ClovaCall: Korean Goal-Oriented Dialog Speech Corpus for Automatic Speech Recognition of Contact Centers

Jung-Woo Ha[1,*] *Kihyun Nam*[1,2*], *Jingu Kang*[1], *Sang-Woo Lee*[1], *Sohee Yang*[1], *Hyunhoon Jung*[1], *Hyeji Kim*[1], *Eunmi Kim*[1], *Soojin Kim*[1], *Hyun Ah Kim*[1], *Kyoungtae Doh*[1], *Chan Kyu Lee*[1], *Nako Sung*[1], *Sunghun Kim*[3]

[1]Clova AI Research, NAVER Corp., [2]Hankuk University of Foreign Studies, [3] The Hong Kong University of Science and Technology

jungwoo.ha@navercorp.com

Abstract

Automatic speech recognition (ASR) via call is essential for various applications, including AI for contact center (AICC) services. Despite the advancement of ASR, however, most publicly available call-based speech corpora such as Switchboard are old-fashioned. Also, most existing call corpora are in English and mainly focus on open domain dialog or general scenarios such as audiobooks. Here we introduce a new large-scale Korean call-based speech corpus under a goal-oriented dialog scenario from more than 11,000 people, i.e., ClovaCall corpus. ClovaCall includes approximately 60,000 pairs of a short sentence and its corresponding spoken utterance in a restaurant reservation domain. We validate the effectiveness of our dataset with intensive experiments using two standard ASR models. Furthermore, we release our ClovaCall dataset and baseline source codes to be available via https://github.com/ClovaAI/ClovaCall.

Index Terms: ClovaCall, Korean call speech corpus, automatic speech recognition, goal-oriented dialog utterance

1. Introduction

Call-based customer services are still prevalent in most online and offline business. In particular, call centers have played a crucial role in most business domains for a few decades and recently extended to contact centers which provide additional functions such as email, VoIP, and text chatting[1]. However, the increasing costs and the harsh working environments of contact centers have brought the necessity to apply artificial intelligence (AI) to contact center operation [1]. AI for contact center (AICC) is an AI agent that communicates with human customers via call, which rapidly increases in B2B markets [1]. Since AICC is based on a telephone environment, automatic speech recognition (ASR) via call is essential for successful AICC operation.

ASR has been one of the tasks remarkably improved by deep learning since early years of 2010s [2, 3]. It is well known that the improvement of ASR results from large-scale speech corpora including Wall Street Journal [4], TIMIT [5], Switchboard [6], CallHome [7], and Librispeech [8] datasets. However, most publicly available call speech corpora are very old-fashioned such as Swichboard, Wall Street Journal, and CallHome because they were released more than 20 years ago. Also, the language of most call corpora is mainly English, and thus call corpora of low resource languages are very scarce. The other issue is is that the utterances of most corpora are based on open domains such as day-life conversation and audiobook contents. Even if small numbers of Korean speech corpora are publicly available such as AIHub [2] and Zeroth project[3], they contain general open domain dialog utterances. Therefore, ASR models trained from these speech data generally show poor recognition performance when applied to domain-specific tasks due to the differences in their data distribution and vocabularies. In particular, AICC requires an accurate ASR model to ensure the precise intent classification or slot extraction [9] from user natural language utterances.

Here we release a new large-scale Korean call speech corpus containing goal-oriented dialog utterances under a restaurant reservation scenario, i.e., ClovaCall speech corpus. The proposed ClovaCall includes 61,000 pairs of short sentences and their utterances recorded via call by more than 11,000 people. In specific, the number of unique sentences is 8,990, and all of them are natural language questions and answers which frequently appear when making reservations. The utterances that each subject read given sentences aloud are recorded over a phone. Because most sentences are designated for reservation and short with at most 10 seconds, our dataset does not suffer from end point detection and alignment problems, dissimilar to Librispeech which is extracted from audiobooks. ClovaCall can be useful for diverse AICC-based reservation services because most words and expressions prevalent in reservations are commonly used regardless of its application domains, including time, people, date, and location.

We demonstrate the effectiveness of the proposed ClovaCall with extensive experiments. We employ two standard ASR models such as Deep Speech 2 (DS2) [10] and Listen, Attend and Speech (LAS) [11] under three training schemes including pretraining-finetuning, from-scratch training, and scratch training with data augmentation. Besides, we use two additional datasets for effective verification. One is an in-house Korean call-based goal-oriented dialog speech corpus on questions and answers for daily company life (QA Dataset) for verifying the necessity of task-specific data. The other is a large-scale Korean open domain dialog speech corpus from AIHub, an online Korean data hub site, for pretraining ASR models. Experimental results show the ASR models trained from large-scale open domain data only provide very poor recognition performances. Thus, the task-specific speech datasets are essential for speech recognition of goal-oriented dialogs like AICC. Interestingly,

The first two authors equally contributed to this work.

[1]https://aircall.io/blog/call-center/contact-center-vs-call-center/

[2]http://www.aihub.or.kr/aidata/105

[3]https://github.com/goodatlas/zeroth

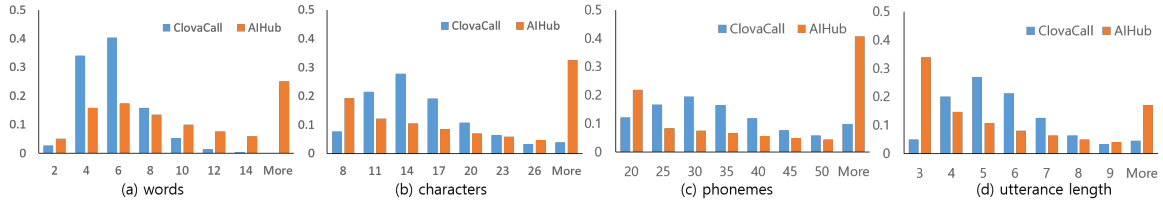

Figure 1: *Distributions of four attributes on the raw set of ClovaCall dataset (goal-oriented) with blue bars compared to those of AIHub dataset (open domain) with orange bars. We can find the different patterns between two datasets for all attributes.*

pretraining with open domain data remarkably improves the ASR accuracy compared to scratch training with task-specific data only even though their sampling rates and frequently used words are different from each other.

2. Related Work

Large-scale speech corpora publicly available allows ASR models to be applied to many valuable real-world applications. Early public speech corpora were released in 1990s, including Wall Street Journal [4], TIMIT [5], Switchboard [6], and CallHome [7]. These datasets are still prevalent as benchmark datasets for evaluating ASR models [10, 12, 13, 14]. More recently, Librispeech [8] is the most popular benchmark speech corpus on which the latest state-of-the-art ASR models are evaluated [15, 16, 17, 18]. Despite their usefulness, existing speech corpora mainly deal with general open domain dialogs. Even if the large-scale corpora are helpful for pretraining ASR models, the models not finetuned with task-specific data are likely to provide poor recognition accuracy when applied to recognize user utterances in goal-oriented scenarios such as call centers and reservations services (See Sec. 4.2). This poor performance results from the distribution difference between open domain and task-specific goal-oriented dialogs. However, compared to open domain dialog speech corpora, goal-oriented speech corpora are rarely released publicly.

3. Clova Call Speech Corpus

3.1. AI for Contact Center

ClovaCall dataset construction is one of main subtasks in *AI for Contact Center (AICC)* project [4] of NAVER Clova [19]. The goal of AICC is to develop an AI agent which can help human contact center employees to communicate with customers via phone. In perspective of technology, the main functionality of AICC contains ASR, natural language understanding such as intent classification and slot filling, goal-oriented style dialog management, response generation, and voice synthesis. Here, we focus on its ASR component and construct a large-scale speech corpora concentrating on a restaurant reservation scenario.

3.2. Data Construction from Humans

ClovaCall contains 60,746 utterance and short sentence pairs on the restaurant reservation scenario via call. The process of data construction was carried out in the following order: 1) making a sentence pool, 2) call-based recording utterances with the sentences, and 3) refining the recorded speech data.

Sentence pool. We utilized Crowdworks[5], a Korean crowd sourcing platform, to make a pool of candidate sentences. First, we defined 10 categories, 86 intents, and 7 multi-turn situations for restaurant call scenarios. 10 categories, which are high-level topics, include `reservation`, `delivery`, and 8 FAQ categories like `working time`, `menu`, and `discount`. 86 intents belong to one of 10 categories and contain `whether the restaurant is opened now`, `closing time`, `recommended menu`, etc. We also defined 7 multi-turn situations, which could be appear in a call to restaurants, include `reservation change` and `delivery call`, etc. The crowd-workers were asked to imagine and generate multiple interrogative or answer sentences for given intents and situations. After quality assurance process was performed by human experts, 8,990 sentences, which are mainly answer sentences, were selected to comprise the candidate pool by eliminating duplicated sentences.

Values	Words	Chars	Phonemes	Utter. time + silence
Voc size	4,704	613	53	-
Mean	4.39	13.79	32.39	2.94s +2.57
Stdev	1.99	5.50	12.99	1.77s +0.79
Max / Min	17 / 1	48 / 3	116 / 5	30s / 0.3s + 0 / 0.7

Table 1: *Statistics of four attributes of ClovaCall-Base dataset. silence means including silence regions in utterances*

Call-based recording utterances. Utterance recording was performed based on crowd sourcing, operated by ourselves. 10 unique sentences are given to each crowd-worker. The crowd-worker reads each of the sentences aloud once or twice via call to make at most 20 utterances, which were transmitted into our server. From 11,000 people, we gathered more than 120k pairs of short sentences and utterances. Compared to Librispeech, there do not exist end point detection and alignment problems in our data because the utterances are short enough considering call-based reservation scenarios. Besides sentences, each utterance also has its anonymous speaker index as one of the labels. This allows our data to be useful for speaker identification task.

Refining data. Data gathered via crowdsourcing are likely to contain many noises, and thus it is essential to refine the gathered data. First, we carried out qualitative evaluation on the gathered data, which was performed by human experts engaged in CrowdWorks so that we could select a total of 82,306 utterance-sentence pairs. This is the *raw* version of ClovaCall-Full. Next, we removed the starting or the ending silence regions below a specific energy level in the raw waveform of utterances. We used `Librosa` [20] with 25db as the threshold for silence elimination. The silence-free data is called *clean* version. Finally, we selected top-30 intents with

[4] https://clova.ai/aicontactcenter
[5] https://www.crowdworks.kr/

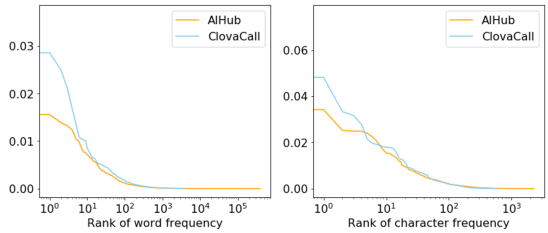

Figure 2: *Comparison between ClovaCall and AIHub datasets in terms of frequency rank-ratio of each word and character. Left and right graphs are for word and character.*

ClovaCall	Ratio	Size	No	Rank in AIHub
Words	0.1	381	56	28,979
	0.25	952	230	37,664
	0.5	1,904	662	45,459
	1.0	3,808	1,714	52,065
Chars	0.1	61	0	93.7
	0.25	151	0	162.8
	0.5	302	0	227.3
	1.0	603	3	388.3

Table 2: *Usage patterns of top frequent words and characters of ClovaCall in AIHub dataset. Ratio and size denote the ratio and the number of the top ranked words and characters. No means the number of words and characters not used in AIHub. Rank in AIHub is the average rank in AIHub. The unique word and character sizes are different from Table 1 because we discarded numerical characters. The sizes of unique words and characters of AIHub are 390,599 and 2,268.*

the most utterance-sentence pairs to be a dataset containing 60,746 pairs. We call the dataset ClovaCall-Base. We release the *clean* version of ClovaCall-Base via https://github.com/clovaai/ClovaCall. From now, ClovaCall denotes to the *clean* version of ClovaCall-Base for convenience.

3.3. Analysis on ClovaCall Dataset

To show the difference from open domain speech corpus, we compared our dataset with AIHub dataset, the largest Korean open domain dialog corpus. Fig. 1 illustrates the frequency histograms of words, characters, phonemes, and utterance length. Overall, most sentences in ClovaCall include more than 4 and less than 8 words , and more than 11 and less than 20 characters. Thus, the length of most utterances is more than 4 and less than 10 seconds. These distributions reveal the characteristics of restaurant reservation scenario. Compared to those of AIHub, the frequencies of each attribute are more concentrated on a specific region. We conjecture this pattern results from that most utterances in ClovaCall are likely to contain information for reservation while open domain dialog covers much more diverse topics and situations including both very short response utterance such as *"Yes"* and *"Sure"*, and long utterances. Table 1 depicts the number of unique elements, mean, standard deviation, maximum and minimum values for word, character, phoneme, and utterance length. Interestingly, the mean values of utterance length and silence time are very similar, which is caused by call-based recording setup.

Fig. 2 depicts the difference of frequency ratio of each word and character frequency rank. As shown in Fig. 2, we can find that higher ranked words and characters of ClovaCall are used much more frequently than those of AIHub. Table 2 shows the differences of usage pattern between ClovaCall and AIHub. In terms of characters, ClovaCall shows the similar pattern to AIHub considering the number of not used characters and their average rank in AIHub. However, the word usage of ClovaCall is significantly different from AIHub. Frequently used words of ClovaCall do not appear in AIHub or are rarely used with very low rank. These difference between two datasets enhances the necessity of goal-oriented dialog corpora. We show the difference causes poor accuracy of ASR models trained from AIHub data on ClovaCall dataset.

Dataset	Type	Number	Hour
ClovaCall	Training-Base	59,662	50
	Training-Full	81,222	67
	Noise Aug	406,110	337
	Test	1,084	1
QA Call	Training	80,984	83
	Noise Aug	404,920	415
	Test	10,000	12.4
AIHub	Pretraining	381,603	510
	Finetuning	80,105	100

Table 3: *Description of three used datasets for experiments. Underlined data were released as ClovaCall dataset.*

4. Speech Recognition Results

4.1. Experimental Setup

Datasets. We use two additional datasets besides ClovaCall to effectively verify the efficacy of our dataset. One is our in-house speech corpus on internal questions and answers about company lives collected via phone calls, called QA Call dataset. The other is a large-scale Korean open domain speech dialog corpus from NIA AIHub, an open data hub site of Korea Government. The AIHub speech is used for pretraining the ASR models. Also, we verify the results on ClovaCall-Full in addition to ClovaCall. While QA Call and ClovaCall are sampled with 8kHz, AIHub contains the speech voices recorded with 16kHz sampling rate. As shown in Table 3 for experiments, we separate 59,662 and 1,084 sentence-utterance pairs from ClovaCall-Base as training and test sets. The training set of ClovaCall-Full contains approximately 22,000 more pairs whose intent is excluded from ClovaCall-Base. There is neither duplicated speaker nor sentence between two separated sets. For QA Call, we extract the same size of sentence-utterance pairs as ClovaCall as the training set. More data were used as test set of QA Call for robust evaluation. For fair comparison, the augmented amount is similar to the pretraining data of AIHub, which is explained in the next section. In addition, the finetuning data size of AIHub is equal to the training data size of two goal-oriented datasets.

Training schemes. For verifying the effectiveness of ClovaCall and the necessity of task-specific speech corpus, we employ three training scenarios: 1) pretraining and finetuing, 2) training from scratch, and 3) training from scratch with data augmentation. AIHub dataset is used for pretraining. Also, almost the same amount of AIHub data to the training portion of ClovaCall and QA Call was used for finetuning to investigate whether call-based goal-oriented utterance data are essential or not for task-specific services. We verify the effectiveness of pretraining with open domain speech corpora by comparing the results to those trained from data enhanced by two data augmentation methods such as noise augmentation and specaugment [17]. We augmented data with noises using our in-house room simulator by adding different types of noise and reverber-

ations [21] that we obtained from daily environmental recordings. We did not perform experiments with a language model to enhance the ASR accuracy because we mainly focus on the efficacy our dataset in goal-oriented scenarios.

Data preprocessing. First, we upsample the 8Khz waveform datasets to 16Khz so that all datasets have the same frequency resolution. The reason of using upsampling instead of downsampling is that we assume an environment where conventional ASR models for 16KHz sampling rate are used for recognizing call-based speech signals. All models use log-spectrograms as input data, which are calculated with 20ms window size and 10ms stride size using `Librosa`. In addition, all spectrograms were normalized by instance-wise standardazation.

ASR Models. We use two standard ASR models such as DS2 [10] and LAS [11] for verifying the effectiveness of our proposed ClovaCall. DS2 consists of a CNN and an RNN. In our setting, the CNN module has two 2D-Convolutional layers with 32 channels, which reduce both the frequency and the time resolution of the input spectrogram with stride 4 and 2 for each layer. The RNN module consists of five bidirectional LSTM layers. All these layers have 800 hidden units per direction, in total, 1600 units per layer. Next, one fully connected layer outputs the softmax distribution over characters. Finally, DS2 is trained with CTC loss [22]. More details of DS2 are described in [10]. LAS is a sequence-to-sequence model consisting of an encoder, decoder, and attention. The encoder includes a CNN module and an RNN module sequentially. The CNN module is identical to that of DS2. The RNN module of LAS encoder consists of three stacked bidirectional LSTMs with 512 units per direction. The decoder has two unidirectional LSTMs with 512 units and one fully connected layer to predict the character probability distribution. The attention learns the alignment between the encoder outputs and the decoder hidden states. Location-aware attention [23] is employed for the attention context of the previous state. All the experiments are performed based on NAVER Smart Machine Learning (NSML) platform [24, 25].

Metrics. We use character error rate (CER) as a metric:

$$D = Distance_{LEV}(X, Y),\ CER(\%) = \frac{D}{L} \times 100$$

where X, Y are a predicted and a ground truth scripts. The distance D is the Levenshtein distance between X, Y [26] and the length L is a length of ground truth script Y.

4.2. Comparison Results on Datasets

Our experiments focus on verifying the effectiveness of task-specific speech corpora for a certain AICC services. Table 4 depicts the results of two popular ASR models under the three training scenarios described in Sec 4.1. In the pretraining and finetuning scheme, despite the largest size of the general domain dataset, AIHub, the performance of ASR models trained from only AIHub is very poor. We conjecture this poor performance results from the differences between open domain and goal-oriented dialog datasets as shown in Fig. 1. On the other hand, when pretrained with AIHub and finetuned with QA Call or ClovaCall, both models show remarkable improvement. This supports the necessity of using task-specific data for ASR models in real-world goal-oriented services.

In from-scratch training, both DS2 and LAS perform much better in the same domain than the different domain. Because, if a domain shifts, its data distribution and vocabulary also change. Moreover, QA Call provides more stable and better ASR performance than ClovaCall as well as in pretraning-finetuning scheme. We conjecture that these results are from larger size of QA Call testset. Also, QA Call contains a little more vocabulary and topics even though both speech corpora belong to goal-oriented dialog category.

Models (Parameters)	DS2 [10] (56M)		LAS [11] (31M)	
Data	QA	CC	QA	CC
Pretraining and finetuning				
$A^{pt} \rightarrow A^{ft}$	54.6	59.5	62.3	69.2
$A^{pt} \rightarrow QA$	12.2	25.6	**10.9**	26.7
$A^{pt} \rightarrow CC$-Base	35.9	9.54	38.7	8.0
$A^{pt} \rightarrow CC$-Full	32.9	8.31	35.3	**7.0**
From-scratch training				
QA	15.2	34.7	15.3	40.0
CC-Base	75.2	16.7	87.7	22.1
CC-Full	62.3	11.4	76.7	15.1
From-scratch training with data augmentation				
QA /w NA	16.5	38.5	14.8	38.3
CC-Full /w NA	64.4	10.7	81.4	18.9
QA /w SA	16.5	39.7	17.1	43.5
CC-Full /w SA	63.4	10.1	88.3	31.1

Table 4: *CER of each ASR model on two datasets under three training schemes. A, QA, CC denote AIHub, QA Call, ClovaCall datasets, respectively. pt and ft mean the data for pretraining and finetunig. NA and SA are noise augmentation [21] and specaugment [17] on ClovaCall-Full.*

In data augmentation experiments, no meaningful gain was found. We conjecture that two Call datasets have already been distorted by noises in the recording stage. In particular, the poor result of LAS on ClovaCall with SA is likely that too enhanced noise-based regularization harms the model capability of LAS with smaller parameter size, even if overall performance patterns of both models are similar to each other.

These results confirm that task-specific speech corpora play a crucial role in improving ASR models for real-world goal-oriented dialog services such as AICC. Therefore, we expect that our ClovaCall can considerably contribute to call-based reservation services. In addition, we can find that it is required to learn effective representation by pretraining with general-domain data to improve task-specific ASR models as well.

5. Concluding Remarks

We release a large-scale Korean goal-oriented dialog speech corpus, i.e., ClovaCall, which is useful for AI for Contact Center (AICC) services. To the best of our knowledge, our dataset is the first Korean goal-oriented dialog speech corpus. Our ClovaCall contains 60,746 short sentence and utterance pairs under a restaurant reservation scenario. We verify the effectiveness of ClovaCall under three training schemes such as pretraining-finetuning, from-scratch learning, and data augmented from-scratch learning with two additional speech corpora. Experimental results support ClovaCall remarkably improves the performance of ASR models, thus being crucial for call-based restaurant reservation services. Furthermore, our ClovaCall can contribute to ASR models for diverse call-based reservation services, considering that many reservation services share common expressions such as working time and availability.

Our contribution might be considered to be marginal because Korean is not a major language. On the contrary, opening low-resource language speech corpora to the public can enhance research diversity. ClovaCall is a goal-oriented dialog speech, which is rare in even major languages, thus contributing to designing goal-oriented speech corpora in various languages.

6. References

[1] P. Delgado-Martínez, A. Nieto-Hernández, V. Pérez-Mira, F. Vidal-Barrero, L. Onieva, and J. A. Carrillo-Castrillo, "Emerging risks in contact center sector: A review," in *Occupational and Environmental Safety and Health II*. Springer, 2020, pp. 765–773.

[2] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 366–369.

[3] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 14–22, 2011.

[4] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[5] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[6] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium, Philadelphia*, vol. 926, p. 927, 1997.

[7] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[9] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," *arXiv preprint arXiv:1902.10909*, 2019.

[10] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[11] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[12] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[13] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[14] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *International Conference on Learning Representations*, 2020.

[15] K. J. Han, R. Prieto, K. Wu, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions," *arXiv preprint arXiv:1910.00716*, 2019.

[16] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Rwth asr systems for librispeech: Hybrid vs attention-w/o data augmentation," *arXiv preprint arXiv:1905.03072*, 2019.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[18] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7694–7698.

[19] H. Jung, H. Kim, and J.-W. Ha, "Understanding differences between heavy users and light users in difficulties with voice user interfaces," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020, pp. 1–4.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[21] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[23] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[24] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim *et al.*, "Nsml: Meet the mlaas platform with a real-world case study," *arXiv preprint arXiv:1810.09957*, 2018.

[25] N. Sung, M. Kim, H. Jo, Y. Yang, J. Kim, L. Lausen, Y. Kim, G. Lee, D. Kwak, J.-W. Ha *et al.*, "Nsml: A machine learning platform that enables you to focus on your models," *arXiv preprint arXiv:1712.05902*, 2017.

[26] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.