SlideAVSR: 視聴覚音声認識のための論文解説動画データセット

王昊 [1] 栗田修平 [2] 清水周一郎 [3] 河原大輔 [1]
[1] 早稲田大学 [2] 理化学研究所 AIP [3] 京都大学
conan1024hao@akane.waseda.jp shuhei.kurita@riken.jp
sshimizu@nlp.ist.i.kyoto-u.ac.jp dkw@waseda.jp

概要

視聴覚音声認識 (AVSR) は音声認識 (ASR) をマルチモーダルに拡張したもので，音声の代わりに動画を入力として与えるタスクである．多くの研究で使用されている読唇データセットは顔追跡動画のみから構成されるため，AVSR モデルの画像理解能力を評価するには不十分だと考える．本研究では，論文解説動画を用いて，視聴覚音声認識データセット **SlideAVSR** を構築する．多くの専門用語が含まれ，スライドを参照しないと正確な文字起こしは困難であるため，モデルの画像理解能力をより多くの側面で評価できると考える．そして，テキスト情報を参照可能な AVSR モデル **DocWhisper** を提案し，SlideAVSR においてその有効性を確認した．

1 はじめに

言語，画像，動画，音声など複数の種類のデータを同時に扱えるマルチモーダルモデルの研究が注目されている．音声認識 (ASR) をマルチモーダルに拡張し，音声の代わりに動画を入力として与える視聴覚音声認識 (AVSR) がその一例である．これまでの AVSR の研究の多くは，読唇データセット [1, 2] における精度向上を目的として行われてきた．これらの研究で構築されたモデルは読唇データに対し高い性能を誇るが，他の種類の動画には対応できない．

本研究では，AVSR モデルの画像理解能力をより全面的に評価するために，多くの専門用語が含まれ，スライド上のテキスト情報を参照しないと正確な文字起こしが困難な視聴覚音声認識データセット SlideAVSR を構築する．具体的には，YouTube から論文の解説動画を収集し，ChatGPT フィルター，BLIP-2 [3] フィルターを含めた複数のデータ洗練プロセス，話者のアクセントを考慮したデータ分割により高品質なデータセットの構築を実現する．

さらに本研究では，OCR でスライドの内容を効率的に参照可能にする AVSR モデル DocWhisper を提案する．SlideAVSR を用いた実験において，DocWhisper は音声入力のみを用いる Whisper [4] と比べ最大 14.3%の精度向上を示した．そして，OCR の結果における long-tail 問題に対し，単語の出現頻度に基づく順位を算出する手法 FQ Ranker (Frequency Ranker) を提案し，DocWhisper に対する効果を評価した．

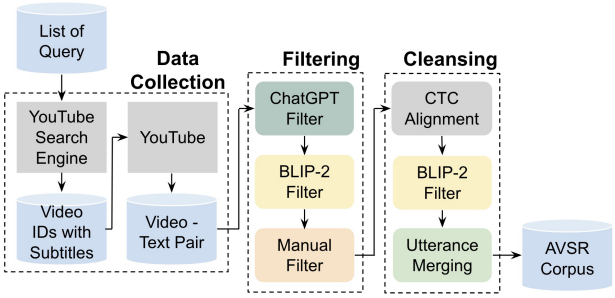


図 1 SlideAVSR の構築手順.

2 関連研究

AVSR のベンチマークとして，多くの読唇データセットが開発されている [1, 2, 5, 6, 7, 8, 9]．その中で多くの研究は LRS2-BBC [1] と LRS3-TED [2] をモデルの評価で使用している [10, 11, 12]．これらの研究で構築されたモデルは読唇データに対し高い性能を誇るが，顔が写っていない動画には対応できない．

豊富な読唇データに比べ，他形式の AVSR データセットは極めて少なく，我々の知る限りでは，作業手順の動画データセット HowTo100M [13] を用いて構築した，映像と音声の整合度が高い動画から含まれる VisSpeech [14] や，一人称視点の動画データセット Ego4D [1)] しかない．より全面的に AVSR モデルの画像理解能力を評価するために，多様なベンチマークデータセットが必要不可欠である．

Whisper を AVSR モデルに拡張する研究として，Peng ら [15] は，CLIP [16] を用いて入力の映像スト

1) https://ego4d-data.org/docs/challenge/

リームを一連の単語トークンに変換し，プロンプトとして Whisper に与えることで，VisSpeech における zero-shot での精度向上を確認している．本研究では，OCR モデルを用いてプロンプトを作成する．また，zero-shot ではなく，fine-tuning を行う．

3 データセット構築

本研究では，多くの専門用語が含まれ，スライドを参照しないと正確な文字起こしが困難な視聴覚音声認識データセット SlideAVSR を構築する．YouTube 動画からの音声コーパス構築法 JTubeSpeech [17] を参考にしつつ，独自のフィルターを加え，動画を対象とし高精度なフィルタリング及びデータ洗練を実施する．本節では，構築の手順について説明する．図 1 に構築手順を示す．

3.1 データ収集

検索フレーズの作成 人工知能分野のトップ国際会議を対象とし，動画検索に使用する検索フレーズを作成する．検索フレーズは，{会議名} {年} {形式} のフォーマットでルールベースで生成する．対象会議のリストを付録 A に示す．コロナ以降オンライン化した会議が多いことを考慮し，対象年を 2020 から 2023 までの四年とする．形式は paper，workshop，talk の三種とする．“ACL 2023 paper” が検索フレーズの例として挙げられる．

字幕付き動画の取得 作成した検索フレーズで動画 ID を取得し，ダウンロード[2)]を実施する．データの質を確保するために，手動字幕がある動画のみを採用する．また，ダウンロードする際には以下の形式を満足する動画に限定する：

- 長さ 5 分以上 20 分以下 (短すぎるあるいは長すぎるものは論文解説である可能性が低いため).
- 動画の形式は MP4，720P，H264.
- 音声の形式は single-channel，16bit，16kHz.

3.2 フィルタリング

フィルタリングによって論文解説ではない動画とスライドが写っていない動画を除去する．

ChatGPT フィルター 動画の概要欄を ChatGPT に与え，以下の二点について確認してもらう．

- この動画は論文を解説する動画である.
- この概要欄は英語で書かれている.

三回生成を行い，一回以上 “Yes” が出力された場合は動画を採用し，それ以外の場合は除去する．モデルとプロンプトの詳細は付録 B に示す．

動画用 BLIP-2 フィルター 動画毎に時間軸の最初，最後，四分位点の計五ヶ所でスクリーンショットを撮り，マルチモーダルモデル BLIP-2 [3] に与え，以下の二点について確認してもらう．

- この画像はスクリーンショットで実写画像ではない.
- この画像は論文解説用スライドの一部である.

スクリーンショット毎に生成を行い，計一回以上 “Yes” が出力された場合は動画を採用し，それ以外の場合は除去する．モデルとプロンプトの詳細は付録 B に示す．

人手フィルター 筆者らによる人手チェックを実施し，自動フィルターでは除去しきれなかった以下の動画を除去する．

- まれにスライドが写っているが，多くの時は作者の顔が写っている動画.
- 論文の PDF が写っている動画.
- 会議のオープニングなどの動画.

3.3 データ洗練

音声と字幕のアライメント，スライドが写ってない発話の除去，発話のマージを行う．

CTC アライメント 多くの字幕のタイミングは不正確であるため，CTC segmentation [18] を用いて音声と字幕のアライメント及びスコアリングを行う．CTC スコアに対し閾値 (-7) を設け，閾値以下の発話を除去する．モデルの詳細は付録 B に示す．

発話用 BLIP-2 フィルター 発話ごとに時間軸の中位点でスクリーンショットを撮り，BLIP-2 でフィルタリングを行う．スクリーンショット毎に三回生成を行い，一回以上 “Yes” が出力された場合は発話を採用する．プロンプトは前述の動画用 BLIP-2 フィルターと同じものを使用する．

発話のマージ 動画作者による字幕は，発話を不自然に極めて短いスパンに分割することがある．CTC segmentation で得られた音声区間を用いて，前の発話の終了時刻と次の発話の開始時刻が一致し，且つ合計の長さが 15 秒を超えない場合，二つの発話を一つにマージする．これにより，Whsiper による音声認識の精度が約 20%向上した．

3.4 データ分割

ASR モデルの性能は話者のアクセント[3)]により大きく変動することが先行研究 [19, 20] で示されてい

2) https://github.com/yt-dlp/yt-dlp

3) 本稿で使用する「アクセント」は，アクセント，イントネーション，音調など総合的なプロソディ情報を指す.

表 1　SlideAVSR の統計量情報.

	#videos	#speakers	#utterances	#hours
Train	195	172	15,803	29.26
Dev	20	20	1,515	3.08
TestA	15	15	1,034	2.21
TestB	15	13	1,111	1.90
Total	245	220	19,463	36.45

る．画像情報は難しいアクセントの認識に貢献するという仮説に基づき，英語のネイティブスピーカーに話者のアクセント分類を依頼し，データセットの分割を行った．インド英語に分類された動画の一部を TestB に，他の動画をランダムシャッフルして Train，Dev，TestA に分割した．同じ話者が複数の分割に属さないように調整を行った．また，機械音声の動画を人手で除去した．

以上の構築作業により，245 個の動画から約 36 時間の AVSR データセットを作成した．表 1 にデータセットの統計量を示す．

4　実験

4.1　提案手法

DocWhisper は，入力の映像ストリームを OCR モジュールに読み込み，認識されたテキストを単語列の形で Whisper にプロンプトとして与え，fine-tuning を行う．Peng ら [15] は zero-shot で CLIP から得られたプロンプトを使用しているが，本研究の予備実験では SlideAVSR における zero-shot での精度向上を確認できなかった．Whisper の事前学習 [4] にプロンプトが使われていないことから，Whisper はプロンプトにロバストではないと推測する．

図 2 に OCR 結果の単語数に対する頻度分布を示す．分布は long-tail であり，プロンプトに 100 単語[4] を入れるとしても全体の七割しかカバーできないことがわかる．この問題に対し，本研究では単語の出現頻度に基づく順位を算出する手法 FQ Ranker を提案する．単語の頻度と親密度 [21] は高い相関を持つことが先行研究 [22, 23] で示されていることから，頻度が低く，難しい単語の順位を上げることで，プロンプトが持つ情報量を増やせると考える．

4.2　実装の詳細

Whisper large-v3[5] を用いて実験を行った．Whisper 及び DocWhisper のチューニングには，AdamW [24] を用いて最適化し，学習率は 2e-5，バッチサイズは 16，warmup step は 1,000 とした．10 epoch の学習を行い，Dev セットで最も精度が高いチェックポイントを評価に用いた．また，学習は異なる 3 つの seed 値を設定して行い，平均値を算出する．DocWhisper は，発話ごとに時間軸の中位点でスクリーンショットを撮り，OCR モジュールに読み込み，認識されたテキストを Whisper にプロンプトとして与える．OCR には Google Cloud Vision API[6] を用いた．FQ Ranker は 2023 年 4 月時点の英語 Wikipedia を用いて単語の出現頻度を計数し，頻度の昇順に OCR の結果をソートした．プロンプトは “word 1, word 2, ..., word n” のような単語列の形でモデルに与えている．プロンプトの最大単語数 ($K \in \{25, 50, 75, 100\}$)，そして FQ Ranker の有無による実験を行った．Whisper 公式実装[7] に従いテキスト正規化を行い，WER (Word Error Rate, 単語誤り率) を用いて評価を行った．全ての実験は一枚の NVIDIA A100 40G で行った．

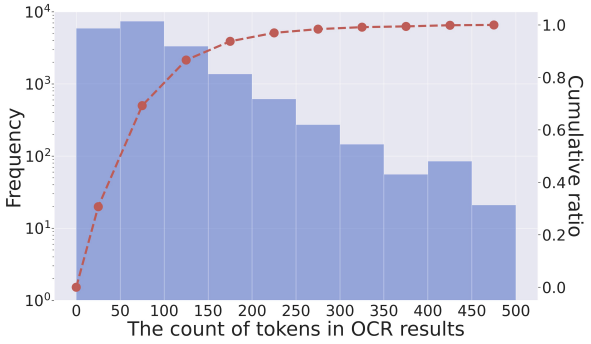


図 2　OCR 結果の単語数に対する頻度分布．500 単語以上のサンプルも存在するが，この図では省略している．

4.3　結果

Whisper 及び DocWhisper の SlideAVSR における定量評価の結果を表 3 に示す．いずれのモデルにおいても，インド英語の動画からなる TestB セットのスコアは TestA セットより劣り，Whisper はインド英語が不得意であることがわかる．Fine-tuning を行うことで，音声入力のみを用いる Whisper は TestA において 1.7%の精度向上を示した．一方，TestB においては精度向上が見られなかった．Train にもインド英語の動画が存在するが，数が少なく，難しいアクセントの克服には不十分だったと考える．

Fine-tuning した Whisper と比べ，DocWhisper は TestA において最大 14.3%，TestB において最大 11%

4)　Whisper がプロンプトに割り当てる最大長は 224 で，100 単語以上の入力は一般的に難しい．
5)　https://huggingface.co/openai/whisper-large-v3
6)　https://cloud.google.com/vision
7)　https://github.com/openai/whisper

表 2 Whisper (W) では置換誤りであったが，DocWhisper (D) では正解した単語の種類及び具体例．

Type	Example
専門用語	W Hyp: we select quantum ~~adhering~~ 2 and nxt as representative of pos protocols
(41%)	D Hyp: we select quantum ethereum 2 and nxt as representative of pos protocols
語形変化	W Hyp: manual ~~transcript~~ we call this setting supervised things we have paired data
(28%)	D Hyp: manual transcripts we call this setting supervised things we have paired data
聞き違い	W Hyp: we can also perform other tasks like ~~normal~~ view synthesis
(24%)	D Hyp: we can also perform other tasks like novel view synthesis
人名	W Hyp: this is a work done at ibm research with ~~gilmoseci chileo~~ and irina ~~rich~~
(7%)	D Hyp: this is a work done at ibm research with guillermo cecchi and irina rish

表 3 SlideAVSR における単語誤り率による定量評価．

Model	Modality	Fine-tune	#Prompt	TestA	TestB
Whisper	A	✗	0	8.23	11.18
		✔		8.07	11.25
DocWhisper + FQ Ranker	A + V	✔	25	7.35	10.82
				7.42	10.59
DocWhisper + FQ Ranker	A + V	✔	50	7.08	10.43
				7.26	10.35
DocWhisper + FQ Ranker	A + V	✔	75	7.02	10.04
				7.26	10.29
DocWhisper + FQ Ranker	A + V	✔	100	**6.91**	**10.01**
				7.04	10.22

の精度向上を示した．スライド上のテキスト情報を参照することで，SlideAVSR における音声認識の性能を大幅に改善することができることがわかり，画像情報は難しいアクセントの認識に貢献する仮説も支持された．さらに，プロンプトの最大単語数が大きくなればなるほど，性能は良くなることがわかった．OCR 経由でスライドを参照しているため，情報量の欠如が少ない方が性能に貢献できると考える．

FQ Ranker は，プロンプト最大単語数が 25 の時に TestB において一定の精度向上を示すが，最大単語数が 50 を上回るとその優劣が逆転する．4.4 節で詳細を示すが，DocWhisper により訂正できた文字起こしは全て専門用語ではないため，親密度が高い単語でも聞き違いを起こす可能性があることが示唆されている．また，頻度に基づく単語のソートは，順序を持つ文脈情報を乱すため，言語モデルである Whisper の Decoder がスライド上のテキスト情報を参照する難易度を向上させたと推測する．

4.4 具体例の分析

Whisper の誤り (削除，置換，挿入) の中では，DocWhisper は置換誤りを一番多く訂正できた．その中身を分析するために，Whisper では置換誤りであったが，DocWhisper では正解した発話を 100 サンプル集計し，専門用語，語形変化，聞き違い，人名の四つのカテゴリに分類した．名詞の複数形，動詞の活用形，動詞の三人称単数など，同じ語彙素からの変形であれば全て語形変化に分類する．同じ語彙素からの変形ではない場合，専門用語，聞き違い，人名にそれぞれ分類する．表 2 に各カテゴリの割合及び具体例を示し，具体例に発話に対応するスクリーンショットを付録 C に示す．予想していた専門用語が一番高い割合を示すが，語形変化と聞き違いの割合も低くないことがわかった．これらの単語の多くは親密度が高く，出現頻度に基づくソートで順位が下げられる可能性があり，FQ Ranker の性能低下に繋がる要因となっていると考える．より効率的に OCR の long-tail 問題を解決でき，情報量の欠如を減らす手法を探るのが今後の課題である．

5 おわりに

本研究では，論文解説動画からなる視聴覚音声認識データセット SlideAVSR を構築した．OCR でスライドの内容を参照可能にする AVSR モデル DocWhisper を提案し，SlideAVSR における有効性を確認すると共に，具体例の分析を行った．また，単語の出現頻度に基づく順位を算出する手法 FQ Ranker を提案し，DocWhisper に対する効果を評価し，性能が出なかった理由について議論を行った．

今後は，OCR をベースとした手法を継続的に改善しつつ，OCR に依存しない end-to-end な AVSR モデルの構築を目指したい．そして，論文解説動画以外にも，スポーツや将棋，ゲームの実況解説動画，料理動画，Vlog など様々な種類の動画を用いて，より多くの側面で AVSR モデルの画像理解能力を評価できるベンチマークの構築を目指したい．最終的には，このようなベンチマークに対応でき，多様な動画入力に高い音声認識性能を持つ AVSR の基盤モデルの構築に取り組みたい．

謝辞

有意義な議論をくださった東佑樹氏に感謝する．SlideAVSR に対するアノテーションに関して理化学研究所 AIP の関根聡氏，鈴木久美氏のサポートを受けた．

参考文献

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 44, No. 12, pp. 8717–8727, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition, 2018.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In **ICML**, 2023.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[5] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, **Computer Vision – ACCV 2016**, pp. 87–103, Cham, 2017. Springer International Publishing.

[6] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 3444–3453, 2017.

[7] Joon Son Chung and Andrew Zisserman. Lip reading in profile. 09 2017.

[8] Brendan Shillingford, Yannis Assael, Matthew W. Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas. Large-scale visual speech recognition. In **Proc. Interspeech 2019**, pp. 4135–4139, 2019.

[9] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In **2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)**, pp. 1–8, 2019.

[10] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. **arXiv preprint arXiv:2201.02184**.

[11] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4491–4503, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[12] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. In **The Eleventh International Conference on Learning Representations**, 2023.

[13] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In **ICCV**, 2019.

[14] Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Avatar: Unconstrained audiovisual speech recognition, 2022.

[15] Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. In **Interspeech**, 2023.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification, 2021.

[18] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. Ctc-segmentation of large corpora for german end-to-end speech recognition. In **International Conference on Speech and Computer**, pp. 267–278. Springer, 2020.

[19] Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 6462–6468, Marseille, France, May 2020. European Language Resources Association.

[20] Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, and Mitesh M. Khapra. Svarah: Evaluating english asr systems on indian accents, 2023.

[21] 藤田早苗, 小林哲生. 令和版単語親密度に基づく大規模語彙数推定調査. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 4N1GS303–4N1GS303, 2022.

[22] Max Coltheart. The mrc psycholinguistic database. **The Quarterly Journal of Experimental Psychology Section A**, Vol. 33, No. 4, pp. 497–505, 1981.

[23] Kumiko Tanaka-Ishii. **Statistical Universals of Language**. Springer Cham, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

A　検索対象の国際会議

表 4 に検索対象として採用した国際会議を示す．

表 4　検索対象の国際会議．

Topic	Conference
NLP	ACL, NAACL, EMNLP
CV	CVPR, ICCV, ECCV
Speech	INTERSPEECH, ICASSP
AI	AAAI, IJCAI
ML	ICLR, ICML, NeurIPS
Data Mining	KDD, WSDM, WWW
Database	SIGMOD, VLDB, ICDE
IR	SIGIR
HCI	CHI

B　フィルタリングとデータ洗練で使用したモデルとプロンプト

3.2 節と 3.3 節で説明した ChatGPT フィルター，BLIP-2 フィルター，CTC アライメントで使用したモデルやプロンプトの詳細について述べる．

ChatGPT フィルター　gpt-3.5-turbo[8] を使用した．使用したプロンプトを表 5 に示す．

表 5　ChatGPT フィルターのプロンプト．

Here is a description of a YouTube video:
{DESCRIPTION}
Using the description, check whether the video meets the following criteria.
- This video is a presentation video of a research paper.
- The description is written in English.
Attention, you can only answer 'Yes' or 'No' and you can only answer one time.

BLIP-2 フィルター　blip2-flan-t5-xl[9] モデルを使用した．使用したプロンプトを表 6 に示す．

表 6　BLIP-2 フィルターのプロンプト．

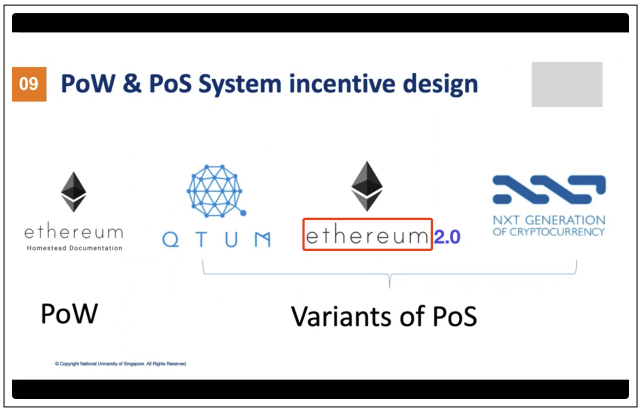
Question: This image is a screenshot of a video,
check whether the image meets the following criteria.
- It is a screen-sharing, not a photo shoot.
- It is a part of a slide for a research presentation.
Attention, you can only answer 'Yes' or 'No' and you can only answer one time.
Answer:

CTC アライメント　ESPnet の kamo-naoyuki_wsj[10] モデルを使用した．

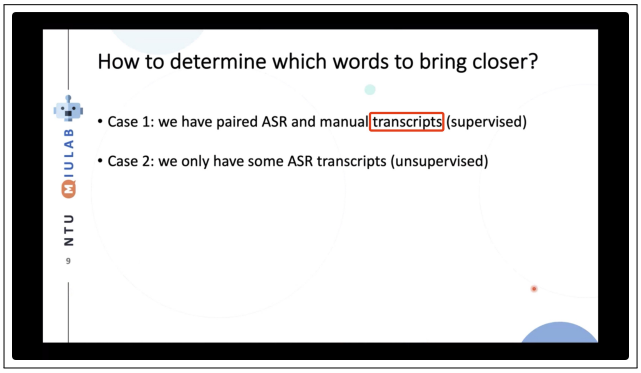
C　具体例

以下に表 2 の発話に対応するスクリーンショットを示す．誤り訂正で参照された部分を赤い枠で囲む．

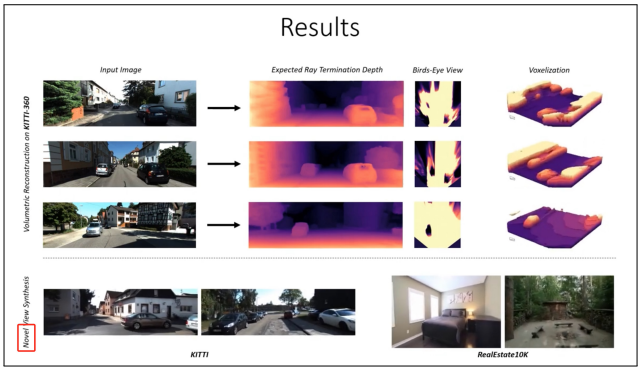
8)　https://openai.com/product
9)　https://huggingface.co/Salesforce/blip2-flan-t5-xl
10)　https://huggingface.co/espnet/kamo-naoyuki_wsj



https://www.youtube.com/watch?v=eepUV9NJxFs



https://www.youtube.com/watch?v=dvUutyo72R4



https://www.youtube.com/watch?v=0VGKPmomrR8

https://www.youtube.com/watch?v=CQBdQz1bmls