2945, pp.416–419, Feb. 2004.
[3] V. Radova and P. Vopalka, “Methods of sentences selection for read-speech corpus design,” Lect. Note Comput. Sci., vol.1692, pp.165–170, Sept. 1999.
[4] NTT-AT, Multi-lingual Speech Database for Telephonometry 1994, 1994.
[5] NTT-AT, Super Wideband Stereo Speech Database, 2007.
[6] ITU-T Recommendation P.800, Methods for Subjective Determination of Transmission Quality, Aug. 1996.
[7] ITU-T Recommendation G.722.1 Annex C, Low Complexity Coding at 24 and 32 kbit/s for Hands-free Operation in Systems with Low Frame Loss, G.722.1 Annex C: 14kHz Mode at 24, 32, and 48 kbit/s, May 2005.
[8] Y.J. Lee, B.W. Kim, and Y. Um, “Speech information technology & industry promotion center in Korea: Activities and directions,” Proc. International Conference on Language Resources and Evaluation (LREC), pp.1851–1854, Canary Islands, Spain, May 2002.
[9] The Ministry of Education of Korea, Korean Grammar Rulebook, Daehan Printing & Publishing, Seoul, Korea, 1995.
[10] S.L. Park, A Study on the Frequencies of Consonants in Adult Conversational Korean, Master Thesis, Ewha Womans University, Feb. 2000.
[11] J.N. Taek, “A quantitative linguistic study on the functional load of phonemes in standard Korean,” Malsori, nos.25-26, pp.65–92, Dec. 1993.
[12] ITU-T Recommendation P.830, Subjective Performance Assessment of Telephone-band and Wideband Digital Codes, Feb. 1996.
[13] IBM Corporation and Microsoft Corporation, Multimedia Programming Interface and Data Specifications 1.0, Aug. 1991.
[14] ITU-T Recommendation P.56, Objective Measurement of Active Speech Level, March 1993.
[15] M. Xie, D. Lindberg, and P. Chu, “ITU-T G.722.1 annex C: A new low-complexity 14 kHz audio coding standard,” Proc. ICASSP, pp.173–176, Toulouse, France, May 2006.
[16] ITU-R Recommendation BS.1534-1, Method for the Subjective Assessment of Intermediate Quality Levels of Coding System, Jan. 2003.
[17] ETRI ETRI Listening Lab. Report for the Qualification Test of G.711.1 and G.722 Super-Wideband Extension - Exp. 1a and 2d, June 2009.
Yoo Rhee Oh received a B.S. degree in Computer Science from Kyungpook National University, Korea in 2004. She then received both M.S. and Ph.D. degrees in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2006 and 2011, respectively. She is now an engineer with Electronics and Technology Research Institute (ETRI). Her current research interest is speech recognition.

Yong Guk Kim received a B.S. degree in electronics and computer engineering from Chonnam National University, Korea in 2006, and an M.S. degree in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2008. From 2008 to 2011, he was a research staff as a Ph.D. Candidate with the Human Computing Lab., GIST. Since 2011, he is a member of technical staff with the Maritime R&D Center, LIG Nex1. His current research interests include HUSS and underwater acoustics.

Mina Kim received a B.S. degree in Electronics Engineering from Busan National University, Korea in 2006, and an M.S. degree in Information and Communications Engineering from the Gwangju Institute of Science and Technology (GIST), Korea in 2008. Since 2008, she is a member of technical staff with the Department of Mobile Communications, LG Electronics. Her current research interests include non-native speech recognition and pronunciation modeling for speech recognition.

Hong Kook Kim received a B.S. degree in Control and Instrumentation Engineering from Seoul National University, Korea in 1988. He then received both M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 1990 and 1994, respectively. He was a senior researcher at the Samsung Advanced Institute of Technology (SAIT), Kiheung, Korea, from 1990 to 1998. During 1998–2003, he was a senior technical staff member with the Voice Enabled Services Research Lab at AT&T Labs-Research, Florham Park, NJ. Since August 2003, he has been with the School of Information and Communications, at the Gwangju Institute of Science and Technology (GIST) as a professor. His current research interests include speech recognition and coding, audio coding and 3D audio, and embedded algorithms and solutions for speech and audio processing for handheld devices.

Mi Suk Lee received both B.S. and M.S. degrees in Electronics Engineering from Hoseo University, Korea in 1991 and 1993, respectively. She then received Ph.D degree in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Korea in 2001. She has been joining with Electronics and Technology Research Institute (ETRI) as a senior engineer since Feb. 2002. Her current research interests include speech and audio coding and quality measure.

Hyun Joo Bae received B.S and M.S degrees in Computer Science from Pusan National University, Korea in 1988 and 1991 respectively. Since joining ETRI in 1991, she has been engaged in research on telecommunication application and converged service platform. She is currently a principal member of engineering staff at the Multimedia Communication Research laboratory, Electronics and Telecommunications Research Institute (ETRI) IT Convergence Technology Research Laboratory, Daegu, Korea.