Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization

Puyuan Peng[1], *Brian Yan*[2], *Shinji Watanabe*[2], *David Harwath*[1]

[1]Department of Computer Science, The University of Texas at Austin, USA
[2]Language Technology Institute, Carnegie Mellon University, USA

pyp@utexas.edu

Abstract

We investigate the emergent abilities of the recently proposed web-scale speech model **Whisper**, by adapting it to unseen tasks with prompt engineering. We selected three tasks: audio-visual speech recognition (AVSR), code-switched speech recognition (CS-ASR), and speech translation (ST) on unseen language pairs. We design task-specific prompts, by either leveraging another large-scale model, or simply manipulating the special tokens in the default prompts. Experiments show that compared to the default prompts, our proposed prompts improve performance by 9% to 45% on the three zero-shot tasks, and even outperform SotA supervised models on some datasets. In addition, our experiments reveal many interesting properties of Whisper, including its robustness to prompts, bias on accents, and the multilingual understanding in its latent space. Code is available here

Index Terms: speech recognition, audio-visual speech recognition, speech translation, zero-shot learning, task adaptation, web-scale speech models

1. Introduction

The study of large scale foundation models [1] has become ubiquitous in many areas of AI, such as large language models for natural language processing [2, 3, 4], vision-and-language models for computer vision [5, 6]. One of the most intriguing aspects of these large scale pretrained models is their **emergent ability** [7], usually invoked by **prompting** [8], to generalize to unseen data or tasks [2, 5]. In addition to its scientific value, the zero-shot generalization capability of large scale models alleviates the burden of collecting specialized datasets or training special-purpose models for new tasks and domains, resulting in tremendous impact on the application of AI.

In the field of audio and speech processing, prompt engineering has only recently started to attract attention. Gao et al. [9] finetuned a wav2vec2 model [10] to produce tokens as prompt for the frozen GPT-2 [11] to do speech and audio classification tasks. Concurrently, Chang et al. [12] studied gradient-based prompt tuning on a pre-trained speech unit language model [13] for speech classification and generation tasks. Kim et al. [14] combined learnable prompts and adapters for efficient finetuning of audio models. Xue et al. [15] is the most similar work to ours. In that paper, the authors trained a Transformer-Transducer model using in-house data on a comparable scale to Whisper, and they ran test time gradient-based adaptation to fine-tune the model for speech translation on unseen language pairs. Our work is different from theirs because our adaptation methods are prompt-based and gradient-free, and we study three different zero-shot tasks instead of just one.

*Our work reveals and analyzes the hidden talent and weaknesses of **Whisper** [16].* It is the first of its kind that studies **gradient-free zero-shot task generalization** abilities of web-scale speech models. We show that **Whisper can be easily adapted to unseen tasks by simply modifying its prompt**. The effectiveness of our proposed prompts are validated on three tasks - **audio-visual speech recognition (AVSR)**, **code-switched speech recognition (CS-ASR)**, and **speech translation (ST) on unseen language pairs**.

2. The Whisper model

Here we briefly describe the Whisper model family [16] with an emphasis on the structure of its default prompt. Whisper is a family of Transformer-based encoder-decoder models [17] with parameters ranging from 39M (Tiny and Tiny.en) to 1.55B (Large and LargeV2). Whisper models can be categorized into two classes based on languages and tasks: English-only models and multilingual models. The multilingual models are trained on 630k hours of web-scraped speech data for multilingual automatic speech recognition (ASR), En→X speech translation (ST), language identification (LID), and timestamp prediction. The English models are trained on the English subset of the data (438k hours) for ASR and timestamp prediction. The encoder of Whisper models takes in log Mel spectrogram, and produces features for the decoder. The decoder consumes encoder features, positional embeddings, and a prompt token sequence. It then produces the transcription of the input speech, or alternatively its translation depending on the prompt. The prompt used in the original Whisper paper is the following: `<|sop|>previous text<|sot|><|language|><|task|><|notimestamps|>`[1]. Those encapsulated in `<||>` are special tokens. `previous text` represents the transcript of the previous utterance, and is optional. For multilingual models, `<|language|>` should be replaced by one of the 99 language tokens that Whisper encountered during training. When the input language is unknown at inference, Whisper will first run LID which results in a probability distribution over the 99 languages, and the language with the highest probability is chosen to fill the `<|language|>` token. `<|task|>` will be replaced by either `<|asr|>` or `<|st|>` depending on whether the model should perform ASR or ST. We keep `<|notimestamps|>` in all prompts as our tasks do not need Whisper to produce timestamps[2].

In all three zero-shot tasks that we consider in this paper, **we only modify the prompt to the Whisper decoder without modifying the model weights or architecture.** See table 1 for a summary of our proposed prompts.

[1]We use `<|sop|>` to abbreviate `<|startofprev|>`, and `<|sot|>` for `<|startoftranscript|>`. Also `<|asr|>` for `<|transcribe|>`, and `<|st|>` for `<|translate|>` later.

[2]and therefore we omit this token in the rest of the paper

Table 1: *Summary of our proposed prompts and relative improvement over the default prompts. The differences between our prompt and the default are in* ***bold****. In the AVSR task,* `CLIP retrie.` *stands for "CLIP retrieved objects", and* `<default>` *stands for* `<|sot|><|en|><|asr|>`*, please find detailed description of our prompt for AVSR in section 3. For each task only one case is shown in the table, and similar improvements are shown across different datasets and languages in the main text.*

Task	Language(s)	Default prompt	Our proposed prompt	Improvement
AVSR	En	`<|sot|><|en|><|asr|>`	**`<|sop|>CLIP retrie.`**`<default>`	9%
CS-ASR	Zh+En	`<|sot|><|zh|>or<|en|><|asr|>`	`<|sot|>`**`<|zh|><|en|>`**`<|asr|>`	19%
ST	En→Ru	`<|sot|><|ru|><|st|>`	`<|sot|><|ru|>`**`<|asr|>`**	45%

3. Audio-visual speech recognition

The first task is using an ASR system to produce transcription for a video, where the on-screen visual content is semantically related to the speech audio and can therefore aid in recognition [18, 19]. This task is related to, but more general than, performing audio-visual speech recognition (AVSR) on speech audio accompanied by a video of the speaker's facial or lip movements [20].

Approach. Our approach is shown in figure 1. To provide Whisper with a visually-conditioned prompt, we utilize the popular vision-and-language CLIP [5] model along with an external vocabulary of common object words to first 'convert' the visual stream into a sequence of word tokens. To do so, we take every word/phrase in the external vocabulary, construct a sentence with template "This is a photo of a { }". Then we use the CLIP text encoder to pre-compute an embedding vector for each sentence in an offline fashion. At inference time, for each video we sample 3 equally-spaced RGB image frames, use the CLIP image encoder to embed them, and calculate the similarity between the image embeddings and the pre-computed text embeddings. We select the top K objects whose embeddings have the highest similarity scores with the image embeddings for the prompt. Next, we concatenate the K selected object names into a comma-separated list of words, and insert this token sequence into the `previous text` slot of the prompt. This method draws inspiration from the idea of Socratic Models [21], where an engineered interface enables large pretrained models to 'talk' to one other to solve a complex task.

Datasets and implementation details. Our main dataset for the AVSR task is the recently proposed VisSpeech [19], which is a subset of the instructional video dataset HowTo100M [22]. VisSpeech consists of those videos where an audio-only baseline ASR system performs badly, and whose visual stream and speech audio are semantically related. Since VisSpeech is proposed as a test set and it only contains 508 examples, we use another instructional video dataset, How2 [18], for hyperparameter tuning. We use a randomly selected 2000 example subset of How2, and add pub noise to the audio to increase the ASR difficulty similar to [19], since the dataset has been shown to be biased towards clean audio [19] preventing the visual modality from offering significant benefit to its ASR task. For the external object vocabulary, we follow [21] and used the label set of Tencent ML-Images [23], which contains around 10,000 common objects. The number of object K used in the prompt is tuned for each Whisper model separately on our version of the How2 dataset with three different noise levels (SNR=5,0,−5dB).

Results. We found that on our How2 tuning set, using very large number of objects (as many as 90 objects) does not hurt performance. Our manual inspection shows that even when using 30 objects, there are already many irrelevant ones that got mis-retrieved by CLIP. For example, in the example shown in figure 1, we found 'yogurt', 'heavy cream', and 'mayonnaise' in the visual prompt. In addition, more than 90% of the utterances

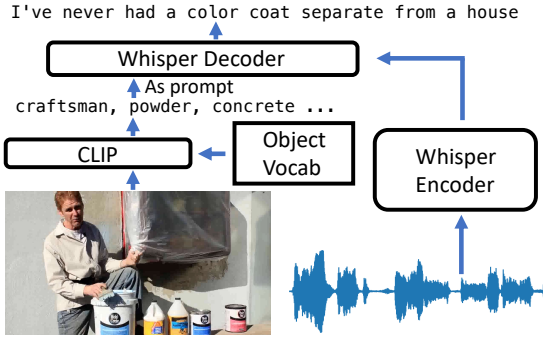

Figure 1: *Framework for visually prompting Whisper. The external object vocab is dataset agnostic.*

in our How2 dataset have a ground truth transcription less than 30 words. *This shows that Whisper is very robust to the noise and length of the prompt.*

For each model, the top 3 best performing number of object choices selected from How2 are used for the experiments on VisSpeech[3], and the average WER is shown in figure 2. We see that the *visually-informed prompt improves the performance for all four English models and three smaller multilingual models, but hurts the performance of the multilingual models Medium, Large, and LargeV2.* In table 2, we compare the previous SotA AVSR results on VisSpeech with the audio-only Whisper performance, and Whisper Medium.en with 50 objects as the visual prompt. We highlight that *visual prompt improve Medium.en by 9%, and even outperforms Large.*

Remarks. We propose a prompting approach that that adapts the audio-only Whisper for audio-visual speech recognition. Based on figure 2, visual prompting helps most of the models with the exception of three larger multilingual models. However, because Large.en and LargeV2.en are not available, it is difficult to draw conclusions on whether it is the model size or multilinguality that hinders the model from benefiting from visual prompting. The fact that visual prompting improves the performance of Medium.en while degrading the performance of Medium suggests that the cause could be multilinguality. If this is the case, multilingual models may benefit from being fine-tuned on monolingual data.

4. Code-switched speech recognition

Code-switched speech refers to the scenario where more than one language is used in the same utterance. With the raise of globalization and democratization of speech recognition technologies, Code-switched ASR (CS-ASR) has become a popular research area [24]. While we cannot know for certain whether Whisper was trained on code-switched data, it is clear that the model's language and task tokens can not explicitly direct the model to do CS-ASR - each language token only represents one

[3]We use 3 number of objects choices to reduce the tuning noise introduced by the mismatch between How2 and VisSpeech.

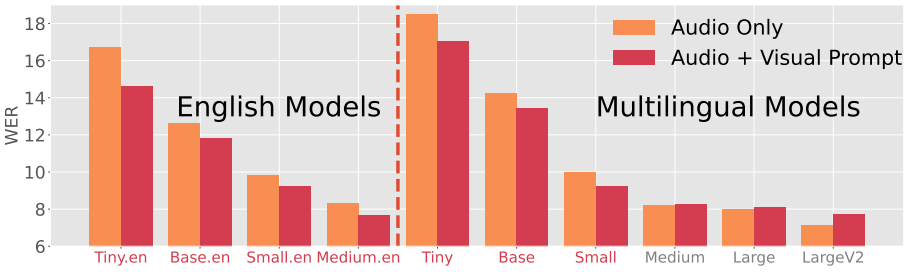

Figure 2: *The effectiveness of visual prompt on VisSpeech across different models.*

Table 2: *Comparison of model performance on VisSpeech. With visual prompt, Medium.en outperforms Large.*

Model	Modality	WER
SotA [19]	A+V	11.28
Whisper Medium.en	A	8.35
Whisper Medium.en	A+V	7.60
Whisper Large	A	8.02
Whisper LargeV2	A	**7.16**

of the 99 training languages, and the task tokens do not convey any information on whether the model should output text in more than one language.

Approach. To test Whisper's CS-ASR capabilities, we use two Mandarin-English code-switched corpora. See table 1 for a quick summary of default and our proposed approach. The default approach (denoted as `default`) is to let Whisper to first run LID to detect the language between the two[4], and then use the detected language in the prompt. While this default approach work to some degree, it relies heavily on Whisper's LID capabilities, but our results show this can sometimes be inaccurate, especially on accented speech. In addition, this approach doesn't explicitly instruct the model to output text in more than one language for intra-sentential code-switched utterances. We propose a simple approach called `concat` that handles the aforementioned issues. The `concat` approach replaces the single language token in the prompt with two language tokens i.e. <|zh|> and <|en|>, as shown in table 1. As will be shown later, despite the simplicity of this approach and the fact that Whisper has never be trained to take two language tokens in the prompt, our approach significantly improves performance.

Datasets and implementational details. We use ASCEND [25] and SEAME [26], which are both Mandarin-English code-switched datasets. Both datasets are spontaneous conversational speech, but ASCEND was recorded from bilingual speakers with different Chinese dialects, while SEAME is recorded from Singaporean and Malaysian speakers. We'll show that despite the fact that both datasets contains the same languages, Whisper performs very differently on them. Tuning is done on the validation sets of ASCEND and SEAME. For our proposed `concat` approach, the hyperparameters we tune are: 1. the order of two language tokens in prompt, and 2. the threshold on Whisper's LID confidence score, above which we use the single detected language's token instead of concatenating two language tokens. Whisper Large with `concat` (<|zh|> first) outperforms all other models and prompts combinations on both datasets, and a threshold of 0.9 works the best for ASCEND, and 1.0 i.e. always concatenating two language tokens, works the best for SEAME.

Results. Table 3 shows the performance of Whisper Large on the validation set of ASCEND and SEAME. In addition to `default` and our proposed `concat` prompts, we also show results when we fixed the language token to be <|zh|> or <|en|> for analysis purposes. *We see that the `concat` method performs the best on both datasets*, and in particular it provides 19% relative improvement on Total MER (mixed error rate) on SEAME compared to `default`. Secondly, with prompt <|zh|>, Whisper performs much better on pure Mandarin utterances on ASCEND (16.3) than SEAME (26.3). Similar results are observed for pure English utterances. This indicates that *Whisper's monolingual ASR performance is much worse on SEAME than on ASCEND.* Next we note that on SEAME, when we use `default` instead of <|en|>, En WER increased from 33.8 to 85.5, while on ASCEND, WER was 31.8 in both cases. This indicates that i.e. *Whisper's LID performance for detecting English is much worse on SEAME than on ASCEND.*

To understand how do different language prompts steer Whisper's output. We manually examined the error modes, and found a common scenario where *the model outputs monolingual translation for code-switched utterances*. This is especially interesting when Whisper does English to Mandarin translation, as the model was only trained to perform X→En translation. This phenomenon inspired us to quantitatively study En→X translation capabilities of Whisper in section 5.

The test set results for CS-ASR are shown in table 4. We see that *with `concat`, Whisper achieves a new SotA for ASCEND*, while on SEAME there is still an considerable gap between zero-shot Whisper and SotA.

Table 3: *Performances of Whisper Large on ASCEND and SEAME validation sets. Zh CER shows results on Mandarin utterances, En WER represents results on English utterances. CS MER shows mixed error rate on code-switched utterance. Total MER is the summarizing metric on the entire dataset.*

Dataset	Lang. prompt.	Zh CER	En WER	CS MER	Total MER
ASCEND	<|zh|>	**16.3**	93.1	33.1	32.6
	<|en|>	90.4	**31.5**	80.1	78.9
	default	17.0	31.8	26.6	22.1
	concat	16.6	31.8	**25.0**	**21.3**
SEAME	<|zh|>	26.3	97.4	43.3	46.7
	<|en|>	99.3	**33.8**	86.9	82.2
	default	27.1	85.5	43.2	45.3
	concat	**25.9**	44.7	**38.4**	**36.9**

Table 4: *Comparison between **zero-shot** Whisper Large and **supervised** SotA models on ASCEND and SEAME test sets*

Dataset	Approach	Zh CER	En WER	CS MER	Total MER
ASCEND	Sup. SotA [27]	-	-	-	25.0
	Whisper+default	19.6	30.3	23.6	22.8
	Whisper+concat	16.8	30.8	22.0	**20.9**
SEAMEDEVMAN	Sup. SotA [28]	-	-	-	**16.6**
	Whisper+default	24.7	76.3	38.2	38.2
	Whisper+concat	23.6	45.8	33.4	32.7
SEAMEDEVSGE	Sup. SotA [28]	-	-	-	**23.3**
	Whisper+default	32.4	82.8	56.4	65.0
	Whisper+concat	31.0	46.7	49.6	47.6

Remarks. Recall that ASCEND is Chinese accented, and SEAME is Singaporean and Malaysian accented, and based on our discussion on table 3, we hypothesize that the performance gap on ASCEND and SEAME is because *Whisper's LID and ASR performance vary drastically on different accents*, even though the underlying languages are the same. We leave a more comprehensive investigation of this hypothesis for future work.

[4]We set the probabilities of the languages other than the two to be 0.

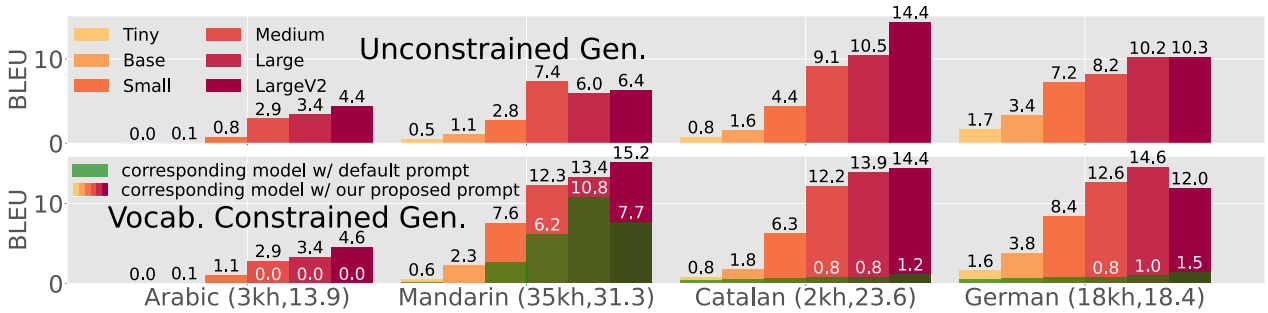

Figure 3: *Whisper Zero-shot En→X ST results. In the labels of x-axis, Arabic (3kh,13.9) is interpreted as in the training of Whisper, Arabic ASR data + Arabic→English ST data amounts to 3k hours, and the supervised ST baseline for English→Arabic is 13.9. The same interpretation applies to other languages. For the default prompt, we only show numbers for larger models for better visualization.*

5. En to X speech translation

In this section, we investigate Whisper's ability to perform En→X speech translation (ST). Note that Whisper is trained on both multilingual ASR and X→En speech translation (ST), but never on En→X ST. Studying Whisper's zero-shot performance on En→X ST could be a way to measure the speech understanding capabilities that emerge from large scale, multilingual, multitask training. we emphasize that the goal of this section is not to achieve SotA performance, but to study the model's emergent, zero-shot translation ability across different language families, amounts of training data, and model sizes.

Approach. The default prompt for ST is to use <|st|> as the task token. However, we found that <|st|> would lead Whisper to only output English no matter what language token is used, unless we constrained the output vocabulary. To instruct Whisper to do En→X ST, we propose to use task token <|asr|> instead, and use language token correspond to the language X. See table 1 for a example of the default and our proposed approach. Counter-intuitive as it might be (using <|asr|> for ST), as we'll show later, our prompt outperforms the default prompt significantly, and even comes close to supervised approaches for some languages.

Datasets and implementation details. We pick Arabic, Mandarin, Catalan and German in CoVoST2 [29], to achieve a resource- and topology-wise diverse evaluation. To be able to compare with supervised, unsupervised, and other zero-shot ST approaches [30], we also evaluate Whisper on En→Ru and En→De from MuST-C V1 [31], and En→Fr from Libri-Trans [32]. As for vocabulary constrain, for Arabic, Mandarin, and Russian, we use the unicode range to constrain the vocab to only contain tokens that belong to their scripts; for German, Catalan, and French we constrain the vocab to only contain tokens that are the top $K\%$ most frequent in their training set text. K is tuned for CoVoST2 on the development set. For MuST-C and Libri-Trans, we set K to be 40% for German and 50% for French based on CoVoST2 tuning results.

Results. In Figure 3, we show different Whisper models' performance on the four CoVoST2 languages. In general, for our proposed prompt, bigger models perform better across languages, and vocabulary constrained generation outperforms unconstrained generation. As for the default prompt (green bars), we didn't show its performance for unconstrained generation as it only output English text, and for constrained generation, it also performs vert poorly except for Mandarin. We compare Whisper's performance with other models in table 5[5]. Whisper performs reasonably on all three directions, and especially well on En→Ru. We note that the comparison in this table should only be treated as a reference. This is because even for the unsupervised and zero-shot approaches, they are particularly designed for ST, and they either leverage machine translation systems [30, 33] or multilingual sentence embedding models [34]. For Whisper, however, we simply adjust its prompt, and the goal is to probe the multilingual understanding of the model.

Remarks. Although Whisper is trained with massive multilingual data, performing En→X might be harder than one expects. Because for the <|st|> task token, the model is never trained to generate non-English text; for the <|asr|> task token the model is never trained to generate text belonging to a different language than the input speech. The fact that Whisper is able to do En→X ST with a simple modification on its prompt reveals that semantically related words and phrases from different languages might be close in the model's latent space. We also expect that we could fine-tune Whisper to boost the performance of ST on new language pairs.

Table 5: *Comparing zero-shot Whisper with supervised and unsupervised approaches for MuST-C (En→De and En→Ru) and Libri-Trans (En→Fr). Zero-shot Whisper performs reasonably on all three directions. *T-Modules [34] relies on strong multilingual sentence embedding models that are trained on bitext.*

Category	Approach	En→De	En→Ru	En→Fr
Supervised	w2v2+mBART [30]	32.4	20.0	23.1
	E2E Transformer [35]	27.2	15.3	11.4
Unsupervised	Chung et al. [36]	-	-	12.2
	Cascaded [30]	22.0	10.0	15.4
	E2E (w2v2+mBART) [30]	23.8	9.8	15.3
Zero-shot	Escolano et al. [33]	6.8	-	10.9
	T-Modules* [34]	23.8	-	32.7
	Whisper w/ default prompt	0.4	8.8	0.8
	Whisper w/ our prompt	18.1	12.8	13.1

6. Conclusion

We investigate the emergent abilities of Whisper through the lens of prompt-based zero-shot task generalization. Our proposed prompts significantly outperform the default prompts in all three tasks that we studied. In addition, we found interesting properties of Whisper - in AVSR, we found that the model is very robust to the length and noisiness of the visual prompt, and the effectiveness of the visual prompt between English models and multilingual models are quite different; in CS-ASR, we identified potential performance gaps between different accents; in ST, we found the surprising results that the <|asr|> task token can be used to instruct the model to do translation and outperforms <|st|>. Many of the above properties are worth further investigating, and can potentially lead to models that are more robust, more generalizable, and have less unwanted bias.

[5]Whisper Large for En→De; LargeV2 for En→Ru and En→Fr.

7. References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *ArXiv preprint*, 2021.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.

[3] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *ICLR*, 2022.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, no. 9, 2022.

[7] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. hsin Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *TMLR*, 2022.

[8] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, no. 9, 2023.

[9] H. Gao, J. Ni, K. Qian, Y. Zhang, S. Chang, and M. Hasegawa-Johnson, "Wavprompt: Towards few-shot spoken language understanding with frozen language models," in *Interspeech*, 2022.

[10] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[12] K.-W. Chang, W.-C. Tseng, S.-W. Li, and H.-y. Lee, "An exploration of prompt tuning on generative spoken language model for speech processing tasks," in *Interspeech*, 2022.

[13] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, 2021.

[14] J. ho Kim, J.-S. Heo, H. seo Shin, C. Lim, and H. jin Yu, "Integrated parameter-efficient tuning for general-purpose audio models," *ArXiv*, 2022.

[15] J. Xue, P. Wang, J. Li, and E. Sun, "A weakly-supervised streaming multilingual speech model with truly zero-shot capability," *ArXiv*, 2022.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv preprint*, 2022.

[17] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[18] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," 2018.

[19] V. Gabeur, P. H. Seo, A. Nagrani, C. Sun, K. Alahari, and C. Schmid, "Avatar: Unconstrained audiovisual speech recognition," in *Interspeech*, 2022.

[20] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[21] A. Zeng, M. Attarian, brian ichter, K. M. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, "Socratic models: Composing zero-shot multimodal reasoning with language," in *ICLR*, 2023.

[22] A. Miech, D. Zhukov, J. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019.

[23] B. Wu, W. Chen, Y. Fan, Y. Zhang, J. Hou, J. Liu, and T. Zhang, "Tencent ml-images: A large-scale multi-label image database for visual representation learning," *IEEE Access*, 2019.

[24] G. I. Winata, A. F. Aji, Z.-X. Yong, and T. Solorio, "The decades progress on code-switching research in nlp: A systematic survey on trends and challenges," *ArXiv preprint*, 2022.

[25] H. Lovenia, S. Cahyawijaya, G. Winata, P. Xu, Y. Xu, Z. Liu, R. Frieske, T. Yu, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. Shi, and P. Fung, "ASCEND: A spontaneous Chinese-English dataset for code-switching in multi-turn conversation," in *LREC*, 2022.

[26] D.-C. Lyu, T. P. Tan, C. E. Siong, and H. Li, "Seame: a mandarin-english code-switching speech corpus in south-east asia," in *Interspeech*, 2010.

[27] T. Nguyen, N. Tran, L. Deng, T. G. da Silva, M. Radzihovsky, R. Hsiao, H. Mason, S. Braun, E. McDermott, D. Can, P. Swietojanski, L. Verwimp, S. Oyman, T. Arvizo, H. Silovsky, A. Ghoshal, M. J. Martel, B. R. Ambati, and M. Ali, "Optimizing bilingual neural transducer with synthetic code-switching text generation," *ArXiv*, 2022.

[28] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Interspeech*, 2018.

[29] C. Wang, A. Wu, and J. Pino, "Covost 2: A massively multilingual speech-to-text translation corpus," 2020.

[30] C. Wang, H. Inaguma, P.-J. Chen, I. Kulikov, Y. Tang, W.-N. Hsu, M. Auli, and J. M. Pino, "Simple and effective unsupervised speech translation," *ArXiv*, 2022.

[31] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *NAACL-HLT*, 2019.

[32] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation," in *LREC*, 2018.

[33] C. Escolano, M. R. Costa-jussà, J. A. R. Fonollosa, and C. Segura, "Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders," *ASRU*, 2020.

[34] P.-A. Duquenne, H. Gong, B. Sagot, and H. Schwenk, "T-modules: Translation modules for zero-shot cross-modal machine translation," in *EMNLP*, 2022.

[35] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *AACL: System Demonstrations*, 2020.

[36] Y. Chung, W. Weng, S. Tong, and J. R. Glass, "Towards unsupervised speech-to-text translation," in *ICASSP*, 2019.

A. Appendix

A.1. Examples of CS-ASR transcription

Table 6: *Examples of how* `concat` *improve transcription over* `default` *on SEAME with Whisper Large. We use ... when transcriptions are the same for all three cases.*

Ground Truth	transcription w/ `default`	transcription w/ `concat`
也不需要做research	也不需要做研究	也不需要做research
这真的是一个very tough question ...	这真的是一个很困难的问题...	这真的是一个very tough question...
每次... 还是choir practice	每次... 还是quiet practice	每次... 还是choir practice
then did you realise the performances...	那你有没有意识到表演	then do you realise the performances...
...你真的是要睡觉了是吗	...你真的是要sweet 小的是吗	...你真的是要睡觉了是吗