Speech Foundation Models and Crowdsourcing for Efficient, High-Quality Data Collection

Beomseok Lee[1,2,3], Marco Gaido[2], Ioan Calapodescu[3], Laurent Besacier[3], Matteo Negri[2]

[1]University of Trento, Italy, [2]Fondazione Bruno Kessler, Italy, [3]NAVER LABS Europe, France

Correspondence: beomseok.lee@unitn.it

Abstract

While crowdsourcing is an established solution for facilitating and scaling the collection of speech data, the involvement of non-experts necessitates protocols to ensure final data quality. To reduce the costs of these essential controls, this paper investigates the use of Speech Foundation Models (SFMs) to automate the validation process, examining for the first time the cost/quality trade-off in data acquisition. Experiments conducted on French, German, and Korean data demonstrate that SFM-based validation has the potential to reduce reliance on human validation, resulting in an estimated cost saving of over 40.0% without degrading final data quality. These findings open new opportunities for more efficient, cost-effective, and scalable speech data acquisition.

1 Introduction

As in any data-intensive domain, collecting high-quality datasets is a fundamental and costly prerequisite for the development of speech-processing applications. Traditional methods heavily rely on human workforce, whose costs, as data collection scales, are hard to sustain. In the quest for scalable solutions to tackle this problem, crowdsourcing emerged as a viable option that also enables the coverage of diverse populations (Cefkin et al., 2014; Poesio et al., 2017). Due to the variable quality of crowd-sourced data, validation methods that discard low-quality contributions are essential to build reliable datasets (Negri et al., 2011; Sabou et al., 2014; Chittilappilly et al., 2016). This need is exacerbated in the collection of speech-text pairs, where various factors, such as recording equipment and conditions, can introduce errors and inconsistencies that compromise data quality (Novotney and Callison-Burch, 2010; Marge et al., 2010). Crowdsourcing this validation process adds substantial overhead, further inflating data collection costs.

Recent advances in foundation models offer new possibilities to enhance the scalability of the process by speeding it up and reducing its inherent costs. When dealing with textual data, Large language models (LLMs) have been successfully used as proxies for human evaluation in tasks like sentiment analysis, machine translation, and text generation (He et al., 2024; Zheng et al., 2023; Kim et al., 2024). Similarly, speech foundation models (SFMs) have shown potential not only for evaluating synthetic and non-synthetic speech (Maiti et al., 2023; Ravuri et al., 2024; Fu et al., 2024) but also for automatizing complex data filtering (Lee and Glass, 2011) and validation tasks (Phatthiyaphaibun et al., 2023). However, all previous works in the area focused on improving automatic speech recognition (ASR) performance rather than optimizing the cost-efficiency of data validation and exploring the relationship between these two objectives, which remain underinvestigated aspects.

To fill this gap, this paper explores the use of SFMs to automatize the validation of crowd-sourced speech data. To this aim, we investigate the employment of off-the-shelf SFMs such as Whisper and SeamlessM4T (Radford et al., 2022; Communication et al., 2023), along with machine translation (MT) models and grapheme-to-phoneme conversion (G2P). Through experiments on French, German, and Korean data, we test the integration of SFMs and crowdsourcing to reduce validation costs while preserving final data quality. Our results show that leveraging SFMs yields a cost reduction by over 40%, while maintaining high data quality, significantly improving the efficiency and scalability of crowd-sourced speech data collection.

2 Context and motivations

This work stems from experiments conducted both during and after the creation of a multilingual speech corpus, `Speech-MASSIVE` (Lee et al., 2024),

Proceedings of the 31st International Conference on Computational Linguistics, pages 6816–6826
January 19–24, 2025. ©2025 Association for Computational Linguistics

covering 12 languages. The corpus was crowdsourced, recruiting native speakers who were instructed to read aloud short sentences and record their voices under controlled conditions.[1] To ensure data quality, each recording was validated by crowdsourced human raters, directed to read the original text, listen to the recording, and label it as valid or invalid. Invalid recordings underwent a second iteration of this two-step recording-validation process, which, to prevent endless cycles, concluded after the second validation regardless of the outcome. As a result, Speech-MASSIVE comprises $84,262$ (t, r, l) triplets for the 12 languages, where t is the original text, r is the acquired recording, and l is the valid/invalid label assigned to r.

With Speech-MASSIVE at hand, the goal of the post-hoc experiments documented in this paper was to assess whether the costs of its creation could have been reduced by automating the validation steps. Specifically, the objective was to assess whether, and to what extent, transcripts generated by existing SFMs could be leveraged to validate the quality of human-recorded speech. Within this framing, we address two key questions: **(1)** Can the distance between SFM-generated transcripts and the original text serve as a reliable proxy for recording quality? **(2)** With comparable final data quality, what are the cost savings of replacing human validation of recorded speech with SFM-based validation?

3 Automated validation methods

Starting from the (t, r, l) triplets of Speech-MASSIVE, our validation method considers the similarity between the original text (t) and the SFM-generated transcripts ($\hat{t}$) of the acquired recordings (r) as a proxy for l. To this end, we explored two policies. The first policy is a **distance-based** method that measures the similarity between t and $\hat{t}$ with two widely used edit-distance metrics—Character Error Rate (CER) and Word Error Rate (WER)—and retains triplets with a distance below a specified threshold. However, this approach may be affected by SFMs' bias towards clean audio or specific accents, potentially invalidating samples with poor recording conditions or strong, distinctive accents.

Our second policy seeks to mitigate this risk by employing a **decision tree** trained on multiple features. In addition to CER and WER scores computed as in the distance-based method, these features include Translation Error Rate (TER) and Phoneme Error Rate (PER) scores, which are also based on edit-distance. TER is computed on the English translations of t and $\hat{t}$, under the assumption that accurate recordings will yield translations that closely match those of the original text. A further advantage of using translations for both t and $\hat{t}$ is the normalization of numbers in the resulting texts. PER is computed by converting t and $\hat{t}$ into phonemes, based on the assumption that this conversion may act as a normalizer for words (e.g. named entities) that were transcribed differently but have similar or identical pronunciations.

4 Experimental setting

Data We experiment with three distant languages—Korean, French, and German—out of the 12 covered in Speech-MASSIVE. **Korean** triplets are used for a preliminary analysis (§5) aimed at comparing our two automated validation methods and selecting the best one. To this end, $5,007$ (t, r, l) triplets were enriched with "gold"[2] quality labels (l^*), produced by two expert linguists, native Korean speakers. A Cohen's Kappa (κ – Cohen 1960) of 0.82 on a subset of 100 common samples indicates 'excellent' agreement (Fleiss et al., 2013) between the two annotators. On the entire annotated set, the κ between the original silver annotations and the gold labels is unsurprisingly lower (0.65), though still within the 'fair to good' range. **French** and **German** triplets are used in our final experiment (§6), which focuses on analyzing the impact of applying the best identified method to quantify the cost savings yielded by SFM-based validation of crowdsourced speech data.

Speech foundation models To generate the transcripts ($\hat{t}$) of the acquired recordings, we considered Whisper-large-v3[3] (Radford et al., 2022) and Seamless-m4t-v2-large[4] (Communication et al., 2023). To identify the better-performing model, we compared their transcription capabilities using the French, German, and Korean test splits of FLEURS (Conneau et al., 2023), computing CER and WER. Whisper-large-v3 exhibited superior performance

[1]Detailed data collection guidelines emphasized the importance of accurate and natural reading, proper recording conditions, and full adherence to the corresponding text.

[2]As opposed to the "silver" ones (l) produced by crowdsourced annotators.

[3]https://huggingface.co/openai/whisper-large-v3

[4]https://huggingface.co/facebook/seamless-m4t-v2-large

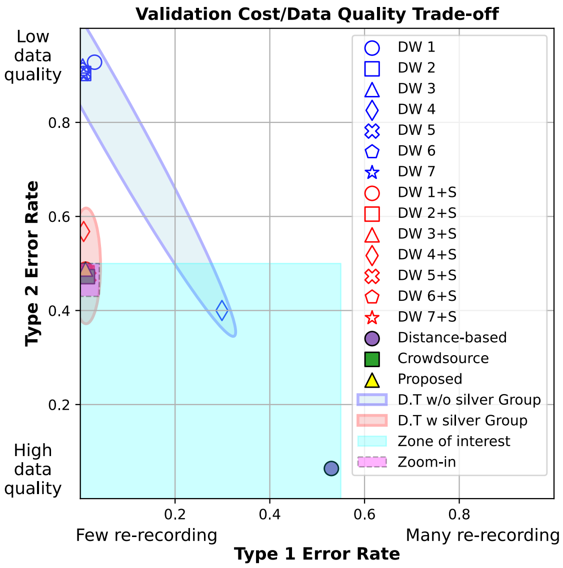

Figure 1: Results of validation methods: DW (decision-tree); DW+S (decision-tree + silver labels); distance-based (simple policy); crowdsource (fully crowd-sourced); proposed (final policy for experiments in §6)

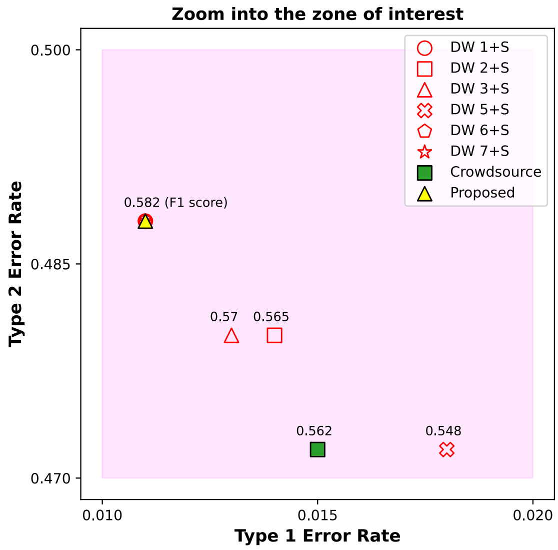

Figure 2: Zoom-in on a specific area of performance (F1 scores displayed above each data point)

as reported in Table 1, justifying its use in our experiments unless otherwise specified. As described in §3, to assess TER and PER as metrics for normalizing the SFMs' transcription outputs, the NLLB-200[5] translation model is employed for translation into English, while a neural G2P model[6] is used to convert graphemes into phonemes.

		Whisper	Seamless-m4t
de-DE	WER	4.22	31.24
	CER	1.48	8.05
fr-FR	WER	5.37	16.24
	CER	1.9	5.73
ko-KR	WER	13.88	26.26
	CER	5.3	11.21

Table 1: CER (↓) and WER (↓) of Whisper-large-v3 and Seamless-m4t-v2-large on FLEURS test data.

5 Distance-based method vs Decision Tree

We evaluated various automatic and semi-automatic validation policies by classifying utterances as valid or invalid. To ensure a fair comparison between the different policies, we use the initial data splits (dev and test) from the `Speech-MASSIVE` Korean subset. The test split, composed of 2,974 samples, is used to evaluate and report performance, while the dev split, composed of 2,033 utterances, is used to train a decision tree when required by the policy. The metrics are calculated by comparing the decisions of the automated validation methods to the gold labels described in §4. We present an overall comparison of the different methods in Fig.1, with the exact values provided in Appendix A.2, Table 6. Crowdsourced annotations (silver labels) are also evaluated against gold labels to position them on the same evaluation plot as the automated methods. Specifically, Fig.1 plots the performance of different methods along *type 1 error rate* (incorrectly classifying valid utterances as invalid, thus requiring re-recording for invalid ones) on the x-axis and *type 2 error rate* (incorrectly classifying invalid utterances as valid) on the y-axis. This visualization highlights the trade-off between the cost of re-recording invalidated utterances (x-axis) and data quality (y-axis). In this context, our zone of interest focuses on regions with moderate type 2 error rate (data quality comparable to or better than crowdsourced annotations) and low type 1 error rate (low re-recording costs). Fig.2 zooms in on a specific area of this zone of interest to better distinguish methods of similar performance.

5.1 Distance-based method

The distance-based method (●) validates recordings only when the ASR output ($\hat{t}$), compared to the reference transcript (t), shows both CER and WER

[5]https://huggingface.co/facebook/nllb-200-distilled-1.3B

[6]https://github.com/lingjzhu/CharsiuG2P

values equal to zero. As shown in Fig.1, this policy has high type 1 error rate (0.53, meaning that it often invalidates valid samples) while at the same time having the lowest type 2 error rate (0.064, crucial for high data quality). In summary: conservative and simple (no additional model training is required) **the distance-based approach prioritizes data quality over reducing re-recording cost**, thus only partially satisfying our requirements.

5.2 Decision tree

For the decision tree model, we experiment with several settings, testing the two SFMs (Whisper-large-v3 and Seamless-m4t-v2-large) using different combinations of CER, WER, PER, and TER features (§3). As shown in Fig.1, basic decision tree methods (DWN, where $N \in [1, 7]$, full details in Appendix A.2) exhibit significantly different behaviour compared to the distance-based method, consistently showing a higher type 2 error rate across all settings. Therefore, targeting a balance between data quality and cost effectiveness, **basic decision tree models may not be ideal**, as only one setting ($DW4$) falls within our zone of interest. To improve the decision tree model and integrate automation with crowdsourcing, we incorporate the silver label (l) from crowd annotators into our feature set ($DWN + S$). The red ellipse region in Fig.1, compared to the blue region, shows the effects of this adjustment in terms of type 2 error rate reduction. Indeed, most settings of hybrid (SFMs + crowdsourcing) approach except for $DW4+S$, fall within our zone of interest. Examining the zoomed-in area of Fig.2 reveals that **hybrid approaches combining SFMs and crowdsourcing perform comparably to the crowdsourcing-only method** (■), suggesting potential cost savings in data validation. The next section discusses our final choice of the optimal policy from these options.

5.3 Proposed method

Our *proposed* validation method (△ in Fig.1-2) relies on a two-step approach. First, we use our distance-based method to validate all utterances, minimizing type 2 error rate. For those still flagged as invalid, we use the silver labels assigned by crowd annotators. Although **the *proposed* approach** may appear as a simplistic decision tree, it achieves performance comparable to $DW\ 1 + S$ (○ in in Fig.1-2) as reported in Table 2. It yields a high F1 score and significantly **reduces the number of re-recordings required,**[7] while maintaining data quality comparable to that achieved through full crowdsourcing.[8] In the next section, we employ this automated method for large-scale data validation in German, while using French with full crowdsourcing as control language.

	Precision	Recall	F1 Score	Type 1 error rate	Type 2 error rate
Distance-based	0.072	0.936	0.134	0.530	0.064
Crowdsource	0.600	0.528	0.562	0.015	0.472
Decision Tree ($DW1 + S$)	0.674	0.512	0.582	0.011	0.488
Proposed	0.674	0.512	0.582	0.011	0.488

Table 2: Evaluation results of distance-based (● in Fig.1), crowdsource (■ in Fig.1-2), decision tree $DW1 + S$ (○ in in Fig.1-2) and proposed method (△ in Fig.1-2).

6 Application to real-world scenarios

We conclude by applying our best automated validation method (△) to a real data collection pipeline involving 11,399 new and yet unlabeled samples of `Speech-MASSIVE` German subset. As a term of comparison to assess cost savings of integrating automated validation into the data collection process, we use `Speech-MASSIVE` French subset, entirely (manually) validated through crowdsourcing.

Table 3 shows the total data collection costs for the French and German `Speech-MASSIVE`. For the German dataset, the SFM validates a large number of utterances with no labor costs. We observe a 43.11% cost reduction in the validation phase, leading to substantial savings in both cost and time by minimizing the need for recruiting and managing human raters. To ensure comparable quality between German and French `Speech-MASSIVE` utterances, validated using different policies, we present WER and CER metrics for all recordings in Table 4. WER as a proxy for data quality shows that our automated validation for German performs similarly to the fully manual process for French.

[7] In addition to the re-recording cost, it is important to note that our proposed method incurs costs only for validating the samples flagged as invalid, whereas the $DWN + S$ method requires validation for all samples.

[8] We consider the 0.016 difference in type-2 error rates between the two methods to be insignificant, especially given that the manual analysis in Appendix C highlights disagreements between human (silver and gold) annotations.

		Cost (# participants)	
		French	German
1st iteration	Recording	£ 694.73 (572)	£ 782.38 (555)
	Automated validation	N/A	£ 0 (Whisper)
	Human validation	£ 333.6 (213)	£ 181.8 (102)
2nd iteration	Recording	£ 36 (30)	£ 36.4 (26)
	Human validation	£ 17.6 (11)	£ 18 (18)

All validations	French	German
cost	£ 351.2	£ 199.8
# participants	224	120

Table 3: Automated validation applied to German data, with French as a control for evaluating cost savings (in parentheses, the number of crowdsourced workers).

	langs	# samples	WER	CER
Speech-MASSIVE	French	11,399	11.09	4.84
	German		11.7	4.19

Table 4: Final dataset quality (WER, CER) comparisons using Whisper-v3-large.

7 Conclusion

We proposed using Speech Foundation Models (SFMs) to reduce the costs of validating speech data collected through crowdsourcing. After exploring various approaches under controlled conditions, we identified a two-step method leveraging Whisper-large-v3 as the most promising. Its application to large-scale validation on German data resulted in a 40% cost reduction without compromising data quality, demonstrating the strong potential of SFMs to enhance the efficiency and scalability of crowd-sourced speech data collection.

8 Limitations

As our proposed method is developed by evaluating only with Korean gold labels, our method lacks language universal development. However, collecting gold annotations for different language for around 5,000 examples is significantly costly. If the proposed methods were developed from various numbers of languages, more universal pattern or method could have been proposed. However, even with this limitation, our proposed method has proven its effectiveness by being successfully applied to German, and further validated through comparison with French to assess dataset quality.

Moreover, this work has limitation with the inherent error in the text corpus which `Speech-MASSIVE` is built upon. As discussed in Appendix C, some incorrect validations from crowdsource workers are likely due to pre-existing errors in the text. Specifically, problematic prompts affect both the recording and validation phases. During recording, workers are instructed to read the prompt exactly as provided, which can lead to confusion when the prompt is erroneous. Additionally, if workers correct errors in the prompt while recording, it may cause confusion for validators, as the recorded audios are correct despite the original prompt being incorrect. However, our manual analysis revealed that such cases are relatively rare (6 instances in total—row E in Table 7 and row GG in Table 8, out of 2,974 examples) and are unlikely to significantly impact the overall findings.

Acknowledgement

The research work presented in this paper has been partially funded by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People) and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The crowdsourcing study was funded by the project UTTER (Unified Transcription and Translation for Extended Reality) funded by European Union's Horizon Europe Research and Innovation programme under grant agreement number 101070631.

References

Melissa Cefkin, Obinna Anya, Steve Dill, Robert Moore, Susan Stucky, and Osariemo Omokaro. 2014. Back to the future of organizational work: crowdsourcing and digital work marketplaces. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW Companion '14, page 313–316, New York, NY, USA. Association for Computing Machinery.

Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. 2016. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2246–2266.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian

Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Szu-Wei Fu, Kuo-Hsuan Hung, Yu Tsao, and Yu-Chiang Frank Wang. 2024. Self-supervised speech quality estimation and enhancement using only clean speech. *Preprint*, arXiv:2402.16321.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing evaluation capability in language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. 2024. Speech-massive: A multilingual speech dataset for slu and beyond. In *Proc. Interspeech 2024*, pages 817–821.

Chia Ying Lee and James Glass. 2011. A transcription task for crowdsourcing with automatic quality control. In *Proc. Interspeech 2011*, pages 3041–3044.

Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. 2023. Speechlmscore: Evaluating speech generation using speech language model. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. 2010. Using the amazon mechanical turk for transcription of spoken language. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5270–5273.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun, Chompakorn Chaksangchaichot, Thanawin Rakthammanon, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Crowdsourced Data Validation for ASR Training. In *Proc. INTERSPEECH 2023*, pages 551–555.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. *Crowdsourcing*, pages 277–295. Springer Netherlands, Dordrecht.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Aditya Ravuri, Erica Cooper, and Junichi Yamagishi. 2024. Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 580–584.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th Conference on Neural Information Processing Systems (Datasets and Benchmarks Track)*.

A Decision Trees

A.1 Training and hyper parameters

To identify the optimal decision tree model, we conducted a 10-fold cross-validation using the F1 score as the evaluation metric. To facilitate interpretation of feature contributions, the tree depth was restricted to 3. Detailed training parameters are provided in the Table 5.

parameters	search space
max depth	[1, 2, 3]
min samples split	[3, 5]
splitter	best, random
criterion	gini, entropy, log_loss
class weight	balanced, none
min samples split	[3, 5, 7, 9, 11, 13]
min samples leaf	[1, 3, 5, 7, 9, 11, 13]

Table 5: Decision tree fitting hyper parameters.

A.2 Performance comparisons between various decision trees

For the evaluation results of the decision tree method discussed in §5.2, Table 6 presents the various settings used for the decision trees, including: 1) the choice of SFMs (Whisper, Seamless-m4t, or both) for feature extraction, 2) the specific features selected from the SFMs' outputs, and 3) whether the silver label (crowd-sourced) is included as an additional feature.

A.3 Decision tree interpretation

We present a decision tree plot to illustrate how features are used to classify samples as valid or invalid. Figure 3 depicts the decision tree for the $DW\ 5 + S$ method, which integrates WER, CER, and silver label features. In this tree, the silver label feature is applied at the root node (level 0) with a threshold of 0.5, meaning that if the silver label equals zero (valid), the samples are further classified by the left child nodes, whereas if the silver label equals 1 (invalid), the right child nodes continue the classification.

When the silver label is valid (left child nodes of the root node), the tree uses a CER threshold of 72.5 at level 1 to classify certain samples beyond the threshold as invalid, resulting in 7 samples being invalidated (Fig. 3). The 1,889 samples where CER is smaller than the CER threshold (72.5) are validated.

Conversely, when the silver label is invalid (right child nodes of the root), the tree utilizes a WER threshold of 13.393. If WER is less than or equal to this threshold, 23 samples are validated, otherwise 114 samples are invalidated.

B Confusion matrices of the policies

To supplement the metrics presented in Table 6, we provide some of the confusion matrix plots in Fig. 4 through Fig. 8 for various policies and settings as additional material.

C Analysis on gold and silver labels mismatch

In this section, we discuss further analysis on mismatch between gold and silver label for the test split of `Speech-MASSIVE` Korean subset.

We begin by analyzing annotation mismatches between gold and silver annotations, focusing on cases where the gold annotation is valid, but the silver annotation is invalid. In Table 7. Overall, we witness 47 mismatch cases (44 unique samples) grouped in 10 different types. To further elucidate the types of mismatches observed, consider the following examples. In the E case, the prompt from the corpus is '철수에세 트윗 남겨', while the correct text should be '철수에게 트윗 남겨'[9]. The mismatch arises from the incorrect use of '세' instead of the correct '게' resided in the prompt. In the J case, the corpus text '최고로 평점이 좋은 록 음악 팟캐스트 보여줘'[10] contains '록', which is a Koreanization of the English term 'rock'. The speaker exhibits 'hyperforeignism', pronouncing '록' according to the English pronunciation ($[rOk]$) rather than the Korean phonology ($[lOk]$).

We further classify the mismatches into four categories: *audio*—where the silver annotation's invalidity may be due to poor audio quality, *speaker*—where the invalidity could stem from characteristics of the speaker, *perfect audio*—where the validation is questionable despite clear and intelligible audio, and *corpus*—where the invalidity may result from a typo in the text prompt. Among the 47 mismatch cases, *audio* category accounts for 16 cases, representing 34%. *speaker* and *perfect audio* each include 14 cases, which corresponds to 30% for each category. *corpus* category represents 6% of the cases.

On the contrary, Table 8 presents an analysis of

[9] En Translation: Send a tweet to Cheolsoo.
[10] Show me the best-rated rock music podcast.

mismatches where the gold is valid, and the silver is invalid, totaling 59 mismatch cases. For AA cases, we observe 'near homophone errors'. For instance, '알람' (alarm, incorrect) is used instead of '알림' (notification, correct) in the prompt '내일 오전 열 시 미팅 관련해서 리마인더 알림 보내줘'[11]. Such errors often arise when characters or words appear similar in both appearance and sound, leading some crowdsource workers to validate the audio with the homophone error. In CC cases, errors involve the omission, addition, or substitution of particles, possibly due to a lack of attention to grammatical rules or simplification of speech. For example, in the prompt '토요일에 비 소식이 있나'[12], one speaker omits the particle '이', resulting in '토요일에 비 소식 있나'. For EE cases, approximation errors are noted where speakers slightly modify pronunciation or spelling while maintaining clarity of meaning. For example, in the prompt '지영이한테서 새로 온 이메일이 있으면 확인해 줘'[13], the speaker says '지영이한테서 새로 온 메일이 있으면 확인해 줘', substituting '이메일' (email) with '메일' (mail).

The mismatch types where the gold standard is valid while the silver label is invalid can be further categorized into two groups. *honest mistake* category ($AA + CC + EE$) includes cases where validators make errors due to the potentially confusing prompts, accounting for 49% of the mismatches. *erroneously generous* category ($BB + DD + FF + GG$) comprises cases where validators are erroneously generous, representing 51% of the mismatches.

[11]Send me a reminder notification for the meeting tomorrow morning at 10 AM.

[12]Is there any news of rain on Saturday?

[13]Check if there is any new email from Jiyoung.

		Method name	Features	Precision	Recall	F1 Score	Type 1 error rate	Type 2 error rate
without silver label	Whisper	DW 1	WER	0.096	0.072	0.082	0.030	0.928
		DW 2	CER	0.387	0.096	0.154	0.007	0.904
		DW 3	PER	0.435	0.080	0.135	0.005	0.920
		DW 4	TER	0.081	0.600	0.143	0.299	0.400
		DW 5	DW 1 + CER	0.387	0.096	0.154	0.007	0.904
		DW 6	DW 5 + PER	0.440	0.088	0.147	0.005	0.912
		DW 7	DW 6 + TER	0.387	0.096	0.154	0.007	0.904
	Seamless-m4t	DS 1	WER	0.086	0.336	0.137	0.156	0.664
		DS 2	CER	0.124	0.248	0.165	0.077	0.752
		DS 3	DS 1 + CER	0.124	0.248	0.165	0.077	0.752
	both	DWS	DW 7 + DS 3	0.387	0.096	0.154	0.007	0.904
with silver label	Whisper	DW 1+S	DW 1 + silver	0.674	0.512	0.582	0.011	0.488
		DW 2+S	DW 2 + silver	0.619	0.520	0.565	0.014	0.480
		DW 3+S	DW 3 + silver	0.631	0.520	0.570	0.013	0.480
		DW 4+S	DW 4 + silver	0.730	0.432	0.543	0.007	0.568
		DW 5+S	DW 5 + silver	0.569	0.528	0.548	0.018	0.472
		DW 6+S	DW 6 + silver	0.681	0.512	0.584	0.011	0.488
		DW 7+S	DW 7 + silver	0.681	0.512	0.584	0.011	0.488
	Seamless-m4t	DS 1+S	DS 1 + silver	0.674	0.480	0.561	0.010	0.520
		DS 2+S	DS 2 + silver	0.635	0.528	0.576	0.013	0.472
		DS 3+S	DS 3 + silver	0.674	0.480	0.561	0.010	0.520
	both	DWS+S	DW 7 + DS 3 + silver	0.681	0.512	0.584	0.011	0.488
		Crowdsource		0.600	0.528	0.562	0.015	0.472
		Distance-based		0.072	0.936	0.134	0.530	0.064
		Proposed		0.674	0.512	0.582	0.011	0.488

Table 6: Results for all the settings of decision tree and other methods (the color-highlighted settings are the best ones displayed in Fig.1).

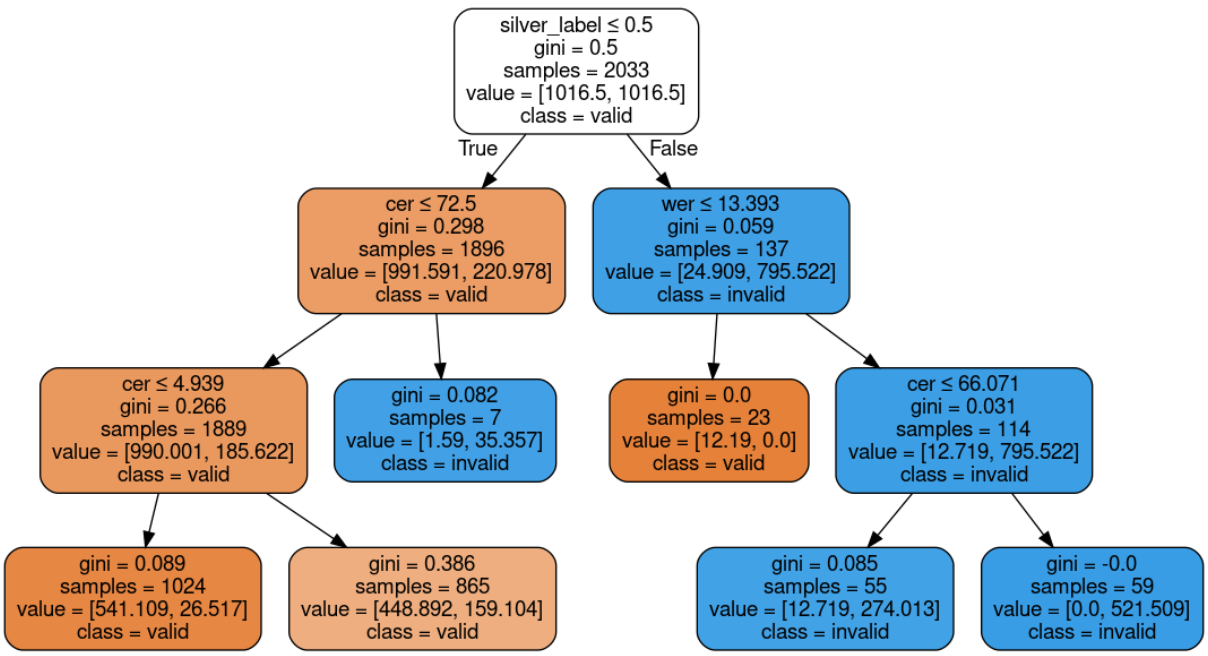

Figure 3: Decision tree graph of DW 5+S method.

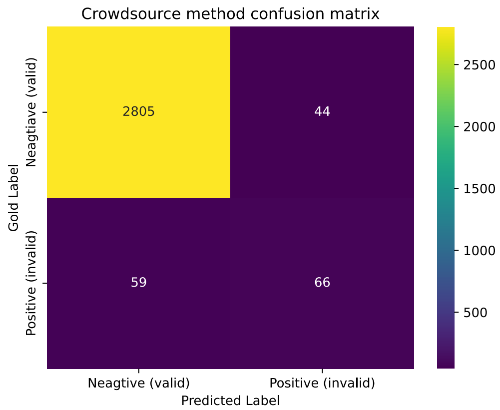

Figure 4: Crowdsource confusion matrix.

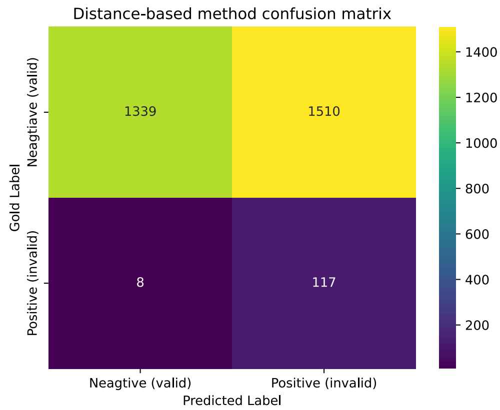

Figure 5: Distance-based method confusion matrix.

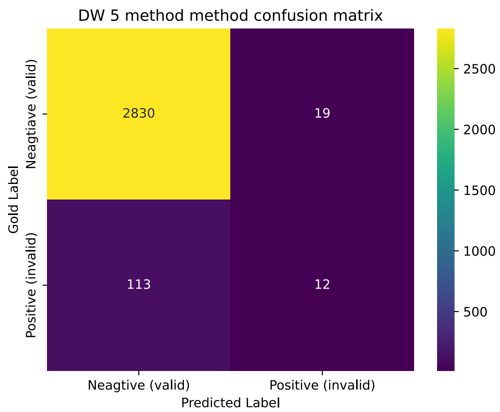

Figure 6: DW 5 method confusion matrix.

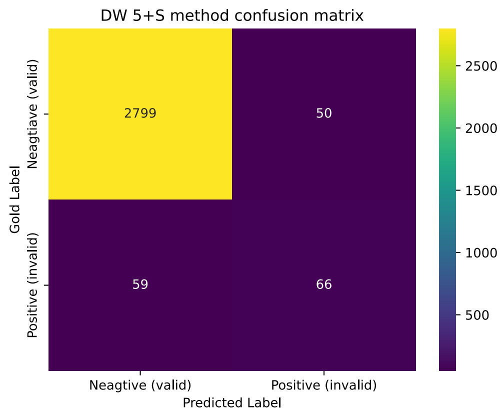

Figure 7: DW 5+S method confusion matrix.

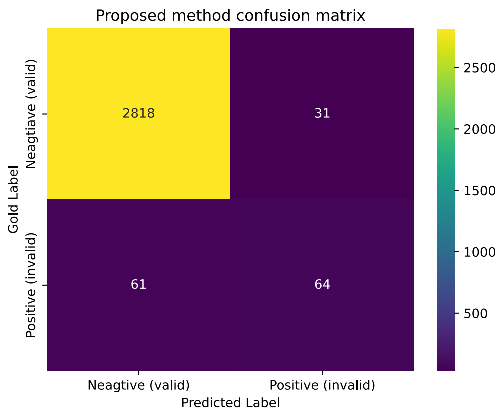

Figure 8: Proposed method confusion matrix.

type	possible reason of silver invalidity	exclusive # samples	non-exclusive # samples	category
A	perfect audio	14		perfect audio
B	non-native accent	7	1	speaker
C	noisy recording condition	5	2	audio
D	rushed speech	4	1	speaker
E	audio-prompt mismatch (*prompt is wrong from the beginning)	3		corpus
F	low volume	3	2	audio
G	some strange (mechanical, noise) sound mixed in the audio	2		audio
H	chopped at the beginning or ending but intelligible	1		audio
I	disfluency in the speech	1		audio
J	hyperforeignism	1		speaker
	# total	41	6	
	total # unique samples	44		

Table 7: Analysis on the mismatched annotations between gold (label=invalid) and silver (label=valid). Exclusive samples refer to those that belong to only one group, while non-exclusive samples represent those that are shared among multiple groups.

type	possible reason of silver validity	# samples	category
AA	near homophone error	11	honest mistake
BB	wrong pronunciation	11	erroneously generous
CC	particle omission/addition/substitution	10	honest mistake
DD	chopped sentence (at the beginning or ending)	9	erroneously generous
EE	approximation error	8	honest mistake
FF	repetition error	7	erroneously generous
GG	wrong prompt from the beginning and audio not matching the prompt	3	erroneously generous
	# total	59	

Table 8: Analysis on the mismatched annotations between gold (label=valid) and silver (label=invalid).