Effectiveness of Whisper's Fine-Tuning for Domain-Specific Use Cases in the Industry

Daniel Pawlowicz[1][a], Jule Weber[2][b] and Claudia Dukino[3][c]

[1] *University of Stuttgart IAT, Institute of Human Factors and Technology Management, Stuttgart, Germany*

[2] *Eberhard Karls University Tübingen, Tübingen, Germany*

[3] *Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany*

Keywords: Whisper, Fine-Tuning, Domain, Speech-to-Text Transcription.

Abstract: The integration of Speech-to-Text (STT) technology has the potential to enhance the efficiency of industrial workflows. However, standard speech models demonstrate suboptimal performance in domain-specific use cases. In order to gain user trust, it is essential to ensure accurate transcription, which can be achieved through the fine-tuning of the model to the specific domain. OpenAI's Whisper was selected as the initial model and subsequently fine-tuned with domain-specific real-world recordings. The fine-tuned model outperforms the initial model in terms of transcription of technical jargon, as evidenced by the results of the study. The fine-tuned model achieved a validation loss of 1.75 and a Word Error Rate (WER) of 1. In addition to improving accuracy, this approach addresses the challenges of noisy environments and speaker variability that are common in real-world industrial environments. The present study demonstrates the efficacy of fine-tuning the Whisper model to new vocabulary with technical jargon, thereby underscoring the value of model adaptation for domain-specific use cases.

1 INTRODUCTION

The integration of machine learning tools into modern business practices has revolutionized workflows by enabling automation, improving efficiency, and reducing administrative burdens. Digital assistants, powered by advanced Speech-to-Text (STT) transcription models, play a central role in these transformations, allowing professionals to perform tasks like scheduling, documentation, and customer relationship management (CRM) through intuitive voice-based interactions (Drawehn and Pohl, 2023).

Although the umbrella term of digital assistant systems is highly flexible (Pereira et al., 2023), the essence of these systems lies in their ability to leverage natural language input for intuitive communication with human users (Budzinski et al., 2019). Modern digital assistants rely on voice input to provide fast, hands-free communication, increasing efficiency and usability in various professional contexts (Hoy, 2018). By automating routine tasks, such as transcribing meeting notes or generating reports, digital assistants enable professionals to focus on core activities like customer engagement or problem-solving, enhancing productivity across industries (Fraunhofer IAO, 2024).

At the heart of these functionalities lies precise STT transcription, which converts spoken language into text that digital assistants can process and interpret. However, while state-of-the-art STT models perform well with everyday language, they often fall short in domain-specific contexts. Challenges arise in accurately transcribing technical terminology, proper nouns, and multilingual content, which are common in specialized fields. This limitation hinders the broader adoption of digital assistants in industries where precise transcription is essential for processing customer data, product information, or technical terms.

The need for reliable transcription accuracy is particularly evident in industries with high levels of customer interaction, where digital assistants can automate tasks like updating CRM systems. In addition, they enable the creation of precise summaries from voice data. These systems help reduce administrative overhead by automating routine tasks, which mini-

[a] https://orcid.org/0009-0004-3090-0787
[b] https://orcid.org/0009-0004-6441-2263
[c] https://orcid.org/0000-0003-2556-3881

Pawlowicz, D., Weber, J. and Dukino, C.
Effectiveness of Whisper's Fine-Tuning for Domain-Specific Use Cases in the Industry.
DOI: 10.5220/0013378100003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1396-1403
ISBN: 978-989-758-737-5; ISSN: 2184-433X

mizes the time employees spend on repetitive activities. As a result, employees can focus more on strategic responsibilities, such as developing new business initiatives and enhancing customer relationships. Additionally, they have more time to engage in technical responsibilities, which involves troubleshooting complex issues and optimizing processes for greater efficiency (Fraunhofer IAO, 2024).

This paper explores the establishment of a comprehensive procedure for optimizing STT for domain-specific applications, which includes both fine-tuning Whisper, a state-of-the-art STT model, and implementing post-processing tailored to domain-specific requirements. Whisper's performance is evaluated in an industrial setting, focusing on its ability to handle complex technical terminology, multilingual content, and noisy environments. By addressing the gap between general-purpose STT capabilities and domain-specific requirements, this study contributes to the broader adoption of digital assistants in specialized fields.

2 METHODS

2.1 Model Selection

Based on an earlier evaluation, three models were picked for an assessment of the state-of-the-art STT performance in domain-specific transcription tasks: DeepSpeech (Hannun et al., 2014), wav2vec 2.0 (Baevski et al., 2020), and Whisper large-v3 (Radford et al., 2022) (Pfeiffer, 2024). These models were selected based on their established performance in prior research and their suitability for multilingual transcription, handling environmental noise, and adapting to diverse accents.

The evaluation focused on three key criteria:

- Accuracy: The accuracy was measured as the Word Error Rate (WER), which is the proportion of words transcribed incorrectly compared to the total number of words spoken. It includes substitutions, deletions, and insertions, calculated as WER $= \frac{S+D+I}{N}$, where S, D, and I are the above errors and N is the total number of words in the reference transcript (Trabelsi et al., 2022; Favre et al., 2013). The WER is popular and widely reported in the literature (Bandi et al., 2023).
- Supported Languages: Non-English languages are underrepresented in training datasets (Milde and Köhn, 2018; Huang et al., 2023a), which is why English models perform better (Bermuth et al., 2021). For this, it is essential to choose a model that performs well in German to ensure good accuracy in use cases.
- Domain Adaptability: Adaptability refers to how easily a STT system can be customized to meet the specific needs of different domains. Customization is particularly crucial for specialized applications, where general models may not perform optimally. Even if a model meets general performance criteria, it may still struggle when applied to the specific requirements of a digital assistant. Therefore, it is essential that the system can be easily adapted and trained to handle the nuances of the new domain.
  For instance, proper nouns are highly dependent on the specific use cases of Automatic Speech Recognition (ASR) systems, highlighting the need for domain-specific adaptation (Lee et al., 2021). Additionally speech patterns, including vocabulary, pronunciation, and tempo, change with age (Fukuda et al., 2020). If a model is trained predominantly on data from younger individuals, its accuracy may decline when used by older speakers. This principle can be generalized across different social groups, as it is well-established that speech varies according to factors such as social class and age.

Whisper stood out for these criteria, especially for its multilingual support and its popularity, which implies extensive resources and high adaptability. In addition, Whisper's architecture, based on large-scale transformer models, is designed to handle different accents and noisy environments, making it particularly suitable for real-world scenarios. Its ability to operate effectively in both pre-trained and fine-tuned configurations provided the flexibility to address domain-specific challenges while maintaining high baseline performance. Although Whisper is the leader in this area at the time of this writing, other models should be considered in the future as this field of research continues to evolve rapidly.

2.2 Whisper Model Evaluation

To assess the performance of the Whisper model in a domain-specific use case, the Whisper large-v3 model (Radford et al., 2022) was employed to transcribe a collection of audio recordings from customer visits in the mechanical engineering field. This assessment included evaluating how Whisper coped with challenging aspects such as environmental noise, accents, and technical terminology specific to the industrial sector. A sample of 21 recordings showed a mean WER of 11% (*min.* = 0.59%, *max.* = 50%). The detailed evaluation revealed that, while the model transcribed common words correctly, it showed some flaws.

1. The out-of-the-box Whisper model struggled with the accurate transcription of proper nouns, particularly customer and product names, which are critical in industrial applications. For instance, product names like ”HyperCut 1530” were occasionally mistranscribed due to the model's lack of contextual knowledge about domain-specific terminology. The accurate transcription of proper nouns not only relies on recognizing the correct word but also on understanding its specific spelling, which can vary significantly between different terms. This was further complicated by the fact that proper nouns were pronounced differently by different speakers in the recording, introducing room for error. This issue extended to other technical terms, where the absence of relevant training data in the default Whisper model led to inconsistent and sometimes nonsensical outputs. These limitations made manual correction necessary, particularly when the transcription involved critical customer or product information.

2. The Whisper model is multilingual, yet it is ill-equipped to handle bilingualism. While the majority of the recordings were in German, English words occasionally emerged in professional settings. This is particularly problematic for technical terms or proper nouns with English-like characteristics. In such instances, the default Whisper model primarily transcribed in German, as this was the language setting, resulting in errors such as ”Focus,” a component of a product name, being transcribed as ”Fokus.”

These issues highlight the importance of domain-specific fine-tuning to enhance Whisper's performance for specialized tasks. While Whisper demonstrated robust general-purpose transcription capabilities, adapting the model to specific use cases is necessary to ensure its reliability in professional environments. For instance, fine-tuning Whisper on domain-relevant data can improve its ability to handle technical terms, multilingual contexts, and structured data such as email addresses and product names.

This study aims to explore the effectiveness of fine-tuning Whisper for such domain-specific applications. By adapting the model to the linguistic and contextual nuances of a particular domain, the goal is to bridge the gap between general-purpose STT performance and the specific requirements of real-world use cases.

2.3 Data Collection

The fine-tuning process utilized a dataset of real-world audio recordings from field service employees of a mechanical engineering company. These recordings were collected to reflect actual usage scenarios, with content focusing on CRM notes, including product names, customer details, and locations. A total of 73 recordings, ranging from 16 seconds to 1 minute and 30 seconds in length, were gathered. To increase variability, some texts were recorded multiple times by different speakers, capturing accents, environmental noise (e.g., cars), and unclear pronunciations. The recordings were primarily in German, with occasional English words for technical terms and product names.

To enhance diversity and address gaps identified during initial evaluations, 14 additional texts were crafted and recorded by a female speaker. These supplemental recordings simulated real CRM notes while introducing domain-specific nuances, such as product terminology. Additionally, the selection of topics for the new texts was carefully made to ensure that they cover relevant scenarios that frequently occur in real-world applications. This not only improves the overall performance of the model but also helps in understanding and processing a broader range of user inquiries. This approach highlights the flexibility of the fine-tuning process, which can be adapted to other domains by tailoring datasets to their specific linguistic challenges.

2.4 Data Annotation

The audio recordings were first transcribed using Whisper's out-of-the-box model. Then an automatic post-processing dictionary was applied, using a rule-based approach to correct common mistakes in the original transcription. While this post-processing dictionary worked well for common misspellings, it was not robust enough to replace the fine-tuning of Whisper. The mean WER was still at 3% after the post-processing for the aforementioned 21 recordings (*min.* = 0%, *max.* = 14%). However it worked well to reduce manual annotation effort by reviewing the original transcript and the corrections instead of transcribing everything by hand.

Background noise, unclear speech, or heavily accented pronunciations were not explicitly annotated or flagged. Instead, such segments were corrected to the best of the annotator's ability. When speakers used unconventional terms or deliberately spelled out words (e.g., “dot” for “.”), the transcription was directly edited to reflect the appropriate punctuation mark. However, this method relied heavily on the assumption that the fine-tuned model would learn to handle such patterns automatically in future iterations. No special markings were used to indicate noisy sections, which might have affected the model's learning in these cases.

The transcriptions and audio paths were formatted into a JSON file as follows.

```
[
    {
    "audio_path": "path\to\audio.m4a",
    "transcription": "This is the
        correct transcription of the
        audio"
    }
]
```

2.5 Training

The recordings, initially in .m4a format, were converted to .wav with a sampling rate of 16 kHz to ensure compatibility with Whisper. Padding and truncation ensured uniform input lengths, with shorter inputs padded and longer ones truncated. An attention mask was used to differentiate meaningful data from padding tokens, preventing unexpected behavior during training as follows:

```
inputs = self.processor(audio,
    sampling_rate=sr, return_tensors="
    pt", return_attention_mask=True)
input_features = inputs.input_features.
    squeeze()
# Tokenize the transcription
labels = self.processor.tokenizer(
    transcription, return_tensors="pt",
     padding = "max_length", max_length
     =448,
truncation = True).input_ids.squeeze()
# Mark padding tokens to ignore loss
    there
labels = torch.where(labels == self.
    processor.tokenizer.pad_token_id,
    -100, labels)
attention_mask = inputs.attention_mask.
    squeeze()
```

The training loop consisted of forward and backward passes to optimize model weights using the AdamW optimizer. Key hyperparameters included:

- Batch size: Due to GPU memory limitations, batch sizes of 1 and 5 were tested. As seen in section 3 batch size 1 demonstrated better generalization and was used for further analysis.
- Epochs: Training was conducted for five epochs, as the validation loss stabilized after the fourth epoch, preventing overfitting.
- Learning rate: The AdamW optimizer adapted the learning rate. The starting learning rate was 5×10^{-5}.

To prevent overfitting, the dataset was split into training (80%) and validation (20%) sets using train_test_split from scikit-learn (Pedregosa et al., 2011), ensuring that the model's performance could be monitored on unseen data. The model's performance was evaluated using the WER and the absolute training and validation loss. This approach provided a clear view of model improvements across epochs.

3 RESULTS

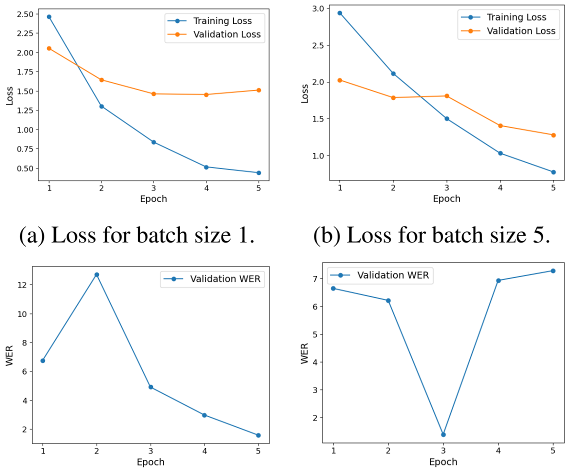
Comparing the results of models trained with batch sizes of 1 and 5 revealed distinct differences in their ability to generalize. For batch size 1, both training and validation losses steadily decreased (Figure 1a), with validation loss stabilizing after the third epoch. This indicates effective learning and a reduced risk of overfitting. In contrast, for batch size 5, while training loss decreased, validation loss fluctuated, suggesting difficulties in generalizing to unseen data (Figure 1b).

Similarly, WER trends (Figures 1c and 1d) show a consistent improvement with batch size 1, whereas batch size 5 exhibited fluctuations, particularly in later epochs. These observations led to the decision to focus on batch size 1 for further analysis, as it provided greater stability and better generalization despite higher computational costs.

Training for five epochs was found to be optimal. The loss graphs for batch size 1 showed continued reduction in training loss, but the stabilization of validation loss after the fourth epoch suggested that additional epochs might lead to overfitting. The training loss decreased from approximately 2.46 to 0.44 by the end of the fifth epoch, reflecting successful model learning. Validation loss initially mirrored the training loss but began to stabilize and increase slightly after epoch three, indicating potential overfitting if training continued beyond the fifth epoch.

WER trends confirmed this pattern. After an initial increase in WER in the first two epochs, the metric dropped significantly from epoch three onward, with the lowest WER of around 2% achieved at epoch five (Figure 1c). This demonstrates that the model progressively improved its transcription accuracy with fine-tuning, particularly for domain-specific terms, validating the choice of hyperparameters.

Fine-tuning the Whisper model addressed key limitations observed in its out-of-the-box performance, particularly with the transcription of proper nouns and bilingual terms. These challenges, crit-



(a) Loss for batch size 1.

(b) Loss for batch size 5.

(c) WER for batch size 1.

(d) WER for batch size 5.

Figure 1: Comparison between batch size 1 and batch size 5.

ical in domain-specific applications, were mitigated through the customization of Whisper to industrial data.

The out-of-the-box Whisper model struggled with the accurate transcription of proper nouns, including customer and product names, often producing inconsistent or incorrect results due to variability in pronunciation among speakers. This issue was compounded by the lack of domain-specific contextual knowledge, introducing further ambiguity.

After fine-tuning, the model showed significant improvements in its ability to transcribe proper nouns accurately, even when pronounced differently by various speakers. The inclusion of domain-specific training data allowed the model to better recognize and adapt to industrial terminology, reducing the need for manual corrections. Even new proper nouns with similar pattern such as product names were transcribed correctly with the fine-tuned model, addressing one of the primary flaws of the out-of-the-box model.

Fine-tuning on bilingual recordings improved the model's ability to handle mixed-language content. While some challenges with English terms persisted, likely cause by the pronunciation of some speakers, the model became better equipped to distinguish between German and English terms within the same transcription, correctly transcribing examples like ”Focus” in product names. This adaptation minimized errors in critical technical terminology, enhancing the usability of transcriptions in professional settings.

3.1 Testing the Final Model

The fine-tuned model was tested on new audio examples to evaluate its performance in real-world scenarios. Fine-tuning effectively addressed significant limitations of the out-of-the-box Whisper model, particularly in the transcription of domain-specific terms, technical jargon, and proper nouns. For example, product names like ”HyperCut 1530” and customer names spelled out in audio were consistently transcribed correctly, even when pronounced differently by various speakers. These improvements highlight the model's enhanced understanding of industrial terminology and its ability to adapt to speaker variability.

However, certain issues persisted. While the transcription of individual terms improved, the model exhibited challenges in maintaining coherence in longer sentences. Repetitive loops, such as repeating phrases unnecessarily, were still observed in complex inputs, indicating difficulty in capturing the semantic structure of extended discourse. Additionally, although the fine-tuned model showed better handling of bilingual content, occasional errors remained in cases where English terms were embedded in predominantly German audio (e.g., ”Focus” being transcribed as ”Fokus”).

Overall, the fine-tuning process significantly improved the model's ability to handle domain-specific vocabulary and context, addressing many of the critical flaws in the out-of-the-box configuration. However, limitations in sentence-level coherence and bilingual adaptability suggest room for further optimization, particularly in handling longer and more complex inputs.

4 DISCUSSION

4.1 Interpretation of Results

The STT transcription component was fine-tuned to a specific domain, focusing on accurately transcribing technical terms, product names, and customer-related information, to optimize the model for economical use in specific branches. A detailed qualitative evaluation of the transcriptions reveals several key strengths and areas of concern, providing valuable insights into the model's capabilities and limitations. The primary focus was on the accuracy of domain-specific vocabulary, the handling of environmental noise, and the system's overall ability to capture the intent and structure of spoken content.

One notable issue encountered during the evaluation of the fine-tuned model was the occurrence of repeated phrases. At certain points during transcription, the model would no longer produce coherent transcriptions of the audio but instead repeat previous or newly generated 2-3 word phrases until the transcription was terminated. This repetitiveness problem makes it challenging to quantitatively analyze and interpret the results. Resolving this issue is crucial for future work, as reliable and accurate transcription is foundational for any downstream NLP tasks.

It is possible that this problem is due to the limitations of Whisper, which is only designed to handle 30 second recordings. This theory is supported by the fact that the repetitive nature of the audio begins after 30 seconds. Although Whisper has a built-in parameter to overcome this, and the original model could transcribe longer sequences, it seems that fine-tuning the model conflicts with this setting. Future work will need to be done to overcome this significant challenge.

As the quantitative evaluation metrics like the WER suffered from the above described problem, they are not very meaningful to analyze the performance of the model, independent of the repetitive phrases. However it can be seen, that the fine-tuned model achieves a lower WER of 2% compared to the out-of-the-box Whisper model and the rule-based post-processing approach.

A qualitative interpretation reveals a notable improvement in correctly transcribing complex technical terms and industry-specific jargon, which is crucial in the context of field service reports. Terms such as product-specific names were consistently transcribed with high accuracy.

This level of precision is essential when documenting technical discussions or generating reports, as incorrect transcriptions could lead to misunderstandings and potential delays in customer support.

A key challenge in transcribing documentation is the accurate recognition of alphanumeric strings, like serial numbers, email addresses or phone numbers, which are often pronounced differently by different speakers. The model generally performed well in this regard, but occasional errors occurred when speakers used abbreviations or non-standard pronunciations. For example, variations in how certain product names were articulated led to inconsistencies in transcription, indicating a need for greater robustness in handling phonetically similar but contextually distinct terms.

One of the most challenging aspects of real-world audio is the presence of background noise, such as conversations, machinery sounds, or traffic noise, which can significantly impact transcription quality. The model exhibited a moderate level of noise robustness, effectively filtering out low-level background disturbances in controlled settings. However, in recordings with high levels of environmental noise or overlapping speech, the model struggled, leading to incomplete or inaccurate transcriptions. This limitation is particularly relevant in the field service domain, where recordings are often made in noisy environments such as workshops or cars.

The same results apply for speaker variation and accents. While the model generally performed well with native German speakers using a standard accent, it showed a decline in accuracy with regional dialects or non-native speakers. This was highlighted in the transcription of uncommon words or of English phrases in a German context, where the pronunciation is highly critical for recognition. Since technical terms or product names often contained English elements, this posed a problem for the model. This points to a need for more diverse training data that can better represent the full range of speaker variations encountered in the field.

Overall, while the STT model shows strong performance in key areas, its practical utility in unstructured and noisy environments remains an open challenge. Addressing these limitations by diversifying the training data would be a critical step in making the digital assistant a more versatile tool for field service professionals working in a variety of real-world conditions.

4.2 Implications

The results presented in this paper demonstrate the potential of leveraging fine-tuned machine learning models to improve the functionality and usability of digital assistant systems, specifically in the context of specific domains, that are underrepresented in the current training data for model training. The STT component showed strong performance in capturing and processing domain-specific language, technical terms, and context-specific entities. The insights gained through the qualitative analysis of the STT model confirm that targeted adaptations of NLP models can significantly enhance digital assistants in specialized industry settings.

The primary goal of this work was to simplify the interaction between users and the digital assistant by enabling intuitive, speech-based communication, thereby reducing administrative workload and allowing employees to focus on technical problem solving rather than laborious documentation tasks. The fine-tuned model achieved this goal by reliably transcribing key technical terms. Despite some limitations in handling environmental noise and non-standard pronunciations in the STT model, the system demonstrated its ability to support technicians in real-world scenarios by capturing essential details from their spoken notes and structuring them into actionable data, especially compared to the out-of-the-box Whisper model.

Another central objective of this project was to create a system that would seamlessly integrate into existing workflows without requiring extensive user training or customization. To achieve this, the digital assistant must be highly accurate, reliable, and able to handle the wide variety of linguistic input encountered in day-to-day operations. This is in line with the need to build trust between the system and its human users. Trust is a key factor in the successful adoption and long-term use of any digital assistant. A digital assistant that consistently performs well and delivers accurate results encourages employees to incorporate it into their daily routines. Conversely, negative experiences–such as low accuracy, misinterpretation of critical terms, or unexpected system behavior–can undermine user trust and make integration difficult. This work has achieved a promising step to this goal.

5 FUTURE STEPS

The immediate priority for future work should be to fully realize the potential of the assistant and address the problems identified in the STT model, particularly the issue of repeated phrases at the end of transcriptions. As highlighted in the discussion, this is likely due to the incorrect handling of the receptive field. Implementing the long-form algorithm in the fine-tuning would be the first step towards solving this problem.

In addition, improving the model's handling of environmental noise and speaker variation should be a high priority. To achieve this, more data augmentation techniques could be employed, such as adding synthetic noise to training samples or using vocal transformation methods to simulate different accents and speech rates. These adjustments would help the model to better generalize to the diverse conditions encountered in real-world scenarios.

In the short term, solving the current problems would make the transcription and therefore the digital assistant more reliable and user-friendly, encouraging adoption and reducing employee skepticism. In the medium term, enhanced models with greater robustness and expanded entity coverage would provide users with a tool capable of handling complex workflows, further reducing manual work and minimizing data entry errors.

Further enhancements should also take into account the fact that the field of ASR is fast-moving and progressive, and therefore domain-specific fine-tuning may be better supported by other models in the future.

In the long term, the development of a multimodal digital assistant with advanced contextual understanding would fundamentally change industrial workflows in a variety of sectors. Such an assistant would act as a true collaborator, capable of understanding nuanced technical discussions, providing real-time support, and integrating seamlessly into complex service environments. This would not only optimize the efficiency of service processes, but also improve customer satisfaction by enabling faster and more accurate resolution of technical issues.

6 CONCLUSION

In summary, this work has demonstrated the effectiveness of fine-tuning machine learning models for domain-specific applications in digital assistant systems within the field service context.

The industrial setting was chosen because it is underrepresented in the literature, where fine-tuning of language models for medical or legal domains is much more common (Yang et al., 2024; Huang et al., 2023b). However, this paper serves as a proof of concept of the efficacy and feasibility of fine-tuning Whisper to any specific domain with real-world recordings and use cases.

The STT component was successfully adapted to capture complex technical terms and industry-specific entities, providing a solid foundation for automating and streamlining documentation processes. The qualitative evaluation highlighted the system's strengths in handling structured language, while also identifying areas for further development to improve robustness and usability.

REFERENCES

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv (Cornell University)*.

Bandi, A., Adapa, P. V. S. R., and Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet*, 15(8).

Bermuth, D., Poeppel, A., and Reif, W. (2021). Scribosermo: Fast Speech-to-Text models for German and other Languages. *arXiv (Cornell University)*.

Budzinski, O., Noskova, V., and Zhang, X. (2019). The brave new world of digital personal assistants: benefits and challenges from an economic perspective. *Netnomics*, 20(2-3):177–194.

Drawehn, J. and Pohl, V. (2023). Einsatz von KI mit Fokus Kundenkommunikation.

Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S., et al. (2013). Automatic human utility evaluation of ASR systems: Does WER really predict performance? In *INTERSPEECH*, pages 3463–3467.

Fraunhofer IAO (2024). DafNe: Digitaler Außendienst-Assistent für Nebenzeitoptimierung. Retrieved December 10, 2024.

Fukuda, M., Nishizaki, H., Iribe, Y., Nishimura, R., and Kitaoka, N. (2020). Improving Speech Recognition for the Elderly: A New Corpus of Elderly Japanese Speech and Investigation of Acoustic Modeling for Speech Recognition. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6578–6585, Marseille, France. European Language Resources Association.

Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv (Cornell University)*.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, 37(1):81–88.

Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., and Wei, F. (2023a). Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., and Feng, Y. (2023b). Lawyer llama technical report.

Lee, T., Lee, M.-J., Kang, T. G., Jung, S., Kwon, M., Hong, Y., Lee, J., Woo, K.-G., Kim, H.-G., Jeong, J., Lee, J., Lee, H., and Choi, Y. S. (2021). Adaptable Multi-Domain Language Model for Transformer ASR. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7358–7362.

Milde, B. and Köhn, A. (2018). Open source automatic speech recognition for german.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pereira, R., Lima, C., Pinto, T., and Reis, A. (2023). Virtual Assistants in Industry 4.0: A Systematic Literature Review. *Electronics*, 12(19).

Pfeiffer, M. (2024). Comparison of existing technologies and models for the design of an AI-supported digital assistance system.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

Trabelsi, A., Warichet, S., Aajaoun, Y., and Soussilane, S. (2022). Evaluation of the efficiency of state-of-the-art Speech Recognition engines. *Procedia Computer Science*, 207:2242–2252. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022.

Yang, Q., Wang, R., Chen, J., Su, R., and Tan, T. (2024). Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise.