Perceptual Synchronization Scoring of Dubbed Content using Phoneme-Viseme Agreement

Honey Gupta
Amazon Prime Video
ghoney@amazon.com

Abstract

Recent works have shown great success in synchronizing lip-movements in a given video with a dubbed audio stream. However, comparison and efficacy of the synchronization capabilities of these methods is still weakly substantiated due to the lack of a generalized and visually-grounded evaluation method. This work proposes a simple and grounded algorithm – PhoVis, that can measure synchronization and the perceived quality of a dubbed video at an utterance-level. The approach generates expected visemes by considering a speaker's lip-pose history and the phoneme in the dubbed audio. A sync distance and a perceptual score is then derived by comparing the generated viseme with the clip's visemes with the help of spatially grounded pose-distances. PhoVis is built upon the most basic audio-video elements i.e. phonemes and visemes to compute agreement, which makes it a domain independent algorithm that can be used to score both original and lip-synthesized videos, allowing measurement and improvement of dubbing quality as well as video-synthesis methods. We demonstrate that PhoVis achieves better language generalization, is aptly tailored for lip-sync measurement and computes audio-lip correlation better than the existing AV sync methods.

1. Introduction

Incoherence between lip-movements and audio of a dubbed video is an aspect of Digital Entertainment Content (DEC) localization that has been widely accepted as passable in the motion picture industry. There are two directions towards easing this discomfort. The first is to alter the dubbing in a way that leads to less incoherence between the audio and facial lip-movements, and the second is to leverage advances in graphics to synthetically modulate the video according to the dubbed audio. The former is still widely opted for in industry due to the risk of introducing synthesis artifacts in *high-value paid* content.

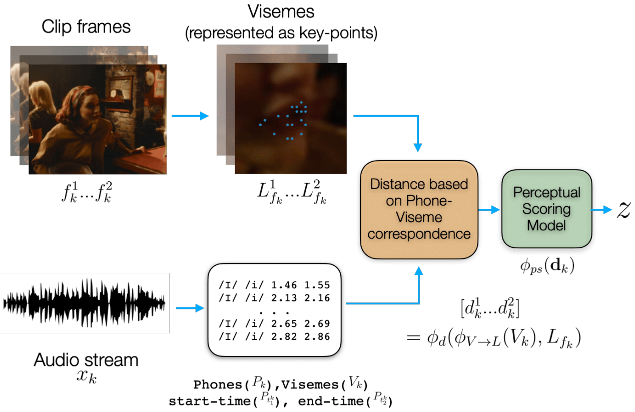

Clip frames
Visemes
(represented as key-points)
$f_k^1 ... f_k^2$
$L_{f_k}^1 ... L_{f_k}^2$
Distance based on Phone-Viseme correspondence
Perceptual Scoring Model
z
$\phi_{ps}(\mathbf{d}_k)$
/I/ /i/ 1.46 1.55
/I/ /i/ 2.13 2.16
. . .
/I/ /i/ 2.65 2.69
/I/ /i/ 2.82 2.86
$[d_k^1 ... d_k^2]$
$= \phi_d(\phi_{V \to L}(V_k), L_{f_k})$
Audio stream
x_k
Phones(P_k),Visemes(V_k)
start-time($P_{t_1^k}$), end-time($P_{t_2^k}$)

Figure 1. Every utterance (*phoneme*) has an associated lip-shape (*viseme*) that forms when the phoneme is spoken. PhoVis uses this correspondence to measure audio-lip correlation in a video. It builds a Euclidean reference dictionary for a set of visemes and uses it to generate sync distances and a perceptual sync score.

With the increasing number of dubbed content being targeted by video streaming players across the industry, dubbing experts need the ability to spot dialogues in dubbed versions where there is low audio-lip sync, and improve the quality of these dialogues by correcting it via re-writing and re-recording the script. This includes dubbing content in over 20 languages. In parallel, lip-synthesis models also require visually and temporally grounded feedback to improve synthesis. Identification of dialogues with poor dubbing quality requires a measure of mismatch between the dubbed audio and lip-movements. In addition, quantifying the experience of watching a dubbed video is also crucial as human-perception is subjective. Therefore, any correction path cannot be triggered or evaluated without a visually and temporally grounded scoring model that can (1) quantify the extent of mismatch, and (2) give a perceptual score quantifying the impact on viewing experience.

For the problem at hand, the existing solutions in the literature are limited. Most of the methods are supervised deep-models and hence, are not easily generalizable across diverse videos spanning multiple genres and languages. Moreover, currently *no method in literature quan-*

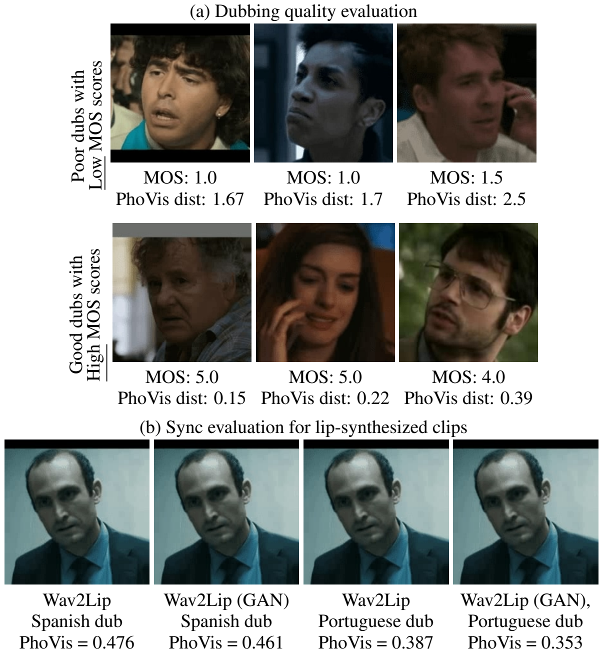


Figure 2. *(Please open in Adobe Reader to play the videos and zoom-in.)* Videos with varying dub quality and synthesized lip-sync. (a) shows poor and excellent dubs as rated by dubbing experts (MOS) along-with the PhoVis distance(↓), (b) uses PhoVis to quantify lip-sync of two lip-synthesis methods on dubbed audios.

tifies human perception of a dubbed video. The closest and most widely used method is SyncNet [13]. It was proposed for identifying time-shift between the audio and video. SyncNet and its derivatives have been adapted for lip-synchronization [40] but they do not directly aim to compute correlation, making them unsuitable for use-cases like *quantifying perceptual degradation when dubbed audio is in a language different from the spoken original.* Moreover, these methods [27] compute an embedding distance between video and audio. Embedding distances are weak visually-grounded metrics and give limited human-interpretable insights into tangible factors/spatial regions that lead to desynchronization.

To address these challenges, we propose a *simple, domain-independent, and visually-grounded perceptual lip-sync scoring model* that measures the correlation between the lip-movements in the video and the dubbed audio. Our method directly compares audio and lip-movements at the *fundamental phoneme* and *viseme* (PhoVis) level and is designed for dubbed content *i.e.* for cases where the video and audio are bound to not match, be it original or morphed. Moreover, we not only compute the audiovisual correlation, we also quantify human viewing experience by converting the correlation scores into human-scaled perceptual scores with the help of our custom perceptual scoring dataset annotated by dubbing experts.

We use correspondence between the phonemes from the audio and visemes in the frame to calculate audio-lip distance. PhoVis generates the expected viseme based on the dubbed phoneme and compares it with the viseme present in the frame with the help of *spatially grounded pose-distances.* These distances are utilized by a perceptual scoring model, which predicts a score in-line with human perception of lip-sync. The proposed method and the generated perceptual scores are *generic* and can be used to score both *original and morphed* video frames, allowing *measurement and improvement of dubbing as well as lip-synthesis quality.* Since the generated score is both visually and temporally grounded, PhoVis's correlation score can be directly utilized for active speaker detection, AV lead/lag detection and other related AV tasks. We perform multiple experiments on 6 *languages* and demonstrate that PhoVis is an aptly tailored solution which is *well correlated with human-perception, generalizes well across languages and performs better than the existing lip sync measurement methods.*

2. Related works

There can be three ways for computing audio-lip movement correlation - (1) distance between audio and face/lip video embeddings [1, 13, 20, 24, 27], (2) distance between video landmarks and the ones corresponding to the audio (Audio2Pose) [12, 18, 19, 52, 53, 58] and (3) distance between visemes corresponding to the video and the audio (Audio2Viseme) [29, 42, 43, 60]. Our proposed method is a combination of the last two approaches. Although not much work has been explicitly done for correlation measurement between audio and lip-movements, methods for multiple AV tasks such as lead/lad detection (SyncNet [13]) and AV speaker diarization (VisualVoice [20], LWTNet [1] and many more [23, 26, 38, 41, 46, 47]) compute generic audiovisual distance as part of the target solution.

AV speaker diarization involves separation of simultaneously spoken dialogue tracks. These works [23, 26, 31, 34, 37, 38, 41, 46, 47, 50, 51, 57, 59] typically utilize video to audio similarity for source separation. Though the models are designed to internally correlate lip-movements to the corresponding audio fragment, they are not designed for explicit audio-lip correspondence measurement. Similarly, most methods for visual-speech grounding [22, 28, 35, 39] and active speaker detection [1, 4, 8, 13, 17, 45, 49, 56] are self-supervised deep-models that extract video and audio embeddings projected in the same subspace. The embeddings generated are pertinent to the specifically targeted task and are not appropriate for dubbing quality scoring. Moreover, embedding distances are weak visually-grounded metrics and they give limited human-interpretable insights into the tangible factors that lead to desynchronization.

PC-AVS [58], TalkingFace [53], Audio2AU [12], Face-Former [18], DFA-NeRF [52] and Joint AT [19] are few

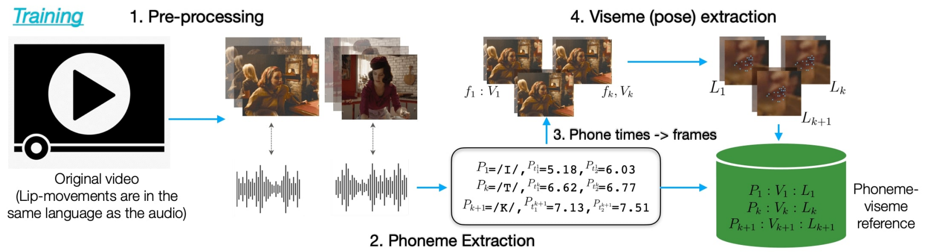

Training
1. Pre-processing
4. Viseme (pose) extraction
$f_1 : V_1$
f_k, V_k
L_1
L_k
L_{k+1}
3. Phone times -> frames
P_1=/I/, $P_{t_1^1}$=5.18, $P_{t_2^1}$=6.03
P_k=/T/, $P_{t_1^k}$=6.62, $P_{t_2^k}$=6.77
P_{k+1}=/K/, $P_{t_1^{k+1}}$=7.13, $P_{t_2^{k+1}}$=7.51
$P_1 : V_1 : L_1$
$P_k : V_k : L_k$
$P_{k+1} : V_{k+1} : L_{k+1}$
Phoneme-viseme reference
Original video
(Lip-movements are in the same language as the audio)
2. Phoneme Extraction

Figure 3. The principle idea behind PhoVis is that given a single speaker clip in the original language, we can build a reference of how the lip-shape should look like when a particular phoneme is spoken. The figure shows the proposed method for building a phoneme-viseme reference that is subsequently used for measuring the audio-lip sync of all the dubbed versions of the clip.

works that perform Audio2Pose *i.e.* their model internally converts the audio stream into facial landmarks. The landmarks can be used for pose-based lip-sync measurement. However, such methods are proposed for video synthesis and are limited to the dataset they have been trained on.

Correspondence between phonemes and visemes has been studied in literature for quite sometime [3, 6, 7, 10, 30, 32, 36, 44]. More importantly, visemes form the base unit in computer graphics and animation industry for synchronizing lip-movements [29, 42, 43, 54, 60]. However, to the best of our knowledge, PhoVis is the first to utilize phoneme-viseme agreement for computing a correlation score between lip-movements and audio, and subsequently compute a human perception score, specifically for dubbed content.

3. PhoVis: Phoneme-Viseme correspondence for audio-lip correlation measurement

In a dubbed video, the frames corresponds to the original language in which the video was created and the audio belongs to the dubbing language. This method focuses on measuring correlation between the utterances, *i.e.* phonemes and the lip-movements or visemes, which are bound to mismatch in this case. Three inputs available for any dubbed content are - original, non-dubbed video (OV), original, non-dubbed audio (OA) and dubbed audio (DA).

3.1. Pre-processing pipeline

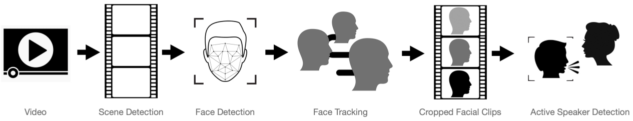


Figure 4. Video pre-processing method

Unlike some methods in literature that are constrained to front-facing videos, this method is expected to tackle *in-the-wild* facial clips as it is targeted for DEC content. Therefore, the first step is to extract viable single-person face tracks with a minimum duration and pixel-area. Figure 4 shows the pre-processing pipeline for extracting clips. We first find the scene boundaries, $\mathcal{S}$ using a simple content based detector from PySceneDetect [11] and perform face detection on each frame $f \in s;\ \forall\ s \in \mathcal{S}$ using standard S3FD [55] model. We track the detected faces *along-with face pose* in each scene, then *align and crop* the tracked faces which have a minimum facial area of 100×100. Note that we assume the lead/lag issues between the audio and video are absent or corrected using other algorithms.

Active Speaker Detection (ASD). There are additional nuances related to cinematographic content that need to be addressed. Based on our study, $\approx$ 50% of the clips generated by the above pipeline have a non-speaking face. To investigate, we manually annotated 174 random OV clips as speaker or non-speaker. Though many advanced algorithms are available [4, 8, 49], for simplicity, we used SyncNet [13]. SyncNet's confidence score performs decently for ASD when the video and audio are in the *original* language. It showed a high precision of 90% with a decent recall at a confidence threshold of 1. We added this as an ASD filter.

3.2. Algorithm details: PhoVis

The principle idea is that for a given clip containing a single speaker in a particular set of imaging conditions, we can derive a reference viseme of how the lip-shape should look like when a particular phoneme is spoken. This expected lip-shape or viseme can be represented as 2d key-points. Using OV clips, we build a reference dictionary R_{PV} for a set of visemes V^* for a given clip x, and during inference, we use this reference to calculate synchronization between the dubbed phonemes and the video, for all available DA versions independent of the language.

Why use Phonemes/Visemes? (1) Being fundamental units, they allow tracking of the exact instances that lead to poor lip-sync. (2) The phoneme-viseme mapping is language agnostic, making PhoVis inherently scalable across languages. A phoneme is the basic unit of spoken sound based on pronunciation, and a viseme represents the lip-pose when a particular phoneme is spoken. Studies suggest a *many-to-one*, static mapping between phonemes and visemes [6]. For *e.g.* the words *pet*, *bell*, and *men* are difficult for lip-readers to distinguish as they have the phonemes /P/, /B/, and /M/, respectively, which map to the same viseme /p/. Therefore, having similar phonemes across

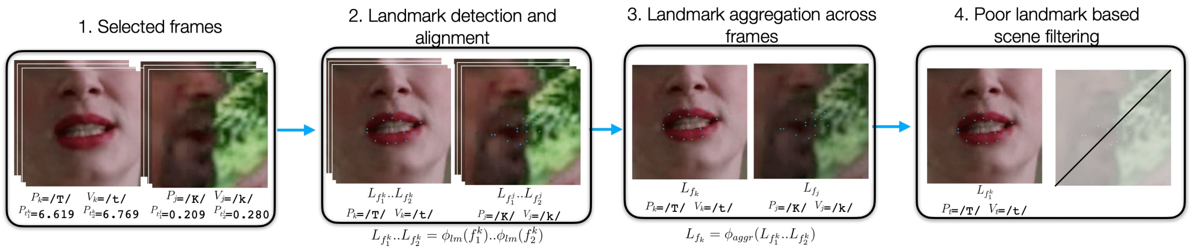
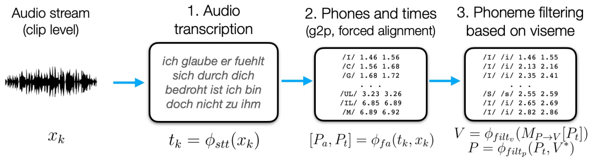

1. Selected frames
2. Landmark detection and alignment
3. Landmark aggregation across frames
4. Poor landmark based scene filtering
P_k=/T/ V_k=/t/
$P_{t_1^k}$=6.619 $P_{t_2^k}$=6.769
P_j=/K/ V_j=/k/
$P_{t_1^j}$=0.209 $P_{t_2^j}$=0.280
$L_{f_1^k}..L_{f_2^k}$
P_k=/T/ V_k=/t/
$L_{f_1^j}..L_{f_2^j}$
P_j=/K/ V_j=/k/
$L_{f_1^k}..L_{f_2^k} = \phi_{lm}(f_1^k)..\phi_{lm}(f_2^k)$
L_{f_k}
P_k=/T/ V_k=/t/
L_{f_j}
P_j=/K/ V_j=/k/
$L_{f_k} = \phi_{aggr}(L_{f_1^k}..L_{f_2^k})$
$L_{f_1^k}$
P_t=/T/ V_t=/t/

Figure 5. Algorithm for extracting visemes (represented as lip-keypoints) from a given set of video stream and phoneme timings.

original and dubbed audios can improve audio-lip sync.

3.3. Phoneme extraction algorithm



```
/I/ 1.46 1.56
/C/ 1.56 1.68
/G/ 1.68 1.72
. . .
/UL/ 3.23 3.26
/IL/ 6.85 6.89
/M/ 6.89 6.92
```

```
/I/ /i/ 1.46 1.55
/I/ /i/ 2.13 2.16
/I/ /i/ 2.35 2.41
. . .
/S/ /s/ 2.55 2.59
/I/ /i/ 2.65 2.69
/I/ /i/ 2.82 2.86
```

Figure 6. Algorithm for phoneme extraction from given audio.

Given a cropped audio x_k, we first run it through a transcription model ϕ_{stt} [2, 14–16, 25] to get the spoken dialogue, as shown in Figure 6. The transcripted dialogue t_k and the audio x_k are fed into a forced aligner ϕ_{fa} that performs grapheme to phoneme (G2P) alignment of the transcripted text with the audio. The outputs of the aligner are phonemes present in the dialogue P_a (in the International Phonetic Alphabet (IPA) format) and their corresponding timings P_t. IPA is used for uniformity and scalability across languages. For G2P, we use Montreal Forced Aligner [33]. Lastly, we filter the extracted phonemes based on the set of languages being considered, as explained below.

Phoneme-viseme mapping. Each language has a set of associated phonemes and each phoneme has a constant mapping to a viseme. The total number of phones can vary depending on the language structure. To compare across languages, we find the visemes that are common across 6 languages we considered – EN, FR, IT, DE, ES and PT. These visemes are V^*={/f/,/i/,/k/,/p/,/s/,/t/}, which map to 16 IPA phonemes. Individual languages might have < 16 phonemes mapping to the 6 visemes. We filter and use only these phones for lip-sync scoring. The phoneme to viseme mapping $M_{P\to V}$ can be found in the supplementary paper.

3.4. Viseme extraction algorithm

For a clip x_i, let there be a phoneme P_k, with start-time $P_{t_k^1}$ and end-time $P_{t_k^2}$. We get frames f_1^k through f_2^k corresponding to phone's duration by using $f_*^k = x_{fps} \times P_{t_k^*}$, where x_{fps} is the frame-rate. We run a 3d facial-landmark detection (LM) model ϕ_{lm} [9] on each frame to extract lip key-points $[L_{f_1^k}..L_{f_2^k}]$, as shown in Figure 5. This gives us a phoneme tag and the viseme for each frame. Visemes are represented by 48 through 68 landmarks of the Multi-PIE face-landmark schema [21], corresponding to the lip-region. Since phonemes typically span over 2-6 frames, we aggregate each landmark across frames and get one viseme representation for each phone $L_{f_k} = \phi_{aggr}(L_{f_1^k}..L_{f_2^k})$. We empirically found taking max of key-points results the best.

Filtering spurious key-points. Some extreme conditions like side pose lead to incorrect key-point predictions, like the second sample of Figure 5. The final step is to filter these spurious visemes. We observed that such erratic behavior is typically accompanied by disarrayed key-points. Hence, we filter out visemes if the 2d variance of key-points is > 200 pixels. This limit was found empirically and is based on the clip-size, minimum face-size and the LM model.

3.5. Building phoneme-viseme reference

Figure 3 summarizes the algorithm to build phoneme-viseme reference from an OV video. To tackle *in-the-wild* clips, we build references at a clip-level. This allows us to bypass the challenge of building a generic reference across imaging conditions. Though *noise* from both phoneme and key-point detection could be an issue. We tackle this using robust aggregations.(1) A clip has $\approx$ 30 phones which are aggregated and mapped to 6 visemes.(2) A phone spans 2-6 frames@25fps, so key-points across frames are also aggregated. Both these steps reduce the impact of noise.

Given a video, we extract frames along-with the original audio segments. We determine the phonemes present in OA $[P_1, ..P_k]$, and their timestamps (Sec. 3.3). For each phoneme, we extract the frames and derive viseme representations $[V_1, ..V_k]$ (Sec. 3.4). For each distinct viseme in V^*, we build a reference represented by a set of key-points. In a clip, there are multiple frames that could have the same viseme. To build one reference set of key-points per viseme, we use aggregation across the viseme samples. The obtained set of key-points $[L_1, ..L_k.]$ have a 2d coordinate for each lip-landmark and represent a particular viseme.

3.6. PhoVis distance measurement

Figure 7 shows the steps involved during inference. For a dubbed video, PhoVis contains seven steps to generate the perceptual lip-sync score. Steps 1 to 4 are similar to the training phase (Fig. 3). The difference between training and inference is that here the audio is the dubbed version (DA). Step 4 involves extracting visemes and lip key-points $[L'_1, ..L'_k]$. For the detected dubbed phones, we use $M_{P\to V}$

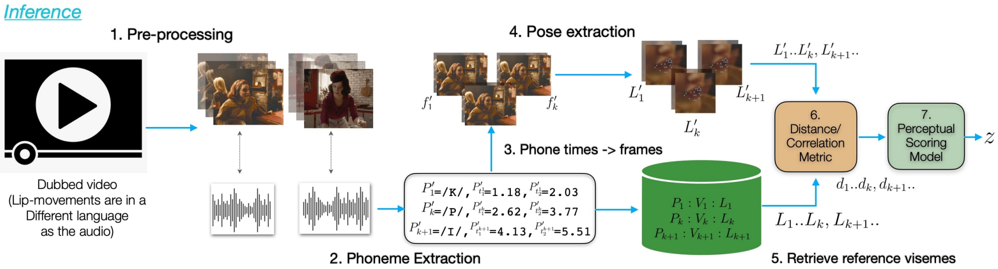


Figure 7. Method for measuring the audio-lip sync and perceptual quality of the dubbed clips using the built references.

to get the *expected* visemes $[V'_1, ..V'_k]$. These visemes indicate the lip-shape that a viewer should've seen if the video was in the dubbed language. If a reference for the *expected* visemes is present in the reference dictionary R_{PV}, then the corresponding reference key-points L_k are fetched.

Step 6 compares the *current* frame's viseme $\mathbf{L}'$ with the reference viseme $\mathbf{L}$. This roughly translates to *how far what we see is from what we hear*. Since the visemes are 2d landmarks, a normalized distance between key-points could give a measure of the desired correlation. We explored multiple metrics (Sec. 4.3.4), and found area-normalized L2 [9] as correlation metric to work the best.

$$\mathbf{d} = \frac{1}{N}\sum_{j=1}^{N}\frac{||L_k^j - L_k'^j||_2}{r} \tag{1}$$

where N is the number of key-points, L'_k is the current frame's landmark and L_k is the reference's landmarks. r is the normalization factor - the max of x and y coordinates.

Perceptual scoring model. To project viseme distances on a human-labeled scale and generate a single perceptual score z, PhoVis takes the sequence of viseme distances for a clip and feeds it to an ML model *with fixed-length padding*. The model is trained to generate a score rating the lip-sync of the entire clip. To build and evaluate this model, we annotated a custom dataset with the help of dubbing experts.

4. Experiments

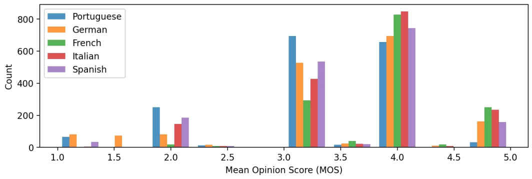


Figure 8. Distribution of the expert annotated perceptual dataset

Dataset. We built an expert annotated dataset containing the Mean Opinion Scores (MOS) for ≈8200 clips in 5 languages. The dataset contains cropped face clips from 815 DEC videos belonging to different genres, taken from the Prime Video catalog. The original versions are in EN and the dubbed clips are in FR, IT, DE, ES and PT. The dataset was annotated by *dubbing experts*, 3 annotations per clip, with the target to score perceptual correlation between audio and the lip-movements in the clip. The scoring was on a scale of [1-5], 1 being poorest. Distribution of the annotated dataset is shown in Figure 8. We removed spurious labels by filtering samples with low inter-annotator agreement.

4.1. Perceptual scoring of dubbed clips

As the primary experiment, the task is to predict a meta score indicating the viewing experience of a dubbed video. Dubbing quality scoring does not have clear boundaries *e.g.what makes a dubbed clip 5 but not 4?* Based on our studies, we also found that humans tend to *not* give low scores [1-3] unless the experience is very disruptive. Using this knowledge, we task a binary scoring objective, where clips with MOS 1-3 are considered as *bad* dubs, and 4-5 are *good*. The question of how the performance alters when a 5-level scoring is used, is discussed in Sec. 4.3.5. The evaluation metrics are weighted precision, recall and F1-score.To generate perceptual scores, we build an RF Classifier that takes a list of the viseme distances for a clip and predicts a binary perceptual score. We fitted 200 random searched models with a 70-30 split and 5-fold cross-validation. Finer training details are discussed in supplementary paper.

4.1.1 Comparison with baseline model

As baseline, we want to analyze how a phoneme-viseme based method compares to an embedding based model trained on the same content, for the same task. For this, we trained an end-to-end embedding model (E2E) with a contrastive learning objective, similar to SyncNet [13]. The video encoder is taken from S3D [48] and pre-trained Wav2Vec2 [5] is the audio encoder. This model is trained on 0.2s clips taken from the same 815 cinematographic videos. The trained model is followed by an RF classifier which takes the embeddings and generates a perceptual score.

In Table 2, PhoVis performs similar or better than the baseline in 4 of 5 languages. Relatively low metrics from the baseline could be attributed to spurious distances derived from deviant frames. Cinematographic content sees a wide diversity of lighting conditions, face poses and angles. Information from lip-regions in case of extreme poses is not reliable. These are filtered out by PhoVis during inference, but an embedding model will give unreliable results.

Table 1. Language-wise performance comparison for perceptual lip-sync scoring.

Language	Method	Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	Error Precision	Error Recall	Error F1 Score
Spanish	PhoVis	0.699	0.710	0.699	0.606	0.314	0.875	0.462
	SyncNet [13]	0.645	0.551	0.565	0.564	0.428	0.331	0.373
	LWTNet [1]	0.616	0.573	0.616	0.582	0.479	0.609	0.536
	VocaLiST [27]	0.583	0.565	0.583	0.570	0.467	0.433	0.449
	SparseSync [24]	0.548	0.587	0.574	0.580	0.409	0.474	0.439
Portuguese	PhoVis	0.555	0.543	0.551	0.547	0.611	0.629	0.620
	SyncNet [13]	0.563	0.534	0.563	0.545	0.595	0.689	0.638
	LWTNet [1]	0.517	0.441	0.526	0.479	0.566	0.635	0.599
	VocaLiST [27]	0.708	0.801	0.708	0.643	0.565	0.613	0.588
	SparseSync [24]	0.595	0.566	0.595	0.485	0.607	0.608	0.607
German	PhoVis	0.635	0.610	0.635	0.620	0.505	0.645	0.567
	SyncNet [13]	0.604	0.570	0.601	0.569	0.465	0.735	0.569
	LWTNet [1]	0.494	0.502	0.494	0.495	0.432	0.422	0.427
	VocaLiST [27]	0.563	0.561	0.563	0.558	0.490	0.313	0.382
	SparseSync [24]	0.508	0.519	0.528	0.508	0.501	0.430	0.463
Italian	PhoVis	0.749	0.708	0.749	0.723	0.362	0.286	0.319
	SyncNet [13]	0.610	0.562	0.610	0.567	0.209	0.157	0.172
	LWTNet	0.526	0.520	0.526	0.522	0.308	0.286	0.296
	VocaLiST [27]	0.646	0.494	0.646	0.516	0.200	0.006	0.011
	SparseSync [24]	0.602	0.606	0.602	0.604	0.621	0.201	0.304
French	PhoVis	0.890	0.850	0.890	0.866	0.176	0.154	0.164
	SyncNet [13]	0.670	0.602	0.670	0.628	0.377	0.258	0.307
	LWTNet [1]	0.473	0.471	0.473	0.472	0.262	0.356	0.302
	VocaLiST [27]	0.778	0.605	0.778	0.681	0.250	0.271	0.260
	SparseSync [24]	0.753	0.652	0.753	0.679	0.217	0.180	0.197

Table 2. Comparison with the Baseline model

Language	Model	Precision	Recall	F1-Score
Spanish	PhoVis	0.710	0.699	0.606
	E2E-Baseline	0.568	0.610	0.588
German	PhoVis	0.610	0.635	0.620
	E2E-Baseline	0.594	0.674	0.632
Italian	PhoVis	0.708	0.749	0.723
	E2E-Baseline	0.672	0.680	0.676
French	PhoVis	0.850	0.890	0.866
	E2E-Baseline	0.782	0.915	0.843
Portuguese	PhoVis	0.543	0.551	0.547
	E2E-Baseline	0.542	0.549	0.545

Perception of sync is not affected when lip-regions are not clearly visible. Since, E2E will emit a distance in all cases, distances from such frames will lead to lower results.

4.1.2 Comparison with existing works

We compare with few relevant AV sync methods with available pre-trained models - VocaLiST [27], SparseSync [24], LWTNet [1] and SyncNet [13]. VocaLiST and SyncNet are the closest models, though none of these are designed for dubbed content, where the audio is bound to not match the video. We add a perceptual scoring head to each model to generate perceptual scores. From Table 1, PhoVis has a better F1 in all languages except PT. It performs consistently well across languages, whereas we notice inconsistent performance from other methods *e.g.* VocaLiST for ES/IT, SyncNet for ES/IT/FR, SparseSync across langs. This indicates that PhoVis generalizes better across languages as compared to embedding based methods, which are influenced by dataset and objective used for training.

Across languages, the range of metrics is large. This could be because (1) quantifying human-perception of audio-lip sync is a complex task, and (2) languages have specific nuances like number of lateral consonants present which influence perception of a dubbed clip. Theoretical AV correlation could be high but perception of the overall clip could be bad, and vice versa, leading to incoherence between the scores and input distances and thus, poor metrics overall (*e.g.* PT). Since this is the first attempt to quantify dubbing quality perception in literature, we believe that the results look promising with a decent scope for future works.

4.2. Applications

4.2.1 Dubbing quality analysis

In this task, PhoVis distance is used to analyze sync quality of dubbed cinematographic content. Figure 9 shows the results. Notice the *in-the-wild* characteristic of speaker-tracks, and the diversity in lighting and face-angles. Columns 1, and 2, 5 show the original and dubbed clips along-with their PhoVis distances. Dubbed clips have a *higher* PhoVis distance than the original, indicating a higher audio-lip mismatch, which is expected. Playing and comparing lip-sync in the two ES and PT dubs and indicates that ES is slightly better for both clips as compared to PT, which is in-line with the difference in the PhoVis distances. Considering all the clips extracted from a given dubbed video, if the estimated distance is higher than a certain threshold for certain clips, like the ones in Figure 2-Row 1, they can be flagged as having bad dubbing quality. Once flagged, the timestamps can be sent to dubbing studios to improve the dubbing quality by rewriting and re-recording the script.

4.2.2 Lip-sync evaluation of synthetic videos

We use PhoVis to measure audio-lip sync in lip-morphed videos and compare two synthesis methods. Columns 3,6

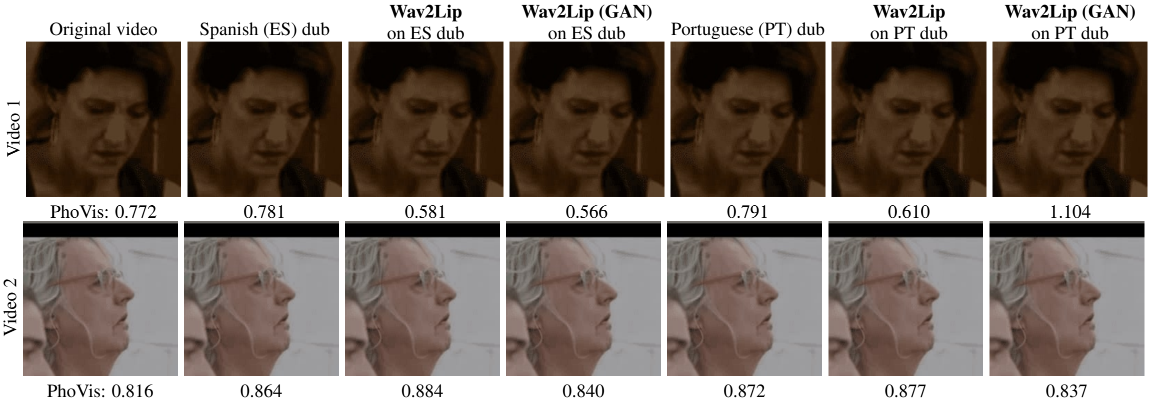


Figure 9. *(Please open in Adobe Reader to play the videos and zoom-in for clarity.)* This figure shows original, dubbed and lip-synthesized clips for two languages along-with the estimated PhoVis distances(↓). Between columns 2 and 5, Spanish dubs seem to have a better sync, which is in-line with the estimated PhoVis distances. Wav2Lip(GAN) seems to have better sync than Wav2Lip, as observed from the videos and the distances. This demonstrates that dubbing quality and synthesis lip-sync comparison can be performed using PhoVis distance.

and 4,7 of Figure 9 show language respective lip-synced versions using Wav2Lip and Wav2Lip (GAN) [40] . Compared to dubs, the synthesized versions have *lower* scores, indicating improvement in audio-lip correlation. Among the two, Wav2Lip(GAN) seems to perform better, which is in-line with the conclusions from the original paper [40].

An interesting observation is that for Video 1, the distance of synthesized clips is lower than the original clip itself. This is because the original clip has non-speaker frames which are also morphed during synthesis, leading to the lower distance. In Video 2, the entire dialogue is spoken with a side face-pose. In this case, synthesis does not lead to a significant reduction in distance as compared to dubs. Note that a limitation of using PhoVis distance is that it does not measure the synthesis artifacts. This could be implicitly captured by the behavior of landmark detection model, but this does not give a direct feedback. Future extension could be to merge PhoVis score with an image quality score for a holistic benchmarking of lip-synthesis methods.

4.2.3 Active speaker detection (ASD)

We test two approaches for ASD on original OV clips. (1) PhoVis distance to flag a clip as having an active speaker if the distance is below a certain threshold (1.0), and (2) Reference dictionary – if the variance of key-points across visemes is low, it means that all visemes have a similar representation, indicating a non-speaker clip. We used this logic to analyze ASD performance on the dataset mentioned in Sec. 3.1. Results in Table 3 indicate that both simple approaches from PhoVis can be decently used for ASD. They also have a better F1-score as compared to the baseline - SyncNet which is used in the pre-processing flow.

Table 3. Active speaker detection using PhoVis

	PhoVis distance	PhoVis references	SyncNet
F1-Score	0.731	0.711	0.614

4.3. Ablation studies and model behavior

4.3.1 Confidence intervals (CI)

We compute 90% confidence intervals for PhoVis distance and scoring probability $P(z)$ in Table 8 for different types of videos. Both distance and $P(z)$ bounds have observable correlation with MOS scores. OV clips, which have synced audio and lip-movements, have distance and score CIs close to MOS $3-5$, indicating them as having good correlation. We also see a distinction in CI between perceptual scores 0 and 1, validating the usability of PhoVis outputs.

Table 4. PhoVis distance/score confidence intervals

	PhoVis Distance↓		$P(z)$ ↑	
	Lower limit	Upper limit	Lower limit	Upper limit
MOS <= 1	1.748	2.559	0.355	0.457
MOS $\in (1, 2]$	1.731	2.076	0.422	0.504
MOS $\in (2, 3]$	2.273	2.659	0.441	0.489
MOS $\in (3, 4]$	1.430	1.931	0.529	0.572
MOS $\in (4, 5]$	1.076	1.818	0.558	0.598
OV clips	0.925	1.705	0.592	0.659
Predicted score: 0	1.268	2.565	0.288	0.338
Predicted score: 1	0.810	1.090	0.606	0.627

4.3.2 Deeper performance analysis

We dive deeper into model performance by calculating *Error precision*, *Error recall* and *Error F1*, shown in last three columns of Table 1. *Error precision* is the precision of the negative class and signifies the model's ability to detect poor dubs. We observe that PhoVis is either comparable or performs better for most languages in terms of Error precision. It is low for FR. Further analysis revealed that French has the highest class-imbalance (1:10), leading to positive class domination and low Error precision across methods. Error analysis in this case would thus be inappropriate.

4.3.3 Input features to the perceptual scoring model

We analyze different approaches for feeding inputs to the perceptual scorer. (1) Directly feeding the obtained distances, (2) Extracting features from the distances based on visemes. Considering 6 visemes gives 36 possible combination of current vs. expected visemes. For each of these, we compute distance statistics like {*mean, median, min, max and percentiles*}, giving a 324-dim input-vector. (3) We know that distance between same visemes should be $\approx$ 0 but we cannot say concretely about inter-viseme distance. We thus feed distances that are computed only between same visemes. Table 5 shows that directly feeding all distances performs the best, while the same-viseme distance has the lowest F1. This is intuitive as feeding raw inputs allows the network to extract relevant features as compared to human engineering. PhoVis uses the first approach.

Table 5. Feature-set for binary perceptual scoring model

	Precision	Recall	F1 Score
1. All-visemes	0.71	0.699	0.606
2. Statistical features	0.619	0.628	0.623
3. Same-visemes	0.608	0.662	0.608

4.3.4 Distance metric for correlation score

Table 6 compares L2, cosine distance, Chebyshev distance and 1d correlation as metrics for computing distance between the current (L') and expected)L) visemes. Considering both F1 and error F1 scores, we observe that L2 seems to have the best performance overall.

Table 6. Performance variation w.r.t. distance metric

Distance	Precision	F1 Score	Error Prec.	Error F1
L2	0.710	0.606	0.314	0.462
Cosine	0.630	0.590	0.354	0.355
Chebyshev	0.621	0.603	0.311	0.448
Correlation	0.606	0.613	0.306	0.432

Intra-viseme distance. A crucial property needed in the distance metric is to have close to zero intra-viseme distance and high variance in inter-viseme distance. This is examined in Figure 10. It shows inter-viseme distances for a sample clip where the y-axis shows the reference viseme and the x represents all visemes that were present in the clip. Most metrics have close to 0 diagonal, depicting that (1) the reference visemes are close to the same current visemes *e.g.*/p/ reference has $\approx$ 0 distance with all the /p/ visemes in the clip. (2) Diagonal has the darkest color in each row indicating that intra-viseme distance is always lower than inter-viseme distance. These observations establish that reference key-points are an effective representation for visemes.

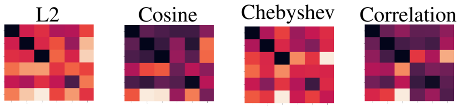


Figure 10. Intra- and inter-viseme scores using different metrics.

4.3.5 Binary vs multi-class perceptual scoring

Expert perceptual dataset contains scores on a scale of 1-5. Table 7 shows results from different scales of perceptual scores. Binary labels are obtained as explained in Sec. 4.1. For 3-class scoring, MOS $\in$ $[1,3)$ are poor dubs, $[3,4)$ are neutral, and $\in$ $[4,5]$ is good. 5-class scoring considers rounded MOS scores as labels. Low results for 5-point scoring could be due to (1) subjectivity of human perception, and (2) ill-defined distinction between scores. This could be alleviated with more data. Due to limited resources, binary is the optimal choice in the current setting.

Table 7. Binary vs multi-class perceptual scoring

	Score class	Accuracy	Precision	F1 Score
PhoVis	Binary	0.699	0.710	0.606
	3 class	0.464	0.450	0.454
	5 class	0.398	0.364	0.377
SyncNet	Binary	0.645	0.551	0.564
	3 class	0.471	0.424	0.428
	5 class	0.329	0.335	0.331

4.3.6 Viseme importance and performance

We analyze performance variation *w.r.t* the size of reference dictionary. The visemes $V^* = \{$/f/, /i/, /k/, /p/, /s/, /t/$\}$ are common across 6 languages. We analyzed two other sets as base for the reference R_{PV} - (1) V^*+ /a/ and /o/, the two other commonly occurring visemes, (2) all visemes (14). 6 visemes map to 16 phonemes, 8 map to 20 and 14 map to 31. Since this is not scalable across languages, we experiment only with ES. Table 8 shows having more visemes in the reference improves detection of bad dubs, both in terms of error precision and error F1. The overall performance of perceptual scoring reduces (in terms of precision and F1). This could be due to reduced distinction between reference visemes. With increased number of visemes, the distinction between visemes decreases and results in poor identification of *good* dubs, leading to reduced performance.

Table 8. Size of viseme-set in reference R_{PV}

Size of Viseme- set in Reference	Precision (weighted)	F1 Score (weighted)	Error Precision	Error F1-Score
6 - {f,i,k,p,s,t}	0.710	0.606	0.314	0.462
8 - {f,i,k,p,s,t,a,o}	0.552	0.620	0.469	0.431
All (14)	0.586	0.602	0.493	0.487

5. Conclusion

We proposed PhoVis, the first method that attempts to quantify audio-lip correlation for *dubbed content* and predict a perceptual score capturing the viewing experience of a dubbed video. PhoVis measures audio-lip correlation at an elementary phoneme and viseme level. We showed that PhoVis is visually and temporally grounded, generalizes well across languages and performs better than the closest lip-sync methods. We also analyzed how PhoVis can be practically used for applications such as helping language experts and lip-synthesis methods in identifying the exact points of desynchronization, allowing them to evaluate and improve viewing experience of dubbed content.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 2, 6

[2] Aashish Agarwal and Torsten Zesch. German end-to-end speech recognition based on deepspeech. In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, pages 111–119. German Society for Computational Linguistics and Language Technology, 2019. 4

[3] Naheed Akhter, Mushtaq Ali, Lal Hussain, Mohsin Shah, Toqeer Mahmood, Amjad Ali, and Ala Al-Fuqaha. Diverse pose lip-reading framework. *Applied Sciences*, 12(19):9532, 2022. 3

[4] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2, 3

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 5

[6] Helen L Bear and Richard Harvey. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, 95:40–67, 2017. 3

[7] Helen L Bear and Richard Harvey. Alternative visual units for an optimized phoneme-based lipreading system. *Applied Sciences*, 9(18):3870, 2019. 3

[8] Otavio Braga and Olivier Siohan. Best of both worlds: Multi-task audio-visual automatic speech recognition and active speaker detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6047–6051. IEEE, 2022. 2, 3

[9] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 4, 5

[10] Luca Cappelletta and Naomi Harte. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)*, pages 322–329. Citeseer, 2012. 3

[11] Brandon Castellano. Pyscenedetect. Technical Report v0.6, http://scenedetect.com/en/latest/, 2022. 3

[12] Sen Chen, Zhilei Liu, Jiaxing Liu, Zhengxiang Yan, and Longbiao Wang. Talking head generation with audio and speech related facial action units. *arXiv preprint arXiv:2110.09951*, 2021. 2

[13] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2, 3, 5, 6

[14] commonvoice-fr Contributors. Common voice stt model. https://github.com/wasertech/commonvoice-fr/releases/tag/v0.8.0-fr-0.3. 4

[15] Coqui. English stt v1.0.0. Technical Report STT-EN-1.0.0, Coqui, https://coqui.ai/models, October 2021. 4

[16] DANBER. Spanish jaco-assistant. https://gitlab.com/Jaco-Assistant/Scribosermo. 4

[17] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022. 2

[18] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 2

[19] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Joint audio-text model for expressive speech-driven 3d facial animation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–15, 2022. 2

[20] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 2

[21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 4

[22] David Harwath and James Glass. Towards visually grounded sub-word speech unit discovery. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3017–3021. IEEE, 2019. 2

[23] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 2

[24] Xie W. Rahtu E. Iashin, V. and A. Zisserman. Sparse in space and time: Audio-visual synchronisation with trainable selectors. In *British Machine Vision Conference (BMVC)*, 2022. 2, 6

[25] Mozilla Italia. Italian stt 2020.8.7. Technical Report STT-IT-2020.8.7, Coqui, https://github.com/coqui-ai/STT-models, April 2021. 4

[26] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10552, 2022. 2

[27] Venkatesh S Kadandale, Juan F Montesinos, and Gloria Haro. Vocalist: An audio-visual synchronisation model for lips and voices. In *Interspeech*, pages 3128–3132, 2022. 2, 6

[28] Khazar Khorrami and Okko Räsänen. Evaluation of audio-visual alignments in visually grounded speech models. *arXiv preprint arXiv:2108.02562*, 2021. 2

[29] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF con-*

ference on computer vision and pattern recognition, pages 2755–2764, 2021. 2, 3

[30] Jiaying Lin, Wenbo Zhou, Honggu Liu, Hang Zhou, Weiming Zhang, and Nenghai Yu. Lip forgery video detection via multi-phoneme selection. 2021. 3

[31] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *European Conference on Computer Vision*, pages 551–569. Springer, 2022. 2

[32] Andrea Britto Mattos, Dario Augusto Borges Oliveira, and Edmilson da Silva Morais. Improving cnn-based viseme recognition using synthetic data. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018. 3

[33] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502, 2017. 4

[34] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. A cappella: Audio-visual singing voice separation. *arXiv preprint arXiv:2104.09946*, 2021. 2

[35] Kayode Olaleye and Herman Kamper. Attention-based keyword localisation in speech using visual grounding. *arXiv preprint arXiv:2106.08859*, 2021. 2

[36] Dário Augusto Borges Oliveira, Andrea Britto Mattos, and Edmilson da Silva Morais. Improving viseme recognition using gan-based frontal view mapping. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2229–22297. IEEE, 2018. 3

[37] Jonah Ong, Ba Tuong Vo, Sven Nordholm, Ba-Ngu Vo, Diluka Moratuwage, and Changbeom Shim. Audio-visual based online multi-source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1219–1234, 2022. 2

[38] Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li. Selective listening by synchronizing speech with lips. volume 30, pages 1650–1664. IEEE, 2022. 2

[39] Puyuan Peng and David Harwath. Fast-slow transformer for visually grounding speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7727–7731. IEEE, 2022. 2

[40] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2, 7

[41] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10493–10502, 2022. 2

[42] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 41–50, 2021. 2, 3

[43] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 2, 3

[44] Julius Richter, Jeanine Liebold, and Timo Gerkamnn. Continuous phoneme recognition based on audio-visual modality fusion. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022. 3

[45] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 2

[46] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Improving on-screen sound separation for open-domain videos with audio-visual self-attention. 2021. 2

[47] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *European Conference on Computer Vision*, pages 368–385. Springer, 2022. 2

[48] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 5

[49] Junwen Xiong, Yu Zhou, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Look\&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *arXiv preprint arXiv:2203.02216*, 2022. 2, 3

[50] Eric Zhongcong Xu, Zeyang Song, Chao Feng, Mang Ye, and Mike Zheng Shou. Ava-avd: Audio-visual speaker diarization in the wild. *arXiv preprint arXiv:2111.14448*, 2021. 2

[51] Chih-Chun Yang, Wan-Cyuan Fan, Cheng-Fu Yang, and Yu-Chiang Frank Wang. Cross-modal mutual learning for audio-visual speech recognition and manipulation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada*, volume 22, 2022. 2

[52] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 2

[53] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[54] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 3

[55] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 3

[56] Yuanhang Zhang, Susan Liang, Shuang Yang, Xiao Liu, Zhongqin Wu, Shiguang Shan, and Xilin Chen. Unicon: Unified context network for robust active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3964–3972, 2021. 2

[57] Dongzhan Zhou, Xinchi Zhou, Di Hu, Hang Zhou, Lei Bai, Ziwei Liu, and Wanli Ouyang. Sepfusion: Finding optimal fusion structures for visual sound separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3544–3552, 2022. 2

[58] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 2

[59] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 2

[60] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018. 2, 3