

V-Align: Visual Forced Alignment via Phoneme to Video Optimal Path Traversal

Souvik Ghosh¹, C. V. Jawahar¹, Vinay Namboodiri²

¹ IIIT Hyderabad, India

² University of Bath, United Kingdom

souvik.g@research.iiit.ac.in, jawahar@iiit.ac.in, vpn22@bath.ac.uk

Abstract

Visual Forced Alignment (VFA) aims to temporally localize linguistic units within silent talking-face videos, enabling alignment when audio is noisy or unavailable. We propose V-Align, a structured framework that formulates VFA as an optimal monotonic path traversal problem over a frame-phoneme compatibility lattice. Given the lip-motion embeddings and transcript-derived phoneme representations, V-Align learns a soft traversal posterior and recovers discrete boundaries via dynamic programming. Training follows a two-stage strategy: Stage 1 encourages optimal path learning without explicit boundary supervision and already achieves competitive performance relative to fully supervised VFA baselines, while Stage 2 refines boundary precision using additional word-level supervision. Experiments on LRS2 and LRS3 demonstrate state-of-the-art MAE and frame accuracy. Code and demo are available at: <https://valign-interspeech.github.io/>.

Index Terms: visual forced alignment, multimodal speech understanding

1. Introduction

With the rapid growth of media consumption through social platforms and streaming services, viewers increasingly rely on subtitles [1] that are tightly synchronized with both spoken dialogue and the visual content. However, manually annotating precise time-alignment information between transcripts and speech is expensive and labor-intensive. To reduce this burden, forced alignment techniques automatically align a given transcription with its corresponding speech signal, producing word or phoneme-level timestamps. Modern audio-based forced aligners, such as WhisperX [2] and Montreal Forced Aligner (MFA) [3] have achieved strong performance under clean recording conditions.

Nevertheless, these systems often fail when audio is noisy, corrupted, or entirely missing [4], as commonly encountered in real-world media, archival footage, or privacy-preserving silent videos. In such scenarios, *Visual Forced Alignment* (VFA) [5, 6] offers an alternative by temporally grounding transcripts directly from lip movements in talking-face videos, enabling alignment in the absence of reliable audio. Research on visual forced alignment remains far less explored than its acoustic counterpart. Early progress in visually grounded temporal localization emerged through visual keyword spotting [7, 8], which aims to identify the temporal occurrence of a single query word within a silent talking-face video. Although such approaches could in principle be extended to full-transcript alignment, repeating keyword localization across all words is computationally expensive and may lead to overlapping or inconsistent boundaries. The first dedicated VFA framework, DVFA [5],

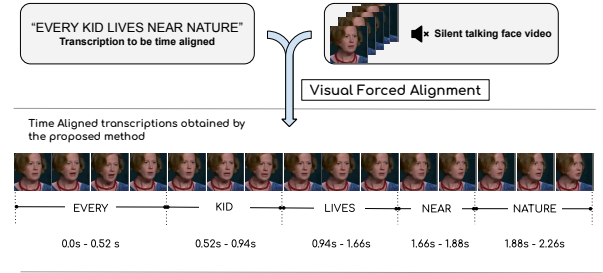


Figure 1: *Overview of Visual Forced Alignment.* Given a silent talking-face video and its transcript, the proposed method predicts temporally aligned word boundaries, producing time-stamped transcriptions grounded in lip motion.

formulates alignment as frame-wise token prediction by jointly integrating video and text representations, together with an auxiliary branch that verifies transcript consistency. More recently, He *et al.* [6] improved upon DVFA by incorporating local context-aware feature extraction and multitask objectives to better capture both global and fine-grained temporal structure. Despite these advances, existing VFA methods remain limited in their ability to recover sharp and temporally consistent phoneme boundaries. Furthermore, these systems rely heavily on explicit ground-truth or audio-derived forced-alignment supervision during training, which restricts their applicability to scenarios where reliable alignment annotations are unavailable.

To address these limitations, we formulate the task of Visual Forced Alignment as an optimal monotonic path traversal problem over a frame-phoneme compatibility lattice. Given a silent talking-face video and its transcript, the objective is to recover a temporally consistent monotonic traversal that segments the visual stream into phoneme-aligned regions. Rather than predicting boundaries directly, we learn a soft traversal posterior over all valid monotonic paths, and obtain discrete phoneme boundaries through structured optimal path decoding. Building on this formulation, we introduce V-Align, a lightweight cross-modal alignment model that constructs dense frame-phoneme compatibility scores between lip-motion embeddings and contextual phoneme representations. A key advantage of V-Align is that strong alignment structure can be learned even without boundary supervision. We adopt a two-stage training strategy: Stage 1 learns monotonic traversal structure without requiring phoneme timestamp supervision using structured objectives that marginalize over all valid paths and sharpen the posterior toward the decoded optimal traversal. It outperforms most prior VFA methods that are trained with explicit audio-derived align-

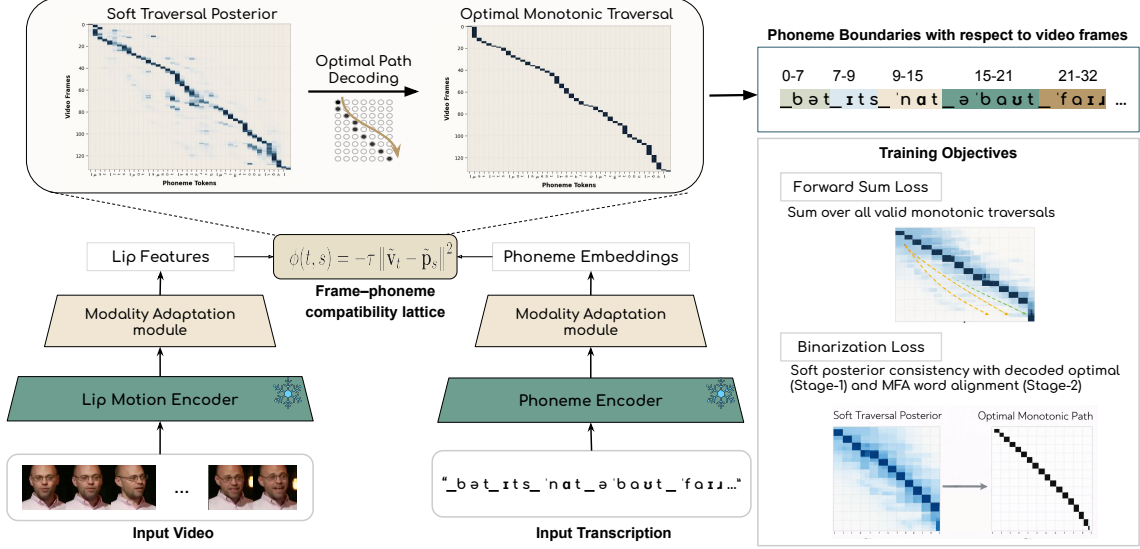


Figure 2: Overview of V-Align for visual forced alignment. Given a silent talking-face video and its transcript, frozen lip-motion and phoneme encoders extract frame- and token-level embeddings, which are mapped into a shared space via lightweight convolutional projections. A frame-phoneme compatibility lattice is constructed using a Gaussian distance kernel, producing a soft traversal posterior over monotonic alignments. Optimal path decoding yields a discrete monotonic traversal and phoneme boundaries in video time. Training is guided by a forward-sum objective that marginalizes over all valid monotonic traversals, and a binarization loss that enforces consistency with the decoded optimal path (Stage 1) and MFA word-level refinement (Stage 2).

ment supervision. Stage 2 further improves boundary precision through word-level refinement using Montreal Forced Aligner supervision, resolving phoneme inventory mismatches, achieving state-of-the-art localization accuracy. Our approach combines boundary-free monotonic path learning, efficient optimal traversal decoding, and word-level MFA refinement to achieve state-of-the-art alignment accuracy.

2. Method

Visual Forced Alignment (VFA) seeks to temporally ground a transcript within a silent talking-face video by localizing the onset and offset of linguistic units along the visual timeline. Given a video sequence

$$\mathbf{x} \in \mathbb{R}^{T \times H \times W \times 3}, \quad (1)$$

and its corresponding phoneme sequence

$$\mathbf{p} = (p_1, \dots, p_{T_p}), \quad (2)$$

our goal is to recover a monotonic segmentation of the video into phoneme-consistent temporal regions.

Since speech progresses sequentially, VFA naturally admits a structured formulation: we view alignment as an *optimal monotonic path traversal* through a frame-phoneme compatibility lattice. Rather than predicting boundaries directly, V-Align learns a distribution over valid monotonic traversals, from which discrete phoneme boundaries are recovered via optimal path decoding.

2.1. Cross-Modal Lattice Construction

We extract frame-level lip motion representations using Visual Transformer Pooling (VTP) [9], a strong Lip encoder used in downstream Visual Speech Recognition tasks. This yields visual embeddings

$$\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_{T_v}] \in \mathbb{R}^{T_v \times d_v}. \quad (3)$$

In parallel, transcript phonemes are mapped into contextual phoneme embeddings using XPhoneBERT [10], a strong phoneme representation model used as a text encoder in Text-to-Speech:

$$\mathbf{P} = [\mathbf{p}_1, \dots, \mathbf{p}_{T_p}] \in \mathbb{R}^{T_p \times d_p}. \quad (4)$$

Both modalities are mapped into a shared alignment space via trainable modality adaptation modules, g_v and g_p , implemented as stacked 1D convolutional projection networks:

$$\tilde{\mathbf{v}}_t = g_v(\mathbf{v}_t), \quad \tilde{\mathbf{p}}_s = g_p(\mathbf{p}_s), \quad (5)$$

2.2. Compatibility Scores and Traversal Posterior

Unlike unimodal TTS alignment [11], the compatibility function must bridge heterogeneous embedding spaces, making lattice sharpness and structural regularization critical. We construct a dense frame-phoneme compatibility lattice by computing similarity scores:

$$\phi(t, s) = -\tau \|\tilde{\mathbf{v}}_t - \tilde{\mathbf{p}}_s\|^2, \quad (6)$$

where τ , the temperature, controls the sharpness of the lattice. Smaller values of τ produce a sharper posterior distribution (discussed in the supplementary material) These scores are normalized to obtain a *traversal posterior* over phoneme positions at each frame:

$$\Gamma_{t,s} = \text{softmax}_s(\phi(t, s)), \quad \Gamma \in [0, 1]^{T_v \times T_p}. \quad (7)$$

Γ serves as a structured distribution over monotonic alignment traversals through the lattice.

2.3. Optimal Monotonic Traversal Decoding

Forced alignment requires discrete boundaries rather than soft posteriors. We therefore recover the most likely monotonic traversal:

$$\pi^* = \arg \max_{\pi \in \mathcal{M}} \sum_{t=1}^{T_v} \log \Gamma_{t, \pi(t)}, \quad (8)$$

where $\mathcal{M}$ denotes the set of valid monotonic paths. This is efficiently solved via dynamic programming (Viterbi-style monotonic decoding), yielding a binary traversal matrix $\hat{\Gamma}$ from which phoneme durations are obtained by summation over frames.

3. Training Strategy

Accurate VFA requires learning both (i) globally consistent monotonic structure and (ii) sharp boundary localization. We adopt a two-stage training paradigm that mirrors this requirement.

3.1. Stage 1: Boundary-Free Path Structure Learning

In the first stage, we train V-Align without any explicit boundary annotations. Instead, we shape the traversal posterior Γ using structured optimal-path objectives [12].

3.1.1. Forward-Sum Objective.

We adopt the forward-sum loss [12], which maximizes the total probability mass over all monotonic traversals consistent with the transcript order:

$$\mathcal{L}_{\text{fs}} = -\log \sum_{\pi \in \mathcal{M}} \prod_{t=1}^{T_v} \Gamma_{t, \pi(t)}. \quad (9)$$

This objective serves as a differentiable relaxation of monotonic path search, encouraging alignment structure to emerge without boundary supervision.

3.1.2. Path Binarization Loss.

To further concentrate probability around a single interpretable traversal, we encourage consistency between Γ and the optimal decoded path $\hat{\Gamma}$:

$$\mathcal{L}_{\text{bin}} = -\sum_{t,s} \hat{\Gamma}(t,s) \log(\Gamma(t,s)). \quad (10)$$

Together, these losses yield a strong boundary-free initialization of the alignment lattice.

3.2. Stage 2: Word-Level Boundary Refinement with MFA

While Stage 1 learns robust traversal structure, achieving frame-accurate boundaries requires supervision. Direct phoneme-level supervision is challenging due to mismatched phoneme inventories between MFA and our tokenizer. We therefore perform refinement at the word level.

Given Γ^{ph} , we aggregate phoneme posteriors belonging to the same word into a word-level traversal posterior:

$$\Gamma_{t,u}^{\text{word}} = \sum_{s \in \mathcal{S}(u)} \Gamma_{t,s}^{\text{ph}}, \quad (11)$$

Model	LRS2		LRS3	
	MAE↓	ACC↑	MAE↓	ACC↑
KWS-Net [7]	171.5ms	53.0%	262.9ms	42.6%
CTC-based [15]	80.4ms	71.7%	124.5ms	60.1%
Transpotter [8]	71.7ms	75.0%	167.3ms	61.8%
DVFA [5]	67.7ms	84.2%	97.7ms	80.2%
He et al. [6]	50.2ms	89.5%	70.5ms	87.0%
V-Align Stage 1 (Ours)	54.6ms	86.13%	82.9ms	79.6%
V-Align Stage 2 (Ours)	32.9ms	91.2%	56.9ms	88.5%

Table 1: Word-level visual forced alignment performance of V-Align under Stage 1 (boundary-free) and Stage 2 (boundary-supervised) training on LRS2 and LRS3.

where $\mathcal{S}(u)$ denotes the phoneme indices corresponding to word w_u . MFA provides word-level boundary targets, and we apply a binarization-style supervision loss:

$$\mathcal{L}_{\text{sup}} = -\sum_{t,u} \hat{\Gamma}^{\text{MFA}}(t,u) \log(\Gamma^{\text{word}}(t,u)). \quad (12)$$

The overall training objective in Stage 2 becomes:

$$\mathcal{L} = \mathcal{L}_{\text{fs}} + \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{sup}}. \quad (13)$$

4. Experimental Setup

Dataset: V-Align is compared against the baselines on LRS2 [13] and LRS3 [14]. The LRS2 dataset is a comprehensive collection of naturally-occurring videos, featuring a variety of news and talk show clips sourced from the BBC. The LRS3 dataset includes TED and TEDx talks, offering a larger and more diverse set of spoken English sequences.

Metrics: To compare with other methods, we used the same experimental setup as our baselines. MAE (Mean Absolute Error) and ACC (accuracy) are used as metrics. MAE is the time differences in milliseconds between predicted and true word boundaries. ACC refers to the frame-level prediction accuracy, which is obtained by dividing the number of correctly predicted frames by the total number of frames in the video.

Implementation details: V-Align is trained for 400k steps in Stage 1, followed by an additional 100k steps of supervised refinement in Stage 2. We train with a batch size of 32 using the Adam optimizer, with an initial learning rate of 1^{-4} annealed to 1^{-6} . The temperature τ is set to 0.0005. All experiments are conducted on a single NVIDIA 2080 Ti GPU. We will release the code and trained weights to facilitate future research in VFA. Montreal Forced Aligner [3], an audio forced alignment tool is used to generate ground truth word alignment labels following the baselines [6, 5]. Both backbone encoders are kept frozen during training, allowing the alignment module to focus solely on structured traversal learning. The losses are weighed equally.

5. Experimental Results

We compare V-Align against representative baselines in visual forced alignment, including KWS-Net [7], Transpotter [8], a CTC-based alignment approach [15], DVFA [5], and the recent method of He et al. [6]. The first two approaches originate from the visual keyword spotting literature, while the CTC-based method leverages monotonic alignment implicitly within a lip-reading framework. DVFA and He et al. are the only prior works explicitly designed for full transcript-level visual

	Loss	MAE↓	ACC↑
Decoding	MAE↓ ACC↑		
Greedy Mono.	72.4	82.0%	
Frame Argmax	59.7	84.3%	
Optimal Path	54.6	89.5%	
Table 2: <i>Decoding ablation.</i>			
	Loss	MAE↓	ACC↑
	Binarization ($\mathcal{L}_{bin}$)	80.1	73.5%
	Forward Sum ($\mathcal{L}_{fs}$)	63.2	80.5%
	Word Alignment ($\mathcal{L}_{sup}$)	50.4	87.5%
	$\mathcal{L}_{bin} + \mathcal{L}_{fs}$	54.6	86.1%
	$\mathcal{L}_{bin} + \mathcal{L}_{fs} + \mathcal{L}_{sup}$	32.9	91.2%
Table 3: <i>Loss ablation.</i>			

forced alignment. Comparisons with KWS-based methods are included for completeness, as done in prior literature.

Quantitative comparisons are summarized in Table 1. We report the publicly available results of each method as presented in the respective papers. Stage 1 of V-Align learns alignment structure without requiring any boundary supervision and already outperforms all existing baselines except the most recent method of He *et al.*, demonstrating the effectiveness of boundary-free optimal traversal learning. Importantly, all baseline methods are trained with explicit audio forced alignment supervision, making Stage 1’s competitive performance particularly noteworthy. This highlights the potential of V-Align for deployment in scenarios where audio-based forced alignment supervision is unavailable or unreliable. Stage 2 further incorporates word-level refinement during training using audio forced alignment targets, improving boundary precision. As a result, V-Align achieves new state-of-the-art performance, reducing alignment error by 13.6ms on LRS3 and 17.3ms on LRS2 compared to prior best results.

We attribute these gains to the structured nature of our formulation. Unlike prior approaches that treat VFA as frame-wise token classification, V-Align explicitly models alignment as a globally consistent monotonic traversal over the frame-phoneme lattice. Qualitative visualizations of the predicted alignments and frame-phoneme compatibility lattices are provided in the supplementary material.

6. Ablation Study

6.1. Effect of Training Objectives

We analyze the contribution of each training objective on the LRS2 dataset, as summarized in Table 3. Using only the binarization loss leads to unstable alignment learning, as sharpening the decoded path without learning global structure is insufficient. The forward-sum objective alone performs substantially better, highlighting the importance of modeling globally consistent monotonic traversals. The joint optimization of forward-sum and binarization further improves alignment quality. Word-level supervision alone yields strong performance due to its direct boundary guidance. Finally, incorporating word-level supervision on top of these objectives yields the best overall performance, indicating that supervised refinement is most effective when built upon a strong monotonic alignment prior.

6.2. Effect of Monotonic Decoding Strategy

To evaluate the importance of globally optimal monotonic traversal decoding, we replace the proposed optimal traversal inference with two alternative decoding strategies applied to the learned traversal posterior: (i) frame-wise argmax decoding and (ii) greedy monotonic decoding. Frame-wise argmax assigns each frame independently to the phoneme with maximum posterior probability, without enforcing temporal consistency. Greedy decoding enforces local monotonic progression by al-

SNR (dB)	MFA		WhisperX	
	MAE (ms)↓	ACC (%)↑	MAE (ms)↓	ACC (%)↑
20	15.6 ms	92.5	55.7 ms	88.1
10	69.5 ms	82.5	74.2 ms	80.9
0	142.3 ms	62.5	230.9 ms	56.7

Table 4: *Impact of audio corruption on audio forced alignment performance on the LRS3 dataset.*

SNR (dB)	MAE (ms)↓	ACC (%)↑
Clean	56.9 / 65.8	88.5 / 85.1
20	63.4 / 82.6	83.9 / 74.2
10	73.7 / 116.4	75.6 / 63.3
0	118.2 / 214.7	68.3 / 44.8

Table 5: *Effect of degraded MFA supervision. Each entry shows $\mathcal{L}_{fs} + \mathcal{L}_{bin} + \mathcal{L}_{sup} / \mathcal{L}_{sup}$ only. Reported on LRS3.*

lowing only stay-or-advance transitions at each frame, but does not optimize the global traversal objective. As shown in Table 2, optimal traversal decoding achieves the best performance, followed by frame-wise argmax, while greedy monotonic decoding yields the lowest results. Results are reported on LRS2.

7. Robustness Analysis

Audio forced alignment performs reliably under clean conditions but deteriorates significantly as audio quality degrades. Table 4 shows the performance of MFA and WhisperX under increasing noise levels, evaluated relative to clean MFA alignment. Here, the signal-to-noise ratio (SNR) controls the severity of corruption, with lower values indicating stronger noise. As SNR decreases, both systems exhibit increasing boundary error and reduced frame accuracy.

We further analyze the impact of such degraded supervision on V-Align. While Stage 1 learns alignment structure without boundary supervision, Stage 2 relies on MFA-generated word boundaries. Table 5 compares the impact of training with degraded supervision, with and without the optimal path objectives. While performance deteriorates in both cases as supervision quality decreases, the degradation is substantially more severe when training relies solely on MFA alignment supervision. These results highlight the practical importance of boundary-free alignment learning, particularly for audio-visual datasets [16] where speech is noisy or unreliable.

8. Conclusion

We introduced V-Align, a structured framework for Visual Forced Alignment that formulates alignment as an optimal monotonic path traversal over a frame-phoneme compatibility lattice. By learning a soft traversal posterior and recovering discrete boundaries via dynamic programming, V-Align achieves precise and temporally consistent localization. A two-phase training strategy enables the model to first learn alignment structure without explicit boundary supervision and subsequently refine boundaries using additional supervision. Our experiments demonstrate state-of-the-art performance on LRS2 and LRS3. These results highlight the effectiveness of structured optimal path learning for visual-only alignment.

9. Acknowledgment

This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

10. Use of Generative AI Disclosure

Generative AI tools were used solely for language refinement and editorial assistance during manuscript preparation. The authors reviewed, verified, and edited all generated content and take full responsibility for the accuracy, originality, and integrity of the work. No generative AI tools were used in the design of the methodology, analysis of results, interpretation of findings, or generation of experimental data.

11. References

- [1] E. Perego, F. Del Missier, M. Porta, and M. Mosconi, "The cognitive effectiveness of subtitle processing," *Media psychology*, vol. 13, no. 3, pp. 243–272, 2010.
- [2] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *arXiv preprint arXiv:2303.00747*, 2023.
- [3] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [4] R. Das, J. Izak, J. Yuan, and M. Liberman, "Forced alignment under adverse conditions," *University of Pennsylvania, CIS Dept. Senior Design Project Report*, 2010.
- [5] M. Kim, C. W. Kim, and Y. M. Ro, "Deep visual forced alignment: learning to align transcription with talking face video," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8273–8281.
- [6] Y. He, L. Yang, and S. Wang, "Enhancing visual forced alignment with local context-aware feature extraction and multi-task learning," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [7] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, "Seeing wake words: Audio-visual keyword spotting," *arXiv preprint arXiv:2009.01225*, 2020.
- [8] K. Prajwal, L. Momeni, T. Afouras, and A. Zisserman, "Visual keyword spotting with attention," *arXiv preprint arXiv:2110.15957*, 2021.
- [9] K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 5162–5172.
- [10] L. T. Nguyen, T. Pham, and D. Q. Nguyen, "Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech," *arXiv preprint arXiv:2305.19709*, 2023.
- [11] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 5530–5540.
- [12] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022-2022*. IEEE, 2022, pp. 6092–6096.
- [13] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.
- [14] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
- [15] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "Ctc-segmentation of large corpora for german end-to-end speech recognition," in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.
- [16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.