A THE DETAILS OF THE BOUNDED DURATION PREDICTOR.

A.1 ALGORITHMIC DETAILS

[Revised Part 6] In Algorithm 1, we present the details of the bounded-duration-predictor algorithm:

1. After normalization, the duration $\mathbf{D}$ can be computed as the allocated duration $\mathbf{D}'$ corresponding to each token based on the target length $\mathbf{T}$:
$$\mathbf{D}' = \texttt{Normalize}(\mathbf{D}) \times \mathbf{T}. \tag{5}$$
2. Following the rounding method, it is converted to an integer predicted duration $\mathbf{PRED}$:
$$\mathbf{PRED} = \texttt{Clamp}(\texttt{Round}(\mathbf{D}'), min = 1). \tag{6}$$
3. Calculate the difference $\mathbf{DIFF}$ between the predicted duration $\mathbf{PRED}$ and the allocated duration $\mathbf{D}'$ for each token:
$$\mathbf{DIFF} = \mathbf{D}' - \mathbf{PRED}. \tag{7}$$
4. Determine whether the predicted duration $\mathbf{PRED}$ still needs adjustment in the number of frames. If an increase is required, select the highest-weighted difference $\mathbf{DIFF}$ from the sequence for its duration+1. Conversely, if a decrease is needed, select the lowest-weighted difference $\mathbf{DIFF}$ for its duration-1.

Input :
T: The desired length of the translation result.
D: The predicted duration sequence for each unit.
Output:
OUT: The duration of each discrete cell after length regulation.

```
//Step1:
D' = Normalize(D) ×T;
//Step2:
PRED = Clamp(Round(D'),min = 1) ;
//Step3:
DIFF = D'-PRED ;
//Step4:
ADD=Zeroes() ;
if Sum(PRED) > T then
    INDEX=TopK(−DIFF,k =Sum(PRED) −T);
    ADD[INDEX]= −1 ;
else
    INDEX =TopK(DIFF,k = T−Sum(PRED));
    ADD[INDEX]= 1 ;
end
OUT=PRED+ADD
```

Algorithm 1: Pseudo-code for bounded-duration-predictor implementation details.

A.2 IMPLEMENTATION SAMPLE

[Revised Part 7] Let's revisit the previous example in section 4.2 for illustration, when $U = \{u1, u2, u3, u4\}$, $D' = \{2.2, 1.8, 2.3, 2.7\}$ and $T = 10$:

- `Step1`: $\mathbf{D}' = [2.2, 1.8, 2.3, 2.7], \mathbf{T} = 10$.
- `Step2`: $\mathbf{PRED} = [2, 2, 2, 3]$.
- `Step3`: $\mathbf{DIFF} = [0.2, -0.2, 0.3, -0.3]$.
- `Step4`: Since $\texttt{SUM}(\mathbf{PRED}) = 9 < 10$, for the largest $\mathbf{T} - \texttt{SUM}(\mathbf{PRED}) = 1$ corresponding token in $\mathbf{DIFF}$, its duration+1, resulting in $\mathbf{OUT} = [2, 2, 2+1, 3] = [2, 2, 3, 3]$.

The sequence of discrete units can be represented as U'=$\{u_1, u_1, u_2, u_2, u_3, u_3, u_3, u_4, u_4, u_4\}$.

Method	Translation	Image	SYNC	Overall	Mean
ST+TTS+Wav2Lip	3.78±0.05	4.03±0.08	3.66±0.04	3.57±0.03	3.76±0.05
ASR+NMT+TTS+Wav2Lip	**4.23±0.06**	4.11±0.07	4.12±0.08	4.18±0.11	4.16±0.08
Translatotron2+Wav2Lip	2.79±0.09	3.98±0.12	3.68±0.07	2.91±0.20	3.34±0.12
S2ST+Wav2Lip	4.03±0.08	4.05±0.04	4.02±0.09	3.86±0.07	3.99±0.07
TransFace(ours)	4.19±0.08	4.08±0.07	**4.28±0.04**	3.93±0.05	4.12±0.06
TransFace(ours)+bounded	4.17±0.06	**4.16±0.06**	**4.28±0.05**	**4.39±0.11**	**4.25±0.07**

Table 5: The detailed MOS (Mean Opinion Score) results for talking head translation. Each dimension is scored individually on a scale of 1 (lowest) to 5 (highest). **Translation**: translation quality, **Image**: image quality, **SYNC**: Synchronization, **Overall**: overall sensation.

Method	Image quality	Synchronization	Mean
Audio(GT)+Wav2lip	4.18±0.32	4.12±0.18	4.15±0.25
Video(GT)	**4.33±0.13**	**4.13±0.11**	**4.23±0.12**
U2S+LipGAN	2.89±0.25	2.39±0.21	2.64±0.23
U2S+Wav2Lip	**4.01±0.20**	3.93±0.24	3.92±0.22
Unit2Lip(ours)	3.95±0.24	**4.01±0.24**	**3.98±0.24**

Table 6: The detailed MOS (Mean Opinion Score) results for unit-based talking head generation. Each dimension is scored individually on a scale of 1 (lowest) to 5 (highest).

B More Implementation Details

B.1 Data Preprocessing.

For visual speech, we extract the facial region from the video for Unit2Lip model training. As in prior research (Prajwal et al., 2020; Shi et al., 2022), we use dlib (King, 2009) to detect 68 facial keypoints, and then isolate a 96x96 region-of-interest (ROI) video segment centered around the face. For the source language audio speech, we extract 80-dimensional mel-filterbank features at 20-ms intervals as input. Regarding the target language audio speech, we apply the k-means algorithm to cluster the representations provided by the well-tuned mhubert into 1000 discrete units for training purposes.

B.2 Model configuration and Training Details.

We adopted the same S2UT model architecture as (Lee et al., 2021), employing 8 attention heads and a embedding size of 512. We select one discrete unit every 20ms to synthesize one audio frame and every 40ms to synthesize one visual frame. The audio-speech vocoder utilizes the unit-based vocoder pre-trained in (Lee et al., 2021), whereas the decoder of the visual-speech synthesizer employs the same architecture as Wav2lip (Prajwal et al., 2020). In the inference process, since the audio speech in language X lacks a corresponding visual speech as reference frames, we utilize the visual speech of the English videos as reference frames for synthesizing the translated talking head. All the cascade models we utilized are publicly available pre-trained systems in Fairseq (Ott et al., 2019). For instance, we employed MMT to convert text from other languages to English, and the FastSpeech2 model to transform text into corresponding audio speech.

B.3 The detailed evaluation process of MOS

[Revised Part 8] Our comprehensive MOS scoring process for talking head translation tasks involves gathering scores across four dimensions: translation quality, image quality, synchronization, and overall sensation. For the unit-based talking head generation, we streamline the evaluation to two dimensions: image quality and synchronization. Translation quality assesses the consistency of the translated content with the original sentence, while image quality evaluates the presence of artifacts in the generated image. Synchronization measures the coherence of audio and visual speech,

and overall sensation indicates the evaluation of the video's authenticity. Each sample is randomly scrambled and presented to 15 participants for scoring. A composite MOS is then calculated by averaging the scores for the corresponding dimensions, with each dimension scored individually on a scale of 1 (lowest) to 5 (highest). Here, we present the MOS (Mean Opinion Score) results for both talking head translation in Table 5 and unit-based talking head generation in Table 6.

C More Experiments

C.1 Samples of Unit-Based Talking Head Generation

We present samples of our unit-based talking head generation method (Unit2Lip). For each discrete unit, we equidistantly selected 6 pairs of corresponding original and generated Talking Head video frames. Notably, the lip shapes between each pair of images are highly consistent, suggesting that the model is adept at reconstructing the lip shapes associated with discrete units while preserving more of the original video information. Furthermore, we conduct experiments in French and the lip shape consistency is also well-maintained, demonstrating that the Unit2Lip can be generalized across languages, not only to English but also to different languages.

(a) Resynthesis sample for English. The top row is the original, the bottom row is the synthesized one.

(b) Resynthesis sample for French. The top row is the original, the bottom row is the synthesized one.

Figure 3: Quantitative comparison of audio and visual speech resynthesized from discrete units on English and French. More experimental results are available on the demo page.

C.2 Can the talking head represent the corresponding content?

In this paper, due to the relatively lower discriminability of visual speech compared to audio speech, lip reading (visual speech recognition) yields significantly lower recognition accuracy. Therefore, most experiments in this paper employ audio-visual speech recognition (AVSR) results for computing BLEU scores. In this subsection, we conduct additional ablation experiments focused on visual speech to investigate its effectiveness in representing the corresponding content. In Table 7, we show the BLEU comparisons for different modalities of speech in different methods. **(1) The talking head can effectively convey the corresponding content.** Comparing with random frames

Table 7: BLEU scores of different methods with distinct modality speech on Es-En and Fr-En of LRS3-T. *Random Frame*: Randomly scrambled video that does not express any information.

	Method	Modality	Es-En	Fr-En
1	Audio(GT)+wav2lip	Visual-Only	13.23	13.23
2	Audio(GT)+wav2lip	Audio-Visual	87.73	87.73
3	Random frame	Visual-Only	0.18	0.68
4	S2ST+wav2lip	Visual-Only	8.59	**7.56**
5	TransFace	Visual-Only	**8.62**	7.53
6	TransFace	Audio-Only	60.76	46.89
7	S2ST+Wav2Lip	Audio-Visual	60.93	45.17
8	TransFace	Audio-Visual	**61.93**	**47.55**

results (#3), the BLEU of the translated video (#5) is improved by 8.46, which proves that it contains some useful information and is not directly random synthesis. Meanwhile, compared with the translated video synthesized by wav2lip (#4), the results of TransFace and S2ST+Wav2Lip are basically the same (8.59 vs. 8.62), which indicates that the synthesized result of our method is basically consistent with that of wav2lip. **(2) The talking head can provide complementary information to audio speech.** Comparing the TransFace results for audio-only (#6) and audio-visual (#8), we can observe a further improvement in the BLEU of translation (from 60.76 to 61.93) with the additional introduction of talking head. This indicates that the talking head, synthesized directly from the discrete units, contains valuable supplementary information to audio speech, further enriching the speech content.

Method	En		Es	
	NMI(↑)	Purity(↑)	NMI(↑)	Purity(↑)
AV-HuBert	**43.7**	**65.8**	12.8	9.1
m-HuBert	42.6	65.1	**41.7**	**63.2**

Table 8: Comparison of clustering effects of m-HuBert and AV-HuBert on En and Es.

C.3 Why m-HuBert instead of AV-HuBert?

[Revised Part 9] Among various discrete unit schemes (encodec (Défossez et al., 2022), av-hubert(Shi et al., 2022), hubert(Hsu et al., 2021), mhubert(Lee et al., 2021), etc.), only mhubert is publicly available and widely used as a multilingual discrete unit extractor. Hence, we chose mhubert to extract discrete unit representations. As demonstrated in other research (Le et al., 2023), if we want the model to encode languages other than English, it must be pre-trained on the speech of these languages.

Here, we also present a comparison of the clustering effect (Shi et al., 2022) of avhubert and mhubert on different languages in Table 8. Notably, there is no difference in performance between the two in English, but their effectiveness varies significantly in other languages. Furthermore, in this paper, we showcase that discrete units based on acoustic-only feature can also be effectively utilized for visual speech synthesis.

C.4 More translation results

[Revised Part 10] In addition to the result of Es-En in Table 4, we further show the translation results of Fr-En in Table 9 to visualize the translation performance of our model on different language pairs. The results indicate that our approach consistently delivers high-quality translation performance across various languages, including Es-En and Fr-En. Additional translation results for other language pairs (En-Es and En-Fr) are available on the demo page.

Table 9: Comparison of translation quality on Fr-En among different methods. ~~Red Strikeout Words~~: mistranslated words with opposite meaning, Blue Words: mistranslated words with similar meaning, Gray Words: the absent words.

Fr-En	Source(Fr)	No fuimos considerados la cosa real.
	Target(En)	we weren't considered the real thing.
	ASR+NMT+TTS+wav2lip	we were not weren't considered the real thing.
	ST+TTS+wav2lip	we were not weren't considered the real ~~ones~~ thing.
	S2ST+Wav2Lip	we were not weren't considered ~~as~~ the real thing.
	TransFace	we were not weren't considered ~~as~~ the real thing.
	TransFace+bounded	we were not weren't considered ~~as~~ the real thing.

D LIMITATIONS AND ETHICAL DISCUSSIONS.

D.1 ISOMETRIC TRANSLATION IS A FUNDAMENTAL REQUIREMENT FOR TALKING HEAD TRANSLATION.

[Revised Part 11] In contrast to speech translation, talking head translation encounters a relatively fixed limit on video frame length, especially evident when dubbing a translated movie. In such instances, the voice actor for translated movies must synchronize the translation to match the original video's duration. The absence of a duration-bounded module in the video results in noticeable frame skipping, leading to a significant loss of realism (refer to the demo page for results without the bounded-duration predictor). This limitation renders the approach unsuitable for professional scenarios like online meetings and movie translating, where the number of generated video frames must align with the original reference video. **The introduction of the `Bounded-Duration-Predictor` becomes imperative in such cases, despite tradeoffs in other factors, as it effectively satisfies the fundamental requirements of talking head translation.** We also acknowledge this approach may cause excessive speedups and slow reads, we plan to address this concern in our future work. Specifically, we intend to investigate the vocabulary length of the generated content to further enhance the realism and authenticity of the translated videos.

D.2 WHY ONLY COMPARE BLEU IN X-EN?

In this work, we only compare BLEU scores for X-En translation results. This is due to the scarcity of current audio/video speech datasets in languages other than English, as well as the notably poorer performance in audio/video speech recognition for languages other than English. These factors contribute to a lack of convincing and credible results in those cases. Nonetheless, we still showcase the relevant translation results on the demo webpage, which you are welcome to review.

D.3 MORE DIFFICULT DATASETS WILL BE ATTEMPTED IN THE FUTURE.

The average length of audio-visual speech data is considerably shorter than that of audio-only speech data, potentially making it easier to train. As part of our ongoing efforts, we will develop longer and more complex audio-visual speech translation datasets, aiming to enhance the robustness of Talking Head Translation.