AV-GRATS: A Parallel Multi-Speed Mandarin Audio-Visual Corpus for Speaking-Rate Robustness Analysis

Dyah A. M. G. Wisnu*, Riyad Ali Ahmad Abuirbaiha†, Stefano Rini†, Hsin-Min Wang‡, and Yu Tsao*
* Research Center for Information Technology, Academia Sinica, Taipei, Taiwan
† Institute of Communications Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan

***Abstract*—This paper introduces AV-GRATS, a naturally produced parallel Mandarin audio-visual corpus for evaluating audio-visual speech recognition (AVSR) under speaking-rate variation. Existing Mandarin AVSR benchmarks do not provide parallel recordings of the same utterances across multiple natural speaking rates, limiting systematic analysis. AV-GRATS currently contains 7,500 utterances from 25 native Taiwanese Mandarin speakers, recorded at five speaking rates from 0.5× to 1.5×. Unlike synthetic time-scale modification, AV-GRATS preserves natural acoustic and articulatory changes across speaking rates. We report benchmark results using AV-HuBERT fine-tuned on Chinese-LIPS. Under a dual-independence protocol with both speaker- and sentence-level separation, zero-shot evaluation achieves 31.2% syllable error rate (SER), while direct fine-tuning on AV-GRATS results in 87.3% SER, indicating transcript memorization and poor generalization. We further analyze Mandarin AVSR errors using tone-aware SER, visual-group SER, and frequency-weighted label-space deflation analysis. Results show a corrected 9.76% tone-related gap and reveal prominent retroflex–sibilant confusion patterns under cross-domain evaluation.**

***Index Terms*—Audio-visual speech recognition; Mandarin; time-scale modification; speaking rate variation.**

I. Introduction

Audio-visual speech recognition (AVSR) has progressed from classical audio-visual integration methods [1] to modern representation-learning models such as AV-HuBERT [2] and Auto-AVSR [3].

However, speaking-rate variation can affect recognition performance in both acoustic and visual speech modalities [4], yet existing AVSR benchmarks largely lack controlled, naturally produced multi-rate recordings. Speaking-rate variability has two complementary roles. First, it occurs naturally: speakers differ in habitual speaking rate, and the same speaker may speak faster or slower depending on speaking style and characteristics [5]. Such variation changes the duration, coarticulation, and synchronization patterns of both the acoustic and visual speech streams, making it a direct robustness challenge for AVSR systems [6]. Second, speaking rate is also an explicit control variable in practical speech technologies. Media playback systems often accelerate speech, while accessibility-oriented systems may slow speech for hearing-impaired users without altering loudness to improve intelligibility. Synthetic speed perturbation [7] does not capture natural acoustic and articulatory changes. This is perhaps more so for Mandarin, a tonal language with distinct articulatory patterns and for which it is still unclear which phonetic units are visually distinguishable in cross-domain settings.

To address these challenges, in this work, we introduce AV-GRATS[1] , a naturally produced parallel Mandarin audio-visual corpus spanning five speaking rates. AV-GRATS builds on the GRATS corpus introduced for natural multi-speed Mandarin speech time-scale modification benchmarking [8], extending it to synchronized audio-visual recordings for AVSR analysis under speaking-rate variation. The current corpus contains 7,500 utterances from 25 native Taiwanese Mandarin speakers, recorded with synchronized audio and video across speaking rates from 0.5× to 1.5×. We further adopt a dual-independence evaluation protocol for limited-sentence AVSR benchmarks, where both speakers and sentences in the evaluation set are excluded from training. Using zero-shot cross-domain evaluation with AV-HuBERT [2] fine-tuned on Chinese-LIPS [9], we perform a preliminary analysis of Mandarin AVSR under speaking-rate variation. We also define nine articulatory visual groups and introduce a three-level syllable error rate (SER) decomposition framework with vocabulary-deflation correction to separately analyze tone-related and visual-group effects.

Our main contributions are as follows:

- **AV-GRATS:** We introduce a naturally produced multi-speed Mandarin audio-visual corpus with 7,500 utterances from 25 speakers across five speaking rates.
- **Dual-independence evaluation:** We adopt speaker- and sentence-independent evaluation and show that direct fine-tuning on limited prompts leads to poor generalization.
- **Mandarin AVSR analysis:** We provide zero-shot AV-HuBERT benchmark results across speaking rates and analyze errors using tone-aware and visual-group-aware SER.
- **Three-level deflation-corrected analysis:** We analyze Mandarin AVSR errors from finer to coarser representations—with-tone SER, no-tone SER, and visual-group SER—and apply frequency-weighted deflation correction. The corrected tone-related gap remains large at 9.76%, while AV

zero-shot visual-group analysis reveals Mandarin articulatory confusion patterns.

II. Related Work

A. Audio-Visual Speech Recognition

Audio-visual speech recognition (AVSR) improves recognition robustness by combining acoustic and visual speech information. AV-HuBERT [2] introduced self-supervised audio-visual pre-training through iterative masked prediction and clustering, achieving strong performance on LRS3 [10] and VoxCeleb2 [11]. Sequence-to-sequence fine-tuning further extended this framework to end-to-end transcription [2], while Auto-AVSR [3] improved recognition robustness using hybrid CTC/attention decoding. Recent studies have examined robustness to corrupted audio-visual inputs, including reliability-aware fusion [12] and corrupted representation learning [13]. However, these works mainly focus on noise, occlusion, or modality corruption, rather than naturally produced speaking-rate variation.

B. Mandarin Audio-Visual Datasets

Large-scale visual speech datasets have supported audio-visual and visual-only speech recognition across multiple languages, including Mandarin [14]. For Chinese speech, the CNVSRC challenge series [15] has promoted VSR benchmarking, while Chinese-LIPS [9] provides a large-scale Mandarin AVSR dataset. However, existing Mandarin audio-visual resources do not provide naturally produced parallel recordings of the same utterances across multiple speaking rates, making it difficult to isolate speaking-rate effects. In addition, limited-prompt benchmarks may be vulnerable to transcript memorization when sentence-level separation is not enforced.

C. Lip Reading and Visual Confusability

Visual confusability has long been studied through viseme modeling, where visually similar phonemes are grouped into shared categories. Prior work has investigated visual feature extraction and phoneme-to-viseme mapping strategies [16]–[19]. In contrast, Mandarin visual speech remains less systematically characterized in neural AVSR settings. Previous Mandarin-related studies have mainly examined auditory and consonant perception [20]. This work analyzes Mandarin visual-group confusion using a neural AVSR model and applies frequency-weighted deflation correction to account for label-space reduction.

III. Datasets

A. Chinese-LIPS

Chinese-LIPS [9] is a large-scale Mandarin audio-visual speech recognition dataset. It contains 28,355 training clips from 207 speakers, covering approximately 100 hours of educational and lecture-style speech. The transcripts are provided in Chinese characters and converted to Pinyin with tone numbers, such as *wo3 men2*, for model training.

[1]The AV-GRATS dataset is available at https://grats.tw/

In this work, Chinese-LIPS is used as the source-domain training corpus for AV-HuBERT fine-tuning. The dataset contains a sentence-independent split, with no transcript overlap between the training, validation, and test sets. This makes it suitable for training a Mandarin AVSR model with stronger sentence-level generalization before evaluating it on AV-GRATS.

B. AV-GRATS

We introduce AV-GRATS (Granular Rate Adaptation for Temporal Scaling – Audio-Visual), a naturally produced parallel Taiwanese Mandarin audio-visual corpus designed to evaluate AVSR robustness under speaking-rate variation. AV GRATS expands on the audio-only version of the dataset, originally presented in [8]. Unlike synthetic time-scale modification, AV-GRATS records speakers naturally producing the same utterances at different target speaking rates. This design preserves natural acoustic and articulatory changes, making the corpus suitable for studying AVSR behavior under realistic temporal variation.

1) Data Collection: The current version of AV-GRATS contains 7,500 audio-visual utterances from 25 native Taiwanese Mandarin speakers, including 13 female and 12 male speakers. Each speaker recorded 60 Mandarin utterances at five speaking rates: 0.5×, 0.75×, 1.0×, 1.25×, and 1.5×. The speakers were young adults aged 20–35 and included Mandarin speakers from both northern and southern Taiwan.

To guide the production of different speaking rates, we used a karaoke-style visual prompting interface. The interface highlighted characters according to the target tempo, allowing speakers to follow the intended speaking rate while still producing natural speech. Utterances containing reading errors, articulation errors, or obvious tempo-following failures were re-recorded to ensure transcript accuracy.

2) Recording Setup: Recordings were conducted in a controlled low-noise studio environment. Video was captured at 1920×1080 resolution and approximately 30 frames per second. Audio was recorded at 44.1 kHz and 16-bit PCM using two microphones: RØDE Wireless GO II and ATR2500x-USB. The dual-microphone setup enables flexible downstream analysis of acoustic robustness. Audio and video streams were synchronized to produce temporally aligned audio-visual pairs. No artificial time-scale modification was applied to either modality.

3) Dataset Organization: AV-GRATS is an ongoing corpus collection effort. The current version contains 7,500 utterances, and future releases will further increase speaker and sentence diversity. In this paper, we use a dual-independence experimental subset to evaluate generalization under strict speaker-level and sentence-level separation.

The subset used in the current experiments contains 5,100 clips. The training portion contains 4,800 distinct recordings, comprising 48 unique sentence prompts produced by 20 speakers at five speaking rates. The test portion contains 300 clips, consisting of 12 different utterances recorded by 5 held-out speakers at five speaking rates. Therefore, the training and

test sets have no speaker overlap and no sentence overlap. This split supports the dual-independence evaluation protocol described in Section IV-E.

The remaining utterances in the current corpus are not used in the main benchmark experiments of this paper and are reserved for future dataset expansion, validation, and additional analysis.

IV. Methodology

A. AVSR Model

We adopt AV-HuBERT Large (24 encoder layers, 1024-dimensional hidden size) pre-trained on the VoxCeleb2 1,326-hour English subset using iterative self-supervised clustering (iter5) [2] as our backbone. For downstream sequence-to-sequence fine-tuning, we use a 9-layer Transformer decoder (1024 hidden size, 8 attention heads, 4096 FFN dimension).

Target labels are represented as space-separated Pinyin syllables and tokenized using SentencePiece (vocabulary size 673, word-level model without dummy prefix) [21]. Optimization uses Adam with a tri-stage learning rate schedule (10k warmup steps followed by 40k decay steps, peak learning rate 3×10^{-4}). The encoder outputs a joint audio-visual representation $\mathbf{z}^{AV} \in \mathbb{R}^{T \times d}$, where T denotes the length of the sequence and $d = 1024$.

The model is trained using label-smoothed cross-entropy loss ($\epsilon = 0.1$):

$$\mathcal{L} = -\sum_{i=1}^{N} \log P\big(y_i \mid y_{<i}, \mathbf{z}^{AV}\big), \tag{1}$$

where y_i is the i-th target Pinyin syllable token and N is the total number of tokens in the utterance. During the initial encoder warm-up stage (first 8,000 steps), only decoder parameters are updated; after warm-up, both encoder and decoder are jointly optimized. All experiments were conducted on 2× NVIDIA RTX 3090 GPUs.

B. Three-Level Error Decomposition

We define a three-level SER formulation to isolate different sources of error:

$$\text{SER} = \frac{S + D + I}{N} \times 100\%, \tag{2}$$

where S, D, and I denote substitution, deletion, and insertion errors, respectively, and N is the total number of reference syllable tokens. We consider three error levels: 1 being the finer and 3 being the coarser.

Level 1 – Pinyin with-tones: Standard SER computed on full Pinyin tokens, where tone is preserved (e.g., *wo3* ≠ *wo2*).

Level 2 – Pinyin without tones: SER computed after removing tone information. We define a mapping $\phi : s_k \mapsto \tilde{s}_k$, where $\tilde{s}_k$ is obtained by stripping tone digits:

$$\text{SER}_{\text{no-tone}} = \frac{S' + D' + I'}{N} \times 100\%, \tag{3}$$

where S', D', and I' are calculated on sequences transformed by $\phi(\cdot)$.

Level 3 – visual group: We collapse the initials into visually confusable groups using a mapping $\psi : \tilde{s}_k \mapsto \hat{s}_k$, where $\hat{s}_k = g(\text{init}(\tilde{s}_k)) + \text{final}(\tilde{s}_k)$, and $g(\cdot)$ maps each initial to a visual group (Table I):

$$\text{SER}_{\text{visual}} = \frac{S'' + D'' + I''}{N} \times 100\%, \tag{4}$$

This level provides a grouped-label diagnostic view of Mandarin visual confusability patterns.

Since visual grouping reduces the target label space, it should not be interpreted directly as an articulatory performance floor without accounting for vocabulary-size deflation. We further quantify error decomposition as:

$$\Delta_{\text{tone}} = \text{SER}_{\text{with-tone}} - \text{SER}_{\text{no-tone}}, \tag{5a}$$

$$\Delta_{\text{visual}} = \text{SER}_{\text{no-tone}} - \text{SER}_{\text{visual}}. \tag{5b}$$

C. Frequency-Weighted Deflation Correction

Removing tone labels and collapsing initials into visual groups both reduce the target vocabulary size. As a result, SER may decrease partly because the recognition task becomes easier, rather than because the model captures meaningful tone or visual structure. To account for this effect, we compute a frequency-weighted chance baseline at each decomposition level using the marginal token distribution of AV-GRATS.

Let $\mathcal{V}_\ell$ denote the target vocabulary at level ℓ, where $\ell \in \{1, 2, 3\}$ corresponds to with-tone, no-tone, and visual-group representations, respectively. Let $p_\ell(v)$ denote the empirical frequency of token $v \in \mathcal{V}_\ell$ in the reference set. The frequency-weighted chance accuracy at level ℓ is computed as:

$$A_\ell^{\text{chance}} = \sum_{v \in \mathcal{V}_\ell} p_\ell(v)^2, \tag{6}$$

and the corresponding chance SER is:

$$\text{SER}_\ell^{\text{chance}} = 1 - A_\ell^{\text{chance}}. \tag{7}$$

The expected deflation caused by moving from level ℓ to level $\ell + 1$ is then:

$$D_{\ell \to \ell+1} = \text{SER}_\ell^{\text{chance}} - \text{SER}_{\ell+1}^{\text{chance}}. \tag{8}$$

Finally, we compute corrected gaps by subtracting the frequency-weighted deflation from the raw gap:

$$\Delta_{\text{tone}}^{\text{corr}} = \Delta_{\text{tone}} - D_{1 \to 2}, \tag{9a}$$

$$\Delta_{\text{visual}}^{\text{corr}} = \Delta_{\text{visual}} - D_{2 \to 3}. \tag{9b}$$

This correction allows us to distinguish meaningful error reductions from reductions that are expected due to target-vocabulary collapse.

D. Visual Confusability Groups

We define nine visual groups for Mandarin initials based on their place and manner of articulation. The main groups include bilabials (BIL: $b/p/m$), labiodental (LBD: f), alveolars (ALV: $d/t/n/l$), velars (VEL: $g/k/h$), palatals (PAL: $j/q/x$), retroflexes (RET: zh/ch/sh/r), and sibilants (SIB: $z/c/s$), following established articulatory phonetic classifications in Mandarin viseme studies [20]. We also include glides (GLI: y/w)

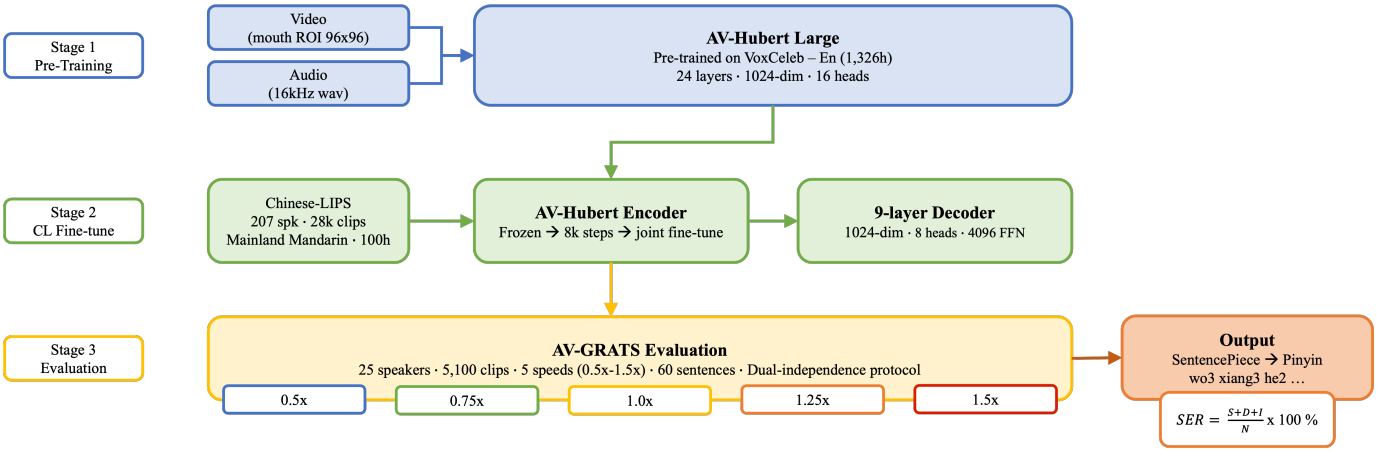


Fig. 1. Cross-domain AVSR evaluation pipeline. Stage 1: AV-HuBERT Large pre-trained on VoxCeleb2. Stage 2: Fine-tuned on Chinese-LIPS (28k clips, Mainland Mandarin). Stage 3: zero-shot evaluation on the 300-clip AV-GRATS dual-independence test set. Direct fine-tuning experiments use the 5,100-clip AV-GRATS experimental subset.

TABLE I
MANDARIN INITIAL VISUAL CONFUSABILITY GROUPS.

Code	Group	Articulation	Initials
ZER	Zero initial	(none)	∅
LBD	Labiodental	upper teeth/lip	f
BIL	Bilabial	both lips	b, p, m
VEL	Velar	soft palate	g, k, h
ALV	Alveolar	tongue tip/ridge	d, t, n, l
SIB	Sibilant	alveolar groove	z, c, s
GLI	Glide	approximant	y, w
PAL	Palatal	hard palate	j, q, x
RET	Retroflex	curled tongue tip	zh, ch, sh, r

and zero initial (ZER: ∅), as both exhibit distinct and consistent visual patterns in Mandarin articulation.

The mapping $g : \mathcal{I} \to \mathcal{G}$ assigns each Mandarin initial to a visual group, as shown in Table I. Two syllables s_i and s_j are defined as visually confusable under the proposed grouping scheme if they satisfy:

$$g(\text{init}(s_i)) = g(\text{init}(s_j)) \quad \text{and} \quad \text{final}(s_i) = \text{final}(s_j). \quad (10)$$

E. Dual-Independence Evaluation Protocol

Standard AVSR evaluation protocols on limited-prompt datasets may not enforce strict independence between training and evaluation splits in both speaker and sentence dimensions. This can introduce two potential sources of shortcut learning: (i) sentence memorization, where models overfit to a limited set of training transcripts, and (ii) speaker-dependent overfitting, where models exploit speaker-specific acoustic or visual patterns that may not generalize to unseen speakers.

In contrast, our dual-independence evaluation protocol explicitly enforces both constraints:

- **Speaker-independent:** all evaluation speakers are excluded from the training set.
- **Sentence-independent:** all evaluation sentences are disjoint from the training sentences, ensuring zero transcript overlap.

For AV-GRATS, the evaluation speakers are spk10, spk16, spk17, spk20, and spk21, and the 12 evaluation sentences are excluded from the 48 training sentences. This protocol is applicable to AVSR benchmarks with limited speaker or sentence diversity, and we use it to evaluate generalization under strict speaker-level and sentence-level separation.

V. EXPERIMENTS AND RESULTS

A. Experimental Setup

Fig. 1 shows our cross-domain AVSR evaluation pipeline. We fine-tuned the AV-HuBERT Large model on the Chinese-LIPS (CL) dataset for Mandarin with a mainland Chinese accent without additional augmentation, and tested it in a zero-shot manner on the AV-GRATS dual-independence test set for Mandarin with a Taiwanese accent (5 speakers, 300 clips, 60 clips per speaking rate). The CL fine-tuned model achieved 7.75% AV SER on the in-domain Chinese-LIPS test set (3,704 utterances), confirming it as a strong and stable Mandarin AVSR model for cross-domain evaluation.

All reported zero-shot results follow the dual-independence evaluation protocol described in Section IV-E. This setup (CL → AV-GRATS) evaluates generalization from Chinese-LIPS to AV-GRATS without in-domain adaptation and provides a controlled setting for analyzing speaking-rate robustness, tone-related errors, and visual-group confusion patterns.

B. Zero-Shot Benchmark Evaluation

Table II reports the zero-shot SER results across five speaking rates on the 300-clip AV-GRATS dual-independence test set. The average SER is 31.20%, matching the aggregate zero-shot result reported in Table III. Performance remains relatively stable from 0.5× to 1.25×, with SER ranging from 26.67% to 28.50%. However, performance degrades substantially at 1.5×, where SER increases to 45.50%. This result

TABLE II
SER (%) OF ZERO-SHOT AVSR AT DIFFERENT SPEAKING RATES ON THE AV-GRATS DUAL-INDEPENDENCE TEST SET(GR).

System	0.5×	0.75×	1.0×	1.25×	1.5×
Base GR ZS (AV)	28.50	26.67	27.33	28.00	45.50

[†]Fast speech exhibits a distinct cross-domain AVSR performance degradation pattern.

TABLE III
THREE-LEVEL SER DECOMPOSITION (%) OF VSR AND AVSR PERFORMANCE ON THE CHINESE-LIPS (CL) AND AV-GRATS (GR) TEST SETS.

Data	Mode	w/Tone	No-Tone	Vis-Grp	Δ_{tone}	Δ_{visual}
CL	VSR	88.9	86.1	82.0	2.8	4.1
GR	VSR	98.2	95.4	94.1	2.8	1.3
GR	AVSR	31.2	21.2	19.6	10.0[†]	1.6

[†]Large tone-related gap under zero-shot cross-domain AVSR evaluation.

suggests that fast speech introduces a distinct recognition difficulty under zero-shot cross-domain evaluation.

C. Three-Level Error Decomposition

Table III summarizes the three-level SER decomposition. VSR means that we perform visual speech recognition using the AVSR model, i.e., only the visual stream is input to the model. We highlight several key findings.

AVSR reduces SER mainly through acoustic information. In cross-domain evaluation, AVSR performs significantly better than VSR (31.2% SER vs 98.2% SER, zero-shot performance on the AV-GRATS dual-independence test set), indicating that the acoustic stream remains the dominant source of recognition capability, even under cross-domain evaluation.

Tone errors dominate residual AV performance. The largest gap in AVSR is the tone-related component ($\Delta_{\text{tone}} = 10.0\%$), while the visual-group gap remains small ($\Delta_{\text{visual}} = 1.6\%$). This suggests that tone prediction is a major source of residual AVSR error under the current cross-domain setting.

Visual-group SER requires careful interpretation. The raw gap between no-tone SER and visual-group SER is small ($\Delta_{\text{visual}} = 1.6\%$). However, collapsing initials into visual groups substantially reduces the target label space. Therefore, the raw visual-group SER reduction should not be interpreted directly as evidence that the model approaches an articulatory confusability floor. We apply a frequency-weighted deflation correction in Section V-D to separate meaningful error reductions from reductions caused by target-vocabulary collapse.

D. Frequency-Weighted Deflation Correction

Since each decomposition level reduces the target label space differently, raw SER gaps may include reductions caused by vocabulary deflation. We therefore compute a frequency-weighted chance baseline using the marginal token distribution of AV-GRATS, following the formulation in Section IV-C.

TABLE IV
FREQUENCY-WEIGHTED DEFLATION CORRECTION FOR SER DECOMPOSITION ON AV-GRATS.

Transition	Raw gap	Deflation	Corrected gap
L1 → L2 (Δ_{tone})	10.00	0.24	9.76
L2 → L3 (Δ_{visual})	1.60	15.16	-13.56

For the transition from Level 1 to Level 2, tone removal reduces the vocabulary from 88 tone-bearing tokens to 75 no-tone tokens. The corresponding frequency-weighted deflation is only 0.24%. Therefore, the corrected tone-related gap remains large at 9.76%, confirming that the observed tone gap is not merely a vocabulary-reduction artifact.

For the transition from Level 2 to Level 3, the deflation effect is much larger because Mandarin initials are collapsed into visual groups. The frequency-weighted deflation for this step is 15.16%, which is larger than the raw visual-group gap of 1.60%. After correction, Δ_{visual} becomes -13.56%. Therefore, the raw visual-group SER reduction should not be interpreted as evidence that the model reaches an articulatory confusability floor. Instead, visual-group SER is used as a diagnostic representation for analyzing Mandarin visual confusability patterns.

E. Visual Group Analysis

Table V shows the visual-group confusion matrix for the Base GR ZS (AV) system on the AV-GRATS dual-independence test set. Unlike the VSR results in Table III, this analysis is based on the audio-visual zero-shot system and is used to examine how Mandarin initial groups are confused after recognition. The model achieves an unweighted average group accuracy of 90.1% across the nine articulatory groups, with per-group accuracy ranging from 71.6% to 100.0%.

Strongly separable articulatory groups. ZER and LBD achieve near-perfect or perfect recognition (98.0% and 100.0%), indicating that zero-initial and labiodental articulations provide highly distinctive visual cues. Bilabial (95.4%), alveolar (95.1%), and velar (92.9%) groups are also robust, suggesting that AV-HuBERT is able to capture global facial motion patterns beyond isolated lip shapes.

Primary confusion: retroflex vs sibilant. The dominant error pattern is confusion between RET and SIB, where 17.3% of RET tokens are misclassified as SIB. This aligns with the known perceptual ambiguity between retroflex and alveolar fricatives in Mandarin and may reflect a visually confusable Mandarin articulatory pattern.

F. Effect of Direct Fine-Tuning on Limited-Prompt Data

We further evaluate direct fine-tuning on the AV-GRATS training set to examine whether the current limited-prompt corpus supports reliable in-domain adaptation. This experiment

TABLE V

VISUAL GROUP CONFUSION MATRIX (%) FOR THE BASE GR ZS (AV) SYSTEM ON THE AV-GRATS DUAL-INDEPENDENCE TEST SET. **BOLD DIAGONAL** = CORRECT RECOGNITION. GREEN ≥80% DIAGONAL ACCURACY; YELLOW = 3–9% OFF-DIAGONAL CONFUSION; RED ≥10% OFF-DIAGONAL CONFUSION. CELLS BELOW 1% ARE SHOWN AS —. AVERAGE = SIMPLE UNWEIGHTED MEAN ACROSS 9 GROUPS.

Reference Group	n	**Hypothesis Group (%)**									**Acc.**
		ZER	LBD	BIL	VEL	ALV	SIB	GLI	PAL	RET	
ZER (∅)	50	**98.0**	—	—	2.0	—	—	—	—	—	98.0
LBD (f)	50	—	**100.0**	—	—	—	—	—	—	—	100.0
BIL (b/p/m)	325	—	1.5	**95.4**	—	2.2	—	—	—	—	95.4
VEL (g/k/h)	350	—	—	—	**92.9**	4.3	—	—	—	—	92.9
ALV (d/t/n/l)	750	—	—	—	—	**95.1**	—	—	—	1.2	95.1
SIB (z/c/s)	225	—	—	—	—	—	**84.4**	—	—	13.8	84.4
GLI (y/w)	325	—	—	—	6.8	3.4	—	**85.5**	—	3.4	85.5
PAL (j/q/x)	475	—	—	—	—	—	—	—	**88.0**	10.7	88.0
RET (zh/ch/sh/r)	450	—	—	—	—	2.9	17.3	—	6.0	**71.6**	71.6

TABLE VI

EXAMPLES OF TRANSCRIPT MEMORIZATION AFTER DIRECT FINE-TUNING ON AV-GRATS. THE MODEL OUTPUTS TRAINING SENTENCES FOR UNSEEN SPEAKERS AND UNSEEN SENTENCES.

Input	**Ground Truth**	**Model Output**	**ED**
spk16_sent021 0.5×	wo3 zuo2 tian1 mei2 neng2 can1 jia1 zhao1 dai4 hui4	wo3 mei3 tian1 zao3 shang4 dou1 yao4 he1 bei1 cha2	0
spk16_sent031 0.5×	da4 duo1 shu4 bei3 fang1 ren2 ai4 chi1 jiao3 zi	da4 jia1 you3 shi4 dou1 ai4 zhao3 ta1 shang1 liang2	1

ED: edit distance to the matched training sentence.

differs from the zero-shot cross-domain experiments above: here, AV-HuBERT Large is fine-tuned directly on AV-GRATS and evaluated on the AV-GRATS dual-independence test set.

Direct fine-tuning on the AV-GRATS training set results in 87.3% SER on the dual-independence test set, which is substantially worse than the zero-shot cross-domain performance of 31.2% SER. The degradation is consistent across speaking rates, ranging from 87.0% to 88.8%. This indicates that the performance decline is not mainly caused by temporal variation, but by poor generalization under the limited-prompt fine-tuning setting.

Dual-independence evaluation. The test set is strictly separated from the training set at both the speaker and sentence levels: the five test speakers are unseen during training, and all 12 test sentences are disjoint from the 48 training sentences. This design eliminates leakage at both the speaker and sentence levels.

Evidence of transcript memorization. As shown in Table VI, the directly fine-tuned model often outputs training sentences verbatim or near-verbatim for inputs from unseen speakers and unseen sentences. In the most striking case, spk16_sent021 produces an exact match to a training sentence, with edit distance 0. Among the 300 test clips, the most frequent output appears 22 times (7.3%), all matching the same training sentence. This suggests that the model maps diverse unseen inputs to a small set of training transcripts rather than performing reliable speech recognition.

Distinct from standard domain overfitting. Unlike typical domain overfitting, the observed failures show strong input invariance, where phonetically distinct utterances can be mapped to the same fixed hypothesis. This highlights the importance of dual-independence evaluation for limited-prompt AVSR benchmarks.

VI. CONCLUSIONS

This paper introduced AV-GRATS, a naturally produced parallel Mandarin audio-visual corpus for evaluating AVSR under speaking-rate variation. The current corpus contains 7,500 utterances from 25 native Taiwanese Mandarin speakers across five speaking rates from 0.5× to 1.5×.

Using AV-HuBERT fine-tuned on Chinese-LIPS, zero-shot evaluation shows stable performance from 0.5× to 1.25×, but substantial degradation at 1.5×. Under the dual-independence protocol, direct fine-tuning on AV-GRATS results in 87.3% SER, indicating transcript memorization and poor generalization. Three-level SER analysis with frequency-weighted deflation correction shows a corrected 9.76% tone-related gap, while visual-group analysis reveals Mandarin-specific articulatory confusion patterns. These results demonstrate the usefulness of AV-GRATS as a diagnostic benchmark for Mandarin AVSR robustness.

ACKNOWLEDGMENT

Generative AI tools were used solely to improve the language and presentation of this manuscript. The authors are responsible for all scientific content, analyses, and conclusions.

REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[2] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=Z1Qlm11uOM

[3] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] L. Pandey and A. S. Arif, "Effects of speaking rate on speech and silent speech recognition," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491101.3519611

[5] M. Kuhlmann, F. Seebauer, J. Ebbers, P. Wagner, and R. Haeb-Umbach, "Investigation into Target Speaking Rate Adaptation for Voice Conversion," in *Interspeech 2022*, 2022, pp. 4930–4934.

[6] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007, intrinsic Speech Variations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639307000404

[7] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech 2015*, 2015, pp. 3586–3589.

[8] G. F. Aghniya, D. A. M. G. Wisnu, S. Rini, and Y. Tsao, "GRATS: A natural multi-speed Mandarin dataset for speech time-scale modification benchmarking," in *Interspeech 2026*, 2026, accepted for publication.

[9] J. Zhao, Y. Jia, S. Wang, J. Zhou, H. Wang, and Y. Qin, "Chinese-LiPS: A Chinese audio-visual speech recognition dataset with lip-reading and presentation slides," 2025. [Online]. Available: https://arxiv.org/abs/2504.15066

[10] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," 2018. [Online]. Available: https://arxiv.org/abs/1809.00496

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech 2018*, 2018, pp. 1086–1090.

[12] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 783–18 794.

[13] S. Kim, S. Cho, S. Bae, K. Jang, and S.-Y. Yun, "Multi-task corrupted prediction for learning robust audio-visual speech representation," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=WEQL5ksDnB

[14] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[15] C. Chen, Z. Liu, X. Li, L. Li, and D. Wang, "CNVSRC 2023: The First Chinese Continuous Visual Speech Recognition Challenge," in *Interspeech 2024*, 2024, pp. 1930–1934.

[16] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, and A. Mashari, "Audio visual speech recognition," IDIAP, Idiap-RR Idiap-RR-35-2000, 2000.

[17] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[18] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40–67, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639317300286

[19] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7, pp. 857–876, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639313000319

[20] M. Xu, J. Shao, and H. Ding, "Viseme identification ability and its effect on auditory consonant perception in Mandarin-speaking younger and older adults," *Speech Communication*, vol. 179, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639326000245

[21] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012