INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Extending an Acoustic Data-Driven Phone Set for Spontaneous Speech Recognition

Jeong-Uk Bang[1], *Mu-Yeol Choi*[2], *Sang-Hun Kim*[2], *Oh-Wook Kwon*[1]

[1] Chungbuk National University, South Korea
[2] Electronics and Telecommunications Research Institute, South Korea
{jubang,owkwon}@cbnu.ac.kr, {mychoi,ksh}@etri.re.kr

Abstract

In this paper, we propose a method to extend a phone set by using a large amount of Korean broadcast data to improve the performance of spontaneous speech recognition. The proposed method first extracts variable-length phoneme-level segments from broadcast data, and then converts them into fixed-length latent vectors based on an LSTM architecture. Then, we used the k-means algorithm to cluster acoustically similar latent vectors and then build a new phone set by gathering the clustered vectors. To update the lexicon of a speech recognizer, we choose the pronunciation sequence of each word with the highest conditional probability. To verify the performance of the proposed unit, we visualize the spectral patterns and segment duration for the new phone set. In both spontaneous and read speech recognition tasks, the proposed unit is shown to produce better performance than the phoneme-based and grapheme-based units.

Index Terms: acoustic units, phone set, spontaneous speech recognition, broadcast data

1. Introduction

The phoneme unit has been long used as the acoustic modeling unit for speech recognition. In recent years, the grapheme unit that does not require grapheme-to-phoneme (G2P) conversion is attracting interests, as end-to-end speech recognition systems are becoming popular. However, the grapheme unit has lower speech recognition performance than the phoneme unit, because the speech signals having various spectral patterns are mapped to one grapheme symbol. These results are commonly observed in both conventional speech recognition systems and recent end-to-end speech recognition systems [1].

In large vocabulary continuous speech recognition, the phoneme unit has a problem that the number is too small to express various acoustic changes. Previous works have complemented this problem using the implicit method [2-3] or the explicit method [4]. Here, we focus on the implicit method of using decision tree [2]. The implicit method is to extend a context-dependent model according to adjacent phonemes, and then uses a decision tree to share parameters of models with similar acoustic characteristics. This method is used as a standard in conventional speech recognition systems.

Spontaneous speech has more pronunciation variation than read speech. The phoneme unit in spontaneous speech has a smaller inter-unit distance and a larger variance than the phoneme unit in read speech [5]. Decision trees show improved speech recognition performance when segmented into acoustically discriminative units, as they better perform when segmented based on the phoneme unit instead of the grapheme unit in read speech. Therefore, if we build a new phone set by clustering common spectral patterns from spontaneous speech, we can expect to improve the performance of spontaneous speech recognition.

We propose a method to improve the performance of spontaneous speech recognition by extending the phone set from a large amount of Korean broadcast data. The proposed unit is extracted in three steps. We first extract variable-length phoneme-level segments, and then converts them into fixed-length latent vectors based on a long short-term memory (LSTM) architecture. Finally, we use the k-means algorithm to cluster acoustically similar latent vectors and then build a new phone set by gathering the clustered vectors. The proposed unit is shown to produce better performance than the phoneme-based and grapheme-based units in both spontaneous and read speech recognition tasks.

The paper is structured as follows. First, in Section 2, we describe Korean broadcast data. Then, we describe a method for building the proposed unit in Section 3, and explain a method for updating the lexicon in Section 4. In Section 5, we show the results of our experiments. Finally, in Section 6, we present our conclusions.

2. Korean Broadcast Data

We use Korean broadcast speech data [6] for about 1,000 hours for the phone set extension experiment. This database is automatically constructed from broadcast audio data and their subtitle text based on the lightly supervised approach [7]. The collected broadcast data has a mixture of background noise and background music, and contains some incorrectly transcribed text that does not match the speech signal.

The broadcast data is largely composed of seven genres: News, current affairs, documentary, culture, drama, children, and entertainment. Here, we note that the news, documentary, and current affairs genres, which include many reading-style utterances, consist of 25%, 12% and 5%, respectively, and account for 42% of the total broadcast data. On the other hand, the culture, drama, entertainment, and children genres, which include many spontaneous-style utterances, consist of 22%, 16%, 12% and 3%, respectively, and account for 54% of the total broadcast data. The remaining 4% consists of sports broadcasts and music programs.

Spontaneous speech is filled with unwanted pauses, word fragments, elongated segments, filler words, self-corrections, and repeated words. If we find and cluster common spectral patterns from broadcast data, we will be able to build new units suitable for spontaneous speech recognition.

http://dx.doi.org/10.21437/Interspeech.2019-1979

3. Proposed Unit

This section describes a method for building the new units from a large amount of Korean broadcast data. Here, we compare various methods of extracting fixed-length vectors from variable-length phoneme-level segments and visualize the spectral pattern and duration of the new units.

3.1. Segment Extraction

To generate a unit with a common spectral pattern from the speech data, we first extract the phoneme-level segments. The previous study [8] searched for an acoustic unit based on frame-level feature vectors of 25.6 msec. However, since the frame-level unit has too short a length, it may not be suitable for expressing various patterns of spontaneous speech signals. For this reason, we extract segments using the phoneme unit that are defined based on phonological knowledge.

The speech segments are extracted using the text-to-speech aligning method commonly used for acoustic model training. To extract the segments, we used an acoustic model of the deep neural network (DNN) structure used in the previous experiment [6] and used feature vectors of 40-dimensional log-Mel filter-bank features spliced for ± 7 frames. The extracted segments have a variable-length duration.

3.2. Segment Representation

We convert the variable-length segments to fixed-length feature vectors. It is difficult to find spectral patterns directly from variable-length segments. To do this, we compare the three methods: 1) Feature vectors of fixed-length obtained by linearly interpolating variable-length segments into fixed length, 2) Latent vectors extracted from the LSTM auto-encoder model [9] using the filter-bank features of each segment as inputs and outputs as shown in Fig. 1 (a), and 3) Latent vectors extracted from the encoder-decoder model using the filter-bank features of each segment as the input of the encoder and the phoneme symbols as the target of the decoder as shown in Fig. 1 (b).

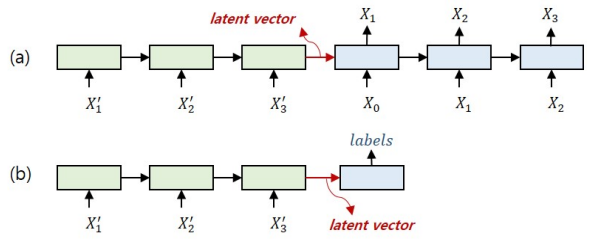

Figure 1: *Structures of LSTM auto-encoder model (a) and encoder-decoder model (b).*

In Figure 1 (a), the X_n represents the n-th 40-dimensional log-Mel filter-bank feature of each segment, and the X'_n represents a feature that contains adjacent ± 1 frames. In Figure 1 (b), the labels are 40 phoneme symbols commonly used in Korean speech recognition. As the model structures, we used a hidden-layer and a memory cell with 80 nodes using the tanh activation function in both models. For training, we used the Adam optimization algorithm with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, a batch size of 64, and 10 epochs with the learning rate 0.01.

We compared the fixed-length vectors obtained in three ways by visualizing them. Visualization experiments are performed on /a/, /c/, and /h/ among 40 Korean phonemes. Here, the pronunciation of Korean phoneme /a/ is similar to /a/ in [p a t], which is the pronunciation of the English word 'pot', Korean phoneme /c/ is similar to /tʃ/ in [tʃ e k] of the English word 'check', and Korean phoneme /h/ is similar to /h/ in [h ae t] of the English word 'hat'. Figure 2 shows the PCA-based visualization [10] results of the feature or latent vectors obtained by the three methods. We also checked the frame length as well as the phoneme symbols of each segment.

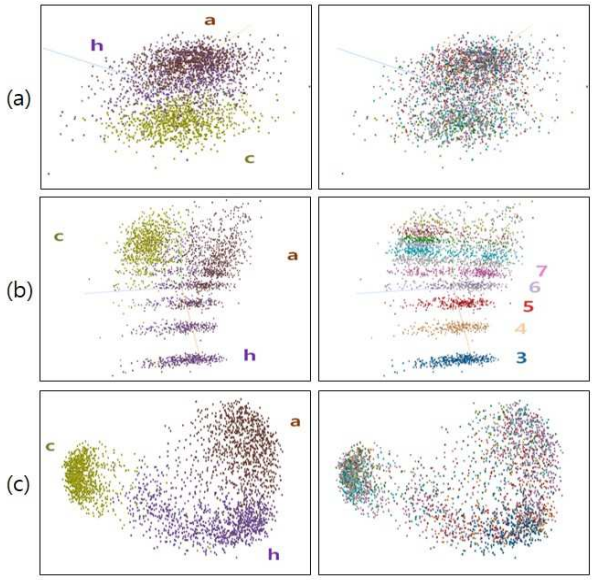

Figure 2: *Scattering plots of fixed-length vectors according to phoneme class (left) and duration (right) obtained by (a) linear interpolation, (b) auto-encoder model, and (c) encoder-decoder model.*

The feature vectors obtained using linear interpolation are in a similar position with different phonemes /a/ and /h/. This simple method did not well model the acoustic difference of each phoneme and is not suitable for finding common spectral patterns of the speech signals.

The LSTM auto-encoder model [9] extracts the high-level acoustic features of each segment from the encoder stage, which is shown as a green block in Figure 1(a), and passes them to the input of the decoder stage, which is shown as a blue block in Figure 1(a). Then, the decoder stage learns to extract the original feature vectors again based on the latent vectors obtained from the encoder stage. Here, the latent vector is learned to be reconstructed as the original feature vector, and has compressed information of the speech signal. In Figure 2 (b), we observed that clusters were created along the length of the frame. This is because the frame length information is most important for restoring the latent vector to the original feature vectors of the segment.

The encoder-decoder model, unlike the LSTM auto encoder model, is learned to represent 40 Korean phonemic symbols as the target of the decoder stage. This is to prevent the segment length information from being emphasized. In Figure 2 (c), we have found that the speech vector extracted by this method better represents the acoustic properties of each phoneme than the vectors extracted by other methods.

We compared the performance of the fixed-length feature vectors by using the Davies-Bouldin score [11], defined as the ratio of intra-cluster distances to inter-cluster distances. Experiments were performed on the vectors and phonemic labels obtained by each method. The scores were 8.8, 9.4, and 5.0, respectively, and the lower the better. As a result, we have found that the encoder-decoder model performs better than the other methods.

3.3. Segment Clustering

To cluster the speech vectors extracted in the previous step, we use the k-means algorithm [12], the most common clustering algorithm. The Euclidean distance used in the k-means algorithm is a significant measure in Manifold space. In clustering experiments, it is physically difficult to use all of the 50 million segments for 1,000 hours. Therefore, we randomly selected 10,000 samples (400,000 total) for each phoneme and performed clustering.

For each of the resulting 100 clusters, we visualized the average spectral pattern of segments, the number of phonemes, and the histogram of the segment length. Here, the spectral pattern was obtained by linearly interpolating segments composed of various lengths into 30 frames, and then outputting them for all phonemes.

Figure 3 shows the segment information belonging the 1st cluster and the 42-th cluster. The segments belonging to the 1st cluster have the largest number of /a/ phonemes as 1,310, followed by /wa/ phonemes as 774. The segments belonging to each cluster have similar spectral patterns and lengths. The segments belonging to the 42-th cluster also have the largest number of segments corresponding to /a/ phoneme. However, the spectral pattern and the length were different from the segments of the /a/ phoneme belonging to the 1st cluster.

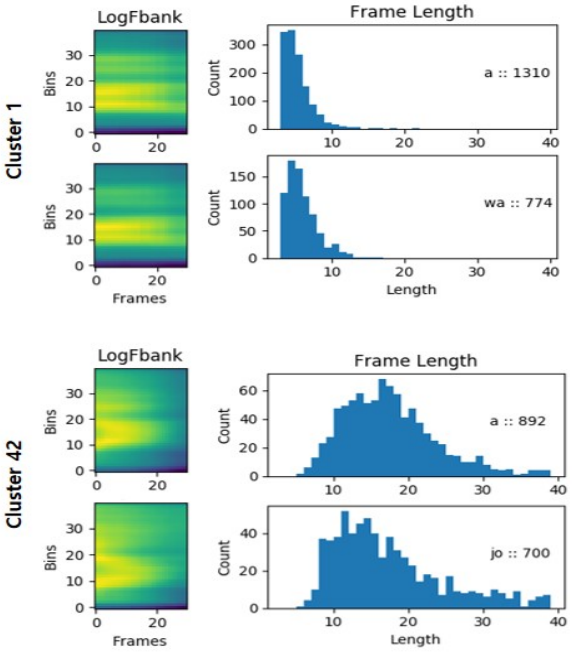

Figure 3: *Segment information of the 1st cluster and the 42-th cluster.*

We observed that segments with similar acoustic patterns clustered together. Moreover, we have also found a cluster of phonemes that are thought to be acoustically different. They mainly consist of segments of short length, whose acoustic characteristics are not apparent by phonetic reduction of spontaneous speech. As a result, we have found that each cluster has discriminability depending on the surrounding phoneme, the phoneme position in the word, and the length of the phoneme.

4. Lexicon Update

We update a phoneme-based lexicon into a new lexicon based on the proposed unit. To do this, we compute the cluster-index for each segment of 1,000 hours from the previously trained k-means model. Then, we update the lexicon by selecting the cluster-index sequence of each word with the highest conditional probability.

Figure 4 shows a histogram of each cluster-index for the phoneme string [g a] of the Korean word '가'. Here, '<bow>' and '<eow>' are dummy symbols that indicate the beginning and end of a word. The first /g/ phoneme has the highest frequency at cluster-index 6 for '<bow>-g+a' considering adjacent phonemes. Similarly, phoneme /a/ has the highest frequency at cluster-index 16 for 'g-a+<eow>' considering adjacent phonemes.

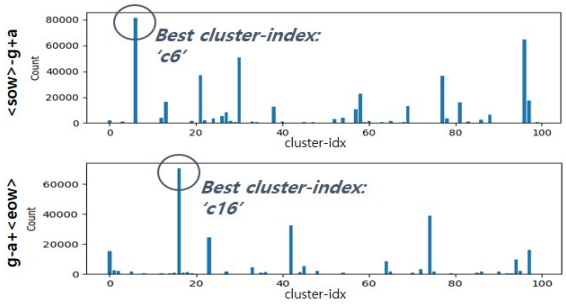

Figure 4: Histogram of each cluster-index for the phoneme sequence [g a].

We search for the cluster-index sequence (C) with the highest conditional probability given the phoneme string (P) of the word as shown in Eq. (1). Here, we have constraints on the context length as shown in Eq. (2) to avoid estimating the conditional probability at low frequency phonemes.

$$logP(\boldsymbol{C}|\boldsymbol{P}) = \sum_{i=1}^{n} logP\left(C_{i-1}, C_i|\widetilde{P}_i\right) - \sum_{i=1}^{n} logP\left(C_{i-1}|\widetilde{P}_i\right) \quad (1)$$

$$\widetilde{P}_i = \begin{cases} P_i^{tri} & if \quad \#\{P_i^{tri}\} > \alpha \times \#\{P_i^{bi}\} \\ P_i^{bi} & elif \ \#\{P_i^{bi}\} > \alpha \times \#\{P_i^{uni}\} \\ P_i^{uni} & else \end{cases} \quad (2)$$

In Eq. (1), the $P\left(C_{i-1}, C_i|\widetilde{\boldsymbol{P}_i}\right)$ is the probability that a segment with new population numbers C_{i-1} and C_i appears at (i-1)-th and i-th, given the i-th context-dependent phoneme ($\widetilde{P}_i$) of a word of n length. In Eq. (2), the $\#\{P_i^{tri}\}$ means the number of segments in the context-dependent unit considering both left and right phonemes of the i-th phoneme such as tri-phone, and the $\#\{P_i^{bi}\}$ means the number of segments in the

context-dependent unit considering the left phoneme. The threshold value (α) is set to 0.2, which is set through the preliminary experiment.

Figure 5 is an example of lexicons constructed with the phoneme unit and the proposed unit. The phoneme-based lexicon uses the same phoneme of /a/. On the other hand, the proposed lexicon uses the extended units depending on the surrounding context.

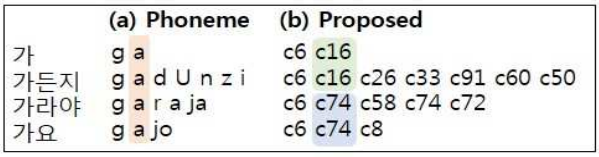

Figure 5: *Example of lexicons constructed with the phoneme unit and the proposed unit.*

5. Experiments and Results

5.1. Experiment Setup

All speech recognition experiments were conducted using the Kaldi toolkit [13]. The input features were equivalent to the globally mean-variance normalized 40-dimensional log-Mel filter-bank features, spliced for ± 2 frames.

The acoustic model used was LSTM-Projection (LSTMP) [14] trained by layer-wise back-propagation supervised with a 3-state left-to-right hidden Markov model. The LSTMP model has 3 layers, where each layer has 1,024 memory cells and 256 hidden nodes in the projection and the output layer with about 8,000 nodes using the softmax activation function. The remaining parameters are set to the default values provided by Kaldi nnet3 scripts [15]. The language model was trained with Kneser-Ney discounting (cut-off 0-3-3) using the SRILM toolkit [16], and included 81 k unigrams, 9.8 M bigrams, and 14 M trigrams from broadcast subtitles excluding the evaluation data set. The decoder was set to have an acoustic weight of 0.125, a beam size of 10.0, and a lattice beam of 5.0. The recognition results were compared by calculating the word error rate (WER) in the decoding unit.

We used about 200 hours of Korean broadcast data as acoustic model training data. The evaluation data set uses 3.5 hours of the read data ('*Read*') composed of broadcast news genres and 7 hours of the spontaneous data ('*Spont.*') recorded directly in the laboratory environment. All evaluation data sets were not used for training acoustic model and building the proposed unit.

5.2. Speech Recognition Experiments

The speech recognition experiments were performed in two steps. We first compared the performance of the grapheme-unit and phoneme-unit. Then, we checked the performance of the proposed unit according to the varying number of clusters (k).

The Korean language has 40 phoneme-units and 52 grapheme-units. Here, the grapheme units are obtained by converting Korean syllables into Roman characters. Table 1 shows the speech recognition performance of the phoneme units and the grapheme units for both read data and spontaneous data.

Table 1: *Word error rate of the phoneme unit and the grapheme unit.*

	Grapheme	**Phoneme**
Read	15.3	14.1
Spont.	46.1	46.0
#Units	52	40

In the read data, the phoneme-units performs 7.8% better relatively than the grapheme-units. Compared to other languages [17], there is little difference between grapheme and phonemic units. This is because Korean language is a phonogram that basically represents speech signal as a symbol. On the other hand, in the spontaneous data, the phoneme unit has similar performance to the grapheme unit, since the inter-phoneme distance becomes close and the intra-phoneme variance increases.

We investigated the speech recognition performance by increasing the number of clusters (k). Table 2 below shows the performance of the proposed unit by increasing the number of clusters from 80 to 200. As the final phone set for acoustic model training, we chose 180 clusters which yielded the lowest word error rate. As a result, we improved speech recognition accuracy by 4.3% for read speech recognition and 4.1% for spontaneous speech recognition compared to the phoneme-based unit.

Table 2: *Results of the proposed unit according to the varying number of clusters.*

	80	**100**	**120**	**140**	**160**	**180**	**200**
Read	14.0	13.6	13.8	13.5	13.7	**13.5**	13.5
Spont.	45.7	44.9	45.0	44.5	44.2	**44.1**	44.3

6. Conclusions

This paper proposed a method for expanding a phone set using a large amount of Korean broadcast data to improve the performance of spontaneous speech recognizer. In the process, we visualized and compared the fixed-length latent vectors obtained by 3 different methods. In addition, we clustered the latent vectors by distance and then analyzed their average spectral patterns and the length distribution of the segments. As a result, we justified that the proposed unit is discriminative according to the surrounding context, the position of the phonemes in the word, and the phoneme length.

We plan to use the entire broadcast speech data to train the acoustic model and verify the performance of the extended unit derived from the grapheme unit rather than the phoneme unit. We also plan to investigate a clustering method where the number of clusters is automatically determined, and a new lexicon update method where multiple cluster sequences are exploited to model multiple pronunciations.

7. Acknowledgements

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korea government [19ZS1140 Development of Core Conversational AI technologies].

8. References

[1] T. N. Sainath, R. Prabhavalkar, S. Kumar, S. Lee, A. Kannan, D. Rybach, V. Schoglo, P. Nguyen, B. Li, Y. Wu, Z. Chen, and C.-C. Chiu, "No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models," *Proc. ICASSP*, 2018, pp. 5859-5863.

[2] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proc. Human Language Technology*, 1994, pp. 307-312.

[3] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Communication*, vol. 46, no. 2, pp. 171–188, 2005.

[4] K.-N. Lee and M. Chung, "Modeling cross-morpheme pronunciation variations for Korean large vocabulary continuous speech recognition," *Proc. EUROSPEECH*, 2003, pp. 261-264.

[5] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech and Language*, vol. 22, no. 2, pp. 171–184, 2008.

[6] J.-U. Bang, M.-Y. Choi, S.-H. Kim, and O.-W. Kwon, "Improving Speech Recognizers by Refining Broadcast Data with Inaccurate Subtitle Timestamps," *Proc. INTERSPEECH*, 2017, pp. 2929-2933.

[7] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–129, 2002.

[8] V. Mitra, D. Vergyri, and H. Franco, "Unsupervised Learning of Acoustic Units Using Autoencoders and Kohonen Nets," *Proc. INTERSPEECH*, 2016, pp. 1300-1304.

[9] Y. A. Chung, C. C. Wu, C. H. Shen, H. Y. Lee, and L. S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," *Proc. INTERSPEECH*, 2016, pp. 410-415.

[10] The TensorBoard repository on GitHub. http://github.com/tensorflow/tensorboard.

[11] D. L. Davies, and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224-227, 1979.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. Mathematical Statistics and Probability*, 1967, pp. 281-297.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," *Proc. ASRU*, 2011.

[14] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," *Proc. INTERSPEECH*, 2014, pp. 338-342.

[15] The Kaldi toolkit repository on GitHub. https://github.com/kaldi-asr/kaldi/blob/master/egs/swbd/s5c/local/nnet3/run_lstm.sh.

[16] A. Stolcke, "SRILM-an extensible language modeling toolkit," *Proc. INTERSPEECH*, 2002, pp. 901-904.

[17] M. Killer, S. Stuker, and T. Schultz, "Grapheme based speech recognition," *Proc. EUROSPEECH*, 2003, pp. 3141-3144.