Toward Practical Automatic Speech Recognition and Post-Processing: a Call for Explainable Error Benchmark Guideline

Seonmin Koo [1 *] Chanjun Park [1 2 *] Jinsung Kim [1] Jaehyung Seo [1] Sugyeong Eo [1] Hyeonseok Moon [1] Heuiseok Lim [1]

Abstract

Automatic speech recognition (ASR) outcomes serve as input for downstream tasks, substantially impacting the satisfaction level of end-users. Hence, the diagnosis and enhancement of the vulnerabilities present in the ASR model bear significant importance. However, traditional evaluation methodologies of ASR systems generate a singular, composite quantitative metric, which fails to provide comprehensive insight into specific vulnerabilities. This lack of detail extends to the post-processing stage, resulting in further obfuscation of potential weaknesses. Despite an ASR model's ability to recognize utterances accurately, subpar readability can negatively affect user satisfaction, giving rise to a trade-off between recognition accuracy and user-friendliness. To effectively address this, it is imperative to consider both the speech-level, crucial for recognition accuracy, and the text-level, critical for user-friendliness. Consequently, we propose the development of an Error Explainable Benchmark (EEB) dataset. This dataset, while considering both speech- and text-level, enables a granular understanding of the model's shortcomings. Our proposition provides a structured pathway for a more 'real-world-centric' evaluation, a marked shift away from abstracted, traditional methods, allowing for the detection and rectification of nuanced system weaknesses, ultimately aiming for an improved user experience.

1. Introduction

Automatic speech recognition (ASR) is a task that recognizes voice and converts it into text, and it is getting more and more attention with the development of voice interface applications and devices such as Alexa, Siri, and Cortana (Williams & Young, 2007; Wang et al., 2018; 2020). Although the performance of ASR models has improved with the development of technology, there are still things to be improved. Since speech recognition results are used as inputs for downstream tasks and affect end-user satisfaction, it is essential to accurately diagnose errors in the ASR model and improve them based on them (Serdyuk et al., 2018; Feng et al., 2022).

However, changing the structure of the ASR model itself or retraining it requires a considerable amount of data and is time-consuming and costly. Moreover, resources to train models are often scarce in the real world (Park et al., 2020). Regarding this situation, post-processing research is emerging as a role to complement the ASR model. The ASR post-processing (ASRP) task aims to correct errors by detecting them in speech recognition results and can improve performance without changing the model structure (Mani et al., 2020b; Liao et al., 2020; Leng et al., 2021). Therefore, effective diagnosis is required to enhance both ASR and ASRP models.

In the real world, there is a trade-off relationship between the ASR model's recognition accuracy and user-friendliness. Even if the ASR model accurately recognizes the input, the user's satisfaction may decrease. This is because humans do not always utter perfect sentences in the real world (e.g., incomplete utterances, sighs, etc.), so readability may deteriorate even if the ASR model accurately recognizes speech. The goal of the ASR model is to recognize the given voice input as text accurately, and the purpose of the ASRP model is to receive the recognition result as input and improve readability to increase end-user satisfaction. To achieve balanced speech recognition results in this trade-off situation, it is recommended to consider both the speech-level for recognition accuracy and the text-level for user-friendliness. To this end, accurate diagnosis and evaluation of errors must be preceded in order to consider both speech- and text-level.

[*] Equal contribution [1] Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea [2] Upstage, Gyeonggi-do, Korea. Correspondence to: Heuiseok Lim <limhseok@korea.ac.kr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

However, there are several challenges in the effective diagnosis and validation of the ASR and the ASRP models. First, the ASR model's existing evaluation methods are insufficient to diagnose the model's performance properly. Most ASR automatic evaluation processes are evaluated through quantitative metrics such as WER (Woodard & Nelson) and CER (Morris et al., 2004). However, since the metrics only deal with text, which is the result of voice recognition, there is a problem that the speech-level is not considered. Also, it is not easy to diagnose each model precisely because the existing benchmark datasets do not classify the characteristics of the collected voice data. In other words, each ASR model with different characteristics can receive the same quantitative score. However, the integrated evaluation score lacks the explanatory power to improve the model's weakness specifically. This leads to the use of human evaluation in the real world, even though many benchmark datasets have been released. Recently, attempts have been made to build data by considering the noisy environment or speaker characteristics to increase explanatory power (Sikasote & Anastasopoulos, 2022; Lakomkin et al., 2019; Gong et al., 2022; Dai et al., 2022). However, these data also lack the explanatory ability to diagnose problems in the model because the number of error types used is limited.

In addition, the existing ASRP research also uses the ASR dataset, not a separate specialized dataset for learning, so those mentioned above low explanatory capability problem remains (Sodhi et al., 2021; Kumar et al., 2022). Although a grammatical error correction (GEC) dataset in which error types are subdivided exists, it is difficult to use it as it is in a speech recognition situation because it does not consider the speech recognition situation (Koo et al., 2022; Yoon et al., 2022). In summary, the diagnosis and verification of both ASR and ASRP are significant, but no proper segmentation benchmark exists.

Therefore, to alleviate this problem, we present an Error Explainable Benchmark (EEB) guideline for diagnosing and validating models by segmenting error types while considering both speech- and text-level. Such research is insufficient even in high-resource English and, in particular, does not exist in Korean. This is the first novel, a well-balanced guideline for both speech- and text-level. In addition, by building a dataset in consideration of the voice environment based on the existing Korean GEC dataset, it provides convenience in construction and enables real-world-centric evaluation.

2. Related Work

2.1. Post-processing Model

Post-processing serves an important role in quality enhancement across various fields by modifying the distorted output into appropriate statements. For instance, in the field of optical character recognition (OCR), traditional approaches such as manual, lexical, and statistical methods have been used (Evershed & Fitch, 2014; Nguyen et al., 2018). More recently, language models like BERT have been employed for error detection in tasks like named entity recognition (NER) and are performed through character-level machine translation (Nguyen et al., 2020).

As another field, machine translation (MT) often utilizes the following methods. Post-processing research is being carried out in automatic post-editing (APE) to improve translation quality by adopting transfer learning (Correia & Martins, 2019). Concurrently, in the grammatical error correction (GEC) field, transformers and the copy mechanism are used to correct spelling and grammatical errors in MT results (Lee et al., 2021). Studies that define error types to construct test sets or utilize an automatic grammatical error annotation system to create datasets also exist to improve Korean GEC studies (Koo et al., 2022; Yoon et al., 2022). Likewise, the study on post-processing is actively explored in a wide range of fields and holds significance in terms of enhancing the quality of output results. This can also be of significant importance in the field of Automatic speech recognition (ASR), which is discussed in the following section.

2.2. ASR Post-Processing Model

ASR post-processing (ASRP) involves the detection and correction of errors in the output of an ASR, distinguishing it from simple error correction in that it considers user-friendliness as an additional aspect. This approach can improve the final quality of statements without modifying the ASR system structure. For instance, in specialized fields like the medical domain, attempts have been made to eliminate punctuation errors in ASR through post-processing (Mani et al., 2020a). Prior research has primarily focused on providing information that allows humans to manually rectify erroneous segments, proposing alternative words for correction or creating an environment conducive to modification (Suhm et al., 2001; Feng & Sears, 2004). External information, such as word alternative hypothesis, noisy context, and accurate context, is provided to assist in post-processing for error correction (Shi & Zhou, 2011). In particular, Bassil & Semaan (2012) use the N-gram dataset for ASR errors to detect and correct errors automatically. Models such as LSTM-based or Transformer-based sequence-to-sequence architectures are adopted to correct the speech recognition results while considering the semantics and spelling (Guo et al., 2019; Hrinchuk et al., 2020).

Recent studies strive to improve ASRP performance by utilizing the results derived from ASR. Gekhman et al. (2022a) introduce the ASR confidence embedding (ACE) layer to the encoder of the ASR model to jointly encode the confidence scores and transcribed text into a contextualized representation. To mitigate the time and cost-related challenges associated with the parallel data required for training, Park et al. (2021) employ Text-to-speech (TTS) and Speech-to-text (STT) technologies to construct parallel data.

2.3. ASR dataset

The availability of suitable datasets is imperative for the active progression of ASRP. Previously, post-processing studies have been conducted with ASR datasets. Panayotov et al. (2015) organize the two labels in the ASR dataset that denote the quality of speech recognition, classified into 'clean' and 'other' categories, providing valuable assistance in the analysis. Ardila et al. (2020) constuct comprehensive ASR dataset that includes demographic metadata such as age, sex, and accent to provide a wider representation.

Transcription hypotheses obtained by decoding audio data using an ASR model are used to align hypothesis words with the reference (correct) transcription. The process of labeling errors and non-errors is facilitated by employing the minimum edit distance (Gekhman et al., 2022b). In the context of Chinese language datasets, a significant dataset is available for speech recognition systems, labeled with audio devices and recording environments (Bu et al., 2017). Gekhman et al. (2022b) build a dataset by aligning hypothesis words with the reference (correct) transcription through a transcription hypothesis obtained by decoding audio data with an ASR model and labeling errors and nonerrors using minimum edit distance. In the context of Chinese, a large-scale dataset is available for speech recognition systems labeled with audio device information and recording environments (Bu et al., 2017).

To mitigate the problem of insufficient training data, methodologies that synthesize data via data augmentation methods have been proposed (Liao et al., 2022). However, the overall quality of the data is more crucial than the size. Specifically, the detailed datasets that consider both speech- and text-level like the real world are absent. Consequently, we aim to construct the ASR Post-Processing dataset, which contemplates audio- and text-level for the first time.

3. Proposed Error Explainable Benchmark (EEB) Dataset

3.1. Why EEB?

Most speech benchmark datasets typically consider error types from either speech-level for recognition accuracy or text-level for user-friendliness. However, speech recognition encompasses both speech and text in its processing pipeline, necessitating consideration of both aspects to mirror real-world situations truly. Our motivation for proposing a guideline for an EEB dataset, which contains both aspects, is detailed below.

Firstly, in real-world speech recognition, it is essential to consider the accuracy of model and end-user satisfaction simultaneously. To facilitate this, we propose to map the accuracy of the ASR model to 'speech-level errors' and end-user satisfaction to 'text-level errors' to mitigate this inherent trade-off.

From the perspective of the accuracy of the ASR model, it should output the recognition results 'as heard,' regardless of the quality of the user-provided input. Conversely, from the standpoint of the end-user receiving the result, satisfaction increases when the output is presented in a refined state, despite any errors in the initial input. For instance, if a speaker stammers during their speech, the ASR model would likely deem its output more accurate if it recognizes and outputs all the words uttered. However, this would likely result in lower readability from the user's perspective.

A post-processor for speech recognition, validated by text-level error types, can be utilized to mitigate this. This post-processor would process speech recognition results to generate and provide user-friendly outcomes. Consequently, a benchmark dataset that can consider both aspects is required to handle the trade-off situation inherent in actual speech recognition scenarios.

Secondly, there are insufficient error types for a detailed diagnosis. Since benchmarks measure performance with quantitative metrics, it's crucial to fine-grain characteristics for a more detailed diagnosis. In industry contexts, communication between model and service teams is critical. When there's an issue with the model, clear criteria for the data flywheel significantly facilitate communication. That is, distinguishing the error type criteria for speech- and text-level aids in detailed diagnosis for model improvement. However, traditional benchmark datasets lack sufficient error types for detailed model analysis, leading to extensive usage of human evaluation in real-world settings. Humans can cope using commonsense, even if the criteria are unclear, but existing benchmarks with limited error types fall short. Hence, to solve the explainability issue, we must define the error type criteria that consider both the speech- and text-level and create benchmarks to achieve human-level explainability.

Thirdly, speech- and text-level errors coexist in real-world speech recognition systems. The ASR system needs to convert voice input into text output, a process often accompanied by simultaneous speech-level errors, like noisy environment issues, and text-level errors, such as spelling mistakes. However, the conventional focus has been chiefly on speech-level errors, making it challenging to handle a wide range of cases that could occur in reality. As a real-world setting requires consideration of both speech and text, we build a more practical benchmark considering both perspectives to enhance the explainability of speech recognition results.

To facilitate this, we define errors at speech- and text-level and propose a guideline for building speech recognition resources and benchmarks, as well as post-processor benchmarks that consider various environments.

3.2. Speech-Level Error Type

Error types at the speech-level refer to factors that trigger inaccuracies in speech recognition situations. For example, identical utterances may be challenging to recognize due to background noise (Sikasote & Anastasopoulos, 2022). Additionally, even in quiet environments, individuals do not consistently articulate perfect sentences and each speaker has unique characteristics that may negatively influence speech recognition (Gong et al., 2022).

Table 1 illustrates the speech-level error type classification criteria considering these characteristics. The speech-level error types allow the classification of two main categories (noisy environments and characteristics of interlocutor) and more detailed error types, with 24 sub-types for noise error and 13 for speaker characteristics.

Considering environments inundated with noise, it does not represent a quiet recording situation but rather a condition intertwined with noise. Real-world scenarios frequently involve inputs replete with ambient noise (Sikasote & Anastasopoulos, 2022). Reflecting on these practical situations where voice interface applications and devices are deployed, we propose an enhanced categorization scheme that closely follows the classification in the AI-HUB's noisy environment speech recognition dataset [1] which are representative Korean data platform. We divide the noisy environment errors into 11 nuanced subcategories, including **home appliances**, where recognition is impaired due to surrounding appliance noise; **individual transportation**, which includes instances with ambient transportation noise; **street**, covering situations with disruptive street noise; **cafe/restaurant**, addressing cases with the cafe or restaurant ambient noise; **market/shopping mall**, indicating instances with market or shopping mall noise; **public transportation**, comprising cases with subway or bus noise; **terminal**, reflecting instances with terminal noise; **construction site**, for cases hindered by construction site noise; **factory**, indicating instances with factory noise; **nature ambient**, for cases disturbed by natural sounds. Lastly, we include an **etc.** category for instances where recognition is affected by external noise types not encompassed in the previous categories.

Considering speaker characteristics, recognition can be hampered due to the individual traits of the recorder. Inspired by studies on idiolectal elements in the field of psycholinguistics (Ha & Sim, 2008; Shin et al., 2005), we propose a nuanced categorization comprising 13 detailed subcategories. **Pause (silent)** category captures instances where silence intervenes mid-utterance before completion—for instance, when 'I am eating' is articulated as 'I am... eating'. **Filled pause** represents cases characterized by the habitual insertion of filler sounds during pauses, as in utterances supplemented by sounds such as 'um... uh... so I'. **Interjection** category encompasses instances where one or more words or phrases irrelevant to the intended message are interjected, evident in utterances like 'Okay I see, but you know'. **Parenthetical** category includes instances where grammatically correct, but semantically neutral phrases are inserted—for instance, utterances incorporating phrases such as 'you know' and 'I mean'. **Unfinished interlocutor** category denotes cases where the utterance concludes prematurely—for instance, when 'I am eating' is truncated to 'I am...'. **Word repetition** category signifies instances where the same word is iterated, as in saying 'Hello' as 'Hello Hello'. **Syllable repetition** category characterizes cases where the same syllable is iterated—for instance, when 'Hello' is articulated as 'He-hello'. **Phoneme repetition** category encapsulates instances where the same phoneme is repeated, such as saying 'Hello' as 'Hel-llo'. **Sustained** category accounts for instances where part of an utterance is elongated, exemplified in 'Is that so—right?'. **Hyperfluency** category represents instances of excessive verbosity. **Mutter** category includes cases where utterances are murmured in an indistinct manner, as in 'That.. is.. like that...'. **Dynamic error** category encompasses instances where syllable articulation strength is incongruous with the intended utterance, or instances that are challenging to comprehend at the human-level. Finally, **speaking rate** category accounts for instances where rapid speech pace hinders comprehension at a human-level.

3.3. Text-Level Error Type

Text-level error types refer to issues that emerge in speech recognition results and must be addressed by post-processing. Since the output of the speech recognizer serves as the input for downstream tasks, it is one of the most significant factors influencing end-user satisfaction. By improving the performance of downstream tasks through quality input and diagnosing the performance of post-processing models through detailed error types, it is possible to enhance end-user satisfaction.

Existing datasets that detail error types, such as GEC datasets, do not consider speech recognition situations (Koo et al., 2022; Yoon et al., 2022). Therefore, we reconfigure the Korean GEC dataset, K-NCT, to suit speech recognition situations. The existing K-NCT dataset includes errors that only occur at the text-level and not in speech situations (Koo et al., 2022). Hence, errors that do not have vocal characteristics are removed.

[1] https://www.aihub.or.kr/

Error Type			Description
Noisy environments	Home appliances	Washer/dryer machine	Difficulty in recognition due to ambient electrical appliance noise.
		Vacuum cleaner	
	Individual transportation	Motorcycle	Difficulty in recognition due to surrounding individual transportation noise.
		Siren	
		Honk	
	Street	Road side	Difficulty in recognition due to the surrounding street noise.
		Crowd	
	Cafe/restaurant	Conversation	Challenges in perception due to the noise in cafes/restaurants.
		Non-conversation	
	Market/shopping mall	Traditional market	Difficulties in perception caused by the noise in markets/shopping malls.
		Shopping mall	
	Public transportation	Subway platform	Difficulty in recognition due to surrounding public transportation noise.
		Inside the subway	
		Inside the train (STR/KTX)	
		Inside the bus	
	Terminal	Train terminal waiting room	Challenges in perception due to the noise at terminals.
		Bus terminal waiting room	
	Construction site	Outdoor construction site	Difficulties in perception caused by the noise at construction sites.
		Indoor construction site	
	Factory	processing process	Difficulties in perception caused by the noise in factories.
		Assembly process	
	Nature ambient	Sound of rain	Challenges in perception due to natural ambient noise.
		Sound of the waves	
	Etc.	Artificial mechanical sound	In cases where external noise is present, although not falling into the aforementioned categories.
Characteristics of interlocutor	Pause (silent)		When there is a presence of pauses between syllables in speech that has not yet concluded.
	Filled pause		When habitual sounds are inserted during moments of silence or break time.
	Interjection		When phrases or longer segments are inserted regardless of their relevance to the intended content being expressed.
	Parenthetical		When grammatically acceptable sentences are inserted without conveying specific meaning or significance.
	Unfinished interlocutor		When speech is terminated without concluding the sentence.
	Word repetition		Repeating the same word or phrase in succession during speech.
	Syllable repetition		Repeating the same syllable in succession during speech.
	Phoneme repetition		Repeating the same phoneme in succession during speech.
	Sustained		When elongating certain parts of words within a sentence during speech.
	Hyperfluency		When excessively verbose speech is employed.
	Mutter		When muttering with an unclear demeanor.
	Dynamic error		When syllabic intonation is inappropriate for the intended speech purpose or difficult for human-level comprehension.
	Speaking rate		When speech rate is excessively fast, making it difficult for human-level comprehension.

Table 1: Proposed novel speech-level error type classification criteria for ASR and post-processing dataset

Error Type			Description
Spacing			Violating the spacing rules.
Punctuation			Punctuation marks are not attached in Korean sentences or are attached in the wrong.
Numerical			Cardinal number indicating quantity and the ordinal number indicating the order are in error.
Spelling and Grammatical	Remove		Some words are not recognized, or endings or suffixes are omitted.
	Addition		Same word is repeated, or an unused postposition or ending is added.
	Replace		Word is replaced by another word.
	Separation		Separating consonants and vowels in characters.
	Foreign word conversion		Writing differently from the standard foreign language pronunciation.
			Instances of Incorrect Conversion of Syllables between English and Korean.
	Spelling	Grapheme-to-phoneme(G2P)	Writing spellings according to pronunciation.
		Consonant vowel conversion	Spelling error in non-speaking alphabet units.
	Post-position		Instances of inconsistent or missing post-position usage in target utterances.
	Syntax		Cases of grammatically accurate yet interpretatively ambiguous meanings.
	Neologism		Instances of discrepancy between target and its similarity in meaning, pronunciation, and absence in Korean lexicon.

Table 2: Proposed text-level error type classification criteria for ASR and post-processing dataset

Table 2 illustrates the text-level error type classification criteria considering speech recognition situations, including 13 text-level errors that can occur in speech recognition situations.

Spacing encapsulates instances contravening standard spacing conventions. **Punctuation** entails cases where punctuation is omitted or misapplied in Korean sentences—for instance, when 'Can I teach?' is interpreted as 'Can I teach.' **Numerical** encompasses cases where number conversion fails, such as when 'Ahead of the three-month schedule' is interpreted as 'Bill 2, 3-month schedule'.

Spelling and Grammar consists of ten detailed subcategories. **Remove** designates cases where some word components are not recognized, or endings or particles are missing—for example, when 'The champion is in the final' is misinterpreted as 'Champion final'. **Addition** involves cases where the same word is repeated or unutilized particles or endings are appended. For instance, when 'World's fruits, fish, and meat' is interpreted as 'World's world's fruits, fish, and meat'. **Replace** refers to instances where one word is substituted with another—for example, when 'Apply the filter.' is interpreted as 'Wear the pizza'. **Separation** refers to instances where consonants and vowels in the target utterance are separated, exemplified when 'The discount applies as it is.' is interpreted as 'Discount app - lise as it is.'. **Foreign word conversion** refers to cases where words deviate from standard foreign word pronunciation or some syllables are incorrectly converted from English to Korean or vice versa. For example, when 'Brazil's Samba Festival' is interpreted as 'Brazil's SsamBap Festival,' or 'I prefer to use ATM.' is interpreted as 'I prefer to use hm.'.

Spelling is bifurcated into two types: Grapheme-to-Phoneme (G2P) and Consonant vowel conversion. **G2P** pertains to instances where a character is recognized per its pronunciation. **Consonant vowel conversion** refers to instances where phonemic units are incorrectly spelled. **Postposition** refers to cases where different particles are used or omitted—for example, when 'Ordinary high school students' is interpreted as 'Ordinary at high school students.' **Syntax** involves cases where the grammatical interpretation remains valid, but the semantic interpretation varies. Finally, **neologism** refers to cases where the target word and its meaning and pronunciation are dissimilar and are not included in Korean vocabulary.

4. Construction Process Design

In this work, we propose a comprehensive data construction guideline for the ASR and ASRP dataset, grounded in the application of a grammatical error correction (GEC) dataset. Our methodology encompasses validation, speech recording, noise synthesis, and difficulty tagging of a GEC dataset featuring text-level discrepancies. For the efficiency of the task, we choose the 'consensus labeling' method (Tang & Lease, 2011), in which a human overseer, who possesses an elevated degree of task completion, serves as a quality controller. During the progression of the task, any outcomes that do not conform to the established guidelines are promptly dismissed and subsequently reconstructed.

4.1. Step 1: Verification of Text

In this study, we employ a human-curated GEC dataset, which encompasses various text-level error types (Koo et al., 2022). Considering the inapplicability of the standard GEC benchmark dataset in a speech recognition setting, we selectively compose text-level error types dataset. In particular, we extract 13 categories that resonate with speech recognition scenarios (e.g., honorific colloquial expression) and reorganize their hierarchy for ease of labeling. Consequently, our refined dataset includes data reflecting 13 error types relevant to speech recognition contexts.

Subsequently, we authenticate the quality of the filtered dataset focusing on the alignment between labels and text, and the inclusion of text-level errors with a specific consideration of the speech recognition context. Validation processes proceed with a human supervisor, priorily trained with each error type. Evaluators are presented with an erroneous sentence, its correct counterpart, and a specified error type with the corresponding error span indicated. They are then tasked with assessing whether the sentence contains the presented error types. Sentences deemed to be incorrect are appropriately amended. This procedural framework ensures the generation of a high-quality dataset.

4.2. Step 2: Speech Recording

In the second phase, we request the recording participants to incorporate characteristics of interlocutor errors into their recordings by presenting them with speech-level errors and transcription relevant to the respective error types. At most 3 error types are presented, which could include an instance of 'no error type', indicating clean data. The placement of the error within the sentence is non-specific, with the ensurance that it includes only the errors specified. The recording environment should be ensured to be quiet without background noise. Each recorder is instructed to speak as naturally as possible, emulating their speech patterns when interacting with a voice interface application in real-world scenarios. After completing the recording, participants have the opportunity to listen to their own voice, and if they determine that the speech does not meet the criteria, they can re-record it. Participants are required to go through the process of listening to their recorded speech in order to complete the recording task.

4.3. Step 3: Synthesis of Background Noise

In the next stage, we incorporate background noise into the recording to reflect the noise environment error in the proposed speech-level. The background noise used for this integration is derived directly from recordings of the identified environments. We ensure that the collected noise spans a duration longer than that of the recording file, fostering noise diversity. To mimic real-world situations, we conduct both single and multiple noise syntheses while filtering out instances that are unlikely to co-occur. During noise synthesis, the noise is integrated as though it is ambient background noise, designed to be audible at the onset of the voice file. Noise is composited into the recording by randomly excising sections, thus ensuring variation within sounds, even when they are categorized under the same noise type.

4.4. Step 4: Difficulty Annotation

We employ a framework that distinguishes between utterances considered easier for ASR and those deemed harder or more noisy for ASR (Breiner et al., 2022). We extend this framework to include the tagging of difficulty using a Likert scale by human annotators. Humans listen to audio file and select score based on evaluation criteria. We ask humans, 'How difficult is it to recognize the presented speech accurately as the same as the transcript?' Scores range from 1 (very easy) to 5 (very difficult). Three evaluators assess each audio file, and the average score is selected as the difficulty level of the data. This allows for a detailed analysis of the model's performance

5. Conclusion

Since speech recognition results are used as inputs for downstream tasks and affect end-user satisfaction, it is important to diagnose and improve the weak types of speech recognition models. Considering the trade-off in real-world scenarios, achieving a balanced ASR environment requires diagnosing and validating both speech-level accuracy and text-level user-friendliness. We propose an Error Explainable Benchmark (EEB) guideline for diagnosing and validating models by segmenting error types while considering both speech- and text-level. To facilitate the construction process, we utilize a GEC dataset that includes text-level errors and structure the process into validation, recording, synthesis of background noise, and difficulty tagging stages, employing consensus labeling within each stage to enhance the efficiency and quality of the task. Through our EEB guidelines, it is possible to build a dataset that can specifically diagnose and verify the performance of ASR models and post-processors. Through this, automatic evaluation close to the real-world situation is possible.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), and this work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2022R1A5A7026673).

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Bassil, Y. and Semaan, P. Asr context-sensitive error correction based on microsoft n-gram dataset. *arXiv preprint arXiv:1203.5262*, 2012.

Breiner, T., Ramaswamy, S., Variani, E., Garg, S., Mathews, R., Sim, K. C., Gupta, K., Chen, M., and McConnaughey, L. Userlibri: A dataset for asr personalization using only text. *arXiv preprint arXiv:2207.00706*, 2022.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5. IEEE, 2017.

Correia, G. M. and Martins, A. F. A simple and effective approach to automatic post-editing with transfer learning. *arXiv preprint arXiv:1906.06253*, 2019.

Dai, W., Cahyawijaya, S., Yu, T., Barezi, E. J., Xu, P., Yiu, C. T., Frieske, R., Lovenia, H., Winata, G., Chen, Q., et al. Ci-avsr: A cantonese audio-visual speech datasetfor in-car command recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6786–6793, 2022.

Evershed, J. and Fitch, K. Correcting noisy ocr: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 45–51, 2014.

Feng, J. and Sears, A. Using confidence scores to improve hands-free speech based navigation in continuous dictation systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(4):329–356, 2004.

Feng, L., Yu, J., Cai, D., Liu, S., Zheng, H.-T., and Wang, Y. Asr-robust spoken language understanding on asr-glue dataset. 2022.

Gekhman, Z., Zverinski, D., Mallinson, J., and Beryozkin, G. RED-ACE: Robust error detection for ASR using confidence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2800–2808, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.180.

Gekhman, Z., Zverinski, D., Mallinson, J., and Beryozkin, G. Red-ace: Robust error detection for asr using confidence embeddings. *arXiv preprint arXiv:2203.07172*, 2022b.

Gong, Y., Yu, J., and Glass, J. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, 2022.

Guo, J., Sainath, T. N., and Weiss, R. J. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5651–5655. IEEE, 2019.

Ha, J.-W. and Sim, H. S. A comparison study of interjectional characteristics between people who stutter and people who do not stutter. *Communication Sciences and Disorders*, 13(3):438–453, 2008.

Hrinchuk, O., Popova, M., and Ginsburg, B. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 7074–7078. IEEE, 2020.

Koo, S., Park, C., Seo, J., Lee, S., Moon, H., Lee, J., and Lim, H. K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria. *IEEE Access*, 10:118167–118175, 2022.

Kumar, R., Adiga, D., Ranjan, R., Krishna, A., Ramakrishnan, G., Goyal, P., and Jyothi, P. Linguistically informed post-processing for asr error correction in sanskrit. *Proc. Interspeech 2022*, pp. 2293–2297, 2022.

Lakomkin, E., Magg, S., Weber, C., and Wermter, S. Kt-speech-crawler: Automatic dataset construction for speech recognition from youtube videos. *arXiv preprint arXiv:1903.00216*, 2019.

Lee, M., Shin, H., Lee, D., and Choi, S.-P. Korean grammatical error correction based on transformer with copying mechanisms and grammatical noise implantation methods. *Sensors*, 21(8):2658, 2021.

Leng, Y., Tan, X., Zhu, L., Xu, J., Luo, R., Liu, L., Qin, T., Li, X., Lin, E., and Liu, T.-Y. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *Advances in Neural Information Processing Systems*, 34:21708–21719, 2021.

Liao, J., Eskimez, S. E., Lu, L., Shi, Y., Gong, M., Shou, L., Qu, H., and Zeng, M. Improving readability for automatic speech recognition transcription. *Transactions on Asian and Low-Resource Language Information Processing*, 2020.

Liao, J., Shi, Y., and Xu, Y. Automatic speech recognition post-processing for readability: Task, dataset and a two-stage pre-trained approach. *IEEE Access*, 10:117053–117066, 2022.

Mani, A., Palaskar, S., and Konam, S. Towards understanding asr error correction for medical conversations. In *Proceedings of the first workshop on natural language processing for medical conversations*, pp. 7–11, 2020a.

Mani, A., Palaskar, S., Meripo, N. V., Konam, S., and Metze, F. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6344–6348. IEEE, 2020b.

Morris, A., Maier, V., and Green, P. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. 01 2004.

Nguyen, T.-T.-H., Coustaty, M., Doucet, A., Jatowt, A., and Nguyen, N.-V. Adaptive edit-distance and regression approach for post-ocr text correction. In *Maturity and Innovation in Digital Libraries: 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19-22, 2018, Proceedings 20*, pp. 278–289. Springer, 2018.

Nguyen, T. T. H., Jatowt, A., Nguyen, N.-V., Coustaty, M., and Doucet, A. Neural machine translation with bert for post-ocr error detection and correction. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pp. 333–336, 2020.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Park, C., Yang, Y., Park, K., and Lim, H. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562, 2020.

Park, C., Seo, J., Lee, S., Lee, C., Moon, H., Eo, S., and Lim, H.-S. Bts: Back transcription for speech-to-text post-processor using text-to-speech-to-text. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pp. 106–116, 2021.

Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., and Bengio, Y. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5754–5758. IEEE, 2018.

Shi, Y. and Zhou, L. Supporting dictation speech recognition error correction: the impact of external information. *Behaviour & Information Technology*, 30(6):761–774, 2011.

Shin, M.-S., Ahn, J.-B., Nam, H.-W., and Kwon, D.-H. A study of dysfluency characteristics in normal adults and children in monologue. *Speech Sciences*, 12(3):49–57, 2005.

Sikasote, C. and Anastasopoulos, A. BembaSpeech: A speech recognition corpus for the Bemba language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7277–7283, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.790.

Sodhi, S. S., Chio, E. K.-I., Jash, A., Ontañón, S., Apte, A., Kumar, A., Jeje, A., Kuzmin, D., Fung, H., Cheng, H.-T., et al. Mondegreen: A post-processing solution to speech recognition error correction for voice search queries. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3569–3575, 2021.

Suhm, B., Myers, B., and Waibel, A. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)*, 8(1):60–98, 2001.

Tang, W. and Lease, M. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*, pp. 1–6, 2011.

Wang, L., Fazel-Zarandi, M., Tiwari, A., Matsoukas, S., and Polymenakos, L. Data augmentation for training dialog models robust to speech recognition errors. *arXiv preprint arXiv:2006.05635*, 2020.

Wang, S., Gunter, T., and VanDyke, D. On modelling uncertainty in neural language generation for policy optimisation in voice-triggered dialog assistants. In *2nd Workshop on Conversational AI: Today's Practice and Tomorrow's Potential, NeurIPS*, 2018.

Williams, J. D. and Young, S. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

Woodard, J. and Nelson, J.T., y. . . j. . W. t. . A.

Yoon, S., Park, S., Kim, G., Cho, J., Park, K., Kim, G. T., Seo, M., and Oh, A. Towards standardizing korean grammatical error correction: Datasets and annotation. *arXiv preprint arXiv:2210.14389*, 2022.