Perceptual Synchronization Scoring of Dubbed Content using Phoneme-Viseme Agreement
— Supplementary —

Honey Gupta
Amazon Prime Video
ghoney@amazon.com

4. Experiments

4.3. Ablation studies and model behavior

4.3.7 Key-points importance and aggregation function

In all the experiments presented in the main paper, we used 48 through 68 lip landmarks for viseme representation. In this experiment, we use a subset of these landmarks for building the reference and computing correlation. In Table 1, landmarks corresponding to (1) inner lip-region are 60-68, (2) outer lip-region are 48-59 and alternate are the alternate landmarks between 48-68. We observe that considering just inner landmarks performs the lowest, mostly due to indistinguishable variation between viseme representations. Performance of remaining sets is similar, suggesting that outer lip-region could be the most relevant factor as it is common across them. Due to similar results and insignificant computational overhead, we considered all lip landmarks in PhoVis. In Sec.3.6 of the main paper, we mentioned that there are multiple ways to aggregate key-points across frames to derive the reference. Table 2 shows the results when different aggregation functions are used. We observe that each function has its trade-off between F1-score and error metrics. Moreover, we observed that using an averaging technique reduced inter-viseme distinction. Therefore, between minimum and maximum, max yields slightly better error F1. Hence, was used max in PhoVis.

Table 1. Lip landmark importance in reference R_{PV}

Lip-keypoints set in Reference	Precision (weighted)	F1 Score (weighted)	Error Precision	Error F1-Score
Inner	0.678	0.560	0.380	0.368
Outer	0.592	0.595	0.324	0.487
Alternate	0.627	0.572	0.326	0.490
All	0.710	0.606	0.314	0.462

Table 2. Key-point aggregation function and performance

Keypoint aggregator	Precision (weighted)	F1-Score (weighted)	Error Precision	Error F1-Score
Mean	0.613	0.618	0.380	0.368
Median	0.610	0.615	0.319	0.484
Min	0.645	0.650	0.381	0.347
Max	0.710	0.606	0.314	0.462

4.3.8 ML model for perceptual scoring.

We experimented with three ML models - MLP, Random Forest Classifier and Support Vector Classifier (SVC) for each approach in the task above. Model-search details can be found in the supplementary. The results for Spanish are shown in Table 3. RF performs better for PhoVis and Baseline, while MLP is better for SyncNet and VocaLiST.

Table 3. Precision of ML models for Binary scoring (Spanish)

Model	PhoVis	E2E	SyncNet	VocaLiST
MLP	0.635	0.521	0.551	0.565
RF	0.710	0.568	0.541	0.500
SVC	0.572	0.547	0.563	0.501

Discussion and future work

The proposed PhoVis model has the potential to act as a fundamental method over which multiple solutions related to various AV problems can be built. However, PhoVis itself can be improved by further tuning different components of the method.

To validate the efficacy of our approach, we used basic face and landmark detection models in the paper. To reduce the impact of any poor performance displayed by either of these models on our predictions and remove edge-cases of extreme face poses, lighting conditions, etc., we included a filtering step in our viseme extraction pipeline. This step scans for any frames that have spurious landmarks (Section 3.4 and Figure 4 of main paper). Sample erratic frames fil-

tered out by our pipeline are shown in Figure 1. We explicitly tackle the extreme pose cases, which other embedding methods would blindly process and give erratic results. This suggests that the results presented in the paper form a base benchmark with the simplest face or landmark detection models and thus can be tuned further.

Figure 1. *Filtered cases of LM model*

Since, there is a temporal factor associated with visemes, as well as landmarks, a basic extension could be to consider temporal consistency while generating the reference viseme. However, the typical duration of a phoneme is 100-250 ms. At 25 fps, a phoneme will have 2–6 frames. Considering an error margin of $\approx$ 50ms [1] in start/end times, the no. of *confident* frames becomes quite less to provide temporal information. Fig.2 shows a sample phoneme /T/ where the last frame seems unrelated to the viseme /t/. It would be interesting to test quantitatively, but we do not expect a significant boost.

Figure 2. *Phoneme: /T/, duration:150ms.*

Another experiment could be to compute distribution of the the key-points for each reference viseme instead of using aggregated 2d landmarks. This could be done by building models like Gaussian mixture model that approximate viseme distribution, which can then be compared against current frame's visemes for computing the sync distance.

A line of potential future work could be to use PhoVis as a base and build solutions for different applications that can be targeted by this technology. PhoVis computes audio to lip distance at the elementary phoneme and viseme level. The extracted phoneme-viseme correspondence or the reference dictionary can be seamlessly used for active speaker detection, AV lead/lag detection and many other AV problems that involve lip-movement to audio matching.

Improving dubbing quality scoring. PhoVis distance does not examine the image quality or synthesis artifacts while generating the perceptual score, as it is designed for audio-lip sync evaluation. This could be implicitly captured by the behavior of landmark detection model, but this does not give a direct feedback. Thus, a future extension could be to merge the PhoVis score with an image quality score for better benchmarking of lip-synthesis methods. It can also be clubbed with audio perceptual quality while accessing the dubbing quality. Incorporating audio as a modality could also help expand our method to better accommodate tonal information, which could help expand the scoring to tonal languages. To sum up, the experience of watching a dubbed video is a variable that is dependent on the different quality aspects of multiple modalities, each of which could be incorporated in future with the audio-lip synchronization score to predict a holistic dubbing score.

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

3.4. Phoneme-viseme mapping

To perform viseme comparison across languages, we find the set of visemes that are common across the 6 P1 languages (English, French, Italian, German, Spanish and Portuguese) we considered. These visemes are $V^* = \{/f/, /i/, /k/, /p/, /s/, /t/\}$. Therefore, we filter out phonemes corresponding to the above 6 visemes and use only these phonemes/visemes for correlation measurement. Below is the mapping M of IPA Phonemes to 6 common viseme for P1 languages that we have utilized in our method. The filtering functions ϕ_{filt_v} and ϕ_{filt_p} give the extracted set of phonemes P and the corresponding visemes V, respectively.

```
fr-FR: {b: p, d: t, f: f, g: k, j: i, k: k, l: t, m: p,
        n: t, n: k, p: p, s: s, t: t, v: f, z: s, i: i},
es-ES: {b: p, d: t, f: f, g: k, j: i, k: k, l: t, m: p,
        n: t, n: k, p: p, s: s, t: t, x: k, z: s, i: i},
de-DE: {b: p, d: t, ç: k, f: f, g: k, h: k, j: i, k: k,
        lm: p, n: t, n: k, p: p, s: s, v: f, x: k, z: s},
pt-BR: {t: t, b: p, d: t, f: f, g: k, j: i, k: k, l: t,
        m: p, n: t, p: p, s: s, t: t, v: f, z: s, i: i},
it-IT: {b: p, d: t, dz: s, f: f, g: k, h: k, j: i, k: k,
        l: t, m: p, p: p, s: s, t: t, ts: s, v: f, i: i},
en-US: {b: p, d: t, f: f, g: k, h: k, j: i, k: k, l: t,
        m: p, n: t, n: k, p: p, s: s, t: t, v: f, i: i}
```

3.5. Building the phoneme-viseme reference

For each distinct viseme present in the clip, we build a reference viseme which is represented by a set of lip key-points. These key-points are 2d coordinates representing a certain point of the lip region. For a given clip, there are multiple frames that correspond to the same viseme. Therefore, to build a reference set of key-points per viseme for the clip, we again use an aggregation logic across the viseme samples. The obtained set of key-points $[L_1, ..L_k, L_{k+1}, ..]$ have a 2d coordinate for each lip-landmark. This set becomes the reference for a particular viseme and forms the reference dictionary R_{PV}. The maximum size of R_{PV} is 6, each representation corresponding to a viseme in V^*.

Reference visemes for few sample clips are shown in Figure 3. In the figure, each color represents the landmarks for one viseme. We observe that not all clips have 6 reference visemes. This is because the number of reference visemes depends on the phonemes present in the original

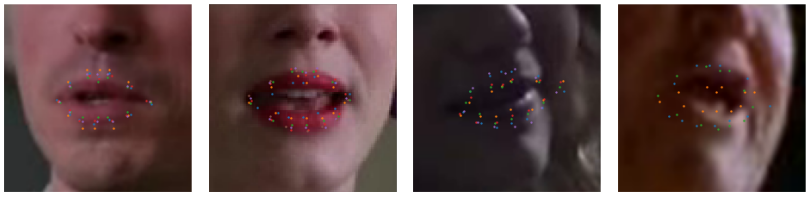

Figure 3. Reference visemes for sample clips. Each color represents the landmarks for one viseme. Note that not all the clips have 6 reference visemes. The number of reference visemes depends on the phonemes in the original dialogue and landmark detection accuracy.

dialogue and landmark detection accuracy. We also observe that the distinction between visemes in each clip is different. For *e.g.*, clip 1 has almost similar landmarks for different visemes, whereas clip 4 has only 3 visemes in its reference dictionary but there is significant distinction between them. This could be due to various factors such as phonemes in the dialogue, actor's speaking style, dialogue delivery, and so on. This variation is captured by our method at a dialogue level, making it robust to these factors. A generic viseme representation would either fail in these cases or would need adaptation/recalibration.

3.5.1 Distance metric for score calculation

We experimented with 4 distance metrics for PhoVis correlation measurement between the expected viseme L and the current frame's viseme L'. Below are the equations for the distance metrics.

Area normalized L2 distance

$$\mathbf{d}_{L2} = \frac{1}{N}\sum_{j=1}^{N}\frac{||L_k^j - L_k'^j||_2}{r} \tag{1}$$

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.

Cosine distance

$$\mathbf{d}_{cos} = 1 - \frac{u.v}{||u||_2||v||_2} \tag{2}$$

where u contains linearly stacked key-points from L'_k from the current frame's viseme, and v contains linearly stacked key-points from L_k, from the reference viseme. Linear stacking is done to compute distance over 2d key-points.

Chebyshev distance

$$\mathbf{d}_{cheby} = \max_i |u_i - v_i| \tag{3}$$

where u contains linearly stacked key-points from L'_k from the current frame's viseme, and v contains linearly stacked key-points from L_k, from the reference viseme.

Correlation

$$\mathbf{d}_{corr} = 1 - \frac{(u-\bar{u}).(v-\bar{v})}{||u-\bar{u}||_2||v-\bar{v}||_2} \tag{4}$$

where $\bar{u}$ is the mean of the elements of u and $x.y$ is the dot product of matrices x and y.

A. Appendix

A.1. Dataset statistics

Language-wise statistical properties for data analysis of our annotated perceptual dataset can be found in Table 5. We annotated ≈8200 dubbed clips in 6 P1 languages, where the original was in English and the dubbed versions were in French, Italian, German, Spanish and Portuguese. Each clip was annotated by 3 dubbing experts who were familiar with the language. Annotation values were from 1-5. (-1) was the expected label in case the annotator was not able to label the clip, mostly due to dark frames, lip-occlusion, etc. The table shows the statistics of each expert annotator - in the fields Label 1, Label 2 and Label 3, respectively. Columns 4, 5 and 6 represent the median, mean and minimum aggregation techniques to generate a label for each clip. For *e.g.* column 5 shows the properties of *mean* aggregation, which therefore corresponds to Mean Opinion Score (MOS).

For French, Italian and Spanish, we see that 50% of clips have 4 or 5 as label. In German and Portuguese, we see that 50% label is 3. Hence, for binary scoring, we chose 0-3 as *Bad* dubs and 4-5 as *Good* dub. For experiments, we divided our annotated dataset in a train-test split of 70-30 ratio. Table 4 shows the number of samples per language in our train and test sets, that were used for all the binary perceptual scoring experiments mentioned in the paper.

A.2. Perceptual scoring model selection details

From the following parameter search range, we randomly sampled 200 configurations for each model and selected the model based on 5-fold cross-validation.

A.2.1 Search parameter range

Multi-layer Perceptron (MLP)

Table 4. PV (annotated) dataset distribution for binary perceptual scoring

Language	Train			Test			Total
	Good	Bad	Total	Good	Bad	Total	
Spanish	404	777	1181	173	334	507	1688
French	764	259	1023	328	111	439	1462
Italian	760	430	1190	325	185	510	1700
German	436	730	1166	187	313	500	1666
Portuguese	416	788	1204	179	338	517	1721

- Hidden layer sizes = $10 - 500$, with a step of 5 + (x,x) , x $\in [10, 200]$ at a step of 5
- Activation = $[relu, tanh, identity]$
- Learning rate = $[0.0001, 0.001, 0.005, 0.0005, 0.00001]$
- Optimizer = Adam
- Loss = Binary cross-entropy

Random Forest (RF)

- No. of estimators = $10 - 1000$, with a step of 10
- Max features = $[auto, sqrt]$
- Max depth = $10 - 50$, with a step of 5 + $[None, 1, 2, 5]$
- Min samples split = $[2, 5, 10, 20]$
- Min samples leaf = $[1, 2, 4]$
- bootstrap = [True, False]
- Optimizer = Adam
- Loss = Binary cross-entropy

Support Vector Classifier (SVC)

- Regularization =$10 - 500$, with a step of 10
- Kernel = $[linear, poly, rbf, sigmoid]$
- Gamma = $[scale, auto]$
- Optimizer = Adam
- Loss = Binary cross-entropy

References

[1] 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. 2

Table 5. Language-wise statistics of PV Annotated dataset used for all the experiments in the paper. Label 1, 2 and 3 are the annotations per annotator. We have 3 annotators per clip. Annotation values were from 1-5. -1 was the label in case the annotator was not able to label the clip, mostly due to dark frames, lip-occlusion, etc. For French, Italian and Spanish, we see that 50% of clips have 4 or 5 as label. In German and Portuguese, we see that 50% label is 3. Hence, for binary scoring, we chose 0-3 as *Bad* dubs and 4-5 as *Good* dub.

		Label 1	**Label 2**	**Label 3**	**Median**	**Mean**	**Minimum**	**Variance**
German	Count	1674	1675	1677	1672	1672	1672	1672
	Mean	3.2736	3.32358	3.18247	3.4079	3.41228	2.76854	0.45624
	Std. Deviation	1.40376	1.3424	1.56715	0.9998	0.92172	0.92748	0.45408
	Min	-1	-1	-1	1	1	1	0
	25%	3	3	3	3	3	2	0.22222
	50%	3	4	3	4	3.66667	3	0.22222
	75%	4	4	4	4	4	3	0.66667
	Max	5	5	5	5	5	5	3.55556
French	Count	1466	1462	1467	1458	1458	1458	1458
	Mean	3.82742	3.6498	3.7621	3.92661	3.90855	3.31824	0.37411
	Std. Deviation	1.06618	0.94355	1.51346	0.66954	0.5458	0.71379	0.36667
	Min	-1	-1	-1	1.5	1.5	1	0
	25%	3	3	4	4	3.66667	3	0.22222
	50%	4	4	4	4	4	3	0.22222
	75%	5	4	5	4	4.33333	4	0.66667
	Max	5	5	5	5	5	5	2.88889
Italian	Count	1705	1706	1704	1699	1699	1699	1699
	Mean	3.75249	3.70281	3.38087	3.69423	3.69521	3	0.49004
	Std. Deviation	1.16521	1.05127	1.12408	0.82596	0.63673	0.78637	0.4221
	Min	-1	-1	-1	1	1.33333	1	0
	25%	3	3	3	3	3.33333	3	0.22222
	50%	4	4	4	4	3.66667	3	0.22222
	75%	5	4	4	4	4	4	0.66667
	Max	5	5	5	5	5	5	2.88889
Spanish	Count	1689	1692	1694	1687	1687	1687	1687
	Mean	3.68384	3.21809	3.22019	3.47985	3.42946	2.73266	0.46537
	Std. Deviation	1.06628	0.98975	1.35249	0.8856	0.81381	1.01835	0.42166
	Min	-1	-1	-1	1	1	1	0
	25%	3	3	2	3	3	2	0.22222
	50%	4	3	3	4	3.66667	3	0.22222
	75%	4	4	4	4	4	3	0.66667
	Max	5	5	5	5	5	5	2.88889
Portuguese	Count	1733	1730	1733	1728	1728	1728	1728
	Mean	3.19042	3.26994	2.92037	3.19647	3.17332	2.2934	0.69655
	Std. Deviation	1.09195	1.12928	1.23016	0.8484	0.683	0.88356	0.52821
	Min	-1	-1	-1	1	1.33333	1	0
	25%	3	3	2	3	2.66667	2	0.22222
	50%	3	3	3	3	3.33333	2	0.66667
	75%	4	4	4	4	3.66667	3	0.88889
	Max	5	5	5	5	5	5	3.55556