Multilingual Motion Memory Adaptation for Turkish 3D Talking Faces

Hüseyin Temiz[1,2], Berk Gökberk[1], and Lale Akarun[1]

[1] Department of Computer Engineering, Boğaziçi University, İstanbul, Türkiye
[2] Yapı Kredi Teknoloji, İstanbul, Türkiye

Abstract. Speech-driven 3D facial animation has largely been developed and benchmarked on English-centric datasets such as VOCASET and BIWI, leaving open whether its motion-memory and stylization mechanisms transfer to languages with non-English phoneme inventories. We address this gap for Turkish, whose vowels /ɯ/, /ø/, and /y/ have no direct English counterpart, through a multilingual adaptation of MemoryTalker. We first introduce SUDFace3D, a set of frame-level FLAME annotations for the full Turkish SUDFace corpus (50 speakers, 150 sequences), and define a benchmark for Turkish talking faces. To our knowledge, this is the first dedicated, controlled Turkish 3D talking-face benchmark. Through a multilingual phoneme-probing analysis, we test four speech encoders: phoneme probing favours mHuBERT-147 and MMS-1b as the most Turkish-aware representations, while Stage-2 mesh adaptation yields XLS-R-300M as an equally strong backbone on the downstream task.

Replacing the audio encoder of MemoryTalker with multilingual encoders and fine-tuning the Stage-2 style pathway and decoder on SUDFace3D, while keeping the audio encoder and motion memory frozen, substantially narrows the cross-language mesh gap: lip-vertex error drops from the 0.83–0.95 zero-shot range down to 0.30–0.32 for the three multilingual backbones. We also find that swapping the audio encoder alone is insufficient for zero-shot mesh generation; the main cross-language bottleneck lies downstream of the encoder. Stage 2 training, even with a small target-language dataset, significantly improves the results.

Keywords: Speech-driven facial animation · Cross-lingual generalization · Multilingual audio encoders · Motion memory · Turkish Talking Face Benchmark · FLAME · SUDFace3D

1 Introduction

Speech-driven 3D facial animation aims to synthesize realistic, audio-synchronized facial motion directly from speech, with applications in virtual avatars, telepresence, and film and game production. The task is challenging: lip and jaw motion must be phonetically accurate at the millisecond scale, while the rest of the face must convey speaker identity and expressiveness consistent with the audio. Recent progress has moved from regression-based models such as VOCA [6]

and MeshTalk [24] to Transformer- and diffusion-based decoders such as FaceFormer [10], CodeTalker [32], and FaceDiffuser [27], and to unified models such as UniTalker [11]. Among these, MemoryTalker [15] stores reusable audio–motion alignments in an external motion-memory bank and produces personalized animation from audio alone, without a one-hot identity at inference.

Yet the field remains strongly English-centric, particularly at the level of 3D supervision. The few high-precision scan-based 3D datasets, such as VOCASET [6] and BIWI [12], are captured with multi-view 3D scanners on English-speaking subjects only. Larger reconstruction-based corpora such as the TFHP dataset introduced by DiffPoseTalk [28] (English-only) and MultiTalk [16] (multilingual but sourced from in-the-wild YouTube video) trade controlled recording conditions for scale (Table 1). The standard audio backbones, HuBERT [14] and wav2vec 2.0 [2], are pre-trained on English speech. Every stage of the pipeline, from the audio encoder to the 3D supervision, is therefore tuned to English, and it is unclear whether a motion memory learned in this regime transfers to languages with different phoneme inventories.

Turkish has the vowels /ɯ/, /ø/, and /y/ [13], which are absent from the standard English phonemic inventory [25]. It therefore provides a natural cross-lingual test case for speech-driven facial animation, since phoneme-to-viseme correspondence is central to lip synchronization [9,29].

We study this cross-lingual setting with SUDFace [26], a Turkish facial video corpus with 50 speakers and three speaking conditions (memorized anthem, formal speech, free speech). For this work, we extract frame-level FLAME [18] parameters for all SUDFace videos and use them to evaluate how MemoryTalker's audio encoder, motion memory, and Stage 2 style head behavior on Turkish speech. We focus on MemoryTalker because its explicit motion-memory bank exposes a clean question: Are the stored articulatory primitives English-specific, or do they transfer to other languages through the mapping of a multilingual audio encoder?

Our contributions are summarized as follows:

- We introduce SUDFace3D: frame-level FLAME annotations for all 150 SUDFace videos and introduce a Turkish language talking face benchmark.
- We compare four speech encoders on Turkish/English phoneme probing and identify two well-performing encoders: mHuBERT-147 as a compact Turkish-aware backbone, and the MMS-1b which performs well at a much higher complexity level. We further analyze which layer yields the highest overall accuracy and use those features.
- We show that swapping the audio encoder alone does not close the zero-shot Turkish mesh gap, but a frozen-encoder Stage-2 contrastive fine-tuning on SUDFace3D reduces lip-vertex error from 0.83–0.95 to 0.30–0.32 for the three multilingual backbones, recovering most of the gap without retraining the motion memory.

Taken together, these results suggest that MemoryTalker's motion memory is articulatorily grounded enough to transfer through shared phonemes, but it is

English-biased, which requires target-language Stage 2 supervision for Turkish-specific articulation.

2 Related Work

2.1 Speech-Driven 3D Facial Animation

Speech-driven 3D facial animation has progressed from regression-based models to Transformer and diffusion architectures. VOCA [6] regresses FLAME mesh deformations from DeepSpeech features conditioned on a one-hot speaker, and MeshTalk [24] disentangles audio-correlated and audio-uncorrelated face motion via a categorical latent space. FaceFormer [10] introduces an autoregressive Transformer decoder conditioned on wav2vec 2.0 [2] audio features and one-hot speaker identity, and CodeTalker [32] learns a discrete motion prior via vector quantization, mitigating over-smoothing. SelfTalk [22] adds a self-supervised regression loss on disentangled speech representations, while EmoTalk [21] explicitly disentangles emotion from speech content for emotional 3D animation. Imitator [31] conditions generation on a reference video for style transfer. More recent work casts the problem as diffusion-based generation: FaceDiffuser [27] synthesises 3D facial animation with a diffusion model, and 3DiFACE [30] adds editing capability and stochastic head-pose generation. ScanTalk [19] extends the diffusion paradigm to operate on unregistered face scans of arbitrary topology. UniTalker [11] proposes a unified model that scales across multiple 3D talking-face datasets. MemoryTalker [15] departs from this lineage by introducing an external motion-memory bank that stores reusable facial-motion primitives indexed by audio, removing the need for a one-hot identity at inference. THUNDER [8] revisits speech-driven 3D talking-head generation from a different direction, introducing an analysis-by-audio-synthesis loop in which the generated 3D facial motion is passed through a mesh-to-speech model and compared with the input audio, improving lip synchronization while preserving stochastic expressive motion. Most of these methods are trained and evaluated on English-centric datasets, and the audio backbones they rely on are pre-trained on English speech. MultiTalk [16] addresses multilinguality at the dataset level with large-scale multilingual training and motion-token-based modeling, but does not study how existing English-trained models transfer cross-lingually. Hallo [34] and its follow-up methods are highly relevant to audio-driven 2D portrait animation, but their primary focus is visual fidelity, long-duration generation, dynamics, and controllability rather than multilingual speech transfer.

It therefore remains an open question whether the learned motion representations, particularly the motion-memory primitives of MemoryTalker, generalize to other languages. Recently, Polyglot [20] has successfully applied mHuBERT-147 [3] embeddings to PolySet, a new dataset derived from the MultiTalk [16] video data. Our SUDFace3D results confirm that multilingual self-supervised speech representations are a more suitable audio interface than English-only features for cross-lingual 3D facial animation.

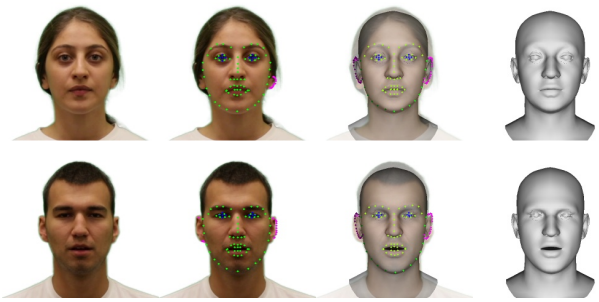


Fig. 1: Two sample subjects from SUDFace3D. From left to right: (a) background-cropped frame, (b) detected facial landmarks, (c) landmarks overlaid on the cropped frame, and (d) reconstructed FLAME mesh.

2.2 Cross-Lingual Speech Representations

Self-supervised speech models such as HuBERT [14] and wav2vec 2.0 [2] are pre-trained on English data but transfer to non-English ASR (Automatic Speech Recognition) with appropriate fine-tuning [4]. A second family is pre-trained directly on multilingual corpora. XLS-R-300M [1] extends wav2vec 2.0 to 128 languages, MMS-1b [23] scales to over 1100 languages, and *mHuBERT-147* [3] retains the exact 12-layer/95 M-parameter architecture of HuBERT-base but is pre-trained on 147 languages, offering a multilingual alternative to HuBERT-base at the same parameter budget. These models have been evaluated primarily on cross-lingual ASR and phoneme recognition. Their use as audio backbones for 3D talking-face generation, where the downstream task is articulatory motion rather than phoneme classification, has not been studied.

3 SUDFace3D Dataset

Speech-driven 3D facial animation requires datasets of temporally aligned audio and 3D facial motion, typically represented as dense meshes or parametric models such as FLAME [18]. Collecting high-quality 4D facial data with synchronized audio is expensive, and few such datasets are publicly available. Existing benchmarks fall into two groups. The first includes high-fidelity but small-scale datasets, such as VOCASET [6] and BIWI [12], which provide accurate geometry captured via multi-view 3D scanning systems but are restricted in subject diversity and to a single language (English). The second group consists of larger-scale datasets constructed via 3D reconstruction pipelines, such as TFHP [28] and MultiTalk [16], which offer greater variability in identity, pose, and speaking styles at the cost of noisier pseudo-ground-truth annotations. Such pipelines build on monocular FLAME reconstruction methods such as EMOCA [7]. These datasets cover at most a handful of languages and trade scale against geometric accuracy. To our knowledge, there are no controlled Turkish 3D talking-face benchmarks with frame-level 3D annotations.

Table 1: Speech-driven 3D facial animation datasets. SUDFace3D is the first dedicated, controlled Turkish 3D talking-face benchmark; existing multilingual corpora cover Turkish only as one language among many in-the-wild recordings.

Dataset	Subj.	Duration	3D annotation	Lang.
VOCASET [6]	12	∼0.5 h	FLAME-topology mesh from 3D Scan	En.
BIWI [12]	14	∼1.4 h	Dense mesh	En.
TFHP [28]	∼588	26.5 h	FLAME pseudo-GT	En.
MultiTalk [16]	N/A	423.2 h	FLAME pseudo-GT	20 langs
SUDFace3D [26,35]	50	2.5 h	FLAME pseudo-GT	**Tr.**

SUDFace [26] provides a Turkish-language alternative, a video corpus of 50 speakers with three recording conditions per speaker: a memorized national anthem (NA), a memorized formal speech (SY), and free speech (FS). Each video is approximately 60 seconds long (yielding ∼2.5 hours in total), recorded at 1920×1080 resolution with controlled illumination. Since SUDFace provides only 2D video, we recover frame-level FLAME parameters for each of the 150 sequences using KaoLRM [35] (Fig. 1 shows two example subjects with landmarks and recovered mesh). KaoLRM repurposes the 3D prior of a pre-trained large reconstruction model by projecting its triplane features into the FLAME parameter space, and couples the recovered mesh with FLAME-based Gaussian rendering. This yields single-view reconstructions with improved 3D consistency under pose variation and self-occlusion. We then apply the same 3D-parameter post-processing used in the TFHP dataset preparation [28], including trajectory cleanup and temporal smoothing before training and evaluation. The resulting corpus, SUDFace3D, is the first dedicated, controlled Turkish 3D talking-face benchmark. SUDFace's studio-grade recording conditions and Turkish-only design separates it from the broader multilingual, in-the-wild corpora that may incidentally cover Turkish. We release the SUDFace3D FLAME annotations for all 150 videos along with the analysis code.

4 Audio Encoder Selection for Turkish

The audio backbone is the first stage where an English-trained talking-face pipeline can lose Turkish articulation. We therefore use a frozen-feature phoneme probe to measure how well each encoder exposes Turkish and English phonetic content, before any mesh model is trained. The probe uses 500 Turkish and 500 English FLEURS [5] utterances at 16 kHz. International Phonetic Alphabet (IPA) frame labels are obtained at 50 Hz by Connectionist Temporal Classification (CTC) forced alignment [17] of eSpeak-phonemized transcripts with `wav2vec2-lv-60-espeak-cv-ft` [33]; CTC blank frames are forward-filled with the most recent non-blank label to increase labelled-frame coverage.

For each encoder–layer pair, we train a multinomial logistic-regression probe on a transcript-disjoint train/test split over 71 frequent phoneme classes, and

Table 2: Best layer (L) per-frame phoneme probe on FLEURS-TR/EN. The best layer per encoder is selected by overall accuracy. Frame-level accuracy is reported overall and separately for Turkish and English frames.

Encoder	Params.	Langs.	Best L	Overall acc.	TR acc.	EN acc.
HuBERT-base [14]	95 M	1	12 / 12	0.512	0.489	**0.535**
XLS-R-300M [1]	315 M	128	18 / 24	0.480	0.460	0.500
MMS-1b [23]	965 M	1100+	34 / 48	0.518	0.532	0.505
mHuBERT-147 [3]	**95 M**	**147**	**9 / 12**	**0.531**	**0.537**	0.525

Table 3: Per-phoneme accuracy on Turkish phonemes of interest at each encoder's best probe layer. The three vowels are absent from the standard English phonemic inventory; the two affricates also exist in English ("ch"/"j").

IPA	TR	In Eng.	HuBERT-base	XLS-R	MMS-1b	mHuBERT-147
/ɯ/	ı	no	0.325	0.292	**0.399**	0.393
/ø/	ö	no	0.580	0.531	**0.620**	0.616
/y/	ü	no	0.575	0.469	0.644	**0.651**
/tʃ/	ç	yes	0.678	0.599	0.712	**0.724**
/dʒ/	c	yes	0.687	0.610	**0.709**	0.648

report Turkish and English frame-level accuracy separately. The comparison covers HuBERT-base and three multilingual alternatives (listed in Table 2).

Table 2 and Fig. 2 report the best probing layer per encoder and show which language each frozen representation supports most strongly. HuBERT-base and XLS-R-300M show a clear English bias, with Turkish accuracy trailing English accuracy by 4–5 points. Multilingual pretraining largely closes this gap: both MMS-1b and mHuBERT-147 reach Turkish accuracy that matches or slightly exceeds their English accuracy. Among them, mHuBERT-147 attains the highest overall and Turkish accuracy among the four encoders, despite having roughly 10× fewer parameters than MMS-1b.

Table 3 confirms this trend on the Turkish phonemes of interest, both for vowels absent from English and for affricates shared with English. On the three Turkish-specific vowels (/ɯ/, /ø/, /y/) absent from English, both MMS-1b and mHuBERT-147 outperform HuBERT-base by 4–8 points, exactly where monolingual pretraining lacks data. On the shared affricates, mHuBERT-147 leads on /tʃ/, while MMS-1b leads on /dʒ/; both multilingual models stay at or above HuBERT-base except for mHuBERT-147 on /dʒ/, showing that multilingual pretraining does not sacrifice English-shared phonemes overall. Within the multilingual pair, mHuBERT-147 leads on /y/ and /tʃ/, stays within 1 point of MMS-1b on /ɯ/ and /ø/, and trails only on /dʒ/, where MMS-1b retains a clear margin. We therefore select mHuBERT-147 at the representational level, balancing per-language phoneme accuracy against parameter count. Whether this representational advantage carries through to mesh-level Turkish lip motion is the question we investigate in Section 6.

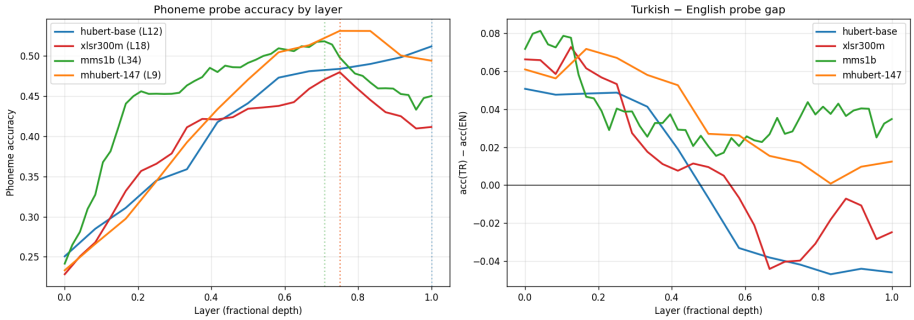


Fig. 2: Per-language phoneme-probe accuracy versus relative encoder depth. mHuBERT-147 reaches the highest overall accuracy and the smallest Turkish–English gap among the four encoders.

5 SUDFace3D-Guided Motion Memory Adaptation

5.1 MemoryTalker

We briefly describe the MemoryTalker pipeline [15] that we adopt as our backbone (Fig. 3). MemoryTalker generates per-frame FLAME vertex offsets from raw audio in two stages, with no one-hot identity input at inference.

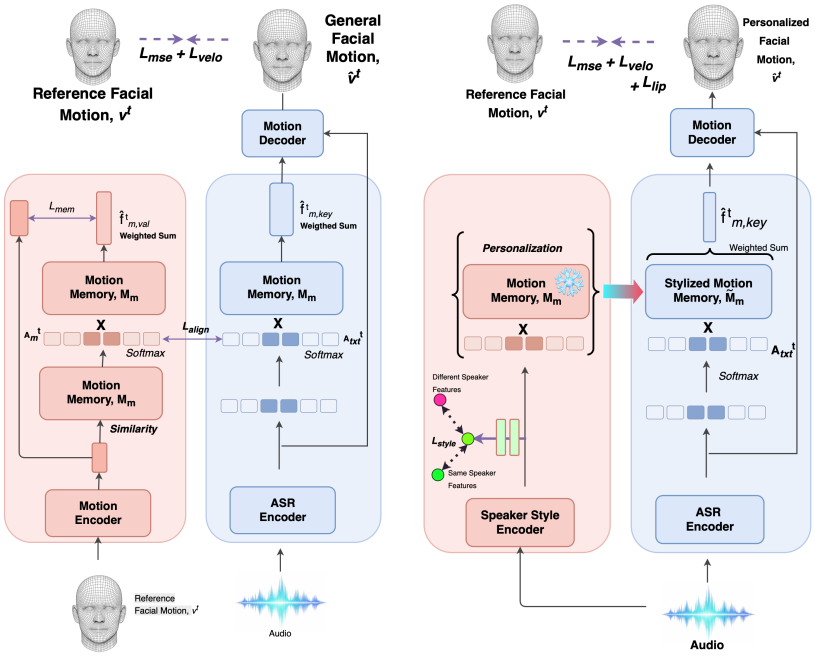
Stage 1 – Audio-driven motion memory (Fig. 3a). Audio is first encoded into per-frame features $\mathbf{f}_a \in \mathbb{R}^{T\times d_a}$ by a frozen HuBERT-base encoder [14] (replaced with multilingual encoders in our adaptation, see Section 5.2). A learned audio-to-text projection produces query features $\mathbf{Q} \in \mathbb{R}^{T\times c}$ that address a learned motion memory bank $\mathbf{M}_m \in \mathbb{R}^{n\times c}$ (n=32 slots) via cosine attention:

$$\mathbf{K}_{\text{txt}} = \text{softmax}\big(\mathbf{Q}\,\mathbf{M}_m^{\top}/\tau\big), \qquad \mathbf{f}_m = \mathbf{K}_{\text{txt}}\,\mathbf{M}_m. \tag{1}$$

The retrieved features $\mathbf{f}_m$ act as articulatory motion primitives that a Transformer decoder maps to FLAME vertex offsets. Stage 1 is supervised by mesh and velocity losses plus auxiliary memory losses (lip-region weighted, slot usage, and audio–memory alignment).

Stage 2 – Speaker-style stylization (Fig. 3b). Stage 2 introduces a speaker-style encoder E_s that extracts a style feature $\mathbf{f}_s \in \mathbb{R}^{32}$ from the mel-spectrogram of a reference utterance and is trained with a triplet loss over speaker identity. $\mathbf{f}_s$ modulates the memory query and the decoder so that the same audio yields different fine-grained motion for different speakers, without requiring an explicit one-hot identity. At inference, E_s is run on the input audio itself, making the model fully audio-only.

We use the VOCASET checkpoint as our starting point for all experiments in this paper, and intervene at two places in the pipeline: the audio encoder (Section 4) and the Stage 2 head (Section 5.2).



(a) Stage 1: Audio-driven motion memory.

(b) Stage 2: Speaker stylization.

Fig. 3: Overview of the MemoryTalker pipeline that we adopt as our backbone. Stage 1 (left) encodes audio with HuBERT-base, replaced by multilingual encoders (MMS-1b, XLS-R-300M, mHuBERT-147), and addresses a 32-slot motion memory M_m via cosine attention to produce per-frame FLAME vertex offsets. Stage 2 (right) adds a speaker-style encoder E_s trained with a triplet loss over speaker identity, modulating the memory query and the decoder so that the same audio yields personalised motion without requiring an explicit one-hot identity.

5.2 SUDFace3D-Guided Adaptation

The released MemoryTalker uses an English-pretrained HuBERT-base audio encoder and a motion memory trained on VOCASET, a small English-only 3D corpus. This setup gives strong English lip synchronization, but it ties the 32-slot motion memory $\mathbf{M}_m$ to English audio–geometry statistics. Rather than committing to a single audio backbone, we evaluate three multilingual encoders, namely MMS-1b (965 M parameters), XLS-R-300M (315 M), and mHuBERT-147 (95 M), where each is required to drive 3D motion synthesis rather than only expose phonetic content to a linear probe.

Starting from the VOCASET-trained checkpoint, we evaluate two distinct adaptation strategies for each candidate encoder. Strategy A (Stage 1 retraining): we replace HuBERT-base with the candidate's best probe layer (Section 4)

and retrain the Stage 1 audio–geometry head on either VOCASET or SUDFace3D, so that the memory is queried with a multilingual audio representation. Strategy B (Stage 2 only): we contrastively fine-tune the Stage 2 speaker-style head on SUDFace3D speaker triplets, while keeping both the audio encoder and the Stage 1 motion memory frozen. This lets Turkish phoneme-to-motion correspondences to be learned without retraining the motion memory.

This comparison tests whether the phoneme-probe ranking of Section 4 carries through to mesh-level motion, or whether downstream behaviour follows a different ordering. The 32-slot motion memory is inherited from VOCASET. For Stage 1 retraining (Strategy A) the objective keeps MemoryTalker's mesh, velocity, and lip-region losses but drops the auxiliary memory-regularization terms (slot usage and audio–memory alignment); for Stage 2 adaptation (Strategy B) training uses MemoryTalker's triplet/contrastive speaker-style loss. Evaluation uses the five mesh metrics defined in Section 6.2.

6 Experiments

6.1 Experimental Setup

We use a gender-balanced, subject-disjoint 35/5/10 SUDFace3D split: 35 subjects for training, 5 for validation, and 10 for test. With three recording conditions per subject (NA, SY, FS), this gives 105 train, 15 validation, and 30 test sequences of about 60s each. All experiments use the same FLAME extraction pipeline (Section 3) and the standard 5023-vertex FLAME topology.

Across the main Stage 1 and Stage 2 experiments, we evaluate four settings: (i) *Stage 1 VOCASET in-domain* trains and evaluates Stage 1 on VOCASET while varying the audio backbone (Table 4); (ii) *Stage 1 SUDFace3D zero-shot* transfers the VOCASET-trained checkpoint to the SUDFace3D test split without Turkish supervision (Table 5, top); (iii) *Stage 1 SUDFace3D in-domain* trains Stage 1 from scratch on SUDFace3D for each backbone (Table 5, bottom); and (iv) *Stage 2 SUDFace3D fine-tuning* adapts the Stage 2 head on top of a VOCASET-trained Stage 1 checkpoint with the audio encoder kept frozen (Table 6).

Stage 2 trains the contrastive style head with an unfrozen decoder, 32 memory slots, and denoised audio. It uses Adam optimizer with learning rates 5×10^{-5} for the style head, 1×10^{-5} for the decoder, and checkpoint selection by validation LVE. The Stage 1 SUDFace3D in-domain experiments use the same 32-slot setup for up to 100 epochs.

6.2 Metrics

We follow MemoryTalker's evaluation code and report five lower-is-better metrics. FVE and LVE measure full-face mean squared vertex and lip vertex errors, respectively. FDD measures the discrepancy in upper-face temporal motion energy between predictions and ground truth. All three use $\times10^{-5}$ mm^2. MOD is the mean absolute vertical mouth-opening error in mm, and LDTW is the unitless FastDTW distance between predicted and ground-truth lip trajectories.

Table 4: In-domain VOCASET Stage 1 results with frozen encoders. Units follow Section 6.2. Best values are shown in bold.

Encoder (layer)	FVE ↓	LVE ↓	FDD ↓	MOD ↓	LDTW ↓
MMS-1b (L34)	**0.0543**	**0.3350**	**0.0078**	**7.39e-4**	**0.0485**
XLS-R-300M (L18)	0.0572	0.3785	0.0080	7.44e-4	0.0515
mHuBERT-147 (L9)	0.0630	0.4180	0.0079	8.60e-4	0.0541
HuBERT-base (CTC)	0.0782	0.5646	0.0089	1.11e-3	0.0625

Table 5: Stage 1 evaluation on SUDFace3D: zero-shot transfer from a VOCASET-trained checkpoint (top) versus in-domain training on SUDFace3D (bottom). Units follow Section 6.2. Best results per column within each setting are shown in bold.

Encoder (layer)	FVE ↓	LVE ↓	FDD ↓	MOD ↓	LDTW ↓
Zero-shot: VOCASET-trained, evaluated on SUDFace3D					
MMS-1b (L34)	0.1195	0.9520	0.0061	1.29e-3	0.0552
XLS-R-300M (L18)	0.1055	0.8437	0.0065	1.15e-3	**0.0506**
mHuBERT-147 (L9)	0.1080	0.8749	0.0044	**1.13e-3**	0.0535
HuBERT-base (CTC)	**0.1044**	**0.8309**	**0.0032**	1.20e-3	0.0524
In-domain: trained on SUDFace3D					
MMS-1b (L34)	0.0385	**0.3230**	0.0070	**5.57e-4**	**0.0280**
XLS-R-300M (L18)	**0.0357**	0.3262	0.0040	5.63e-4	0.0282
mHuBERT-147 (L9)	0.0403	0.3554	0.0057	5.83e-4	0.0292
HuBERT-base (CTC)	0.0375	0.3450	**0.0030**	6.02e-4	0.0293

6.3 Stage 1: Audio Encoder Comparison

We compare audio backbones in three Stage 1 settings: VOCASET in-domain, SUDFace3D zero-shot, and SUDFace3D in-domain. This separates representational quality from the downstream audio–geometry head learned for a particular dataset. For each multilingual backbone, we feed Stage 1 the layer that maximized phoneme-probe accuracy in Section 4: layer 34 for MMS-1b, layer 18 for XLS-R-300M, and layer 9 for mHuBERT-147 (Table 2). For HuBERT-base we follow the original MemoryTalker setup [15] and use the features after the CTC head rather than a hidden Transformer layer. This keeps the HuBERT-base run directly comparable to the published baseline and reflects the way the encoder is actually consumed by MemoryTalker's audio–text projection. Table 4 illustrates that on VOCASET, MMS-1b performs the best across all metrics, indicating that large multilingual features can improve Stage 1 when the downstream head is trained in-domain.

Table 5 shows Stage 1 evaluation results on SUDFace3D. In the zero-shot block, HuBERT-base leads the other encoders on FVE, LVE, and FDD, but every encoder's metrics are far worse than its in-domain counterpart. LVE more than doubles between the best in-domain encoder (MMS-1b, 0.3230) and the best

Table 6: Forward transfer from VOCASET Stage 1 to SUDFace3D Stage 2. Stage 2 is contrastively fine-tuned on the 35-speaker SUDFace3D train split and evaluated on the 30-sequence held-out test split, with the audio encoder kept frozen throughout; the Stage 1 source is always a VOCASET-trained *frozen*-encoder checkpoint. Units follow Section 6.2; best values are bold.

Encoder	FVE ↓	LVE ↓	FDD ↓	MOD ↓	LDTW ↓
MMS-1b (L34)	0.0347	0.3094	0.0035	**5.65e-4**	0.0448
XLS-R-300M (L18)	**0.0336**	**0.3006**	0.0036	5.68e-4	**0.0447**
mHuBERT-147 (L9)	0.0358	0.3223	0.0031	5.85e-4	0.0460
HuBERT-base (CTC)	0.0402	0.3747	**0.0029**	6.43e-4	0.0499

zero-shot encoder (HuBERT-base, 0.8309), and FVE shows similar degradations across all four encoders. Notably, HuBERT-base, the English-aligned encoder, outperforms the multilingual ones in the zero-shot setting. This suggests that the VOCASET-trained motion memory is tuned to the specific feature distribution of HuBERT-base, and a multilingual encoder, despite producing better Turkish phonetic features, queries the memory in a representation space the memory has never seen. The motion memory and audio–geometry head learned on English VOCASET are therefore not sufficient for zero-shot Turkish synthesis, regardless of which audio encoder is substituted.

The second block of Table 5 illustrates that when Stage 1 is trained directly on SUDFace3D, MMS-1b remains best on overall mesh performance (LVE, MOD, LDTW), while XLS-R-300M leads on FVE and HuBERT-base on FDD. Encoder choice is coupled to the downstream training domain: phoneme probes identify useful multilingual representations, but Stage 1 mesh quality also depends on the learned audio–geometry head. This motivates the SUDFace3D-guided Stage 2 adaptation (Section 6.4), rather than expecting an encoder swap alone to bridge the cross-lingual gap.

6.4 Stage 2: SUDFace3D-Guided Fine-Tuning

Stage 2 in MemoryTalker is a contrastive-style adaptation step that learns a per-speaker style modulation on top of the frozen Stage 1 motion memory $\mathbf{M}_m$ (Section 5). We use Stage 2 to adapt the VOCASET-trained Stage 1 checkpoint to Turkish: the 35 SUDFace3D training speakers supply the contrastive cohort, and the 10 held-out test speakers measure cross-lingual generalisation. We compare all four Stage 1 backbones identified in Section 4—MMS-1b (L34), XLS-R-300M (L18), mHuBERT-147 (L9), and HuBERT-base (CTC)—and keep the audio encoder frozen throughout Stage 2; only the contrastive style head E_s, the style-modulation parameters, and the decoder are trainable. The 32-slot motion memory is inherited unchanged from VOCASET; the contrastive style head is trained from scratch on SUDFace3D speaker triplets.

Table 6 shows that SUDFace3D Stage 2 transfer sharply improves the VOCASET Stage 1 zero-shot results in Table 5, bringing LVE from the 0.83–0.95

Table 7: Mean per-vertex L2 error (mm) on Turkish-vowel frames of the SUDFace3D test split. Stage 1: in-domain training on SUDFace3D from scratch; Stage 2: contrastive FT on a frozen VOCASET Stage 1. *Macro* = unweighted mean over the three vowels; best per column across all rows in bold.

Encoder (layer)	/ɯ/ (ı)	/y/ (ü)	/ø/ (ö)	Macro
Stage 1: in-domain training on SUDFace3D				
MMS-1b (L34)	0.646	0.627	0.628	0.634
XLS-R-300M (L18)	0.558	**0.538**	**0.542**	**0.546**
mHuBERT-147 (L9)	0.639	0.625	0.626	0.630
HuBERT-base (CTC)	**0.549**	0.541	0.556	0.549
Stage 2: contrastive FT on a frozen VOCASET Stage 1				
MMS-1b (L34)	**0.544**	0.533	0.550	0.542
XLS-R-300M (L18)	0.546	**0.530**	**0.543**	**0.540**
mHuBERT-147 (L9)	0.549	0.544	0.568	0.554
HuBERT-base (CTC)	0.561	0.570	0.590	0.574

zero-shot range down to 0.30–0.32 for the three multilingual encoders (0.37 for HuBERT-base). XLS-R-300M (L18) achieves the best FVE, LVE, and LDTW; MMS-1b (L34) is best on MOD; HuBERT-base (CTC) is best on FDD; mHuBERT-147 (L9) is not the column-winner on any metric.

All four runs in Table 6 keep the audio encoder frozen and update only the style head, the style-modulation parameters, and the decoder. Despite this minimal update, the three multilingual encoders close most of the cross-language gap: LVE drops from the 0.8437–0.9520 zero-shot range (Table 5, top block) to 0.3006–0.3223 after Stage 2 adaptation, confirming that the Stage 1 motion memory is articulatorily reusable for Turkish once a multilingual audio encoder is placed underneath. XLS-R-300M takes the lead on the main mesh metrics (FVE, LVE, LDTW), with MMS-1b slightly better on MOD and HuBERT-base on the dynamics metric FDD; mHuBERT-147 is competitive across the board at roughly 10× fewer parameters than MMS-1b. HuBERT-base, the only English-only encoder, lags the multilingual group on every lip/mesh metric (LVE 0.3747), the same English-bias signal that recurs in the Turkish-vowel analysis (Table 7).

6.5 Spatial Error Analysis on Turkish-Specific Vowels

To analyze performance on Turkish-specific vowels, we localise each Turkish-specific vowel via CTC-forced phoneme alignment and, to account for coarticulation, evaluate per-vertex L2 mesh error over a temporal window that extends 0.1s before and 0.2s after each phoneme, yielding 17,800 test frames in total. Table 7 reports the macro and per-phoneme means for all four encoders at both stages. Stage 2 contrastive fine-tuning improves on in-domain Stage 1 for the three multilingual backbones—macro drops by 0.006 mm for XLS-R-300M, 0.076 mm for mHuBERT-147, and 0.092 mm for MMS-1b—but *worsens*

HuBERT-base on every vowel (0.549 → 0.574 mm macro), exposing the same English-bias signal as in Table 6: a VOCASET-trained motion memory pulled into the Turkish domain through an English-only encoder is harder to retarget than through a multilingual one. XLS-R-300M in Stage 2 holds the overall best macro at 0.540 mm and the best per-phoneme score on /y/, while Stage 1 XLS-R-300M is marginally best on /ø/ (0.542 vs. 0.543 mm); MMS-1b in Stage 2 is essentially tied at 0.542 mm macro and wins on /ɯ/—confirming that, for multilingual encoders, the cheap Stage 2 route generally matches or improves an in-domain Stage 1 retraining at the macro level, and is the preferred option when the target-language data budget is small.

7 Conclusion

We studied how speech-driven 3D facial animation transfers from English to Turkish, and proposed a parameter-efficient approach for closing the gap. We introduced SUDFace3D, the first dedicated, controlled Turkish 3D talking-face benchmark, with frame-level FLAME annotations for all 50 SUDFace speakers (Section 3). A multilingual phoneme probe on FLEURS-TR/EN identified *mHuBERT-147* as the strongest compact backbone at the representational level, with the highest overall and Turkish phoneme accuracy at ∼95 M parameters, an order of magnitude smaller than the next-best alternative, MMS-1b (Section 4). This advantage does not carry to the mesh level: after Stage-2 adaptation the main mesh metrics go to XLS-R-300M (FVE, LVE, LDTW) and MMS-1b (MOD), while mHuBERT-147 wins none yet stays competitive at ∼10× fewer parameters, making it the preferred choice on an accuracy/cost trade-off rather than on raw mesh quality. Sweeping these encoders through both MemoryTalker stages, we found that swapping the audio encoder alone does not close the zero-shot mesh gap (LVE 0.83–0.95 across all four backbones); the bottleneck lies in the VOCASET-trained motion memory and audio–geometry head. A frozen-encoder Stage-2 contrastive fine-tuning on SUDFace3D speaker triplets, with the motion memory inherited unchanged, drops LVE to 0.30–0.32 for the multilingual encoders (a ∼3× reduction), recovering most of the gap at minimal trainable-parameter cost. Training Stage 1 directly on SUDFace3D yields comparable mesh quality but requires retraining the motion memory. The Stage 2 route is therefore preferable when the target-language data budget is small.

More broadly, English-trained motion memories appear to be articulatorily reusable for non-English target languages, provided the audio encoder underneath is multilingual and a small target-language Stage-2 supervision is added. This offers a practical path to extending speech-driven 3D facial animation beyond English without scan-quality 3D data in the target language.

We chose MemoryTalker for two reasons: it reaches top results on VOCASET, and it keeps the motion memory as a separate module. This lets us change the audio encoder and the Stage 2 head one at a time, and see which part is responsible for the English bias. A natural next step is to test cross-language

transfer in models without an explicit motion memory, such as CodeTalker [32] or diffusion-based decoders.

References

1. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., Auli, M.: XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale. In: Interspeech, pp. 2278–2282 (2022)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: NeurIPS, pp. 12449–12460 (2020)
3. Boito, M.Z., Iyer, V., Lagos, N., Besacier, L., Calapodescu, I.: mHuBERT-147: A Compact Multilingual HuBERT Model. In: Interspeech, pp. 3939–3943 (2024)
4. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In: Interspeech, pp. 2426–2430 (2021)
5. Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., Bapna, A.: FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech. In: SLT, pp. 798–805 (2023)
6. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, Learning, and Synthesis of 3D Speaking Styles. In: CVPR, pp. 10101–10111 (2019)
7. Daněček, R., Black, M.J., Bolkart, T.: EMOCA: Emotion Driven Monocular Face Capture and Animation. In: CVPR, pp. 20311–20322 (2022)
8. Daněček, R., Schmitt, C., Polikovsky, S., Black, M.J.: Supervising 3D Talking Head Avatars with Analysis-by-Audio-Synthesis. In: 3DV (2026). arXiv:2504.13386
9. Edwards, P., Landreth, C., Fiume, E., Singh, K.: JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. ACM Trans. Graphics **35**(4), 127:1–127:11 (2016)
10. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In: CVPR, pp. 18770–18780 (2022)
11. Fan, X., Li, J., Lin, Z., Xiao, W., Yang, L.: UniTalker: Scaling up Audio-Driven 3D Facial Animation through a Unified Model. In: ECCV (2024)
12. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-D Audio-Visual Corpus of Affective Communication. IEEE Trans. Multimedia **12**(6), 591–598 (2010)
13. Göksel, A., Kerslake, C.: Turkish: A Comprehensive Grammar. Routledge, London (2005)
14. Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. Audio, Speech, Lang. Process. **29**, 3451–3460 (2021)
15. Kim, H.K., Lee, S., Kim, H.G.: MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization. In: ICCV, pp. 11241–11251 (2025)
16. Kim, S.-B., Lee, C.-Y., Son, G., Oh, H.-B., Ju, J., Nam, S., Oh, T.-H.: MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset. In: Interspeech, pp. 1380–1384 (2024)
17. Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., Rigoll, G.: CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. In: SPECOM, pp. 267–278 (2020)

18. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a Model of Facial Shape and Expression from 4D Scans. ACM Trans. Graphics **36**(6), 194:1–194:17 (2017)
19. Nocentini, F., Besnier, T., Ferrari, C., Arguillère, S., Berretti, S., Daoudi, M.: ScanTalk: 3D Talking Heads from Unregistered Scans. In: ECCV (2024)
20. Nocentini, F., Seo, K., Liu, Q., Ferrari, C., Berretti, S., Ferman, D., Kim, H., Garrido, P., Caliskan, A.: Polyglot: Multilingual Style Preserving Speech-Driven Facial Animation. arXiv:2604.16108 (2026)
21. Peng, Z., Wu, H., Song, Z., Xu, H., Zhu, X., He, J., Liu, H., Fan, Z.: EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. In: ICCV, pp. 20687–20697 (2023)
22. Peng, Z., Luo, Y., Shi, Y., Xu, H., Zhu, X., Liu, H., He, J., Fan, Z.: SelfTalk: A Self-Supervised Commutative Training Diagram to Comprehend 3D Talking Faces. In: ACM MM, pp. 5292–5301 (2023)
23. Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., Auli, M.: Scaling Speech Technology to 1,000+ Languages. J. Mach. Learn. Res. **25**(97), 1–52 (2024)
24. Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y.: MeshTalk: 3D Face Animation from Speech Using Cross-Modality Disentanglement. In: ICCV, pp. 1173–1182 (2021)
25. Roach, P.: English Phonetics and Phonology: A Practical Course. 4th edn. Cambridge University Press, Cambridge (2009)
26. Şentürk, Y.D., Tavacıoğlu, E.E., Duymaz, M.İ., Sayım, B., Alp, N.: The Sabancı University Dynamic Face Database (SUDFace): Development and Validation of an Audiovisual Stimulus Set of Recited and Free Speeches with Neutral Facial Expressions. Behavior Research Methods **55**(6), 3078–3099 (2023)
27. Stan, S., Haque, K.I., Yumak, Z.: FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. In: ACM SIGGRAPH MIG, pp. 1–11 (2023)
28. Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.-H., Yu, M., Liu, Y.-J.: Diff-PoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models. ACM Trans. Graphics **43**(4), 46:1–46:9 (2024)
29. Taylor, S.L., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J.K., Matthews, I.: A Deep Learning Approach for Generalized Speech Animation. ACM Trans. Graphics **36**(4), 93:1–93:11 (2017)
30. Thambiraja, B., Aliakbarian, S., Cosker, D., Thies, J.: 3DiFACE: Diffusion-Based Speech-Driven 3D Facial Animation and Editing. arXiv:2312.00870 (2023)
31. Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized Speech-Driven 3D Facial Animation. In: ICCV, pp. 20621–20631 (2023)
32. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.-T.: CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In: CVPR, pp. 12780–12790 (2023)
33. Xu, Q., Baevski, A., Auli, M.: Simple and Effective Zero-Shot Cross-Lingual Phoneme Recognition. In: Interspeech, pp. 2113–2117 (2022)
34. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical Audio-Driven Visual Synthesis for Portrait Image Animation. arXiv:2406.08801 (2024)
35. Zhu, Q., Cao, X., Wang, Z., Zheng, Y., Taketomi, T.: KaoLRM: Repurposing Pre-trained Large Reconstruction Models for Parametric 3D Face Reconstruction. In: 3DV (2026). arXiv:2601.12736