Assessing the Visual Speech Perception of Sampled-Based Talking Heads

Paula D. Paro Costa, José Mario De Martino

Department of Industrial Automation and Computer Engineering
School of Electrical and Computer Engineering
University of Campinas, Campinas, São Paulo, Brazil
paula@dca.fee.unicamp.br, martino@dca.fee.unicamp.br

Abstract

Focusing on flexible applications for limited computing devices, this paper investigates the improvement on the visual speech perception obtained by the implicitly modeling of coarticulation on a sample-based talking head that is characterized by a compact image database and a morphing visemes synthesis strategy. Speech intelligibility tests were applied to assess the effectiveness of the proposed context-dependent visemes (CDV) model, comparing it to a simpler model that does not handle coarticulation. The results show that, when compared to the simpler model, the CDV approach improves speech intelligibility in situations in which the audio is degraded by noise. Moreover the CDV model achieves 80% to 90% of visual speech intelligibility of video of a real talker in the tested cases. Additionally, when the audio is heavily degraded by noise, the results suggest that the mechanisms that explain visual speech perception depends on the quality of the audible information.

Index Terms: facial animation, sample-based, 2D, speech intelligibility

1. Introduction

Speech synchronized facial animation systems, or talking heads, represent a compelling technology to obtain more natural and efficient human-computer interfaces for mobile computing devices, that give users access to powerful services and resources through small displays and limited input mechanisms.

Our research focuses on the development of a customizable and compact videorealistic talking head that can be deployed on small mobile devices with limited memory and processing capabilities like wrist gadgets, home appliances, toys and mobile phones.

In order to avoid the higher computational costs of manipulating and rendering sophisticated 3D polygonal face meshes, we propose an inherently photorealistic image-based, or 2D, facial animation synthesis approach. Implementations of 2D talking heads can be typically divided in three different synthesis strategies: morphing of visemes, concatenative synthesis and machine learning based algorithms.

Morphing-based systems are characterized by a tiny image database that spans most relevant mouth configurations of a language (visemes). Visemes selected from the database define key-poses of the final animation and the frames between adjacent key-poses are synthesized through an image morphing algorithm. In order to reduce the size of the image database, the modeling of visible speech articulatory movements is overly simplified and, typically, coarticulation effects are not properly modeled. A representative work of this approach is *Miketalk* [1].

The concatenative synthesis approach is characterized by improved levels of videorealism at the cost of storing a large database of video fragments and numerous pre-processing and labeling steps. The synthesis of a new animation, in a process that can be called "video rewrite" [2], is realized by concatenating and stitching together small sequences of frames selected from the database based on their visual features and the phonetic context they reproduce.

In systems that implement the machine learning strategy, the corpus of sample images is the training set used to parameterize the image database space and to learn the allowed trajectories in this space. The facial animation delivered by these systems typically presents smooth transitions and videorealistic results [3], [4], however, the training phase must be repeated for each new face model, or avatar.

Focusing on the flexibility of easily generating new avatars and applications for limited computing devices, our research strives to improve the videorealism, and therefore the visual speech perception, achieved by a pure 2D morphing-based implementation while keeping its image database small and applying a straightforward synthesis algorithm that does not require a training phase.

To improve the videorealism, we propose the use of an image database of context-dependent visemes. Visemes, or "visual phonemes", can be defined as the typical articulation pattern that carries the visual cues to distinguish the various phonemes of a language. Based on an extension of this definition, De Martino et al. [5] identified that, due to coarticulation, some visemes may present small perceptible variations in their dynamics depending

on their phonetic context. The mapping of these variations for the most common phonetic contexts of a language gives origin to the definition of context-dependent visemes (CDVs).

Our study is performed comparing two different approaches to model visible speech. The first approach is characterized by an image database of 20 simple visemes (SV model) that implements a one-to-one mapping between a set of Brazilian Portuguese phonemes and a set of visemes represented by pictures of the mouth and chin regions of a real face (see Figures 1 and 2). The second approach adapts the context-dependent visemes (CDV) modeling to create a 34 context-dependent visemes database for Brazilian Portuguese that is capable of implicitly modeling the most common speech coarticulation effects for this language.

For talking heads, there is no universally accepted criteria to measure videorealism, which is the capacity of a facial animation to be confused with the video of a real face. Both verbal and non-verbal communication aspects are typically assessed by subjective tests and the so called, "Turing tests". However, poor results in subjective tests typically do not provide useful feedback to improve the synthesis methodology. On the other hand, objective tests based on the comparison between original and synthesized articulators trajectories provide results that cannot be directly related to the human observers perception. Speech intelligibility tests provide objective measurements while taking the user perception into consideration. We argue that this type of evaluation provides useful feedback to compare and track the progress of different visible speech models as shown, for example, by Benoît and Le Goff [6], Beskow et al. [7] and Ouni et al. [10].

The performance of the SV and CDV models was assessed through speech intelligibility tests and their scores were compared to the unimodal auditory and bimodal (audiovisual) real talker stimuli in different conditions of audio degradation. The results obtained so far show that the CDV modeling is able to improve speech intelligibility in all tested situations. Additionally, the comparison between CDV and SV models revealed that in the situation where the speech audio is heavily degraded by noise, the intelligibility scores for both models are similar. These results contribute with new evidence that suggests that the visual speech perception mechanisms are influenced by the quality of audible information.

The carried out evaluation also shows the importance of performing objective test protocols to evaluate and compare different visible speech models. As can be shown by our results, this type of test is capable of pointing new directions for improvements on the model.

2. Coarticulation Modeling based on Context-Dependent Visemes

In [5], De Martino et al. devised a context-dependent viseme study for Brazilian Portuguese. Using motion capture techniques, the visual motor pattern of different homophenous groups[1] were measured and analyzed under different phonetic contexts. Applying a clusterization algorithm, the study objectively identified different visemes that can be associated to the same homophenous group, depending on the phonetic context in which they are produced. Despite the fact the study was conducted for Brazilian Portuguese, its underlying principles can be applied to any language.

Tables 1 and 2 summarize the context-dependent visemes, or CDVs, identified in [5]. The first column of each table presents the homophenous groups considered and, for simplicity, the context-dependent visemes are named according to the first phone of each homophenous group. The third column shows the phonetic contexts associated to each context-dependent viseme.

In the present work, the CDVs are represented by an image database of 34 photographic images corresponding to 22 consonantal context-dependent visemes (second column of Table 1, see Figure 1) added to 11 vocalic visemes (Table 2, see Figure 2) and a viseme representing the posture of lips when no speech is uttered (called silence viseme).

In order to evaluate the effective contribution of the implicitly coarticulation modeling provided by the CDV image database, we reduced the original database of 34 visemes to a subset of 20 visemes that univocally associates each homophenous group of Tables 1 and 2 to a viseme. In other words, in this simple visemes (SV) database, the influence of phonetic context is not taken into account and the database does not model coarticulation. The SV model was built with visemes that cover the greater number of phonetic contexts, which correspond to visemes most resistant to coarticulation: $< p_1 >$, $< f_1 >$, $< t_1 >$, $< s_2 >$, $< l_1 >$, $< \int_1 >$, $< \Lambda_1 >$, $< k_1 >$, $< \gamma_1 >$, $< i_1 >$, $< e >$, $< \varepsilon >$, $< a >$, $< \text{ɔ} >$, $< o >$, $< u >$, $< \text{ɪ} >$, $< \text{ɐ} >$, $< \text{ʊ} >$. The twentieth viseme corresponds to the silence viseme. In Figs. 1 and 2, the SV database is represented by the visemes with a bold frame around them. It is important to note that this database implementation follows a strategy similar to the one implemented in MikeTalk [1].

Consider, for example, the synthesis of the logatome "pupu", or [pupʊ]. When applying the CDV model, the corresponding sequence of speech animation key-visemes is $< p_2 u p_2 \text{ʊ} >$. Considering the SV model, the animation is synthesized using the key-visemes: $< p_1 u p_1 \text{ʊ} >$.

[1] Homophenous sounds are phones that share the same place of articulation and that are not distinguishable by visual cues alone.

Homophenous Group	Visemes	Phonetic Contexts
[p,b,m]	$< p_1 >$	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]
	$< p_2 >$	[pu] [upɪ] [upʊ]
[f,v]	$< f_1 >$	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]
	$< f_2 >$	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]
[t,d,n]	$< t_1 >$	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]
	$< t_2 >$	[ta] [atɐ]
[s,z]	$< s_1 >$	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]
	$< s_2 >$	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]
[l]	$< l_1 >$	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]
	$< l_2 >$	[la] [ilɐ] [alɪ] [alɐ]
	$< l_3 >$	[lu]
	$< l_4 >$	[ilʊ] [ulʊ]
[ʃ,ʒ]	$< ʃ_1 >$	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]
	$< ʃ_2 >$	[ʃu] [uʃʊ]
[ʎ,ɲ]	$< ʎ_1 >$	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]
	$< ʎ_2 >$	[ʎu] [uʎɪ] [uʎɐ]
	$< ʎ_3 >$	[iʎʊ] [aʎʊ] [uʎʊ]
[k,g]	$< k_1 >$	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]
	$< k_2 >$	[ka] [akɐ]
	$< k_3 >$	[ku] [ikʊ] [akʊ] [ukʊ]
[ɣ],[ɾ]	$< ɣ_1 >$	[ɣi] [ɣa] [iɣɪ] [iɣɐ] [aɣɪ] [aɣɐ] [uɣɐ]
	$< ɣ_2 >$	[ɣʊ] [iɣʊ] [aɣʊ] [uɣɪ] [uɣʊ]

Table 1: Consonantal context-dependent visemes (adapted from [5])

2.1. Animation Synthesis

The timed phonetic transcription of the speech to be visually animated provides the information to select visemes from database and to associate them as key-frames of final animation. The articulatory targets represented by the key-frames were, for simplicity, assumed to be at the center of duration of each speech segment. The animated face is synthesized stitching together the selected visemes from the database to a base-face image and using a non-linear morphing process to generate the intermediate frames between adjacent key-visemes of animation. The morphing process is guided by five anchor points: two points in the center of upper and lower lips, two points on each corner of the mouth and one point on the chin. Considering the small number of feature points that guide the morphing process, radial basis functions (RBF) were adopted as the warping function since they are proven to be an effective tool in multivariate interpolation problems of scattered data [8].

Homophenous Groups	Visemes	Phonetic Contexts
[i,ĩ]	$< i_1 >$	All contexts except [tit] and [ʃiʃ].
	$< i_2 >$	[tit] and [ʃiʃ].
[e,ẽ]	< e >	All contexts.
[ɛ]	< ɛ>	All contexts.
[a,ɐ̃]	< a >	All contexts.
[ɔ]	< ɔ>	All contexts.
[o,õ]	$< o >$	All contexts.
[u,ũ]	$< u >$	All contexts.
[ɪ]	< ɪ>	All contexts.
[ɐ]	< ɐ>	All contexts.
[ʊ]	< ʊ>	All contexts.

Table 2: Vocalic context-dependent visemes (adapted from [5])

3. Visual Speech Perception Assessment

A perceptual evaluation based on speech intelligibility tests was carried out with four types of test stimuli: unimodal auditory, bimodal synthetic talking head with no coarticulation modeling (SV), bimodal synthetic talking head with coarticulation modeling based on context-dependent visemes (CDV) and a bimodal real talker.

3.1. Test Stimuli

Twenty-seven logatomes conveyed in a carrier phrase uttered by a native female speaker of Brazilian Portuguese were used as test stimuli. The carrier phrase has the following structure: "Ela fala <logatome>." ("She says <logatome>."). The words were constructed following the structure $'CVCV$ (stress on the first syllable), with C=/p, f, t, s, l, ʃ, ʎ, k, ɣ/ and V=/i, a, u/, resulting in a total of 27 logatomes. The concatenation of two CV syllables aims at stimulating the production of coarticulation effects during its utterance, in a frequent Brazilian Portuguese bisyllabic, penultimate stress structure. The set of consonants have elements of each homophenous group of Table 1.

The audio track of each recorded phrase was degraded by white noise in order to produce new audio files with three different signal-to-noise ratio (SNR) conditions: -12 dB, -18 dB and -24 dB. The generated audio track was used as unimodal auditory stimuli. The au-

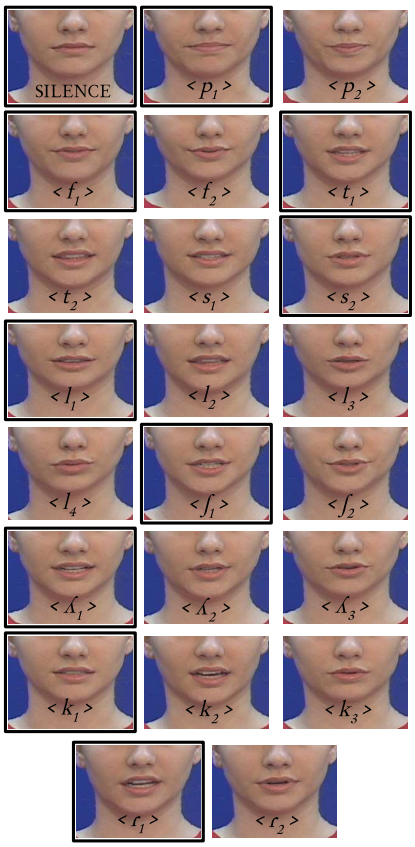


Figure 1: Silence and consonantal context-dependent visemes . The set of visemes with a bold frame form a simpler database that does not include coarticulation modeling.

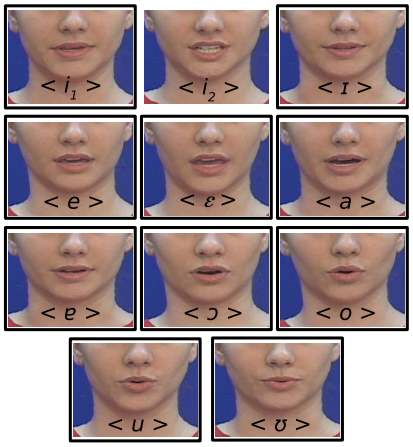


Figure 2: Vocalic context-dependent visemes. The $< i_2 >$ viseme is not part of the simple visemes (SV) image database.

diovisual real talker stimuli were constructed through the resynchronization of the degraded audio track with the original video track of the female speaker. Afterward, the timed phonetic transcription of the recorded audio was used as input to our sample-based talking head system. The facial animations synthesized under the two different versions of image database, were synchronized with the three versions of audio files degraded by noise. The entire test consisted of presenting 324 sentences (4 types of stimuli $\times$ 3 levels of audio degradation $\times$ 27 sentences).

3.2. Method

The evaluation was guided by an automated test application running on a desktop computer in a quiet room dedicated to the experiment. Each participant used a high-quality headphone to hear the audio and, for audiovisual stimuli, the height of the face on the screen was about 7 cm. For each subject, the test application was designed to randomly select a presentation sequence of 12 different types of stimuli. For each type of stimulus, the order of presentation of the 27 sentences was also randomly selected.

After the presentation of an utterance, the subject was asked to indicate the understood logatome from the 28 available options on the screen, consisting of 27 words and a “none of the above” option. After confirming their selection, the participant was able to proceed to the next stimulus. Prior to the beginning of the evaluation session, each subject was informed that the logatomes were of the $'CVCV$ type and words distinct from those shown as options could be presented. The participants were encouraged to choose an option even if they were in doubt among different words and only choose the “none of the above” option when they had absolutely no clue about what they had heard or if they believed they heard something else not provided as one of the options. The participants were allowed to interact with the test application before starting the evaluation to familiarize themselves with its operation.

The perceptual evaluation was carried out with 40 volunteer employees and students of the University of Campinas, with ages ranging from 20 to 62 years old, all native speakers of Brazilian Portuguese. The participants had no previous contact with this research. They all reported normal hearing and normal sight abilities. The duration of each test session averaged 40 minutes.

4. Results

Figure 3 plots the average percentage of correct identification of logatomes for the 4 types of test stimuli, organized according to the three different levels of audio degradation (SNR = -24 dB, -18 dB and -12 dB). The boxplots of Figure 4 depict the variation of the percentage of correct answers among the participants. The au-

diovisual synthetic talking head stimuli are identified by the acronyms CDV to identify the results from the model based on context-dependent visemes and SV to identify the results from the simple visemes model, that does not include coarticulation effects.

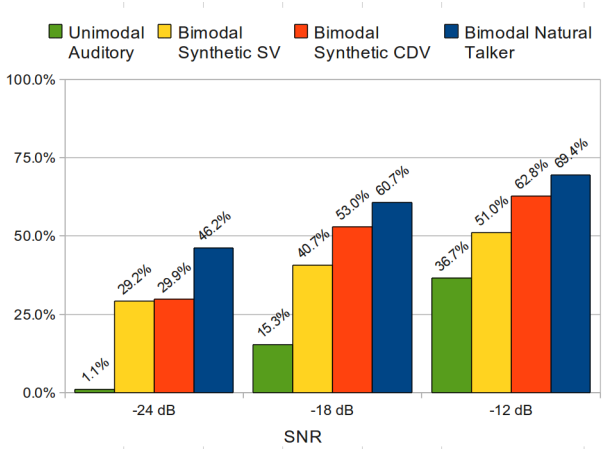


Figure 3: Average percentage of correct identification of logatomes for the 4 types of stimuli, organized according to the three different levels of audio degradation (SNR = -24 dB, -18 dB and -12 dB).

As also observed in [9], three observations can be drawn from the graph of Figure 3. First, speech intelligibility is a function of the noise level present in the audio, with a decrease in performance for lower SNR conditions. Second, the results show the contribution of visual information from the face to speech intelligibility in all tested cases. This contribution is especially evident at the SNR = -24 dB noise level where almost no intelligible audio was present, making the participants exercise lipreading. Third, at SNR = -24 dB, the synthetic talking head was able to provide a gain of approximately 30% in intelligibility compared to a gain of approximately 46% provided by the real talker.

The Wilcoxon Mann-Whitney (WMW) U test was used to perform the pairwise comparison between unimodal audio and the other types of audiovisual stimuli, resulting in a significance level $p < 0.001$ (the alternative hypothesis was that the scores from audiovisual stimuli were greater than the obtained with audio only).

On the other hand, the WMW results for the pairwise comparison of bimodal synthetic CDV model with the bimodal real talker shows that the scores of the video tend to be greater with statistical significance ($p < 0.001, U = 1414$) only at the SNR = -24 dB noise condition. In fact, it is possible to observe in Figures 3 and 4 that the speech intelligibility level of the CDV model tends to be closer to the real talker at SNR = -18 dB and -12 dB noise levels than at SNR = -24 dB.

The same behavior was not observed when comparing the bimodal synthetic SV model and the real talker, whose experimental data shows statistical significance ($p < 0.001$) for all tested SNR levels.

It is important to note that, when compared to the SV model, the synthetic CDV model presented a superior performance in the tested cases where a residual intelligible audio was present but, at SNR = -24 dB, its performance decayed more intensively than the other stimuli, turning its performance comparable to the SV model (see boxplot on Figure 4c).

As an alternative way to interpret the results from speech intelligibility tests, Ouni et al. [10] proposed a relative visual contribution (RVC) metric to measure the speech intelligibility improvement provided by a synthetic animated face relative to the improvement provided by a natural face when the acoustic channel is degraded.

The metric is defined as follows:

$$RVC = 1 - \frac{C_N - C_S}{1 - C_A} \quad (1)$$

where C_S, C_A, and C_N are the intelligibility scores of bimodal synthetic talking head, unimodal auditory and bimodal real talker respectively.

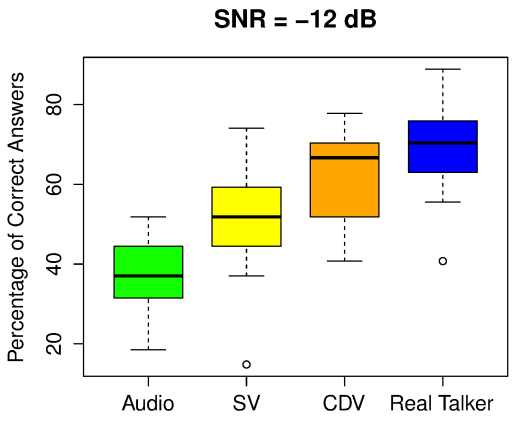
In Table 3 the average values of RVC for all participants are shown, for both the CDV and SV models for the three levels of SNR tested. Again, while the CDV model reaches approximately 90% of the visual performance of the natural face in the presence of residual audio, its performance is comparable to the SV model at SNR = -24 dB.

SNR	Simple Visemes Model	Context-Dependent Visemes Model
-12 dB	0.70	0.89
-18 dB	0.76	0.91
-24 dB	0.83	0.83

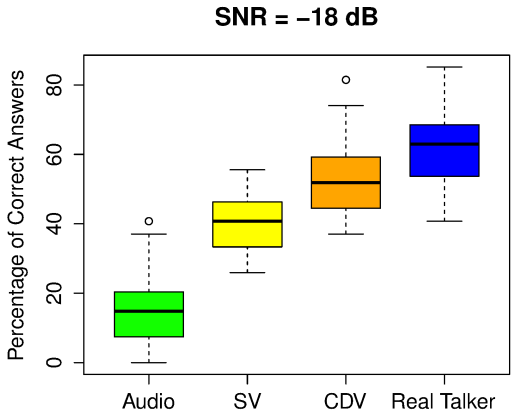
Table 3: Average Relative Visual Contribution Metric

5. Discussion

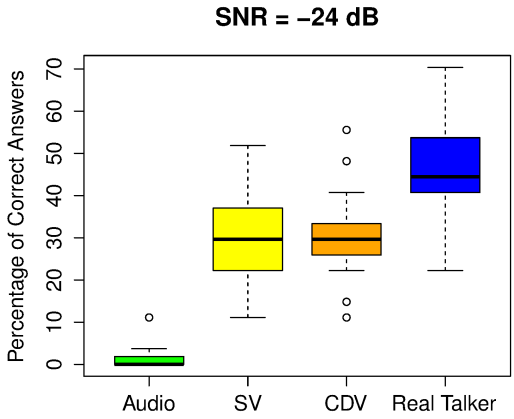
The reported perceptual evaluation was carried out with the objective of determining the contribution of the context-dependent visemes coarticulation model to the improvement of visual speech perception of a compact sample-based talking head. When comparing the CDV model with the SV model, the former clearly shows its superiority in situations where the neural supra-additive integration of visual and audio sensory inputs is taking place [11]. As pointed out by Calvert and colleagues, this integration results in enhanced responsiveness and allows the reduction of perceptual ambiguity. However, the results for the -24 dB condition (unintelligible audio information) suggest that different visual perception mechanisms are taking place and the advantage of the CDV model over the SV model cannot be observed anymore. Since the synthetic animation is generated stitch-



(a) The WMW test between the SV and CDV models, with the alternative hypothesis of CDV scores being greater than SV results is $p < 0.001, U = 1260$.



(b) The WMW test between the SV and CDV models, with the alternative hypothesis of CDV scores being greater than SV results is $p < 0.001, U = 1309$.



(c) The WMW test between the SV and CDV models, with the alternative hypothesis of CDV scores being greater than SV results is not statistical significant.

Figure 4: Boxplots of the percentage of correct answers for the different types of test stimuli.

ing visemes to a base face, additional experiments are necessary to investigate whether and how the visual information observed outside the lips and chin region affects speech intelligibility. Considering the application on small display devices, further investigation is also needed to evaluate how the screen size display may affect visual speech perception.

The evaluation provides encouraging results for the CDV model because, compared to a real face, it presents a relative visual contribution above 80% for all tested cases. The presented synthesis strategy makes it suitable for applications on platforms with limited memory and processing capabilities. Compared to the machine learning techniques that require a training phase for each new face model, the morphing visemes synthesis strategy enables the creation of new avatars through automated image processing steps of a small set of static images of a real face. The underlying principles of this model can be adapted to any language.

While the facial animation research field lacks universally adopted methods to evaluate the new proposed synthesis methodologies that arise each day, results provided by assessments like the one presented in this paper show that comparing different models may provide clear directions to improve the videorealism of talking heads and to better understand the mechanisms of visual speech perception.

6. References

[1] Ezzat, T. and T. Poggio, “Miketalk: A talking facial display based on morphing visemes”, Proc. Computer Animation’98, 96–102,1998.

[2] Bregler, C., M. Covell and M. Slaney, “Video rewrite: Driving visual speech with audio”, Proc. 24th conf. on Comp. Graphics and Interactive Techniques, 353–360, 1997.

[3] Ezzat, T., G. Geiger and T. Poggio, “Trainable videorealistic speech animation”, Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, 57–64, 2004.

[4] Wang, L. J., X. J. Qian, W. Han and F. Soong, “Synthesizing photo-real talking head via trajectory-guided sample selection”, Proc. Interspeech, 446–449, 2010.

[5] De Martino, J. M., L. P. Magalhães and F. Violaro, “Facial animation based on context-dependent visemes”, Computer & Graphics, 30(6): 971–980, 2006.

[6] Benoît, C. and B. Le Goff, “Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP”, Speech Communication, 26(1): 117–129, 1998.

[7] Beskow, J. et al.,“The Teleface project-multimodal speech communication for the hearing impaired”, Proc. Eurospeech, 97: 2003–2006, 1997.

[8] Arad, N., N. Dyn, D. Reisfeld and Y. Yeshurun, “Image warping by radial basis functions: Application to facial expressions”, CVGIP, 56(2),161–172,1994.

[9] Sumby, W. and I. Pollack,“Visual contribution to speech intelligibility in noise”, JASA 26: 212, 1954.

[10] Ouni, S., M. M. Cohen, H. Ishak and D. W., Massaro, “Visual contribution to speech perception: Measuring the intelligibility of animated talking heads.” EURASIP Journal on Audio, Speech, and Music Processing 2007, 2006.

[11] Calvert, G. A., M. Brammer and S. D. Iversen,“Crossmodal identification”, Trends in cognitive sciences 2(7): 247–253, 1998.