Evaluating Automatic Speech Recognition Systems for Korean Meteorological Experts

ChaeHun Park **Hojun Cho** **Jaegul Choo**
KAIST AI
{ddehun,hojun.cho,jchoo}@kaist.ac.kr

Abstract

Automatic speech recognition systems often fail on specialized vocabulary in tasks such as weather forecasting. To address this, we introduce an evaluation dataset of Korean weather queries. The dataset was recorded by diverse native speakers following pronunciation guidelines from domain experts and underwent rigorous verification. Benchmarking both open-source models and a commercial API reveals high error rates on meteorological terms. We also explore a lightweight text-to-speech-based data augmentation strategy, yielding substantial error reduction for domain-specific vocabulary and notable improvement in overall recognition accuracy. Our dataset is available at https://huggingface.co/datasets/ddehun/korean-weather-asr.

1 Introduction

Meteorologists rely on vast, complex databases for forecasting, yet crafting precise SQL queries demands specialized expertise. Natural language interfaces address this gap by translating user questions into database queries (Zhong et al., 2017; Kim et al., 2020; Deng et al., 2022). Notably, Jo et al. (2023) developed an integrated search system for Korean weather data, enabling users to query extensive meteorological information using natural language. This system simplifies data retrieval and thereby improves operational efficiency.

Incorporating Automatic Speech Recognition (ASR) into these systems can further enhance usability by enabling voice-driven queries, a practical advantage for busy forecasters. However, off-the-shelf ASR models—often trained on general-domain or English-centric corpora—tend to misrecognize Korean meteorology terms due to both the language's agglutinative structure and specialized vocabulary (Lee et al., 2019; Cho et al., 2020; Li et al., 2021; Yadav and Sitaram, 2022; Radford et al., 2023; Ferraz et al., 2024; Song et al., 2024).

Answer	11월 상순 화순과 전라도의 풍속 평년값 보여줘 11wol sangsun hwasungwa jeonlado-ui pungsog pyeong-nyun-gab boyeojwo (Show the average wind speed of Hwasun and Jeolla-do in early November)
Zero-shot	11월 상순 화순과 전라도의 ~~풍석평령값~~ 보여줘 11wol sangsun hwasungwa jeonlado-ui ~~pungseok pyeong-nyeong-gab~~ boyeojwo
Fine-tune w/ General	11월 상순 화순과 전라도~~애~~풍석평령값 보여줘 11wol sangsun hwasungwa jeonlado-~~ae pungseog pyeong-nyun-gab~~ boyeojwo
Fine-tune w/ Weather	11월 상순 화순과 전라도의 풍속 평년값 보여줘 11wol sangsun hwasungwa jeonlado-ui pungsog pyeong-nyun-gab boyeojwo
Answer	일기도 앙상블 EPS 평균 편차 KIM 동아시아 500hpa 고도 ilgido angsangbeul ipieseu pyeong-gyun pyeoncha KIM dong-asia obaeg-hectopascal go-do (Weather Map Ensemble EPS Average Deviation KIM East Asia 500 hpa Elevation)
Zero-shot	일기도 앙상블 EPS 평균 편차 ~~큄~~ 동아시아 500~~hp 스카왈~~ 구도 ilgido angsangbeul ipieseu pyeong-gyun pyeoncha KIM dong-asia obaeg ~~hp seukawaol~~ go-do
Fine-tune w/ General	~~4귀도~~ 앙상블 EPS 평균 편차 ~~큄~~ 동아시아 500 ~~쾌도~~ 파스칼 구도 ilgido angsangbeul ipieseu pyeong-gyun pyeoncha KIM dong-asia obaeg-~~kwaedo~~pascal go-do
Fine-tune w/ Weather	일기도 앙상블 EPS 평균 편차 KIM 동아시아 500hPa 구도 ilgido angsangbeul ipieseu pyeong-gyun pyeoncha KIM dong-asia obaeg-hectopascal go-do

Figure 1: **Qualitative comparison of different models for Korean meteorological queries.** All models are based on `Whisper large-v2` (Radford et al., 2023). The models compared include `Zero-shot`, fine-tuned on a `General` domain dataset (i.e., KsponSpeech (Bang et al., 2020)), and fine-tuned on a `Weather` domain dataset synthetically generated by a text-to-speech API. Wrongly predicted words are manually highlighted in ~~strikethrough~~ by the authors.

To quantify these challenges, we created an evaluation set of 5,500 Korean weather queries. Eleven native speakers recorded queries sourced from Jo et al. (2023), following expert pronunciation guidelines and a thorough manual validation process to guarantee correctness. We then benchmarked several popular multilingual ASR models (e.g., Whisper variants) and observed consistent misrecognitions of domain-specific terms—such as *weather map*, *average wind speed*, *KIM* (*Korean Integrated Model*), and units like *hPa*—even after fine-tuning on a large general-domain Korean speech corpus (Bang et al., 2020) (Fig. 1). We also explored lightweight approaches on our dataset, including TTS-based data augmentation (Zheng et al., 2021) and LLM-based post-processing (Hu et al., 2024a).

Our contributions are as follows: (1) We construct a domain-specific ASR evaluation dataset by

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10619–10627
November 4-9, 2025

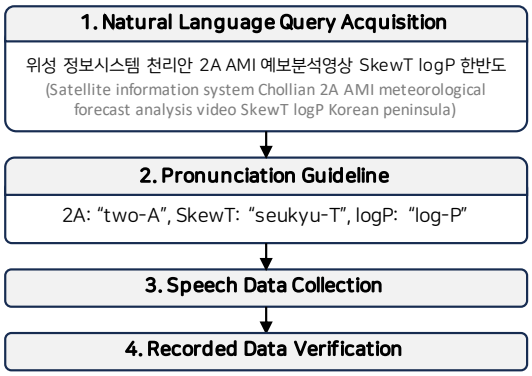

Figure 2: A construction pipeline of ASR evaluation dataset for Korean meteorological domain.

recording and validating weather-related queries from native Korean speakers; (2) we evaluate various ASR configurations and highlight the need for domain-specific adaptation; and (3) we explore TTS-based data augmentation to improve recognition of specialized meteorological terms.

2 Related Work

Speech Recognition for Specialized Domains Speech recognition technology has been extensively studied in various specialized domains, such as medical (Le-Duc, 2024) and financial fields (O'Neill et al., 2021). One major challenge is accurately handling domain-specific terminology. Zheng et al. (2021) demonstrated that synthetic speech data enhances the recognition of out-of-vocabulary words in domain-specific contexts. Shamsian et al. (2024) propose a keyword-guided adaptation to improve the recognition accuracy of models for specialized terms. Our research specifically focuses on the Korean meteorological domain, creating an evaluation dataset tailored to assess ASR performance in this specialized field.

Speech Recognition for Korean Research on ASR for non-English languages, including Korean, presents unique challenges. Multilingual ASR models, such as Whisper (Radford et al., 2023), often underperform with languages like Korean due to unique linguistic characteristics and limited representation in training datasets (Li et al., 2021; Yadav and Sitaram, 2022; Ferraz et al., 2024; Song et al., 2024). This performance drop is often attributed to the lower proportion and diversity of non-English data in training sets. Efforts to improve ASR for Korean include datasets like KsponSpeech, a spontaneous speech corpus, and ClovaCall, a goal-oriented dialog speech corpus, provide valuable resources for developing and evaluating Korean ASR systems (Bang et al., 2020; Ha et al., 2020). We evaluate ASR performance specifically within the Korean meteorological domain, using a tailored evaluation dataset to identify and address domain-specific performance limitations.

Num. samples	5,500
Utterance time (sec)	$7.05_{2.55}$
- min / max time	0.92 / 29.98
Avg. chars / words	$24.49_{15.62}$ / $7.59_{4.34}$
Unique words	4,955
Absent ratio (%)	24.86

Table 1: **Dataset Statistics**. *Absent ratio (%)* refers to the percentage of unique words not presented in the general Korean ASR dataset (Bang et al., 2020), divided by the total number of unique words in our dataset.

3 Dataset Construction

This study aims to develop a specialized dataset for ASR systems within the Korean meteorological domain. The construction process includes several key steps: Acquiring the natural language questions, obtaining pronunciation guidelines for specialized terms from domain experts, collecting speech data from native speakers, and verifying the correctness of the recorded audio. The detailed data construction pipeline (§3.1, Fig. 2) and the analysis of the resulting dataset (§3.2) are as follows.

3.1 Data Construction Pipeline

Acquiring Natural Language Questions We used the natural language question dataset released by Jo et al. (2023), which covers diverse meteorological queries. From its URL and SQL subsets, we selected 3,575 and 1,925 questions, respectively, to ensure broad topic coverage. While the original dataset includes structured query mappings (e.g., SQL), we used only the natural language questions, focusing on how forecasters would verbally express their information needs. For more on the original dataset, see Jo et al. (2023).

Obtaining Pronunciation Guidelines from Domain Experts To ensure the correct pronunciation of domain-specific terms, we collaborated with meteorological experts. They provided detailed guidelines—for example, pronouncing `"BUFR"` as `"buffer"` and `"AMI"` as `"A-M-I"`. Additionally, terms such as `"33009 station"` could be pronounced as `"three three zero zero nine station"` or `"thirty-three thousand and nine station."` These guidelines helped speakers produce consistent and intelligible renditions of

General	thing (거) but (근데) not (안) that (그) just (그냥) I (내가) such (그런) uh (어) what (뭐)
Weather	show (보여줘) AWS (AWS) inform (알려줘) standard normals (평년값) surface (지상) highest temperature (최고 기온) ocean (해양) radar (레이더) weather chart (일기도)

Table 2: Nine most frequently occurring words in Korean ASR datasets in open-domain dialog (`General`) and weather (`Weather`) domains. The KsponSpeech (Bang et al., 2020) is used as a general domain dataset.

specialized terms. Full details are in Appendix A.

Collecting Speech Data from Native Speakers We recruited eleven native Korean speakers (7 male, 4 female, all in their twenties) to record the queries. Participants were provided with the pronunciation guidelines and detailed instructions. Each was compensated at a rate of 16,000 KRW (approx. 11.6 USD) per 100 queries. All recordings were conducted in quiet and controlled environments.

Verifying the Recorded Voices Following the collection of speech data, a rigorous verification process was implemented. Human verifiers listened to each recording to ensure accuracy and adherence to the pronunciation guidelines. They also checked for clarity and the absence of background noise or errors. Two of the authors are employed for this verification process. This process was essential to guarantee the quality and reliability of the dataset for subsequent ASR evaluation.

3.2 Dataset Analysis

Table 1 shows the overall statistics of our dataset, which includes 5,500 spoken queries. The average utterance length is 7.05 seconds, with an average of 7.59 words and 24.49 characters per query. Notably, 24.86% of the unique words in our dataset are absent from the general-domain ASR corpus (Bang et al., 2020), highlighting the specialized vocabulary of meteorological speech.

Table 2 lists the most frequent words in general-domain and weather-domain datasets. While general speech contains common conversational terms, our dataset prominently features specialized meteorological vocabulary. This contrast underscores the need for domain-specific datasets to improve ASR performance in technical applications.

4 Experimental Setup

4.1 Evaluation Datasets

We use our Korean ASR dataset for the weather domain to evaluate different ASR models. The dataset comprises 5,500 spoken queries, which we randomly split into 500 samples for the development set and 5,000 samples for the test set. Additionally, we utilize the *eval-clean* test set from the *KsponSpeech* dataset (Bang et al., 2020) to evaluate the models in a general conversation domain. This dataset consists of 3,000 audio files and their corresponding answer transcriptions.

4.2 Metrics

We use Character Error Rate (CER) and Word Error Rate (WER) as evaluation metrics to measure the distances between the prediction and ground-truth transcription. Additionally, we employ a space-normalized Word Error Rate (sWER) (Bang et al., 2020), which accounts for the flexibility and variations of space rules in Korean.

4.3 Models and Training Datasets

Zero-shot Multilingual Models For our experiments, we used the multilingual Whisper model family (Radford et al., 2023), evaluating four different model sizes: `tiny` (39M), `small` (244M), `medium` (769M), and `large-v2` (1550M). These pre-trained models were assessed in a zero-shot manner to evaluate their performance without additional fine-tuning on our dataset. The target language and task in Whisper's prefix tokens were set to *Korean* and *transcribe*, respectively.

Fine-tuning on General Open-domain Dialogues To explore the impact of adapting a multilingual ASR model to the Korean language, we fine-tuned an ASR model using the KsponSpeech (Bang et al., 2020) dataset. The dataset contains 619k training instances about open-domain dialogue utterances from Native Korean annotators. The Whisper large-v2 is fine-tuned on this dataset. More implementation details are in Appendix 4.4.

Data Augmentation for Meteorological Domain Inspired by Zheng et al. (2021), we hypothesize that teaching the model the pronunciation of specialized weather terms is crucial for accurate transcription. To this end, we used a TTS system to generate audio for 10k and 9.8k natural language queries from the URL and SQL subsets of Jo et al. (2023), respectively. Queries in our evaluation set were excluded to prevent test leakage. We employed the Google TTS service[1] and converted English words to Korean pronunciations based on expert guidelines (Section 3.1) before synthesis. The

[1] `https://gtts.readthedocs.io/`

Model	Params.	KsponSpeech$_{\text{Eval}}$ (3k)			Weather$_{\text{Dev}}$ (0.5k)			Weather$_{\text{Test}}$ (5.0k)		
		CER	WER	sWER	CER	WER	sWER	CER	WER	sWER
(1) Zero-shot Evaluation of Whisper model family (Radford et al., 2023)										
Tiny	39M	32.67	53.69	47.78	33.53	74.93	54.19	38.24	81.99	58.61
Small	244M	16.75	32.57	26.2	16.39	46.62	27.69	22.92	57.4	34.65
Medium	769M	14.33	29.73	21.21	12.65	38.67	19.31	19.26	50.38	26.54
Large-v2	1550M	14.66	29.97	20.75	11.78	36.31	19.03	18.35	46.74	25.46
(2) Whisper-Large-v2 finetuned on Different Datasets										
General	1550M	**10.48**	24.05	**16.49**	16.92	49.59	21.75	24.57	63.28	28.22
Weather*	1550M	15.81	32.48	25.28	**6.93**	**17.99**	**9.13**	**10.64**	**31.81**	**13.61**
General+Weather*	1550M	10.54	**23.94**	16.52	8.68	22.58	12.50	17.46	43.41	20.25
(3) Commercial API										
Google STT	-	-	-	-	20.38	57.09	29.45	23.45	71.89	31.10

Table 3: **ASR evaluation results.** We report CER, WER, and sWER, where lower values indicate better performance. The table consists of three model groups: (1) zero-shot Whisper models of varying sizes, (2) Whisper-large-v2 fine-tuned on different datasets, and (3) a commercial ASR system (i.e., Google STT). In Group (2), *General* refers to fine-tuning on KsponSpeech, and *Weather** denotes fine-tuning on the TTS-generated weather dataset. The lowest and second-lowest scores of each column are highlighted in bold and underlined, respectively.

resulting TTS audio was used to fine-tune Whisper large-v2, either alone or combined with KsponSpeech, yielding two ASR variants.

Post-processing with Unimodal LLMs To assess whether large language models (LLMs) can improve ASR outputs without acoustic input (Chen et al., 2023; Hu et al., 2024a), we used GPT-4o-mini (Achiam et al., 2023) to revise the top-1 transcription from Whisper-large-v2. The model was evaluated in both zero-shot and 20-shot settings, using examples from the development set. We adapted the prompt from Chen et al. (2023) to the Korean meteorological domain to better handle domain-specific terms. This setup allows us to test how effectively an unimodal LLM can refine ASR transcriptions based solely on textual input.

4.4 Implementation Details

All fine-tuned models are parameter-efficiently trained using LoRA (Hu et al., 2022) with r=32. The fine-tuning process involved 3 epochs of training with a batch size of 48, a learning rate of 1e-3, and a warmup ratio of 0.1 with AdamW optimizer (Loshchilov and Hutter, 2017). For fine-tuned models, we saved the checkpoint after each epoch and selected the best one based on the lowest CER score from the development set of our dataset. The greedy decoding (Holtzman et al., 2020) is used as a decoding algorithm for all models. All models are implemented with the Transformers framework (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). For the post-processing with unimodal LLMs, we used the following text prompt: *"Below is the best-hypotheses transcribed from speech recognition system for Korean Meteorological experts. Please try to revise it and write the response for the true transcription."*

5 Results and Analyses

In this section, we describe the key observations. Table 3 summarizes results across different setups.

Zero-shot ASR models struggle with domain-specific terminology Whisper models show high CER and WER when transcribing meteorological queries, particularly for specialized terms. Larger models (e.g., Whisper-large-v2) reduce errors relative to smaller variants but still struggle with domain-specific vocabulary, highlighting the need for further adaptation.

Fine-tuning improves general performance, but domain-specific adaptation is essential Fine-tuning Whisper-large-v2 on KsponSpeech lowers CER and WER overall, but improvements on meteorological queries are limited. This suggests that general-domain data alone is insufficient for handling specialized terms.

Synthetic data (TTS) aids domain adaptation Training with TTS-generated weather-domain queries improves transcription of meteorological terms. The model fine-tuned on both KsponSpeech and synthetic data achieves balanced performance across general (CER 10.54%) and weather domains (CER 17.46%), while the model trained only on synthetic data achieves the lowest error rates in the weather domain (CER 10.64%).

LLM-based post-processing yields inconsistent gains As shown in Figure 3, LLM-based post-processing does not consistently improve recogni-

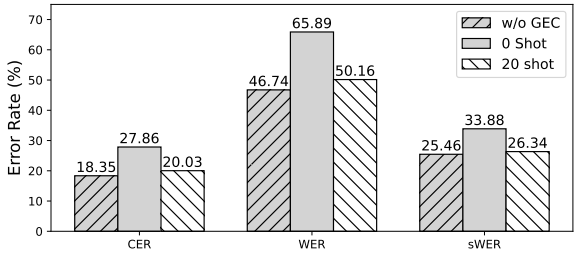

Figure 3: **Effect of LLM-based post-processing.** We compare three conditions: without LLM-based grammatical error correction (*w/o GEC*), with zero-shot (*0 Shot*), and with 20-shot (*20 Shot*) setups.

tion of domain-specific terms. Error rates remain higher than the zero-shot baseline, suggesting that text-only refinement is insufficient without acoustic adaptation. While training a dedicated GEC model or using N-best hypotheses (Chen et al., 2023; Hu et al., 2024b) may help, these approaches require large corpora and increase inference costs.

Commercial ASR systems perform poorly on specialized data Google's STT model yields the highest error rates (e.g., 20.38% on the Weather-test set), underperforming even the zero-shot Whisper-large-v2. This underscores the limitations of generic ASR systems in specialized domains and the need for targeted adaptation. Overall, these results highlight the importance of fine-tuning with domain-specific data. While LLMs can support transcription refinement, substantial improvements require acoustic-level adaptation.

Case Study To illustrate common ASR errors, we present examples from Table 4. One issue involves domain-specific term misrecognition. In a wind forecast transcription, "지상 시계열 바람장미 풍속계급별 판촉횟수 CALM 0 5m/s 순별 하순 211 인제" *("Ground-level wind rose classified by wind speed, CALM 0–5m/s, late-stage 211 Inje")*, both zero-shot Whisper and Google STT misinterpret "CALM", distorting critical numerical information. Another case involves unintended style shifts. In "8월 30일 증가압과 기온 평년값은 뭐니" *("What are the historical averages for pressure and temperature on August 30?")*, the post-processed output rephrases it more formally as "8월 30일 증가압과 기온 평년값은 무엇인가요?". While fluency improves, such shifts may be undesirable in contexts requiring a colloquial tone. These cases reflect two key challenges: (1) preserving domain-specific and numerical terms, and (2) refining grammar while maintaining intended speech style. Addressing these requires domain-aware and context-sensitive correction strategies.

Answer	8월 30일 증기압과 기온 평년값은 뭐니
ZS	8월 30일 증기압과 기~~온 평광~~값은 뭐니
$FT_{General}$	8월 30일 ~~관~~기압과 기온 평년값은 뭐니
FT_{Both}	8월 30일 증기압과 기온 평년값은 뭐니
Google STT	8월 30일 증기압과 기온 평년 값은 뭐니
LLM_{20shot}	8월 30일 증기압과 기온 평~~광~~값은 ~~무엇인가요~~
Answer	관측횟수 CALM 0 5m/s 순별 하순 211 인제
ZS	관측~~해수컴~~ 0 5m~~f~~ 순별 하순 211 인~~재~~
$FT_{General}$	관측 ~~해수 칸 영~~ 5m ~~퍼셈커~~ 순별 하순 211 인제
FT_{Both}	관측~~해서~~ CALM 0 5m~~%~~ 순별 하순 2 1 1 인제
Google STT	관측~~에서 컴~~ 0 5 m ~~버서커~~ 순별 ~~화~~순 2~~일~~ 인제
LLM_{20shot}	관측~~해~~수~~컴~~ 0~~~~5 m/s 순별 하순 211 인~~재~~

Table 4: **Qualitative Results** Incorrectly recognized characters are marked with ~~strikethrough~~, while missing words are indicated with a wavy underline. Space errors are omitted for better readability.

6 Conclusion

We introduce a specialized evaluation dataset for assessing ASR performance in the Korean meteorological domain, addressing the lack of domain-specific benchmarks. Our experiments reveal that while fine-tuning on general-domain data improves overall accuracy, specialized terminology remains a major source of error. By releasing this dataset, we aim to support further research in developing robust, domain-adapted systems that better reflect the demands of real-world forecasting scenarios.

Limitations

While our study provides a targeted benchmark for ASR in the Korean meteorological domain, several limitations remain. First, the dataset consists of scripted queries recorded in clean environments, which may not fully represent the **acoustic variability found in real-world settings**. In practice, meteorologists often speak spontaneously, with disfluencies, hesitations, or overlapping speech, especially during live broadcasts or team discussions. Our dataset does not yet capture such spontaneous or noisy conditions. Second, while we explored text-only post-processing using unimodal LLMs, we did not investigate more **advanced correction strategies with acoustic cues** or multiple ASR hypotheses. These approaches could further improve recognition of subtle or ambiguous terms but require additional data and computational resources. Lastly, our work is focused on a single domain (meteorology) and a single language (Korean). The findings may not generalize to **other specialized domains**, such as healthcare or law, or to other low-resource languages. Expanding this research

to cross-domain and cross-lingual settings remains an important direction for future work.

Ethical Statement

All spoken queries in our dataset were recorded with the informed consent of native Korean speakers, who were compensated fairly for their participation. The recordings were collected in controlled environments to ensure quality and to minimize any unintended background content. To protect speaker privacy, no personally identifiable information (PII) was included in the dataset. All utterances were manually reviewed to avoid harmful, offensive, or culturally insensitive content. While the dataset aims to represent meteorological language use in Korean, we acknowledge that it may not fully capture regional or situational variations in speech patterns.

Acknowledgement

This work is supported by a grant of a Developing Intelligent Assistant Technology and Its Application for Weather Forecasting Process (KMA2021-00123), Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10(19):6936.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and EngSiong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.

Won Ik Cho, Seok Min Kim, and Nam Soo Kim. 2020. Towards an efficient code-mixed grapheme-to-phoneme conversion in an agglutinative language: A case study on to-Korean transliteration. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 65–70, Marseille, France. European Language Resources Association.

Naihao Deng, Yulong Chen, and Yue Zhang. 2022. Recent advances in text-to-sql: A survey of what we have and what we expect. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2166–2187.

Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. 2024. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Jung-Woo Ha, Kihyun Nam, Jingu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Hyeji Kim, Eunmi Kim, Soojin Kim, Hyun Ah Kim, et al. 2020. Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, page 409.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuchen Hu, Chen Chen, Chengwei Qin, Qiushi Zhu, EngSiong Chng, and Ruizhe Li. 2024a. Listen again and choose the right answer: A new paradigm for automatic speech recognition with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 666–679, Bangkok, Thailand. Association for Computational Linguistics.

Yuchen Hu, CHEN CHEN, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EngSiong Chng. 2024b. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations*.

Jinkyung Jo, Dayeon Ki, Soyoung Yoon, and Minjoon Seo. 2023. An integrated search system for Korea weather data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 765–774, Singapore. Association for Computational Linguistics.

Hyeonji Kim, Byeong-Hoon So, Wook-Shin Han, and Hongrae Lee. 2020. Natural language to sql: Where are we today? *Proceedings of the VLDB Endowment*, 13(10):1737–1750.

Khai Le-Duc. 2024. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17365–17370.

Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee. 2019. Adversarially trained end-to-end korean singing voice synthesis system. *Interspeech 2019*.

Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W Ronny Huang, Min Ma, and Junwen Bai. 2021. Scaling end-to-end models for large-scale multilingual asr. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1011–1018. IEEE.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 1081–1085. International Speech Communication Association.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Keyword-guided adaptation of automatic speech recognition. *arXiv preprint arXiv:2406.02649*.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. Lora-whisper: Parameter-efficient and extensible multilingual asr. *arXiv preprint arXiv:2406.06619*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Terminology	Pronunciation (Korean)	Pronunciation (English)
TD	티디	T-D
AMEDAS	아메다스	A-me-da-seu
MOS	엠오에스, 모스	M-O-S, Mos
GOCI	고씨, 지케이투비	Go-ssi, G-K-2-B
KIM	킴	Kim
CAPE	케이프	Cape
900	구영영, 구공공, 구백	Gu-yeong-yeong, Gu-gong-gong, Gu-baek
APS	에이피에스	A-P-S
WTEM	원템	Won-tem

Figure 4: Samples of pronunciation rules for meteorological terminologies. The `Pronunciation (English)` are manually written by the authors for clarity and were not provided at the data annotation phase.

A Pronunciation Guideline Details for Speech Data Collection

To ensure accurate pronunciation of meteorological terms, we consulted domain experts from the National Institute of Meteorological Sciences (NIMS). The focus was primarily on English words and abbreviations, along with various unit and number expressions. We note that a single word or expression can be pronounced in different ways. Experts provided detailed pronunciation guidelines for 256 words and expressions frequently used in the Korean weather domain. These guidelines were given to annotators as an initial reference and were available throughout the recording process. Selected examples from the pronunciation guidelines are shown in Fig. 4.

B Qualitative Results

We present further prediction results of different open-source ASR models on Figures 5 and 6.

Answer: 레이더 지점 연직 BRI 백령도 PhiDP 16일 4km (Radar vertical profile at BRI station, Baengnyeongdo, PhiDP, 16th day, 4km range)			
Tiny: 레이더 지점 연직 ~~비열 아아~~ 백~~령~~도 Phi dp 16일 ~~사클로마터~~	**Small**: 레이더 지점 연직 BRI ~~100~~도 PHIDP 16일 4km	**Medium**: 레이더 지점 연직 BRI 백~~영~~도 PHIDP 16일 4km	**Large**: 레이더 지점 연직 BRI ~~1000~~도 PHIDP 16일 4km
Answer: 해양 연안방재 문숫자 229 북격렬비도 5분 파장 (Marine coastal disaster monitoring number 229 Bukgyeongryeolbido)			
Tiny: 해양 연~~한~~방~~째~~ 문~~수~~자 ~~아구~~ 북~~경~~렬 위도 ~~오븐~~ 파장	**Small**: 해양 연~~암~~방~~째~~ 문~~수~~자 2~~2~~9 북경~~열~~비도 5분 파장	**Medium**: 해양 연안방~~째~~ 문~~수~~자 2~~2구~~ 북경~~열~~비도 5분파장	**Large**: 해양연안방~~째~~ 문-숫자 229 북경~~열~~비도 5분 파장
Answer: 지상 지구대기 감시 시계열 울산 자외선 전자기유도입자계수기 L2H 시간값 (Surface global atmospheric monitoring time series Ulsan ultraviolet electromagnetic induction particle counter)			
Tiny: 지상 지구~~-~~대기 감시 시계~~엘~~ 울산 자외선 전~~작아~~ 유도 입~~장해두~~기 L2H 시간값	**Small**: 지상 지구대기, 감시 시계열 울산, 자외선 전자기유도 입장~~에~~수기 L2H 시간~~끝~~	**Medium**: 지상 지구대기 감시 시계열 울산 자외선 전자기 유도 입자 계수기 L2H 시간값	**Large**: 지상 지구대기 감시 시계열 울산 자외선 전자기유도 입자 ~~과~~수기 L2H 시간값
Answer: 1993년 이전 2월 이후 8일부터 29일까지 고성의 습도와 최심신적설을 순서대로 보여줘 (Show the humidity and maximum snow depth of Goseong in order from February 8 to 29 before 1993.)			
Tiny: ~~청구의~~ 93년 이전 2월 이후 8일부터 29일까지 고~~상~~의 습도와 최심신~~작서를~~순서대로 보여줘	**Small**: 1993년 이전 2월 이후 8일부터 29일까지 고성의 습도와 최심신적~~서~~를 순서대로 보여줘	**Medium**: 1993년 이전 2월 이후 8일부터 29일 까지 고성의 습도와 최심신적~~서~~를 순서대로 보여줘	**Large**: 1993년 이전 2월 이후 8일부터 29일까지 고성의 습도와 최~~심~~신 신적~~서~~를 순서대로 보여줘

Figure 5: **Qualitative results of pre-trained Whisper models with different model sizes.** The answer queries in English are manually translated from the original Korean answer. Incorrect model predictions are highlighted with ~~strikethrough~~.

Answer: 낙뢰 관서별 130 울진 오전 오후 반경 (Lightning occurrences by agency, Uljin, morning and afternoon radius)		
ZS: 낙~~래~~ 관서별 130 울진 오전 오후 반경	**FT (General)**: ~~낭래~~ 관서별 130 울진 오전 오후 반경	**FT (Both)**: 낙뢰 관서별 130 울진 오전 오후 반경
Answer: 8월 30일 증기압과 기온 평년값은 뭐니? (What are the average vapor pressure and temperature values on August 30th?)		
ZS: 8월 30일 증기압과 기~~운 평량~~값은 뭐니?	**FT (General)**: 8월 30일 ~~진~~기압과 기온 평년값은 뭐니?	**FT (Both)**: 8월 30일 증기압과 기온 평년값은 뭐니?
Answer: 1980년 이전 4월 이전 하순 충청남도에 가장 최대순간풍속이 강한 곳이 어디니? (Where was the highest maximum instantaneous wind speed in Chungcheongnam-do before April 1980?)		
ZS: 1980년 이전 4월 이전 ~~화~~순 충청 남도에 가장 최대 순간 풍속이 강한 곳이~~었더~~니	**FT (General)**: 1980년 이전 4월 이전 ~~화~~순 충청 남도에 가장 최대 순간 풍속이 강한 곳이~~었더니.~~	**FT (Both)**: 1980년 이전 4월 이전 하순 충청 남도에 가장 최대순간풍속이 강한 곳이 어디니?
Answer: 지상 시계열 바람장미 풍속 계급별 관측횟수 CALM 0.5 m/s 순별 하순 211 인제 (Surface time series, wind rose observation frequency by wind speed class, CALM 0.5 m/s late order 211 Inje)		
ZS: 지상 시~~괴~~열 바람장미 풍속계급별 관측 ~~해수권~~ 0.5m/s 순별 하순 211 인~~재~~	**FT (General)**: 지상 시~~괴~~열 바람 장미 풍속 계급별 관측 ~~해~~수 ~~관영~~ 5m ~~퍼셀컥~~ 순별 하순 211 인제	**FT (Both)**: 지상 시~~괴~~열 바람장미 풍속 계급별 관측~~해서~~ CALM 0~~.~~5 m% 순별 하순 2 1 1 인제

Figure 6: **Qualitative results of Whisper models finetuned on different datasets.** All indicators are the same with Figure 5.