IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 23, 2021, accepted September 11, 2021, date of publication September 24, 2021, date of current version October 12, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3115608*

# Linguistic-Coupled Age-to-Age Voice Translation to Improve Speech Recognition Performance in Real Environments

**JUNE-WOO KIM, HYEKYUNG YOON, AND HO-YOUNG JUNG**
Department of Artificial Intelligence, Graduate School, Kyungpook National University, Daegu 41566, Republic of Korea

Corresponding author: Ho-Young Jung (hoyjung@knu.ac.kr)

This work was supported in part by the Information Technology Research Center (ITRC) Support Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by Korean Government under Grant IITP-2021-2020-0-01808, in part by the Ministry of Science and ICT (MIST), and in part by the MIST through the IITP Program (Development of Semi-Supervised Learning Language Intelligence Technology and Korean Tutoring Service for Foreigners) under Grant 2019-0-00004.

**ABSTRACT** We address a low-performance problem of the elderly in automatic speech recognition (ASR) through feature adaptation agnostic to the ASR. Most of the datasets for speech recognition models consist of datasets collected from adult speakers. Consequently, the majority of commercial speech recognition systems typically tend to perform well on adult speakers. In other words, the limited diversity of speakers in the training datasets yields unreliable performance for minority (e.g., elderly) speakers due to the infeasible acquisition of training data. In response, this paper suggests a neural network-based voice conversion framework to enhance speech recognition of the minority. To this end, we propose a voice translation model including an unsupervised phonology clustering to extract linguistic information to fit the minority's speech to a current acoustic model frame. Our proposal is a spectral feature adaptation method that can be placed in front of any commercial or open ASR system, avoiding directly modifying the speech recognizer. The experimental results and analysis demonstrate the effectiveness of our proposed method through improvement in elderly speech recognition accuracy.

**INDEX TERMS** Speech recognition, voice translation, spectral feature transform, age-on-demand speech recognition.

## I. INTRODUCTION

Automatic speech recognition (ASR) technology refers to perceiving a given sequential input speech as a corresponding word or character and converting them into a complete text sentence. Sophisticated acoustic and language models are required to construct text from acoustic characteristics in the recognition phase. Among several trials to enhance ASR performance, the deep neural network - hidden Markov model (DNN-HMM)-based ASR system has led to notable improvements in the existing recognition systems [1]–[5]. The two modules in this hybrid mechanism have different roles respectively; DNN computes probabilities of the observation for all tri-phone states and HMM computes the sequential properties of phoneme information obtained from the DNN model. This attribute of the DNN is possible due to direct distribution estimation when modeling the posterior probability of the acoustic properties of the speech.

Unlike the HMM-based existing speech recognition system, which was mainstream at the time [6], the end-to-end ASR methods were released and gradually updated [7]–[16] which enabled high-performance speech recognition, even with a large vocabulary speech dataset scale. Moreover, in the general unsupervised settings, recent state-of-the-art ASR models have shown better performance than that of supervised learning in downstream tasks [17]–[26]. As such, the ASR system has gained huge attention from various domains due to its practical use and is developing rapidly. In this paper, we will focus on solving data bias using real commercial speech recognizers that are trained with large-scale vocabulary.

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

While the majority of speech recognition systems typically perform well on the adult speakers, we claim that current the ASR systems are prone to provide unreliable performance for the minority (e.g., elderly people), owing to differences in the diversity of acoustic and verbal correlation [27]–[31]. What's worse, most of the datasets that are used to train the ASR systems mainly consist of speech collected from ordinary adults. Therefore, the small fraction of the majority of data is vulnerable to a typical imbalance problem [32], which hinders the ASR models from comprehending the utterances of the elderly accurately.

Although fine-tuning a pre-trained ASR model would contribute to enhancing the speech recognition performance of ASR systems, datasets that rarely contain the elderly's voice still restrict the range of the improvements. Even if we can tune the model with the voice of the elderly, it is hard to solve the rooted problems that are mostly attributed to the various characteristics in the utterances of the elderly. Moreover, constructing a large elderly speech dataset is relatively laborious and expensive compared to that of adults. In addition, building the ASR model from scratch would be a difficult task that requires a lot of knowledge, computing power, and considerable data resources.

To cope with this problem, we propose a new paradigm to improve the performance of general ASR models that would typically show poor recognition of elderly utterances. To this end, we suggest an age-to-age voice conversion framework, which introduces an unsupervised phonology clustering method to bridge the respective phoneme characteristics of adults and the elderly. In practice, our method can be lightly joined in front of any commercial ASR model as a spectral feature adaptation approach. Our proposed approach elevates the elderly speech recognition performance and does not directly affects the original ASR system performance. In that respect, our goal is to yield high ASR performance for the elderly by converting their voices into that of adults.

To implement our proposed model, we use an open elderly dataset and we utilize another adult speech dataset (200 hours) to bridge phoneme characteristics of the two age groups in phonology clustering. To train the voice conversion model, we have collected the data that contain command-like short sentences that are generally used to operate smart home devices with adults' voices in practical usage. The voices of the two groups of elderly and adults are paired correspondingly.

To verify the solid accuracy of our suggested framework, we use an open ASR-system.[1] Experimental results validate the effectiveness of our proposed age-to-age voice translation model through an improvement in speech recognition accuracy.

Previous studies on the voice conversion method are described in Section II. The proposed algorithm and network are described in Section III, and experimental settings, results, and analysis are presented in Section IV and V. In Section VI, the conclusion of this paper is drawn.

[1]This Open ASR API is available at https://aiopen.etri.re.kr/guide_recognition.php

## II. RELATED WORKS

Voice Conversion (VC) refers to a process of converting input voice style into that of target maintaining the linguistic information of the input. VoiceGAN [33] suggested a style transfer framework for VC based on generative adversarial networks. This model has learned to transform the speaking style of utterances of the input speaker into that of the target without the additional linguistic information embedding. Parrotron [34] successfully converted the speech from disabled speakers into the ordinary's utterance and elevated the speech recognition performance for the people who are physically limited in their articulation. To improve the signal-to-signal conversion performance, an auxiliary speech recognition network was additionally connected to the encoder, which showed that recognition of multitask training is effective for the robustness of the model. In our previous work [35], an end-to-end VC was conducted at the spectrogram level of input speech without a linguistic clustering.

Some previous research has strived for acquiring powerful representations via the Transformer [36] network capable of computing both text-to-speech (TTS) and VC. Huang *et al.* [37] conducted the VC via a pre-trained TTS model, which is the transformer-based network trained with a large-scale corpus to transfer knowledge for the conversion process. Since this approach inherited the pre-trained weights from the TTS model, generated speech from the decoder is limited in terms of diversity. Liu *et al.* [38] focused on the transformer network-based context preservation mechanism and a pre-trained single-speaker TTS model in the perspective of model adaptation for one-to-one VC. While dual learning composed of TTS and the ASR is only focused on capturing latent representations for the text and speech [39], [40], our approach tries to bridge the linguistic latent values in the heterogeneous speech of the adult and elderly.

We are aimed at improving the low-performance of the ASR on the elderly speech as described in Parrotron [34]. However, instead of auxiliary tasks, we couple linguistic information expressed from similar phonemes in respective age groups by the unsupervised phonology clustering method.

## III. LINGUISTIC-COUPLED VOICE SIGNAL TRANSLATION

In this section, we describe our method, voice signal translation combined with linguistic information. This is devised to extract inherent characteristics in elderly utterances which are sparse in benchmark datasets. We also present a method that enables successful voice translation between speakers who have different speech styles and enunciation due to their aged vocal organs. As a new approach, we apply a simple unsupervised method to connect a phonology relation between the two group's speech features that are caught in each Mel-spectrogram frame. In other words, Mel-spectrogram frames obtained from the same phonemes are linked to having

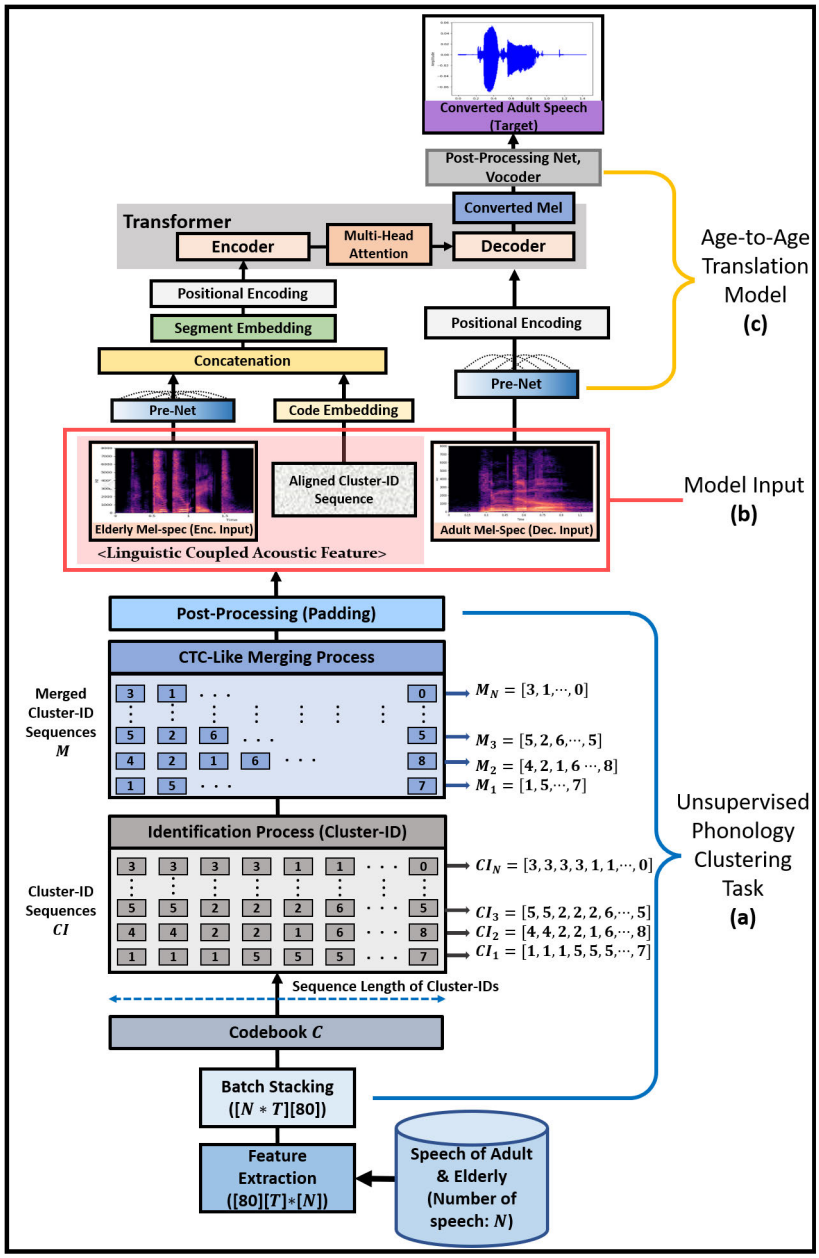


**FIGURE 1.** The full schematic of linguistic-coupled age-to-age voice translation model.

similar properties in respective groups. We extract speech-based linguistic information using $K$-means clustering to combine linguistic information for corresponding feature frames from speech data. Fig. 1 shows our overall proposed method. The main contributions are as follows:

- We propose an unsupervised phonology clustering method to induce a relation between the same phonemes found in both adults and elderly utterances and to bridge the homogeneous characteristics in speech signals of the respective group.
- We then explain how to acquire quantized clustered-ID from the feature frame using the codebook which is obtained from $K$-means clustering. In this process, due to the different pace of utterances, the cluster-ID between two speakers articulating the same sentence may not match frame-to-frame. To this end, the merging process is conducted to make a unique cluster-ID sequence.
- Lastly, we demonstrate how to use the linguistic information as input of our proposed voice translation model and to train it with a code embedding method.

### A. UNSUPERVISED PHONOLOGY CLUSTERING

The difficulties of collecting the elderly and children's voices have led most of the open dataset to be composed of adults' utterances. In that respect, we want to solve the problem of relatively scarce elderly speech data by using adult data. For this purpose, first, we bridge the homogeneous characteristics

from utterances of each age group. Fig. 2 presents the step-by-step process of our proposed phonology clustering method.

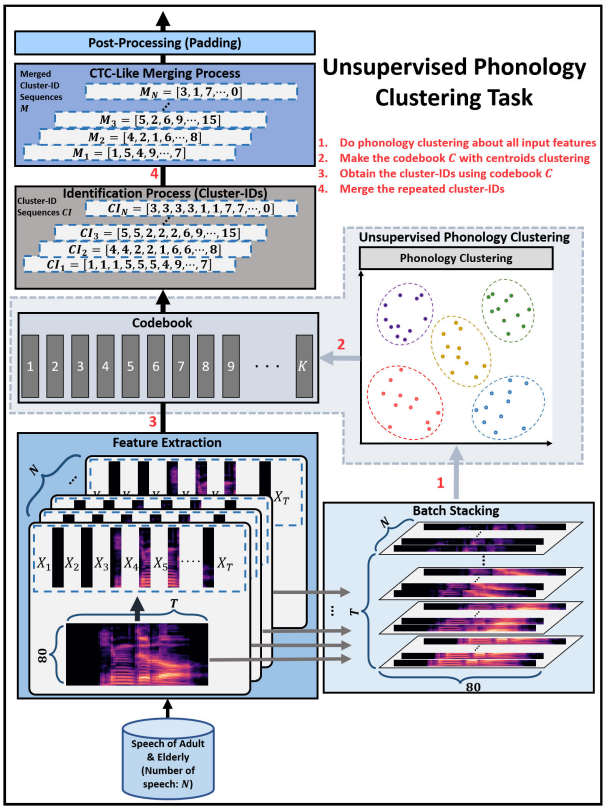


**FIGURE 2. Unsupervised phonology clustering procedure to extract linguistic information.**

Here, we organize every utterance in both adult and elderly dataset like adult $A = \{a_1, a_2, \cdots, a_n\}$ and elderly $E = \{e_1, e_2, \cdots, e_n\}$. The silent segment in every utterance is processed through a voice activity detection algorithm. All the speech features are extracted via an 80-dimensional log Mel-filterbank and the shape of each feature becomes $(80, T)$. We then stack the respective Mel-spectrograms to compose batches. Consequently, the whole database can be expressed as $(80, N * T)$, where $N$ is the number of data and $T$ is the sequence length of each Mel-spectrogram, and $N * T$ is $\sum_{i=1}^{N} T$.

After the batch stacking procedure, we get $\mathbf{X} = \{\mathbf{x}\}_{i=1,t=1}^{N,T}$ that contain all separated Mel-spectrogram feature vectors for all time step from $A$ and $E$ (see "**1**" in Fig. 2). We then quantize $\mathbf{X}$ by partitioning $\mathbb{R}^{80}$ into $K$ regions $r_1, \cdots, r_k$. This computation can be processed by constructing a codebook $C = \{c_1, \cdots, c_k\}$, where the $c_k$ indicates a codeword that is set of centroids using the $K$-nearest neighbors. Thus, the $K$ codebook IDs are used to represent $\mathbf{X}$ through phonology clustering (see "**2**" in Fig. 2). Note that the codebook of our method is not trained by the Gumbel-Softmax approach described in [41], [42].

We initialize the codebook $C$ and discover the optimal partitions $\mathcal{R}_k$ as

$$\mathcal{R}_k = \left\{\mathbf{x} : d(\mathbf{x}, \mathbf{c_k}) \leq d(\mathbf{x}, \mathbf{c_j}), 1 \leq k \neq j \leq K\right\} \quad (1)$$

where $d\{(a, b)\}$ is the euclidean distance between $a$ and $b$, $\mathbf{c}$ is the codebook index, $j$ is a different centroids, and $K$ is the number of codewords where $k = 1, \cdots, K$. We then extract phonology information using the $k$-th codewords $\mathbf{c}_k$ as

$$\mathbf{c_k} \leftarrow \underset{\mathbf{c_k}}{\operatorname{argmin}} \mathbb{E}\left[d(\mathbf{x}, \mathbf{c_k}) | \mathbf{x} \in \mathcal{R}_k\right], \quad k = 1, \cdots, K \quad (2)$$

Based on this clustering all feature vectors of $\mathbf{X}$ are represented as cluster-ID by its closest codeword (see "**3**" in Fig. 2). Through this procedure, it is possible to efficiently bridge the framewise phoneme which represents linguistic similarity in both the elderly and the adult datasets. In practice, the $A$ and $E$ are transformed to the $CI$ sequences like described in Fig. 2.

However, since the speech tempo for elderly speakers is relatively slower than that of adults, the $CI$ from $A$ and $E$ may not match frame-to-frame. To be precise, we need not the number of each cluster-ID, but a sequence composed of a unique cluster-ID to represent the utterance. In order to address the above problems, we propose a merging technique derived from Connectionist Temporal Classification (CTC) [43]. If the current time, maximum time length and $i$-th values are denoted respectively as $i$, $T$ and $value_i$, we merge repeated cluster-IDs from $i$ to $i + t$ until the $value_i \neq value_{i+t}$ to get non-repeated cluster-ID sequence. In the merging process, at the encounter of non-repeating cluster-ID, the time step moves to the next time step increasing $i$ and this interval process continues until the $T$. In this manner, it becomes feasible to minimize the error between the $CI$s of the two age groups (see "**4**" in Fig. 2).

We expect that applying the aforementioned procedure is able to enhance the voice translation performance. Furthermore, it can be used to extract meaningful representation between different two age generations. We concatenate these speech-based linguistic features extracted from the clustering and the log Mel-spectrogram features to form the final linguistic coupled acoustic feature. We will demonstrate how to use these feature vectors in III-B.

### B. AGE-TO-AGE VOICE TRANSLATION (A2AVT)

Transformer-based [36] network capable of fast parallel computing is the main component of our proposed model. Recent studies have verified that transformer-based networks are suitable for voice conversion [35], [37], [38]. Unlike typical usage of transformer models, our proposed voice conversion model does not require pre-trained weights. Also, we propose a different encoder structure than the decoder by coupling the linguistic information with speech features utilizing additionally obtained cluster-IDs.

Fig. 3 illustrates our age-to-age translation model. In the encoder module, the log Mel-spectrograms of the elderly speech and the corresponding cluster-ID sequence are fed

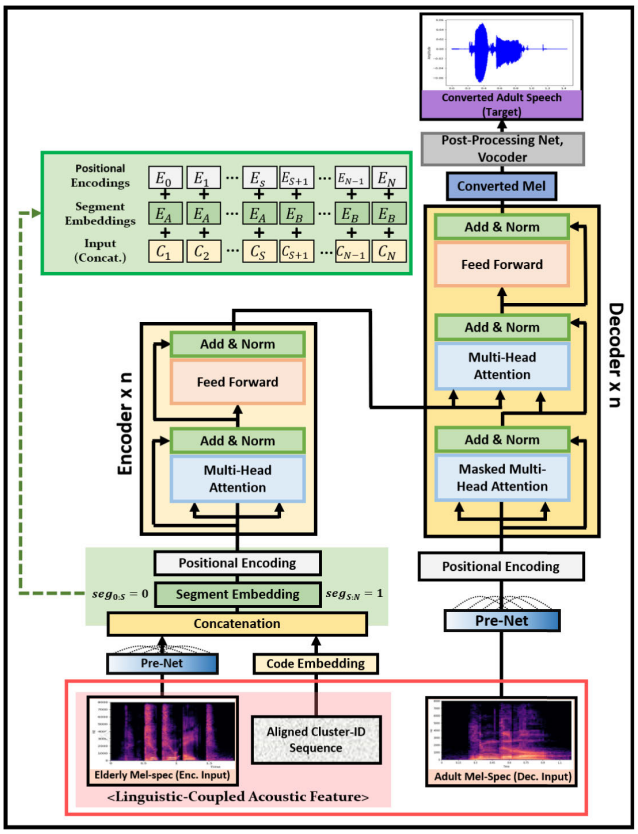


**FIGURE 3.** The end-to-end architecture of voice translation composed of encoder-to-decoder.

as the input linguistic-coupled acoustic feature. The features then pass through the Pre-Net [44] and the cluster-ID proceeds through a code embedding with the class number of $C + 1$ where padding value is added for matching the fixed maximum length. We conjecture that using Pre-Net helps to convert a more clear voice that connects both cluster-ID and speech features.

To concatenate both the Mel-spectrogram feature and cluster-ID, we borrow the segment embedding concept from BERT [45] and modify it to fit our input feature. While the "[SEP]" token is placed between the pair of sentences in BERT, we just pair our speech feature and the sequence of cluster-IDs without the token. Different segment position indexes are assigned to distinguish the speech feature and cluster-ID. The value of "0" is fed to the position of the speech feature ($seg_{0:s} = 0$) and "1" is assigned to the location of the cluster-ID ($seg_{s:N} = 1$). Consequently, the vocabulary size of our segment embedding is "2". The positional encoding then is added for the entire concatenated input to get the absolute position values. In the decoder module, only the Mel-spectrogram of the target adult is fed as input in the training phase.

Pre-Net is composed of two fully connected layers which have 0.5 dropout probability as described in [44]. The hidden size of the Pre-Net used in our architecture is 256. Every single encoder layer consists of two sub-layers: multi-head self-attention and a feed-forward layer. Two sub-components are connected via a residual network and layer normalization [46] is applied to each of them.

The decoder module is almost the same as the encoder except for its applying the look-ahead masking on multi-head attention. In time step $t$ of the decoder module, the look-ahead masking prevents the future speech features $m_{t+1}, m_{t+2}, \ldots m_T$ which are the target Mel-spectrogram frames to be exposed to the model. When the $t$ is increasing, the look-ahead masking gradually gets narrowed due to the autoregressive mechanism. In addition, the zero-padded part to fill the maximum length is ignored in the calculation of the self-attention phase.

In order to evaluate the quality of the converted voice, we need a post-processing network [44] to convert the Mel-spectrogram to linear spectrogram, and a synthesizer [47] to restore the linear spectrogram into a waveform. This reconstruction process is applied in the same condition as [44], except the reduction factor $\tau$.

Our objective function $\mathcal{L}_{sum}$ includes the two losses $\mathcal{L}_{mel}$ and $\mathcal{L}_{spec}$ as

$$\mathcal{L}_{sum} = \mathcal{L}_{mel} + \mathcal{L}_{spec} \tag{3}$$

where the $\mathcal{L}_{mel}$ loss is derived between the target adults' Mel-spectrogram and the converted result which is the output from the last decoder block. The $\mathcal{L}_{mel}$ loss is defined as

$$\begin{aligned}\mathcal{L}_{mel} &= L_1\lambda_1 + L_2\lambda_2 \\ &= \sum_{i=1}^{N}\sum_{k=1}^{T}|m_{ik} - \hat{m}_{ik}|\lambda_1 + \sum_{i=1}^{N}\sum_{k=1}^{T}(m_{ik} - \hat{m}_{ik})^2\lambda_2 \end{aligned} \tag{4}$$

where $N$ is the number of data, $T$ is the time length of each dataset, $m$ is the target Mel-spectrogram frame, and $\hat{m}$ is the predicted Mel-spectrogram frame, respectively. $\mathcal{L}_{mel}$ is back-propagated and affects weights in the proposed voice conversion model which manages conversion between the speeches of the two generations.

On the other hand, $\mathcal{L}_{spec}$ loss described in Equ.(3) is yielded between the target adults' linear spectrogram and the predicted linear spectrogram of the converted Mel-spectrogram (obtained from the last decoder block). The mathematical expression of the $\mathcal{L}_{spec}$ loss is as below:

$$\begin{aligned}\mathcal{L}_{spec} &= L_1\lambda_1 + L_2\lambda_2 \\ &= \sum_{i=1}^{N}\sum_{k=1}^{T}|s_{ik} - \hat{s}_{ik}|\lambda_1 + \sum_{i=1}^{N}\sum_{k=1}^{T}(s_{ik} - \hat{s}_{ik})^2\lambda_2 \end{aligned} \tag{5}$$

where $s$ is a target linear spectrogram frame and $\hat{s}$ is a predicted linear spectrogram frame of the $\hat{m}$ in Equ.(4). In other words, $\mathcal{L}_{spec}$ influences the post-processing network which is trained to speculate the tendency of linear spectral magnitude obtained by estimating sampled spectral characteristics in the Mel-frequency scale with a linear frequency scale [44]. Based on this procedure, we can reconstruct the linear spectrogram as the waveform using the synthesizer. We use the Griffin-Lim [47] algorithm as the synthesizer to reconstruct

the converted linear spectrogram as the waveform. The $\lambda_1$ and $\lambda_2$ used in both Equ.(4) and Equ.(5) are the hyper-parameters to adjust the ratio of each logit, and the mathematical formulation of the ratio is described as below:

$$\lambda_1 + \lambda_2 = 1.0 \tag{6}$$

## IV. EXPERIMENTAL SETUP

### A. DATASET

We use a VOTE400 speech dataset,[2] an AIHub speech dataset,[3] and an adult reference dataset (ARD) that we collected from two speakers. The VOTE400 dataset totally consists of 400 hours of the recorded utterance of the elderly whose average age is 79.47 seniors. The ratio of the gender in this dataset is 5.29:1 (women-vs-men). We focus on utterances that have short command words and are mainly spoken to operate smart home devices. Based on this condition, we sample 1330 training files recorded by 103 speakers and 161 test files recorded by 62 speakers (9:1 ratio).

The AIHub dataset is composed of around 1,000 hours of conversations uttered by 2,000 adults. This dataset was recorded in a daily dialogue environment where noise, laughter, and breathing are included. Among the whole dataset, we only select 200 hours of the AIHub dataset to perform only for the performance of phonology cluster modeling, not for age-to-age translation. In other words, we combine these 200 hours of data with the VOTE400 elderly dataset to obtain the aligned cluster-IDs to prepare the voice conversion training.

To make a target data that corresponds to elderly input data, we had to record adults' voices. We gathered 2 speakers, a thirty-year-old male, and a twenty-seven-year-old female. The recording was conducted in a small office room, and the man recorded his voice with his cell phone, which has a 48KHz sampling rate. Also, the woman recorded her voice with her cell phone which has a 44.1KHz sampling rate. As a result, we use ARD and corresponding elderly utterances with the same transcription to train the proposed voice conversion model.

Note that the above three datasets were recorded in different conditions. This is significant in that the proposed method can work in real environments.

### B. IMPLEMENTATION DETAILS

In detail, the hidden dimension of the model in Fig. 3 is 256, the number of the multi-head attention is 8, the dropout rate is set as 0.1 and the size of the feed-forward hidden dimension is 1024. Our model has 6 encoder layers and 4 decoder layers respectively. At the input module, masking is applied to hinder the attention to focus the zero-padded parts of linguistic-coupled acoustic features including cluster-IDs and log Mel-spectrograms. Masking positions are applied when self-attention is performed in the multi-head attention block.

[2]https://ai4robot.github.io/mindslab-etri-vote400/
[3]https://aihub.or.kr/aidata/105

We set the batch size as 16 and train the model until 85,000 steps ≈1060 epochs with a single NVIDIA RTX, which has 24GB memory. We use the Adam optimizer [48] with $\beta_1$ as 0.9, $\beta_2$ as 0.98, and $\epsilon$ as 1e-9 and 4000 warm-up steps. Each of both $\lambda_1$ and $\lambda_2$ is 0.5 in Equ.(4) and Equ.(5) explained in Section III-B.

Since there are various sampling rates in the datasets, we fixed all of the sampling rates as 16kHz and performed a resampling on all of them. We use the 80-dimensional log Mel-spectrogram with a window size of 32ms and an overlap size of 16ms and 512-point Fourier transform, and they are all normalized.

For constructing the unsupervised linguistic clusters, the codebook size is 32. Korean alphabets contain 19 consonants, 10 monophthongs, and 11 diphthongs. The phoneme set can be divided into 32 to 36 units if transitional sounds are not taken into account. The performance for the different numbers of $K$ would be demonstrated in V-E.

### C. UTILIZING OPEN ASR SYSTEM

To objectively compare the recognition performance of original elderly voices and our converted results, we use a commercial open recognizer that is able to handle large-scale vocabulary. Since the open ASR system only allows the waveform files as a system input, we added the post-processing network and the Griffin-Lim synthesizer for restoring the converted Mel-spectrograms into the waveforms to evaluate our model.

## V. RESULTS

The result of ASR performance on the original elderly utterance and ARD is demonstrated in this section. We then show the improved performance of the ASR from converted voice yielded from the proposed model. In addition, the distribution of the cluster-IDs from the utterances, which are spoken in the same transcriptions by the elderly and adults, is compared. Furthermore, the results of the mean opinion score (MOS) evaluation on the voice are presented.

**TABLE 1. Speech recognition results for the utterances spoken by adult and elderly speakers.**

| Group | Gender | CER (%) | Average CER (%) |
|---|---|---|---|
| Adult | Male | 11.08 | 12.80 |
| | Female | 14.51 | |
| Elderly | Male | 28.82 | 27.13 |
| | Female | 25.44 | |

### A. ASR PERFORMANCE ON ORIGINAL SPEECH

We set the character error rate (CER) as the criterion to evaluate speech recognition accuracy. First, we evaluate the performance of both the elderly test set and the target adult datasets. As shown in Table 1, the CER of the average elderly test set is 27.13% while male adult CER was 11.08% and female adult 14.51%, and the average adult CER was 12.80%.

According to the results, the commercial speech recognizer is more accustomed to ordinary adult speakers than the elderly speaker.

Before applying the learning-based approach, in addition, we use the conventional vocal tract length normalization (VTLN) method to evaluate the performance improvement rate. The speech data of elderly males and females are normalized according to adult male and female vocal tracts, respectively. Table 2 shows the performance of the VTLN method. The VTLN is somewhat effective for elderly male speech, but has little improvement for female speech. This shows that the normalization method has limitations in improving the elderly speech recognition performance.

**TABLE 2.** Speech recognition results after VTLN of the elderly data.

| VTLN | | CER (%) | ERR (%) |
|---|---|---|---|
| Source | Normalization Target | | |
| Elderly Male | Adult Male | 20.71 | 28.14 |
| Elderly Female | Adult Female | 25.11 | 1.30 |

### B. ASR PERFORMANCE ON CONVERTED SPEECH

The main interest of our approach is to improve elderly speech recognition beyond the generation of elderly speech via age-to-age voice translation. From this point of view, we try to compare the recognition performance of elderly speech using the conventional voice generation technique.

To this end, we conduct the evaluation using a denoising autoencoder (DAE). DAE filters out the differences from the speech of adults that occur in elderly speech due to aged vocal organs, defined as ''noise''. The converted voices of the elderly can be obtained from the DAE model under the same experimental conditions as our A2AVT model was trained.

We report the results of four algorithms (DAE, A2AVT, A2AVT + Linguistic-ID, and A2AVT + merged_Linguistic-ID) on CER metrics. The error reduction rate (ERR) in Table 3 indicates the relative improvement ratio between the CER from the converted result and the CER from the original elderly speech in Table 1.

**TABLE 3.** Performance comparison among DAE and A2AVT methods.

| Method | Conversion Target | CER (%) | ERR (%) |
|---|---|---|---|
| DAE | Adult Male | 20.53 | 28.76 |
| | Adult Female | 18.76 | 26.26 |
| A2AVT | Adult Male | 22.96 | 20.33 |
| | Adult Female | 20.97 | 17.57 |
| A2AVT + Linguistic-ID | Adult Male | 24.50 | 14.99 |
| | Adult Female | 22.51 | 11.52 |
| A2AVT + merged_ Linguistic-ID | Adult Male | 19.21 | 33.34 |
| | Adult Female | 14.35 | 43.59 |

We get 28.76% of the average ERR through the conventional DAE method. We obtain 18.95% of the average ERR via the A2AVT model. On the other hand, evaluation results of the A2AVT + Linguistic-ID show 13.25% of the average ERR, which is a little higher than the original speech of the elderly. Based on the recognized results, we expect that due to the different speaking tempos of adult and elderly speakers, segment aligning of the linguistic-ID sequence is required.

Thus, we acquire the best improvement in the A2AVT + merged_Linguistic-ID method of merging repeated IDs from the same linguistic segments into a single ID. While the average CER of the original elderly speech is 27.13%, our proposed A2AVT model + merged_Linguistic-ID shows 16.78% of the average CER and 38.47% of the average ERR. This shows that the merging technique proposed in this paper can connect homogeneous properties in the heterogeneous speech fashion in our phonology clustering.

When the elderly voices are translated into adult male speech, the ERR was relatively lower than in the case where those are translated into adult female speech. We guess this attribute of this tendency is owing to the scarcity of elderly male speakers' utterances. This implies that voice translation performance between the same gender is more effective, and it is expected that the best results can be obtained if the proposed model is trained for each gender.

In addition, we evaluate speech recognition performance by applying the voice conversion to the vocabulary not used in the A2AVT learning. We applied the proposed method to random 30 elderly utterances, and the results are presented in Table 4.

**TABLE 4.** The performance of the A2AVT model for arbitrary vocabularies.

| Method | Conversion Target | CER (%) | ERR (%) |
|---|---|---|---|
| No Conversion | | 30 | |
| A2AVT + merged_ Linguistic-ID | Adult Male | 25 | 16.67 |
| | Adult Female | 12.5 | 58.33 |

Table 4 shows that the proposed method works for arbitrary vocabularies not used as transform pairs in the A2AVT learning.

As a result, the improved recognition performance verifies the effectiveness of our A2AVT and linguistic-coupled information method, and shows that the proposed method can be adopted without any modification on the commercial ASR system.

### C. QUALITY MEASUREMENT OF CONVERTED VOICE

To evaluate the quality of our converted voices, we employ a mean opinion score (MOS) and a Mel-cepstral distortion (MCD) measurement.

In the MOS evaluation settings, we recruited 60 native Korean subjects whose ages are between 20 to 39. The evaluation was conducted in a quiet environment and the participants were suggested to wear a headset to concentrate

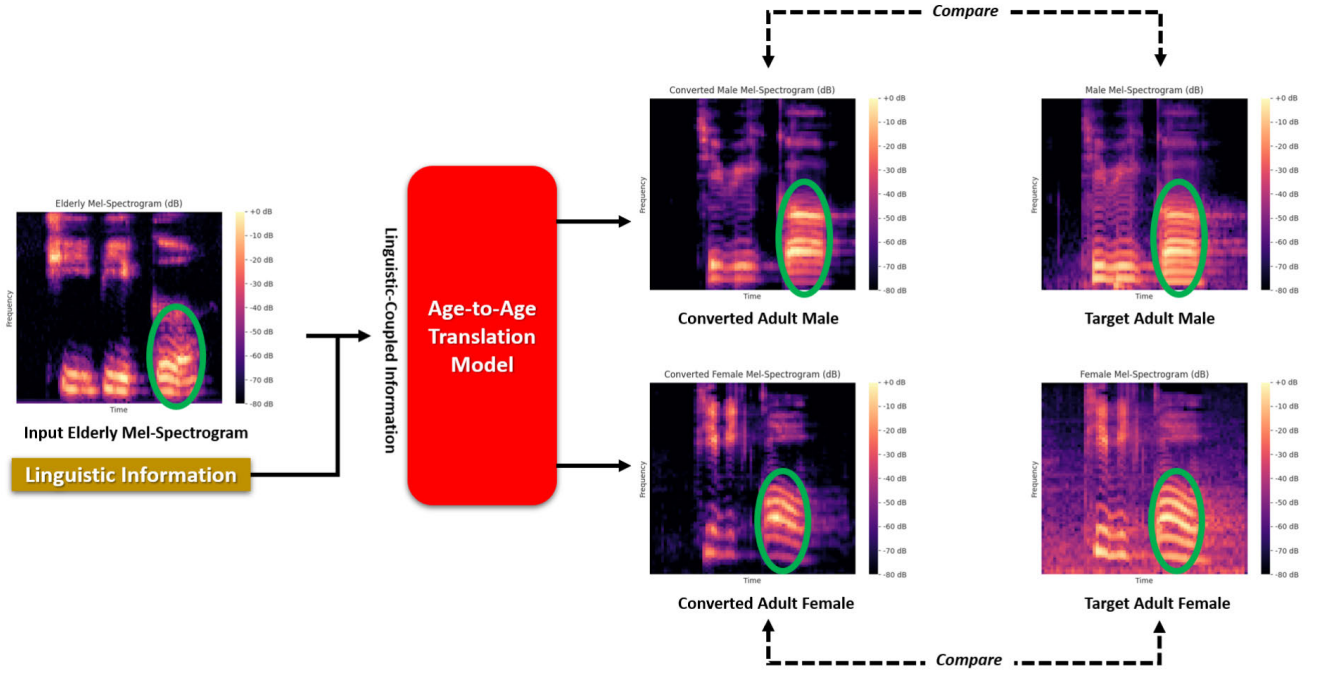


**FIGURE 4.** Comparison of elderly input Mel-spectrogram, converted results, and the target adult Mel-spectrograms.

on the objective judgment. The participants mainly scored the two phases of converted voice quality. The first aspect is naturalness. Score criterion for the degree of naturalness depends on how much the converted voices are natural compared to the given script. The second aspect is similarity. Evaluation of the degree of similarity depends on how similar the converted elderly voices with ground truth male and female adults' speech. Participants scored on a 1-5 points scale for the two aspects and the higher score is the best.

**TABLE 5.** MOS evaluation for the converted voice.

| Task | Target | | |
|---|---|---|---|
| | Male | Female | Average |
| Naturalness | 4.19±0.17 | 4.63±0.11 | 4.41±0.15 |
| Similarity | 4.14±0.18 | 4.69±0.10 | 4.41±0.16 |

Table 5 describes the results of MOS evaluation. The naturalness score of the converted elderly to male adult speech is 4.19 ± 0.17 and the similarity score is 4.14 ± 0.18. The naturalness score of speech converted from elderly to female is 4.63 ± 0.11 and the similarity score is 4.69 ± 0.10. Like the results of ASR, the latter converted voices get higher scores than its counterpart. The overall average naturalness score is 4.41 ± 0.15 and the average similarity score is 4.41 ± 0.16.

In addition, we present the results of the MCD measurement with the adult voice using the original elderly voice and the converted elderly voice. To measure the MCD, we use the 80-dimensional log Mel-spectrogram instead of Mel-frequency cepstral coefficients. If the two voices have different lengths, zero paddings are added to match the longer Mel-spectrogram. Table 6 shows that the converted elderly voice is closer to the adult voice compared to the original elderly voice.

**TABLE 6.** Results of MCD measurement with the adult voice.

| MCD Pair | Distortion Value (dB) |
|---|---|
| Adult - Original Elderly | 13.77 |
| Adult - Converted Elderly | 6.78 |

### D. ANALYSIS

We analyze the elderly voices converted into adults via our proposed model. In addition, analysis of cosine similarity between linguistic-ID sequence and distribution of the linguistic cluster-ID sequence of the same utterance from two groups are given.

Fig. 4 exhibits the age-to-age translation results using the proposed method. The input elderly Mel-spectrogram, its converted results, and the target adult Mel-spectrogram are represented from the left to right order. Among the top two Mel-spectrograms, the left one represents the converted result and the right one shows the target adult male voice respectively. At the bottom, the Mel-spectrograms of the converted result and its target female adult voice are presented.

As shown in Fig. 4, the formants of each Mel-spectrogram are colored green. In comparison, while the formants in the converted result and the target Mel-spectrogram have similar shapes, the input elderly Mel-spectrogram is different from its

results. Specifically, both target Mel-spectrograms show horizontal formants. The female adults' Mel-spectrograms show a slight increase and gradual decrease, while the original Mel-spectrogram shows wave-shaped formants.

**TABLE 7. Cosine similarity of cluster-IDs from different genders and ages.**

| Group A | Group B | Cosine Similarity (A, B) | |
|---|---|---|---|
| | | Not Merged | Merged |
| Adult Male | Adult Female | 0.9549 | 0.9596 |
| Elderly | Adult | 0.8522 | 0.8833 |

Here, we report the results of cosine similarity between the two linguistic-ID sequences. As shown in Table 7, we achieve the same similarity for the two genders in the adult group. It is speculated that the comparable cosine similarities of the two adult groups are attributed to their analogous speaking tempo. On the other hand, we find the application of the merging technique shows different results in comparison to elderly and adult voices. The merged IDs have shown higher similarity, therefore, our merging technique can generate homogeneous properties in speech features from the two different age groups.

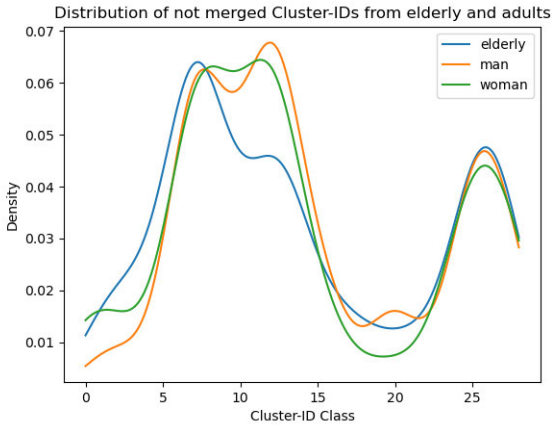


**FIGURE 5. The distributions of cluster-IDs of the elderly and the adults obtained without the merging process.**

Further, we analyze the two distribution functions obtained with and without the application of the merging process. As described in Fig. 5, the linguistic-IDs of all speakers draw a similar envelope shape. The distribution functions indicate that the proposed phonology clustering method well maps the phonemes from the two age groups. While most of the classes show similar envelopes, the shape of 9 to 14 classes of the elderly is different from that of adults. We conjecture that it is due to the difference in speaking tempo of the elderly and adults.

Fig. 6 shows the distribution functions of merged cluster-IDs. Thanks to merging technique, we can observe more similar distributions than shown in Fig. 5. Therefore, the merging technique can provide the linguistic bridging

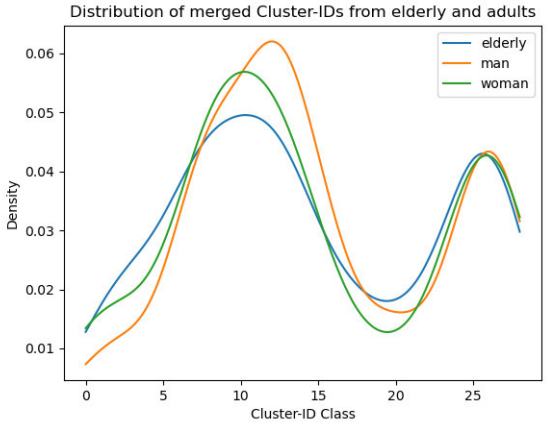


**FIGURE 6. The distributions of cluster-IDs of the elderly and the adults obtained with the merging process.**

of the two groups by leveraging the proposed phonology clustering method with the merging technique.

### E. PERFORMANCE COMPARISONS ACCORDING TO THE NUMBER OF LINGUISTIC CLUSTERS

We demonstrate the effect of the number of linguistic clusters, $K$ for A2AVT + merged_Linguistic-ID model. When $K$ is 8, 16, 32, and 64, respectively, the CER of converted voices are compared.

**TABLE 8. Performance comparison according to the number of linguistic clusters.**

| $K$ | Conversion Target | CER (%) | ERR (%) |
|---|---|---|---|
| 8 | Adult Male | 20.75 | 28.00 |
| | Adult Female | 24.50 | 3.69 |
| 16 | Adult Male | 20.31 | 29.53 |
| | Adult Female | 15.45 | 39.27 |
| 32 | Adult Male | 19.21 | 33.34 |
| | Adult Female | 14.35 | 43.59 |
| 64 | Adult Male | 34.22 | -18.74 |
| | Adult Female | 24.06 | 5.42 |

In Table 8, the best performance is obtained when $K$ is 32. This indicates that the clustering based on Korean phoneme sets is effective as the linguistic information for the A2AVT method.

## VI. CONCLUSION AND FUTURE WORKS

In this paper, we have focused on improving the performance of the commercial ASR system which is weak at recognizing the outlier voices like the elderly. To this end, we introduced linguistic-coupled information through the unsupervised phonology clustering method and proposed the age-to-age voice translation using the linguistic-coupled information to enhance the speech recognition performance for the elderly. In that respect, the proposed method is the

alternative adaptation methodology that can be placed in front of any commercial or open ASR system. We demonstrated the effectiveness of our proposed A2AVT method and merged linguistic-coupled information through the improvement in speech recognition accuracy from the commercial ASR system.

As future work, including elderly speech recognition, we also plan to verify the method can be applied for the recognition of speech of children, disabled people with stammer articulation, as well as different accents and dialects. In addition, we will design our method to be able to operate within non-pairwise conditions using the utterances from different sentences.

## REFERENCES

[1] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learn. Speech Recognit. Rel. Appl.*, Vancouver, BC, Canada, 2009, vol. 1, no. 9, p. 39.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2011.

[3] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 131–136.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[5] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8614–8618.

[6] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.

[7] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*. [Online]. Available: http://arxiv.org/abs/1412.5567

[8] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764–1772.

[9] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 167–174.

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[11] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.

[12] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and J. Chen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.

[13] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.

[14] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[15] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4774–4778.

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036–5040.

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: http://arxiv.org/abs/1807.03748

[18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 3465–3469.

[19] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech*, Sep. 2019, pp. 146–150.

[20] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[21] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: From supervised to semi-supervised learning with modern architectures," 2019, *arXiv:1911.08460*. [Online]. Available: http://arxiv.org/abs/1911.08460

[22] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[23] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6419–6423.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.

[25] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020, *arXiv:2010.10504*. [Online]. Available: http://arxiv.org/abs/2010.10504

[26] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-Y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 344–350.

[27] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 349–352.

[28] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. EUROSPEECH*, 1997.

[29] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1999, pp. 145–148.

[30] A. Potamianos, A. Potamianos, S. Narayanan, and S. Member, "Robust recognition of children's speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[31] S. Kwon, S.-J. Kim, and J. Y. Choeh, "Preprocessing for elderly speech recognition of smart devices," *Comput. Speech Lang.*, vol. 36, pp. 110–121, Mar. 2016.

[32] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[33] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2506–2510.

[34] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, Sep. 2019, pp. 4115–4119.

[35] J.-W. Kim, H.-Y. Jung, and M. Lee, "Vocoder-free end-to-end voice conversion with transformer network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998–6008.
[37] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, Oct. 2020, pp. 4676–4680.
[38] R. Liu, X. Chen, and X. Wen, "Voice conversion with transformer network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, p. 7759.
[39] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5410–5419.
[40] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2802–2812.
[41] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, *arXiv:1611.01144*. [Online]. Available: http://arxiv.org/abs/1611.01144
[42] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," 2016, *arXiv:1611.00712*. [Online]. Available: http://arxiv.org/abs/1611.00712
[43] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.
[44] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.
[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Minneapolis, MI, USA, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423
[46] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[47] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

**JUNE-WOO KIM** received the B.S. degree from the Department of Information and Communication Engineering, Mokwon University, Daejeon, South Korea, in 2017, and the M.S. degree from the Department of Artificial Intelligence, Kyungpook National University, Daegu, South Korea, in 2021, where he is currently pursuing the Ph.D. degree.

From 2017 to 2018, he was a Researcher with the Institute for Artificial Intelligence, Korea Advanced Institute of Science and Technology (KAIST), Daejeon. His research interests include speech recognition, unsupervised learning, self-supervised learning, and voice conversion.

**HYEKYUNG YOON** received the B.S. degree in humanities from Geumgang University, Nonsan, South Korea, in 2018. She is currently pursuing the master's degree with the Department of Artificial Intelligence, Kyungpook National University. Her main research interests include spoken language understanding and intent classification.

**HO-YOUNG JUNG** received the B.S. degree in electronics engineering from Kyungpook National University, Daegu, South Korea, in 1993, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 1995 and 1999, respectively. His Ph.D. dissertation was on robust speech recognition. He joined the Electronics and Telecommunications Research Institute, Daejeon, as a Senior Researcher, in 1999, and was a Project Leader of the Spoken Language Intelligence Research Group, from 2002 to 2019. Since 2019, he has been an Assistant Professor with the Department of Artificial Intelligence, Kyungpook National University. His current research interests include machine learning, interactive learning, autonomous agent systems, speech understanding, and natural language processing.