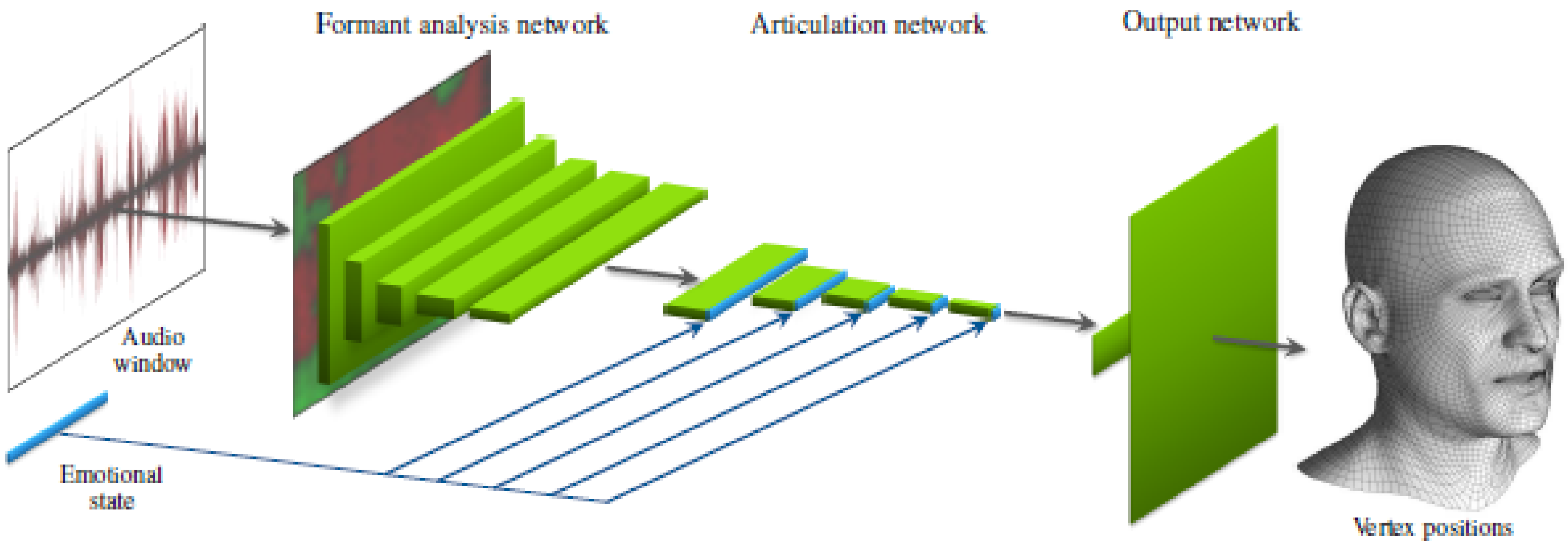
Speech-Driven 3D Stylish Avatar Animation by End-to-End Learning of Visemes Coefficients with Weak Loss Function Based on Landmarks Motions

Nurlanov Zhakshylyk, Ilya Krotov, Ivan Glazistov, Victor Lempitsky
MIPT, Scoltech, Samsung SRR, Samsung AI Center
2018

Plan:

1. Introduction and Related Works
2. Available Datasets
3. Our approach
4. Short report
5. Results



Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion, Karras et al., NVIDIA, 2017

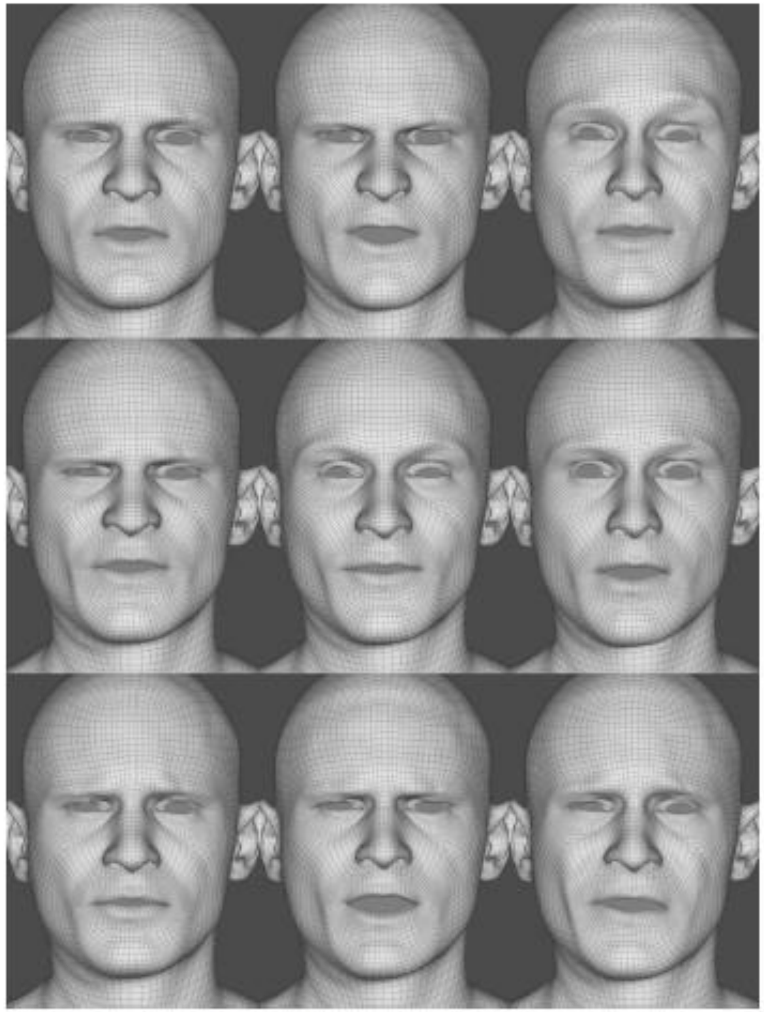


Fig. 6. The emotional state has a large effect on the animation, as shown on the accompanying video. These nine poses are inferred from the same audio window using different emotion vectors.

Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion, Karras et al., NVIDIA, 2017

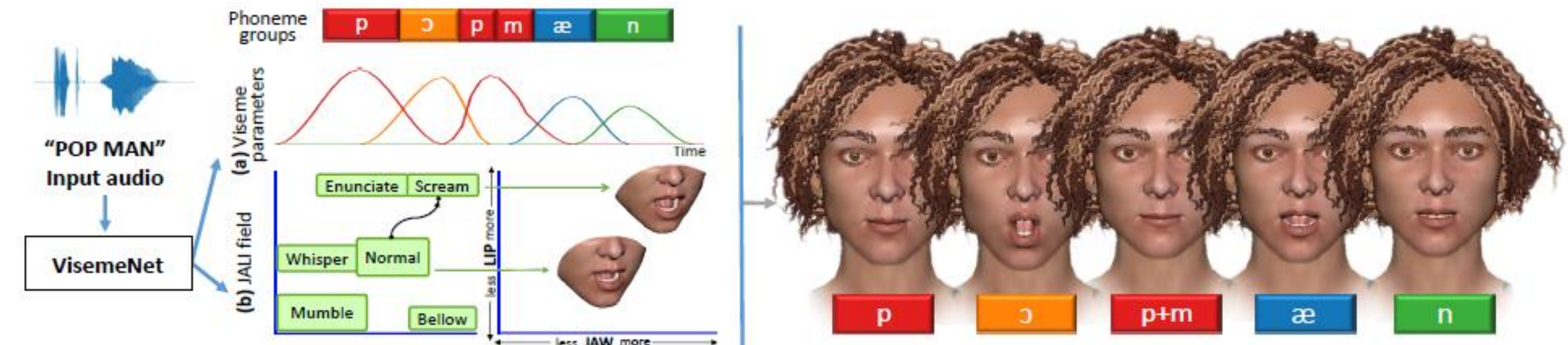


Fig. 1. VisemeNet is a deep-learning approach that uses a 3-stage LSTM network, to predict compact animator-centric viseme curves with proper co-articulation, and speech style parameters, directly from speech audio in near real-time (120ms lag).

VisemeNet: Audio-Driven Animator-Centric Speech Animation, Zhou et al., University of Massachusetts Amherst, 2018

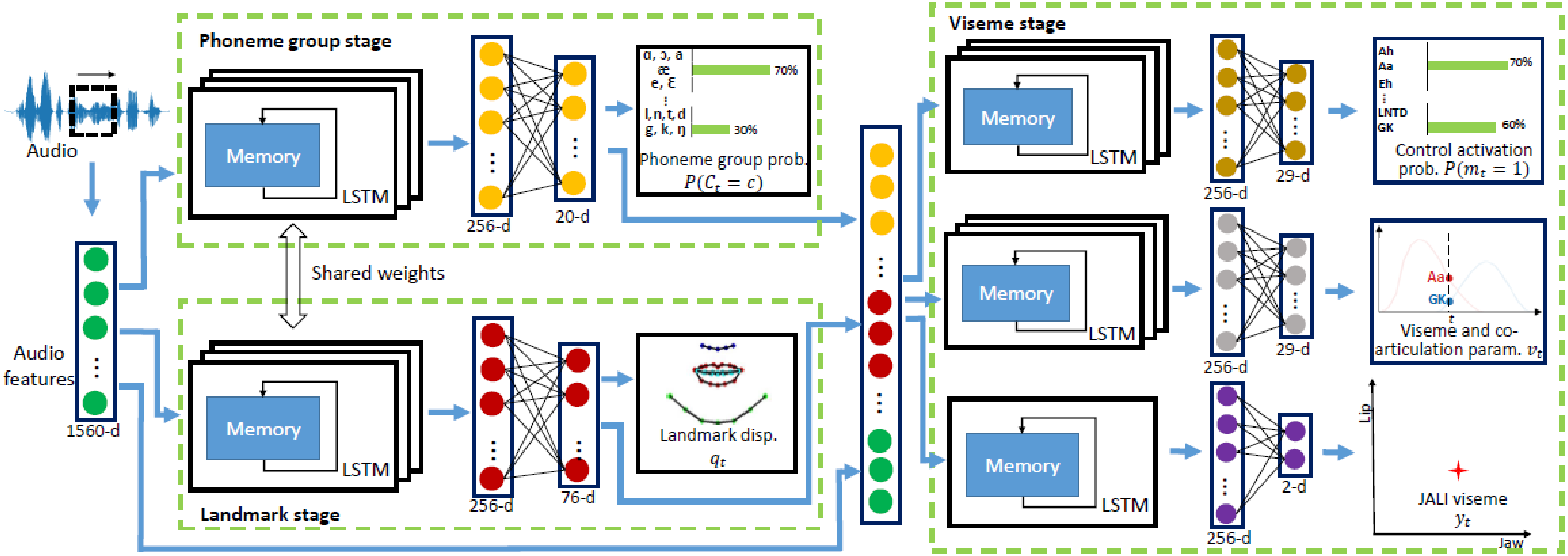
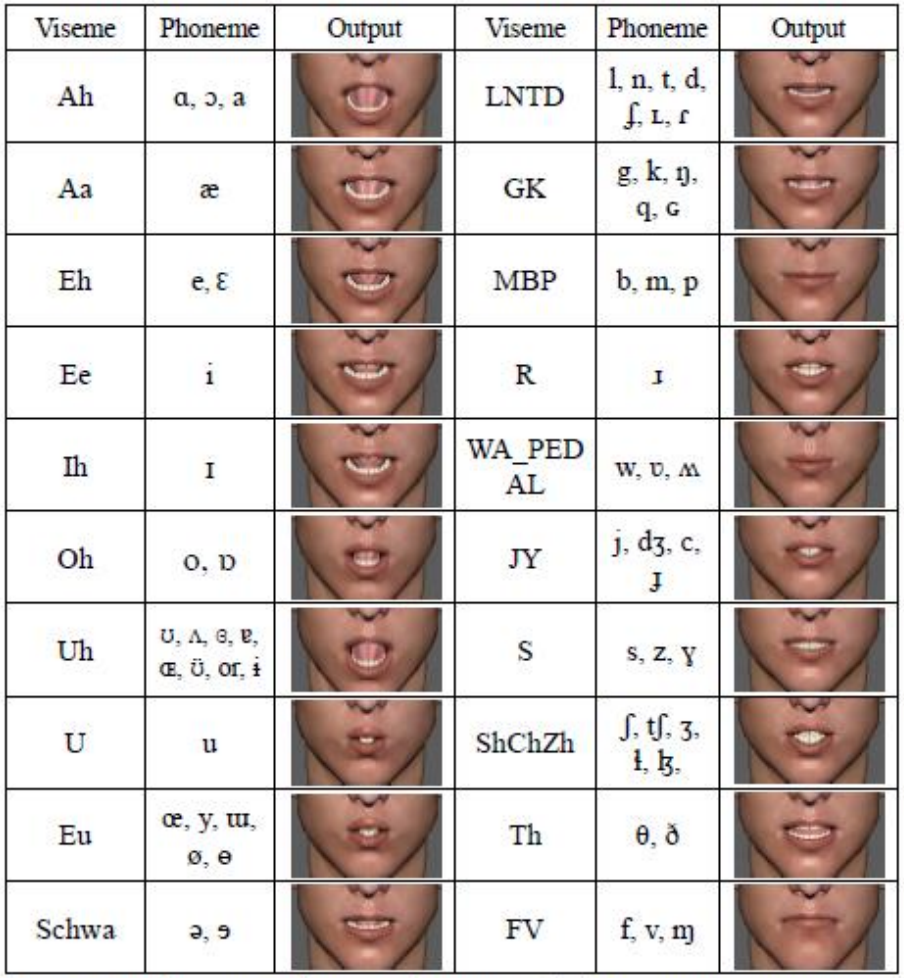


Fig. 3. Our architecture processes an audio signal (left) to predict JALI-based viseme representations: viseme and co-articulation control activations (top right), viseme and co-articulation rig parameters (middle right), and 2D JALI viseme field parameters (bottom right). Viseme prediction is performed in three LSTM-based stages: the input audio is first processed through the phoneme group stage (top left) and landmark stage (bottom left), then the predicted phoneme group and landmark representations along with audio features are processed through the viseme stage.

$$L_1(\theta, \phi, \omega) = w_c \; L_c(\theta, \phi) + w_q L_q(\theta, \omega) + w'_q L'_q(\theta, \omega) \qquad (1)$$

where weights of the three losses are set as $w_c = 0.75$, $w_q = 0.25$, $w'_q = 0.1$ in all our experiments, computed via hold-out validation.

VisemeNet: Audio-Driven Animator-Centric Speech Animation, Zhou et al., University of Massachusetts Amherst, 2018



Viseme	Phoneme	Output	Viseme	Phoneme	Output
Ah	ɑ, ɔ, a		LNTD	l, n, t, d, ʃ, ʟ, ɾ	
Aa	æ		GK	g, k, ŋ, q, ɢ	
Eh	e, ɛ		MBP	b, m, p	
Ee	i		R	ɹ	
Ih	ɪ		WA_PEDAL	w, ʋ, ʍ	
Oh	o, ɒ		JY	j, dʒ, c, ɟ	
Uh	ʊ, ʌ, ɵ, ɐ, ɶ, ö, oɾ, ɨ		S	s, z, ɣ	
U	u		ShChZh	ʃ, tʃ, ʒ, ɬ, ɮ,	
Eu	œ, y, ɯ, ø, ɵ		Th	θ, ð	
Schwa	ə, ɘ		FV	f, v, ɱ	

Fig. 2. List of visemes along with groups of phonemes (in International Phonetic Alphabet format) and corresponding lower face rig outputs that our architecture produces.

VisemeNet: Audio-Driven Animator-Centric Speech Animation, Zhou et al., University of Massachusetts Amherst, 2018

Figure 1: The proposed end-to-end face synthesis model, capable of producing realistic sequences of faces using one still image and an audio track containing speech. The generated sequences exhibit smoothness and natural expressions such as blinks and frowns.

The accuracy of the spoken message is measured using the word error rate (WER) achieved by a pre-trained lip-reading model. We use the LipNet model [2], which surpasses the performance of human lipreaders on the GRID dataset. For bot h content metrics lower values indicate better accuracy.

Method	PSNR	SSIM	FDBM	CPBD	ACD	WER
Ground Truth Videos	N/A	N/A	0.121	0.281	$0.74 \cdot 10^{-4}$	21.40%
L_1 loss	28.47	0.859	0.101	0.210	$0.90 \cdot 10^{-4}$	27.90%
$L_1 + Adv_{img}$	27.71	0.840	0.114	0.274	$1.04 \cdot 10^{-4}$	27.94%
$L_1 + Adv_{img} + Adv_{seq}$	27.98	0.844	0.114	0.277	$1.02 \cdot 10^{-4}$	**25.45%**

End-to-End Speech-Driven Facial Animation with Temporal GANs, Vougioukas et al., iBUG Group, Imperial College London, Samsung AI Centre, Cambridge, UK, 2018

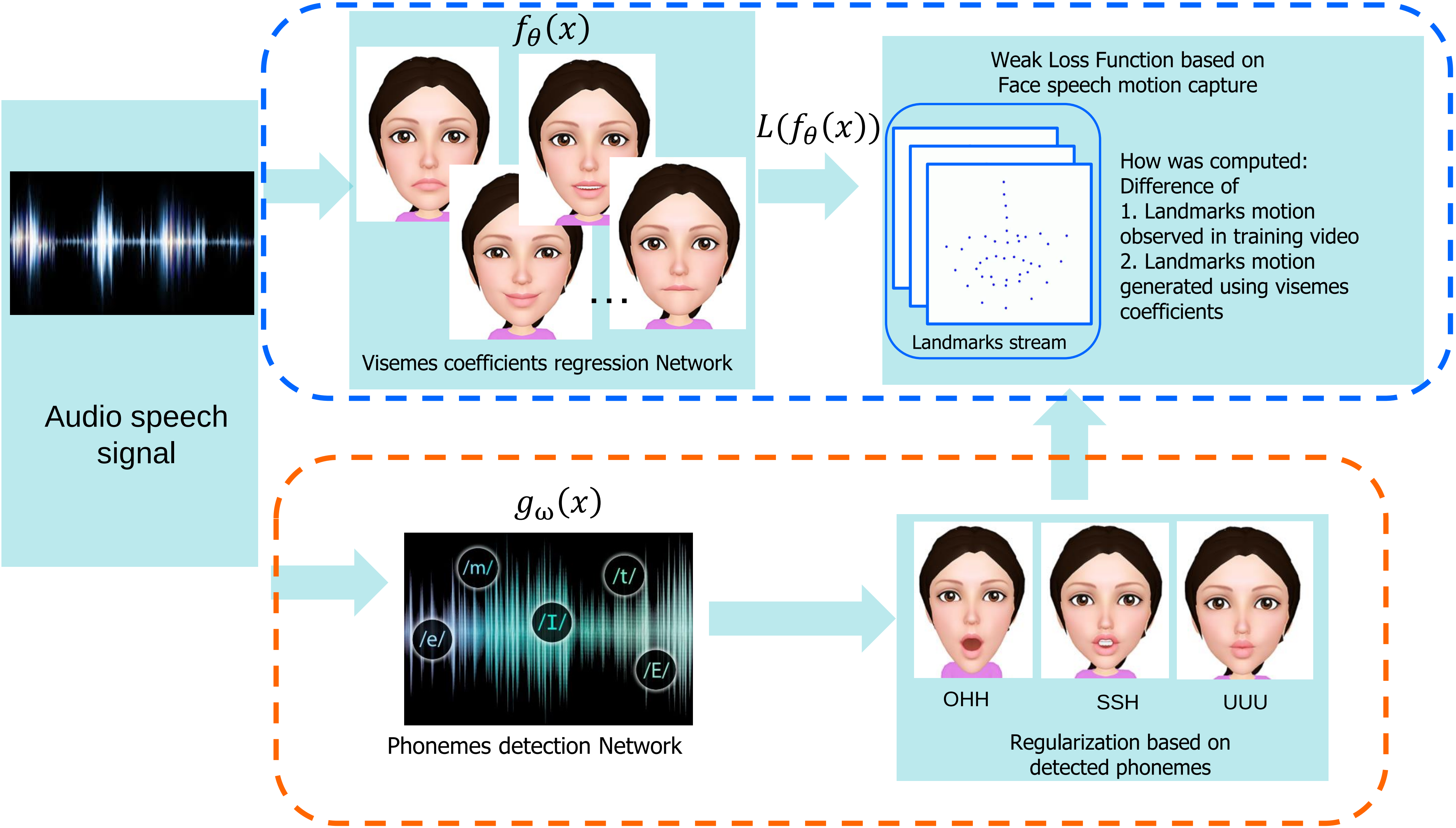
Datasets

GRID	SAVEE	BIWI 3D	TCD-TIMIT	FaceWarehouse	RAVDESS	eNTERFACE'05
Cooke 2006	Wang 2010	Fanelli 2010	Harte 2015	Chen Cao 2012	Livingstone 2018	Martin 2005
t + a + v	t + a + v	t + a + v + RGBD	t + a + v	RGBD + emotions + a set of facial feature points	t + a + v (speech, song)	Audio-Visual Emotion Database
1000 sentences	480 sentences	1109 sentences	6913 sentences		2 sentences	
34 speakers (18 m, 16 f)	4 male speakers	14 speaker (6 m, 8 f)	62 speakers		24 speaker (Professional actors)	42 speakers
14 hours	30 min	1 hour			7356 recordings	
neutral style	5 emotions	various emotional and neutral style	Neutral style, 3 lipreaders, straight on and 30* angle		8 emotions with emotional intensity	6 emotions
freely available for research use	available free of charge for research purposes. Register.	The database can be obtained upon request, **for research purposes only**. A license agreement must first be signed (no students) and sent to .	available to members of collaborating academic institutions.	available for academic research purposes. We only send the data to approved researchers	free of charge and without restriction from the open access repository Zenodo	available, free of charge, for research purposes only

TABLE I
LIST OF ENGLISH-LANGUAGE AVSR DATABASES (SOME INFORMATION TAKEN FROM TABLES IN [7], [11], AND [12]) (SR = Speech Recognition)

Database Acronym	Speakers # (# Female)	Content e.g. isolated words	Video Resolution, FPS	Stated Purpose
AMP/CMU [13]	10 (3 F)	78 isolated words	720x480	N/A
AVletters [14]	10 (5 F)	Alphabet set	80x60, 25fps	Letter recognition
AV-TIMIT [15]	223 (106 F)	TIMIT-SX sentences	720x480, 30fps	Continuous SR
AVICAR [16]	86 (40 F)	Digits, TIMIT sentences	720x480, 30fps	SR in a car
AVOZES [4]	20 (10 F)	Digits, continuous speech	720x480, 30fps	Continuous SR
BANCA [17]	208 (104 F)	Numbers, names, addresses	720x576, 25fps	Speaker verification
CUAVE [18]	36 (17 F)	Digits	720x480, 30fps	Speaker-independent digit recognition
DAVID [19]	258 (126 F)	Digits, alphabet, syllables and phrases	560x480, 25fps	Speaker/SR
GRID [20]	36 (16 F)	Command sentences	720x576, 25fps	Small-vocab CSR
IBM LVCSR [21]	290	Continuous speech	740x480, 30fps	LVCSR
VidTIMIT [22]	43 (19 F)	TIMIT sentences	512x384, 25fps	AVCSR
Valid [23]	106	Digit strings + sentence	720x576, 25fps	Speaker/SR
TULIPS1 [24]	12 (3 F)	First 4 English digits	100x75, 30fps	Isolated digits
XM2VTS [25]	295	Digit strings + sentence	720x576, 25fps	Speaker/SR
QuLips [8]	N/A	Digit strings + sentence	720x576, 25fps	
CMU-AVPFV [26]	10	150 isolated words	640x480, 30fps	Profile vs front view lip features
HIT-AVDB-II [12]	30 (15 F)	Digits, English and Chinese phrases	N/A	View angle for speaker and SR
LiLIR [27]	1	200 sentences	N/A	View angle for SR
WAPUSK20	20 (9 F)	Command sentences	640x480, 32fps	
BAVCD [28]	15	Connected digits	640x480, 20fps	Visual and depth feature examination
UNMC-VIER [29]	123 (49 F)	digits, TIMIT sentences	708x640, 25fps	Environments and SR
AusTalk [30]	1000	Digits, isolated words, SCRIBE sentences	640x480, 48fps	Speaker/SR

Our approach: FACS (Visemes) training with weak loss function based on landmarks motions



Short report

Our solution consists of “Phonemes-Landmarks Predictor” and “Visemes Post-processing module”.

“Phonemes-Landmarks Predictor” takes audio features as input. If given an audio signal, we extract a feature vector for each frame. Our feature vector concatenates 13 Mel Frequency Cepstral Coefficients (MFCCs) that have been widely used for speech recognition, 26 raw Mel Filter Bank (MFB) features, and finally 26 Spectral Subband Centroid. Features are extracted every 10 ms. The frequency analysis is performed within windows of size 25 ms in the input audio. Finally, input vectors for each frame overlay 250 ms, and has dimensions 24x65.

“Phonemes-Landmarks Predictor” outputs visemes coefficients (which are then transformed into landmarks) and phoneme group probabilities for each frame.

In next stage “Visemes Post-processing module” mixes phonemes and raw visemes and outputs final visem mesh coefficients.

Short report

As ground truth we used GRID dataset, for landmarks detection we used SRR-design detector, for phonemes estimation from text we used Montreal Aligner.

For all samples in dataset we estimated landmarks shifts from base frame to current frame. And using predefined matrix, we converted predicted visemes coefficients to these shifts.

On training stage we construct multi term loss function. First term calculated as mean square error for predicted landmarks and ground truth landmarks. Second and third terms calculates as finite differences of the first and second orders respectively.

$$\text{Overall\_loss} = \left||l - l^*|\right| + \lambda_1 \sum_{k=0}^{n=18} w_i (1 - p_i) + \lambda_2 MSE(\dot{l}, 0) + \lambda_3 MSE(\ddot{l}, 0)$$

Results

- Phonemes accuracy is equal to 85%.
- The Inter Ocular Distance error is equal to 1.5%. (< 5% is normal)
- The network infers with a lag of 130 ms (~650 ms for Visemenet)
- Comparison with other state-of-the-art models is being prepared