Refining Transcripts With TV Subtitles by Prompt-Based Weakly Supervised Training of ASR

1st Xinnian Zhao
Department Electrical Engineering ESAT-PSI
KU Leuven University
Leuven, Belgium
xzhao1@esat.kuleuven.be

2nd Hugo Van Hamme
Department Electrical Engineering ESAT-PSI
KU Leuven University
Leuven, Belgium
hugo.vanhamme@esat.kuleuven.be

***Abstract*—This study proposes a novel approach to using TV subtitles within a weakly supervised (WS) Automatic Speech Recognition (ASR) framework. Although TV subtitles are readily available, their imprecise alignment with corresponding audio limits their applicability as supervised targets for verbatim transcription. Rather than using subtitles as direct supervision signals, our method reimagines them as context-rich prompts. This design enables the model to handle discrepancies between spoken audio and subtitle text. Instead, generated pseudo transcripts become the primary targets, with subtitles acting as guiding cues for iterative refinement. To further enhance the process, we introduce a weighted attention mechanism that emphasizes relevant subtitle tokens during inference. Our experiments demonstrate significant improvements in transcription accuracy, highlighting the effectiveness of the proposed method in refining transcripts. These enhanced pseudo-labeled datasets provide high-quality foundational resources for training robust ASR systems.**

***Index Terms*—speech recognition, weakly supervised training, subtitle prompts, weighted attention**

I. Introduction

Recently, foundation models have emerged as a dominant force in the field of artificial intelligence, offering extensive knowledge bases for various tasks and modalities [1]. These large-scale pre-trained models have significantly enhanced the performance of many downstream applications by leveraging vast datasets [2]. However, the datasets used in these models are not uniform across all tasks or domains, leading to discrepancies in performance [3], [4]. Whisper, a widely known foundation speech model, faces the same challenge [5]. It is trained on a large amount of web data along with a small portion of labeled data with weak supervision. Imbalances in data sizes and quality pose great challenges in generalizing the Whisper model effectively to underrepresented domains and languages [6]–[8]. Fine-tuning on labeled data from the target domain or language is a common strategy to bridge the gap [9]. However, in low-resource scenarios, obtaining sufficient labeled data for real-world applications remains a challenge. To address this, we explore methods to fine-tune the pre-trained Whisper in a weakly supervised (**WS**) framework, and aim to refine the generated transcripts with TV subtitles.

Foundation Flanders (FWO) under grant S004923N of the SBO programme.

TV subtitles are a readily available resource, featuring concise and clear text specifically designed for readability [10]. They are often used as weak labels for ASR training due to their lack of precise alignment with audio [11], [12]. However, these methods typically require subsequent fine-tuning with verbatim transcripts to meet transcription accuracy objectives.

In this study, we use subtitles as prompts for the Whisper text decoder instead of taking them as training targets. We experiment on Flemish, a Dutch dialect. The primary weak labels for training are pseudo transcripts generated by the pre-trained Whisper model. Subtitles, acting as prompts, provide additional contextual cues for transcription, facilitating the iterative refinement of transcripts. To further enhance performance, we introduce a weighted attention (**WA**) mechanism during inference. This mechanism selectively emphasizes relevant words in subtitle prompts, aligning them more closely with the spoken audio while disregarding irrelevant words. The present work on integrating subtitle prompting (**SP**) training and WA inference yields measurable improvements in word error rate (WER), demonstrating the effectiveness of our approach in creating higher-quality datasets for weakly supervised ASR applications without requiring additional labeled data. The key contributions of this paper are:

- Development of a training methodology that incorporates SP within a WS frame work, handling the misalignment between subtitles and speech.
- Introduction of a WA mechanism for inference, enhancing the model's ability to extract relevant information from subtitles.
- Demonstration of how this approach improves dataset quality for WS ASR, with a focus on underrepresented and low-resource domains.

II. Related Work

The generative text decoder in Whisper supports prefix prompting, which has inspired several studies exploring its potential across various tasks. [13], [14] designed particular prompts to adapt Whisper for zero-shot tasks, while [15]–[17] fine-tuned pre-trained Whisper models using domain-specific prompts. Our work differs from these studies in several key aspects. First, the lack of precise verbatim transcripts necessitates training in a weakly supervised setting. Second, subtitles

are unsuitable as direct prompts for a pre-trained Whisper model, as their overlap with audio content can mislead the model, causing truncated outputs. Finally, a WA mechanism is introduced during inference, which addresses the unique challenges of using subtitle prompts.

III. Methods

A. SP Training

The objective of the task is to condition transcript generation on subtitle prompts as contextual information. Any model with a generative decoder is suitable for this purpose. We select Whisper due to its robust weakly supervised training on large-scale, multilingual web-sourced data, which enhances its transferability to our spontaneous Flemish broadcast audio.

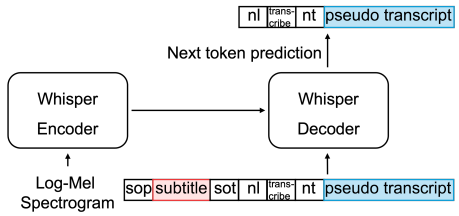


Fig. 1. Framework of subtitle prompting Whisper.

Figure 1 illustrates the framework for Whisper SP, which integrates subtitles as context prompts and pseudo transcripts as training targets to fine-tune the model. To leverage transfer learning from generic Dutch, we fine-tune the <|nl|> language token using the Flemish training data. The final input to the decoder during training is structured as <|sop|>subtitles <|sot|><|nl|><|transcribe|><|notimestamps|>.

We denote the subtitled dataset as X, Y_s, where X is the audio and Y_s is the paired subtitles. The pseudo transcripts Y_{pt_0} are first generated for X using a pre-trained Whisper model. Whisper is known to hallucinate by repeating the same word in its output [18], especially when the audio is heavily overlapped. We remove these rare errors by filtering based on transcript length. The filtered transcripts serve as the starting point for the iterative training process. With each iteration, the model refines the pseudo transcripts by leveraging the contextual information provided by the subtitle prompts. After training over the entire dataset once, the updated pseudo labels are denoted as Y_{pt_1}. This process is repeated over t iterations, gradually evolving the pseudo labels from Y_{pt_0} to Y_{pt_t}.

It should be noted that the proposed iterative training differs from self-training, where only pseudo transcriptions are used for training, requiring error control strategies to prevent error propagation [19]. In our approach, subtitle prompts provide additional information, while the loss is computed on the transcript. This allows the model to progressively refine the pseudo transcripts until the information in subtitles is fully extracted.

B. WA inference

Subtitles often overlap with speech content while also containing extraneous information not reflected in the audio. The relevance of subtitle tokens to speech can be evaluated using cross-attention weights. Tokens highly relevant to speech are expected to exhibit strong and concentrated cross-attention weights to specific speech frames. In contrast, distraction tokens tend to have relatively uniform attention weights spread across all frames. Intuitively, relevant tokens are better suited to guide transcription through self-attention. Based on this principle, we aim to select relevant tokens while minimizing the influence of irrelevant ones.

To achieve this, we introduce a WA mechanism that employs the Gini coefficient [20] as a measure of relevance. The Gini coefficient, commonly used in economics to measure inequality in a distribution, is adapted here to quantify the distribution of cross-attention weights over input speech frames. The formula for the Gini coefficient is defined as:

$$g_i = \frac{\sum_{k=1}^{N}(2k - N - 1) \cdot \mathcal{CA}[i,k]}{N \cdot \sum_{k=1}^{N} \mathcal{CA}[i,k]}, \tag{1}$$

where $\mathcal{CA}$ represents the cross-attention weight matrix, i denotes a specific token in the subtitle prompt, k indexes speech features of the input frames, and N is the total number of speech frames. The cross-attention weights $\mathcal{CA}$ are derived from the first cross-attention layer, where higher-level text representations have not yet been computed, providing a clearer and more direct mapping from text tokens to audio inputs. At this layer, the attention map exhibits the most straightforward monotonic alignment between text and speech. Before computing g_i, the values of $\mathcal{CA}[i,k]$ are sorted cross N frames to ensure a positive value for g_i. The final value of g_i ranges from 0, indicating that $\mathcal{CA}$ has a uniform distribution, to 1, signifying that $\mathcal{CA}$ exhibits a highly unequal distribution, resembling a sharp focus.

The Gini coefficient calculated over the prompt sequence in length T_p is denoted as $\mathcal{G}$, which is then utilized to weight the prompt values in self-attention layers as follows:

$$K' = \mathcal{G}K_p \oplus K_t; \quad V' = \mathcal{G}V_p \oplus V_t, \tag{2}$$

Here, K and V represent the key and value matrices, respectively, with subscripts p and t denoting the prompt and transcript targets. The matrices K_p and K_t, with p and t sequences in lengths of T_p and T_t, are concatenated along the time dimension using $\oplus$, resulting in a final sequence of length T that matches with the input length. Thus, the dimensions of the final K' and V' matrices remain unchanged. This weighting mechanism ensures that subtitle tokens most relevant to the speech input are emphasized during the attention computation.

The final self-attention output is then calculated as:

$$H = \text{softmax}\left(\frac{QK'^{\top}}{\sqrt{d}}\right)V', \tag{3}$$

where Q represents the query derived from the prompt and the transcript, and d denotes the model dimension. Through Eq. 3,

the model is guided to prioritize speech-relevant tokens in subtitle prompts while ignoring distractions, thereby enhancing transcription accuracy.

IV. Experiments

A. Data

The subtitled dataset used in this study consists of 760 hours of multi-genre recordings from 16 TV programs broadcast by the Vlaamse Radio- en Televisieomroeporganisatie (VRT, Flemish Radio and Television Broadcasting Organisation) between September 2020 and November 2022. These recordings span four primary genres—news, talk shows, entertainment, and drama—covering a broad spectrum of topics, including politics, economics, education, culture, and sports. Most audio segments in the dataset are around 15 seconds long, with a maximum duration of 30 seconds, aligning with the input window requirements of Whisper. These segments are typically longer than the corresponding subtitle timings, which are kept short for readability. No additional pre-processing was performed on the audio; non-speech sounds, such as music, applause, and ringtones, are preserved within subtitle intervals. This simplifies the speech recognition pipeline while retaining the inherent challenges posed by spontaneous speech.

For reliable evaluation, 6 hours of speech (approximately 2,600 utterances) from each genre were manually annotated with verbatim transcripts. This annotated subset, referred to as *subs-annot*, was excluded from the training set. To evaluate the suitability of the subtitles for the ASR task, the WER of the subtitles in *subs-annot* was computed using the verbatim transcripts as the reference. The resulting WER of 34.3% highlights the misalignment between subtitles and the speech content, motivating our decision to use subtitles as prompts rather than direct training targets.

B. Training details

The models were implemented using the HuggingFace Transformers library. Training was conducted on a single Nvidia H100 GPU with an effective batch size of 64, using the Adam optimizer with a learning rate of 1×10^{-5} and 1,000 warmup steps. The models were trained iteratively across multiple epochs.

V. Results and Discussions

A. SP Training

Experiments were conducted to enhance transcripts using SP without requiring verbatim data. To demonstrate that SP provide contextual clues for transcript generation, we compared the WERs (%) on the *subs-annot* set across pre-trained Whisper models (medium and large variants) and models fine-tuned with or without SP. The results are presented in Table I.

The pre-trained Whisper models exhibit strong zero-shot performance on our Flemish broadcast dataset, achieving WERs below 20% for both the medium and large models. These results indicate that the transcripts are nearly intelligible and highlight Whisper's impressive generalization capabilities. We did not report results for directly prompting the pre-trained Whisper models with subtitles, as this configuration consistently produced truncated transcripts or no output at all. This behavior is likely due to significant overlap between subtitle prompts and speech content. The model tends to align the subtitle prompts with the audio and avoids repeating the prefix, resulting in incomplete outputs. This observation underscores the limitations of Whisper's text decoder in handling prompts without careful design and emphasizes the importance of tailoring prompts to prevent disruptions in transcript generation.

TABLE I
Comparison of WERs (%) on the *subs-annot* set between pre-trained, no-prompting (NP) fine-tuned and SP fine-tuned Whisper medium or large models.

Model	Whisper-medium	Whisper-large-v3
Pre-trained	18.75	13.07
NP Fine-tuned	13.05	21.49
SP Fine-tuned	**11.54**	**11.37**

We transcribed all the audio in the training set using the pre-trained Whisper-large-v3 model. The resulting pseudo transcripts Y_{pt_0} were used in the first iteration of fine-tuning. Not surprisingly, self-training the Whisper-large-v3 model with Y_{pt_0} resulted in a significant increase in WER, as shown in the second row of Table I. This suggests error propagation, which is a well-known challenge in self-training [19]. We did not apply any error control strategies, as our goal was to demonstrate the effectiveness of SP in refining errors in pseudo transcripts. Analyzing the errors in the outputs reveals that deletions account for more than 80% of the total errors. Specifically, the outputs show that the model often fails to transcribe entire utterances or their beginnings. This issue is particularly pronounced in utterances that initially had deletions at the start in Y_{pt_0}, potentially caused by inaccuracies in segmenting speech. Consequently, the model struggles with determining when to begin transcribing. As a result, the large model, being highly parameterized, may inadvertently reinforce its own biases and errors during self-training, leading to degraded performance.

In contrast, fine-tuning the Whisper-medium model with Y_{pt_0} led to improved performance. This is likely because the medium model had more capacity to benefit from Y_{pt_0}, considering the WER gap between the pre-trained medium and large models. Incorporating SP into fine-tuning effectively addresses the self-training challenge in large models. The WER decreased for both the medium model and the large model compared to fine-tuning without prompts, even though the models were trained for only one iteration. This demonstrates that subtitles provide additional information to guide generation, leading to more refined transcripts as a result.

Although subtitles do not provide the exact speech content, they often include the most informative words necessary for comprehension, such as specific named entities. These words are typically rare in the training set and therefore challenging to transcribe. To gain deeper insight into the information

TABLE II
RWERS (%) AND OWERS (%) TESTED ON RARE WORDS AND OUT-OF-VOCABULARY WORDS FROM THE TEST SET.

Model	Whisper-medium		Whisper-large-v3	
	rWER	oWER	rWER	oWER
Pre-trained	38.42	77.34	29.74	71.94
NP Fine-tuned	31.04	74.94	31.90	72.32
SP Fine-tuned	**24.63**	**70.22**	**24.23**	**69.64**

provided by subtitles, we break down the WERs in Table I into rare word error rate (rWER) and out-of-vocabulary (OOV) word error rate (oWER), with the results presented in Table II. Here, rare words are defined as words with frequency lower than 10 in the training set, while OOV words are completely absent from the training set. There are 1,762 rare words and 1,333 OOV words out of 76,684 words in the reference transcripts of *subs-annot* set.

From Table II, OOV words exhibit higher WERs than rare words overall, which aligns with intuition. SP consistently benefits the transcribing in terms of both rWER and oWER. When integrating SP to fine-tuning, the rWER decreasing significantly from 31.04% to 24.63%. This is notable because fine-tuning without prompts already exposed the model to rare words, resulting in an visible improvement in rWER compared to the zero-shot results from the pre-trained model. Similar decreases are observed in oWER, suggesting that SP not only help the model recall rare words but also improve its generalization ability. The large model exhibits slightly better rWER and oWER compared to the medium model after incorporating SP. Notably, SP effectively address the challenges revealed by self-training. By introducing additional information and providing guidance for predictions, subtitles help the model focus on refining its transcription output. Consequently, SP enhances robustness and mitigate the potential risks associated with weakly supervised training.

B. WA inference

After fine-tuning with SP, the WERs of the medium and large models converge to similar levels. Therefore, subsequent experiments are conducted using the medium model for convenience. During inference, Gini coefficients are computed from the first layer's cross-attention weights, as defined in Eq. 1. These coefficients are then applied to the self-attention mechanism at each or all layers to identify the most active one. The resulting WERs (%) are presented in Fig. 2.

Since WA is applied solely during inference and only the test set (*subs-annot*) contains verbatim transcripts, we divide this set into five folds to simulate different test scenarios and ensure robust layer selection. Each contains approximately 500 utterances, represented by different colors in the figure. Testing is performed on each fold to ensure consistent results across the dataset. As shown in Fig. 2, WER decreases across all folds when Gini attention weights are applied, demonstrating the robustness of this mechanism across diverse subsets. Notably, the best WER for each fold is consistently achieved when the

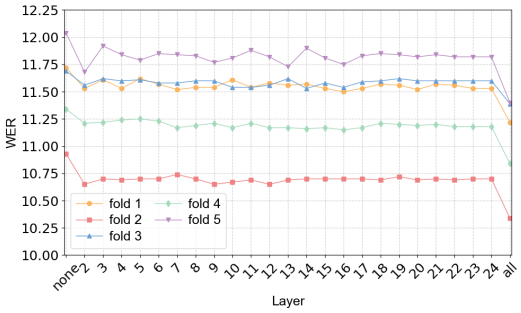


Fig. 2. WERs (%) on the 5 folds of *subs-annot*, obtained by applying Gini attention weights to individual layers or all layers of transformers.

Gini weights are applied to all layers. Under this configuration, the WER on the entire *subs-annot* set improves from 11.54% to 11.02%.

For ablation, we also experiment with max- and entropy-based weights derived from the cross-attention weights between prompt tokens and speech frames. The maximum weight relies heavily on a single frame and also assigns higher weights for irrelevant tokens on average. This makes it highly sensitive to noise and increases the risk of corruption. The results show that the max weighting leads to a high proportion of hallucinations. Entropy values vary significantly, with fully relevant tokens having an entropy of zero and most other tokens exhibiting large values. When normalized to $[0, 1]$, many tokens with a large entropy are compressed to nearly zero, and thus lose their sensitivity. WER results indicate that only Gini weights effectively refine the transcript, highlighting the specificity of the association between prompt tokens and speech frames. Table III presents an example of output transcripts generated using Gini-, max-, and entropy-based weighted attention strategies, with errors underlined. The results show that subtitle prompts can easily refine the format and spelling of named entities in the pseudo labels. However, errors such as "kenden" and "kennen", which both are syntactically and semantically correct, are more challenging to correct. Gini weights provide a strong cue in such cases, whereas both max- and entropy-based weights fail to refine these errors. Moreover, the max weighting tends to overlook certain speech frames, indicating its instability.

C. Iterative training

As training progresses, the model progressively improves its ability to generate accurate transcripts. The training targets are iteratively updated with newly generated transcripts at the end of each training cycle. The WERs achieved over three iterations of SP fine-tuning, followed by WA inference, are presented in Table IV.

The WERs consistently decrease across iterations, although the rate of improvement diminishes with each subsequent iteration in both fine-tuning and inference processes. This trend suggests convergence in the iterative learning process, driven by the refinement of pseudo transcripts and subtitle prompts. To balance performance and efficiency, training is halted after

TABLE III
AN EXAMPLE OF OUTPUT TRANSCRIPTS BY USING SP FINE-TUNING, INCORPORATED WITH GINI, MAX OR ENTROPY-BASED WA INFERENCE.

reference	als je dan dit ziet. we kenden natuurlijk Operatie Zero. maar als je dan dit ziet hoe pijnlijk is dat dan?
subtitle prompt	we kenden natturlijk Operatie Zero, maar als je dit ziet, hoe pijnlijk is dat?
pseudo label	als je dan dit ziet, we kennen natuurlijk operaties zero, maar als je dan dit ziet, hoe pijnlijk is dat dan?
SP fine-tuning	als je dan dit ziet, we kennen natuurlijk Operatie Zero, maar als je dan dit ziet, hoe pijnlijk is dat dan?
SP fine-tuning + Gini WA inference	als je dan dit ziet, we kenden natuurlijk Operatie Zero, maar als je dan dit ziet, hoe pijnlijk is dat dan?
SP fine-tuning + Max WA inference	als je dan dit ziet _ _ _ _ _ _ _ _ _ _ hoe pijnlijk is dat dan?
SP fine-tuning + Entropy-based WA inference	als je dan dit ziet, we kennen de natuurlijk Operatie Zero, maar als je dan dit ziet, hoe pijnlijk is dat dan?

TABLE IV
WERS (%) ACROSS ITERATIONS AFTER SP FINE-TUNING AND WA INFERENCE.

	SP fine-tuning	WA inference
iter1	11.54	11.02
iter2	10.82	10.66
iter3	10.52	**10.34**

three updates to the training targets. Ultimately, a WER of 10.34% is achieved on the *subs-annot* set, highlighting the effectiveness of the iterative approach.

VI. CONCLUSION

In this paper, we propose a method to refine transcripts from existing models using subtitles, eliminating the need for any verbatim dataset. We demonstrate the effectiveness of fine-tuning models with subtitles as prompts, achieving significant improvements in WER. Analysis of rare word and OOV WERs further confirms that subtitle prompts not only enhance the model's recall of low-frequency words but also improve generalization to OOV words. Additionally, the weighted attention mechanism we introduce further boosts transcription performance. Our results also show that the transcripts can be refined iteratively through successive training cycles. This work provides a promising direction for enhancing data quality in weakly supervised ASR systems without requiring additional labeled data.

REFERENCES

[1] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al., "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, p. 100017, 2023.

[2] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[3] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[4] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–40, 2023.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[6] Riefkyanov Surya Adia Pratama and Agit Amrullah, "Analysis of whisper automatic speech recognition performance on low resource language," *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 1, pp. 1–8, 2024.

[7] Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass, "Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages," in *Proc. Interspeech 2023*, 2023, pp. 2268–2272.

[8] Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain, "Benchmarking whisper for low-resource speech recognition: An n-shot evaluation on pashto, punjabi, and urdu," in *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, 2025, pp. 202–207.

[9] Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu, "Exploring native and non-native english child speech recognition with whisper," *IEEE Access*, vol. 12, pp. 41601–41610, 2024.

[10] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1026–1033.

[11] Kritika Singh, Vimal Manohar, Alex Xiao, Sergey Edunov, Ross Girshick, Vitaliy Liptchinsky, Christian Fuegen, Yatharth Saraf, Geoffrey Zweig, and Abdelrahman Mohamed, "Large scale weakly and semi-supervised learning for low-resource video asr," pp. 3770–3774, 2020.

[12] Mengli Cheng, Chengyu Wang, Xu Hu, Jun Huang, and Xiaobo Wang, "Weakly supervised construction of asr systems with massive video data," *arXiv preprint arXiv:2008.01300*, 2020.

[13] Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," *arXiv preprint arXiv:2305.11095*, 2023.

[14] Mohan Li, Simon Keizer, and Rama Doddipatla, "Prompting whisper for qa-driven zero-shot end-to-end spoken language understanding," in *Proc. Interspeech 2024*, 2024, pp. 1330–1334.

[15] Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, and Da-shan Shiu, "Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[16] Hao Ma, Zhiyuan Peng, Mingjie Shao, Jing Li, and Ju Liu, "Extending whisper with prompt tuning to target-speaker asr," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12516–12520.

[17] Jinpeng Li and Wei-Qiang Zhang, "Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11211–11215.

[18] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane, "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.

[19] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov, "Self-training: A survey," *Neurocomputing*, vol. 616, pp. 128904, 2025.

[20] Robert Dorfman, "A formula for the gini coefficient," *The review of economics and statistics*, pp. 146–149, 1979.