EUROGRAPHICS 2015 / O. Sorkine-Hornung and M. Wimmer
(Guest Editors)

Volume 34 (2015), Number 2

VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track

P. Garrido [†1] and L. Valgaerts [‡1] and H. Sarmadi[1] and I. Steiner[2] and K. Varanasi[3] and P. Pérez[3] and C. Theobalt[1]

[1]MPI for Informatics [2]Saarland University and DFKI GmbH [3]Technicolor

(a) (b) (c) (d) (e)

Figure 1: *We modify the lip motion of an actor in a target video (a) so that it aligns with a new audio track. Our set-up consists of a single video camera that films a dubber in a recording studio (b + c). Our system transfers the mouth motion of the voice actor (d) to the target actor and creates a new plausible video of the target actor speaking in the dubbed language (e).*

Abstract

In many countries, foreign movies and TV productions are dubbed, i.e., the original voice of an actor is replaced with a translation that is spoken by a dubbing actor in the country's own language. Dubbing is a complex process that requires specific translations and accurately timed recitations such that the new audio at least coarsely adheres to the mouth motion in the video. However, since the sequence of phonemes and visemes in the original and the dubbing language are different, the video-to-audio match is never perfect, which is a major source of visual discomfort. In this paper, we propose a system to alter the mouth motion of an actor in a video, so that it matches the new audio track. Our paper builds on high-quality monocular capture of 3D facial performance, lighting and albedo of the dubbing and target actors, and uses audio analysis in combination with a space-time retrieval method to synthesize a new photo-realistically rendered and highly detailed 3D shape model of the mouth region to replace the target performance. We demonstrate plausible visual quality of our results compared to footage that has been professionally dubbed in the traditional way, both qualitatively and through a user study.

1. Introduction

Dubbing is the process of replacing the original voice of an actor in a video with a voice that has been recorded off-camera in a studio. The new voice can say the exact same text in the original language, but with improved in-studio quality. However, in most cases the voice of the original actor is substituted with the voice of a different *voice actor* or *dubber* speaking in another language. Dubbing of foreign productions into the locally spoken language is common in countries where subtitling is not widely accepted, such as Germany, France and many Spanish speaking countries.

Dubbing has the advantage over subtitling that it does not draw the attention away from the action on screen. On the other hand, it has been shown that viewers are very sensi-

† e-mail: pgarrido@mpi-inf.mpg.de
‡ e-mail: valgaerts@mpi-inf.mpg.de

tive to discrepancies between the auditory signal and the visual appearance of the face and lips during speech [SP54]. In fact, mismatches between mouth motion and audio can drastically impair comprehension of the spoken language; hearing-impaired people in particular exploit this correlation [OB86, Sum92]. It is thus imperative that the dubbed language track is adjusted well to the visual performance. This requires an expensive and time consuming three-stage process performed by special production companies:

1. *Translation:* Certain mouth shapes are manually annotated in the video, such as the lip closure of the bilabial consonants /m/, /p/, and /b/. Then a transcript, which is semantically close to the original script and yet produces bilabials at roughly the same time, is made in the new language. Consequently, the translation may not be literal.
2. *Recording:* A voice actor in a studio reads out the dubbed transcript in pace with the original performance. Even recording a single sentence may need several trials until alignment with the video is satisfactory.
3. *Editing:* The temporal alignment of the new language track and the mouth motion in the video is improved by manually time-shifting and skewing the new audio.

Despite the efforts of trained professionals, traditional dubbing is unable to produce dubbed voice tracks that match the mouth movements in the target video perfectly. The reason is that spoken words differ between languages, yielding different phoneme sequences and lip motions. Hearing and seeing different languages proves very distracting for many viewers [SP54] and causes even stronger distraction for hearing-impaired persons who rely on lip reading [OB86].

In this paper, we propose a system that visually alters the lip motion as well as the facial appearance of an actor in a video, so that it aligns with a dubbed foreign language voice. We thus take a step towards reducing the strong visual discomfort caused by the audio-visual mismatch in traditionally dubbed videos. Our method takes as input the actor's and the dubber's video as well as the dubbed language track, and then it employs state-of-the-art monocular facial performance capture to reconstruct both performances. This gives us parameters describing the facial performances based on a coarse blend shape model. Via inverse rendering, we additionally reconstruct the incident scene lighting in the target video and the high-frequency surface geometry and dense albedo of the target actor. The 4D facial performance of the actor (3D geometry over time) is modified fully automatically by using a new space-time optimization method that retrieves a sequence of new facial shapes from the captured performance, such that it matches the blend shape sequence of the dubber, yet is temporally coherent, also in its fine-scale surface detail. A phonetic analysis of the dubbed audio finds salient mouth motion events, such as lip closures, which are explicitly enforced in the synthesized performance. The synthesized face sequence is plausibly rendered and lit, after which the lower half of the face is seamlessly blended into the target video to yield the final result.

In summary, our *contributions* are:

- A system for video-realistic model-based resynthesis of detailed facial performances in monocular video that aligns the visual channel with a dubbed audio signal.
- A spatio-temporal rearrangement strategy that uses the input facial performances and the dubbed audio channel to synthesize a new highly detailed 3D target performance.
- The reconstruction of a realistic target face albedo and the synthesis of a plausible mouth interior based on a geometric tooth proxy and inner mouth image warping.

Our system is one of the first to produce visually plausible and detailed, synthetically altered and relit facial performances of an actor's face. We compare our results with traditionally dubbed, unmodified video, both visually and by means of a user study. Since we synthesize the entire mouth region, we do not require that the dubbed audio perfectly aligns with the original target video. Our approach thus simplifies the dubbing pipeline, since the translation into the foreign language can now stay closer to the original script.

2. Related Work

2.1. Visual Cues in Speech Perception

Visual cues, such as *visemes*, are essential for speech perception [Sum92], both for people with normal hearing ability [OB86], but in particular for hearing-impaired persons [LK81]. In fact, under noise, one third of the speech information is conveyed visually through lip gestures [LGMCB94] and a discrepancy between sound and facial motion clearly disturbs perception [SP54]. The discrepancies between the visual and auditory cues can significantly alter the sound perceived by the observer [MM76] and this may explain why many people dislike watching dubbed content [Kil93]. Taylor et al. [TMTM12] report that a direct mapping from acoustic speech to facial deformation using visemes is simplistic and realistic synthesis of facial motion needs to model non-linear co-articulation effects [SC00]. The problem is that the statistical relationship between speech acoustics and facial configurations accounts for approximately 65% of the variance in facial motion [YRVB98], and thus the speech signal alone is not sufficient to synthesize a full range of realistic facial expressions. In view of these findings, we build the mapping from the dubber to the actor primarily using the visual signal obtained through facial performance capture. We thus achieve audio-visual coherence implicitly, which we reinforce by using the acoustic signal as a guide to enforce salient mouth motion events, like lip closures.

2.2. Speech- and Capturing-driven Animation

Our work is related to speech-driven animation of virtual CG faces or bodies. [LTK09] use a hidden Markov model to drive non-semantic body gestures from prosody features. [Bra99] proposed one of the first systems for voice puppetry,

i.e., animating a virtual avatar's face directly from speech, by modeling the joint distribution of acoustic and visual speech. However, predicting the entire range of facial motion from the speech signal is not possible [YRVB98] and the facial performance of an actor is often a better guide for motion retargeting [Wil90, XCXH03, CFKP04, PSS99]. Synthesizing 3D speech animation with learned audio-controlled activation curves for a virtual face muscle system has also been tried [SSRMF06], but plausible facial expressions and lip movements have proven to be hard, even for cartoon avatars, as human visual perception is highly attuned to facial motion and unforgiving of errors. [TMTM12] learn a set of dynamic units of visual speech from a large corpus of motion-captured examples to properly render co-articulation effects. New speech animations can also be synthesized by matching a given phoneme stream to variable- or fixed-length units in a database of motion capture data [KM03, MCP*06], sometimes guided by user-defined emotion specifiers [DN06]. We show that a strong coupling of high-quality performance capture and speech analysis also leads to plausible results with co-articulation effects. Professional film productions drive believable CG avatars, e.g., Gollum in the *Lord of the Rings* movies, with the facial performance and dialog of a dedicated actor captured with complex studio-based multi-camera systems [BHB*11]. This would be unfeasible for us, as we need to capture detailed 3D face models directly from monocular video. Motion and audio capture and transfer to a virtual CG face in real-time has been made possible using cheaper depth cameras [WBLP11] or even monocular video [CHZ14, BWP13]. However, these methods only capture models of low shape detail, do not always work on a given monocular target video and often require person-specific algorithm training under controlled conditions. They are thus not suitable for our task, for which we resort to the monocular, high-quality facial performance capture approach by [GVWT13], which captures a blend shape model and high shape detail under general uncontrolled lighting.

2.3. Video-based Face Animation And Rewriting

Our approach is related to expression synthesis methods that reorder video frames, sometimes in a model-guided way [LO11]. [KSSS10] transfer the facial poses of a source video character to a different character in a target video by rearranging and roughly aligning target video frames in a stop-motion-like fashion. Similarly, [TMCB07] use active appearance models (AAM) for real-time expression cloning, while [SDT*07] perform facial expression cloning between faces in simple static frontal poses by using a differential coordinate representation of a triangle mesh overlaid with the image. [ASWC13] used an extended AAM for image-based text-to-speech synthesis that separates global pose and local variation. None of these systems produces the spatial and temporal quality required for a detailed rendering of the target face, as well as audio-visual coherence. In this paper, we use AAM tracking [SLC11] as an initialization, but then perform dense 3D motion and detail estimation on a full 3D model [GVWT13] along with a global spatio-temporal optimization for synthesizing highly realistic new video animations of the face aligned to a dubbed audio track. [BCS97] rewrite the facial dialog in a monocular video by synthesizing new mouth movements through image warping and rearrangement of video frames. The approach learns a mapping between phonemes and static visemes for one specific actor and one language, but produces medium-quality results, only succeeds for simple head poses, and is not applicable to dubbing between different languages and different individuals. [KIM*14] synthesize the inner mouth for a given frontal 2D animation by employing a tooth and tongue image database and a syllabic decomposition of the speech. In this paper, we generate a convincingly rendered inner mouth by using a textured 3D tooth proxy and mouth cavity that are connected to the tracked blend shape model.

A trained multidimensional morphable model can be used for expression cloning [NN01], e.g., across video recordings of different people [CE05], or to separate the effects of emotion and dialog [EGP02], which is an important problem in facial motion retargeting. [BBVP03] learn a statistical morphable model from a database of human 3D face scans and use it to reanimate faces in images and video. [VBPP05] learn a multilinear model from a database of facial expressions that model visemes, identity and emotions. [DSJ*11] use this model to replace the face region of a target video with that of a different person. Although they show impressive results, applying their approach to dubbing requires that both the dubber and the actor are the same and share a similar frontal head pose. This can not be guaranteed in a general dubbing scenario, where the visual dialog of a given actor has to be completely synthesized and inpainted, which is what our system achieves. Our method is also related to video texture synthesis [SSSE00], where input video frames are rearranged to create a new output. In [XLS*11] *video-based characters* were proposed, which can create novel poses of a human actor by tracking a 3D mesh model, and rearranging and interpolating frames from a database of multi-view video under model guidance. Recently, [LXW*12] and [GVR*14] proposed similar systems for synthesizing novel facial expressions by using either an existing database of facial images or a short sequence of an arbitrary actor performance. Both systems are limited in their ability to render facial dynamics, especially in the mouth region, which is crucial for dialog. Moreover, the first method only works for simple frontal head poses, while both approaches can not handle lighting changes or transfer between different individuals. To the best of our knowledge, our system is one of the first to enable facial video rewrite that meets the visual quality needed in realistic film dubbing, i.e., source and target actors being shot in different surroundings, source and target dialogs in different languages having different phonetic content, and videos recorded with audio from a single camera at standard frame rates.

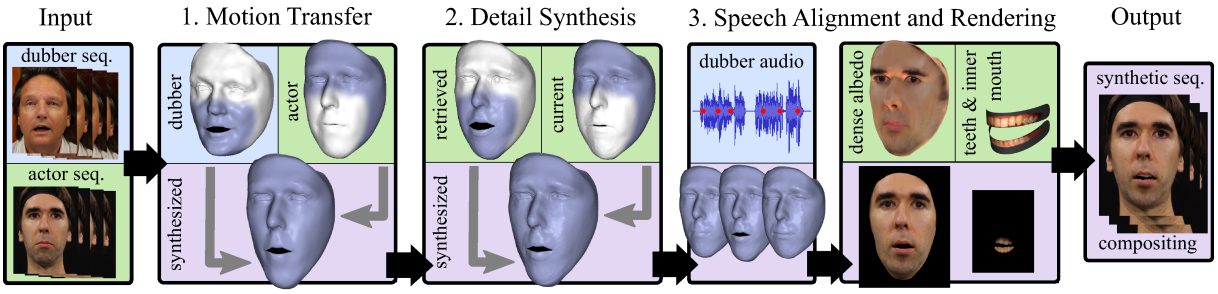

Figure 2: *Overview of our method*

3. Method Overview

Our method takes as input two video recordings with sound. The first recording is the original movie segment of the *actor* performing in the original language. We refer to this as the *target sequence* $I_{\mathcal{T}}$, as it will be modified later. The second recording is the *dubbing sequence* $I_{\mathcal{D}}$, showing the *dubber* reading a translation of the original text, which will serve as the source to synthesize a new target performance.

Our method uses the dubbed language track as the new voice track for the target sequence and modifies the mouth motion of the actor such that it matches the dubbed words. It does this while preserving the appearance and head pose of the actor, as well as the original background and scene lighting. We assume that the dubber reads his text roughly in pace with the actor's performance, but strict alignment of the dubbed language track with the actor's lip movements, as in traditional dubbing, is not necessary because we generate a completely new performance that is in sync by construction. We further assume that the dubber is able to reenact the facial expressions of the actor well, i.e., the target and dubbing sequences bear a similar emotional content. Our algorithm consists of three major steps (see Fig. 2):

1. **Motion Transfer (Sec. 4).** The facial performances of the actor and the dubber are captured using a personalized blend shape model. The target lighting is estimated and high-frequency detail, such as wrinkles and folds, are captured. The blending weights pertaining to the mouth motion of the dubber are transferred to generate a new blend shape sequence for the actor.
2. **Detail Synthesis (Sec. 5).** Actor-specific high-frequency face detail is added to the synthesized blend shape sequence by globally searching for frames with similar detail in the target sequence. We only transfer detail in the lower face region around the mouth, preserving the original detail elsewhere.
3. **Speech Alignment and Rendering (Sec. 6).** Lip closure is enforced by detecting bilabial consonants in the dubbed language track. By using the estimated target lighting and the dense skin reflectance of the actor, we render the synthesized face into the original video. The mouth interior is rendered separately and blended in with the target to produce the final composite.

We will denote by $I^t_{\mathcal{T}}$ and $I^t_{\mathcal{D}}$ the frame at time t in the target and dubbing sequence, with t running from 1 to the number of frames f. For simplicity, we assume that the target and dubbing sequence have the same number of frames and are temporally aligned such that corresponding spoken sentences coincide in time. This can be achieved as a preprocessing step or by recording the dubber in sync with the actor. The final result is the *synthesized sequence* $I_{\mathcal{S}}$, showing the actor speaking in the dubbed language. In the following, we will provide more details on the different steps.

4. Motion Transfer

To capture the facial performances of the actor and the dubber, we employ a state-of-the-art facial performance capture method on monocular video that utilizes an underlying blend shape model and produces a sequence of space-time coherent face meshes with fine-scale skin detail. The parameters of the tracked blend shape model will be used to transfer the mouth motion from the dubber to the actor.

4.1. Monocular Facial Performance Capture

Both the actor's and the dubber's performance is captured using the method of [GVWT13], which uses a personalized blend shape model. This model is a prior on the face shape and describes a basis of variation in facial expressions:

$$\boldsymbol{e}(\beta_1,\ldots,\beta_b) = \boldsymbol{n} + \sum_{j=1}^{b} \beta_j \boldsymbol{b}_j \ . \qquad (1)$$

In this model, $\boldsymbol{n} \in \mathbb{R}^{3n}$ is a vector containing the n 3D vertex coordinates of the face at rest, $\boldsymbol{b}_j \in \mathbb{R}^{3n}$, $1 \leq j \leq b$, are the blend shape displacements at each vertex, and $\boldsymbol{e} \in \mathbb{R}^{3n}$ is the facial expression obtained by linearly combining the blend shape displacements using the *blending weights* $0 \leq \beta_j \leq 1$, $\forall j$. We create a personalized blend shape model of the actor and the dubber by registering a generic blend shape model to a static stereo reconstruction of the face at rest, as described in [GVWT13]. Thus, the actor's blend shape model differs from that of the dubber's in face shape, but their b blend shapes correspond to the same canonical expressions and therefore have the same semantic meaning. For our models $b = 78$ and in our experiments we chose $n = 50000$.

The method of [GVWT13] tracks a sparse set of facial landmarks (eyes, eyebrows, nose, mouth, face outline) in the image sequence by employing 2D feature detection and optical flow. These features are used to estimate the head pose (3D rigid transformation) and the facial expression (blending weights). The result of this *blend shape tracking* step is a sequence of coarse face meshes that are spanned by the blend shape model, but lack fine-scale detail, such as wrinkles and folds. Skin detail is produced in a subsequent *shape refinement* step, which better aligns the facial geometry with the video and adds detail as a per-vertex surface displacement. This step effectively lifts the face geometry out of the blend shape space, while at the same time it estimates the scene lighting and a coarse, piece-wise constant approximation of the face albedo. The final result is a sequence of temporally coherent triangular face meshes $\mathcal{M}_{\mathcal{T}}^t$ for the target sequence and $\mathcal{M}_{\mathcal{D}}^t$ for the dubbing sequence, with $1 \leq t \leq f$.

4.2. Blending Weight Transfer

The blend shape model encodes most of the speech-related motion, such as the movement of the jaw, lips and cheeks, whereas the detail layer mainly encodes person-specific skin deformation, such as emerging and shifting wrinkles. The blend shape models of the actor and dubber are derived from the same generic model and thus share the same semantic dimensions, including those related to speech. We can therefore make the actor utter the same words as the dubber by transferring the temporal curves of the blending weights that activate the mouth region from the dubber to the actor. As explained in Sec. 4.3, these activation curves need further actor specific adjustment during transfer.

We manually identify the $m=49$ blend shapes responsible for the mouth motion as those components that displace vertices on the jaw, lip or cheeks. We quantify a region of influence for these mouth blend shapes by assigning a value between 0 and 1 to each vertex, where 1 means highly affected by mouth motion and 0 not affected at all. These values are found by accumulating the m blend shape displacements at each vertex and mapping them to $[0,1]$, where 0 corresponds to zero displacement and 1 to the median displacement over all vertices. The obtained mask is depicted in Fig. 3 and is used for detail synthesis and image blending (see Sec. 5 and Sec. 6). The mask is extended to include the nose tip, since it is often influenced by the mouth motion in practice.

We transfer the mouth motion of the dubber to the actor at a time t by combining the actor's blend shapes as follows:

$$\underbrace{\boldsymbol{e}_{\mathcal{S}}^t}_{\text{synthesized actor expression}} = \boldsymbol{n}_{\mathcal{T}} + \underbrace{\sum_{j=1}^{m} \beta_{\mathcal{D},j}^t \boldsymbol{b}_{\mathcal{T},j}}_{\text{captured dubber expression}} + \underbrace{\sum_{j=m+1}^{b} \beta_{\mathcal{T},j}^t \boldsymbol{b}_{\mathcal{T},j}}_{\text{captured actor expression}} . \quad (2)$$

Here, $\beta_{\mathcal{T},j}$ and $\beta_{\mathcal{D},j}$, $1 \leq j \leq b$, are the captured blending weights of the actor and the dubber, and $\boldsymbol{n}_{\mathcal{T}}$ and $\boldsymbol{b}_{\mathcal{T},j}$ denote the rest pose and the j-th blend shape of the actor. The

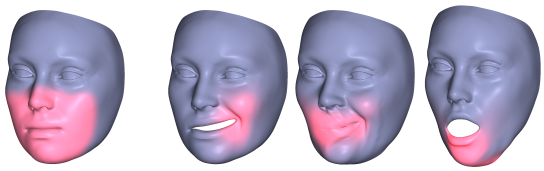

Figure 3: *The region of influence of the blend shapes responsible for the mouth motion (left) and three example blend shapes that activate the mouth (right, color encodes the magnitude of the displacement w.r.t. the rest pose).*

synthesized target expression $\boldsymbol{e}_{\mathcal{S}}^t$, $\forall t$, is identical to the original target expression, except in the mouth region shown in Fig. 3, where it is the same as the expression of the dubbing actor. The synthesized expression $\boldsymbol{e}_{\mathcal{S}}^t$ and the captured head pose can be used to build a sequence of synthetic, coarse face meshes $\mathcal{M}_{\mathcal{S}}^t$ for the actor, which exhibits the same mouth motion as the dubber. This is illustrated in Fig. 4 for the example of Fig. 1. Note that $\mathcal{M}_{\mathcal{S}}^t$ still lies within the blend shape space and therefore lacks any fine-scale detail, such as wrinkles and folds. This detail is necessary for a faithful rendering of the actor and will be added in Sec. 5.

4.3. Blending Weight Correction

The blending weight transfer described in Eq. (2) works well if the blending weight combinations for the actor and dubber have the same meaning. In practice, this is not guaranteed since both blend shape models are manually constructed from independently selected scans of a face at rest. As a result, they share the same semantic dimensions, but do not necessarily agree on the rest pose, i.e., the two models span the same semantic space but might have a different origin.

If there is a small systematic offset in the model origin, we can get an estimate of the true rest pose by selecting the blending weight combination that has the smallest Euclidean norm over all f captured frames, provided that there is at least one neutral expression in the sequence. This blending weight combination with minimum norm is then taken as the true model origin and is used to correct the transferred weights. To this end, we replace $\beta_{\mathcal{D},j}$ in Eq. (2) by

$$\beta_{\mathcal{D},j}^* = \beta_{\mathcal{D},j} - \beta_{\mathcal{D},j}^{\min} + \beta_{\mathcal{T},j}^{\min} \quad \text{for} \quad 1 \leq j \leq m \ , \quad (3)$$

where $\beta_{\mathcal{T}}^{\min} = \arg\min_{(\beta_{\mathcal{T}}^1,\ldots,\beta_{\mathcal{T}}^f)} \| \beta_{\mathcal{T}}^t \|_2$, $\forall t$, is the blending weight combination with the minimum Euclidean norm over all f target frames and $\beta_{\mathcal{D}}^{\min}$ has the same meaning for the dubbing sequence. We observed that this correction step significantly improved the quality of the expression transfer between different individuals.

5. Detail Synthesis

We add fine-scale skin detail to the synthesized target meshes $\mathcal{M}_{\mathcal{S}}^t$ by assuming that wrinkles and folds are correlated to the underlying facial expression, which in turn are

to the blending weights. Detail in the top part of $\mathcal{M}_{\mathcal{S}}^t$ is not influenced by the blending weight transfer and can thus be assumed identical to that of the captured mesh $\mathcal{M}_{\mathcal{T}}^t$. Detail in the mouth region, on the other hand, changes under the effect of the new blending weights and must be synthesized appropriately. This detail has to be actor-specific and will be generated by searching for similar expressions in the captured target sequence and transferring the high-frequency detail layer from the retrieved target geometries.

5.1. Target Frame Retrieval

We wish to retrieve a captured target mesh $\mathcal{M}_{\mathcal{T}}^{i(t)}$ with a similar mouth expression and motion as the current synthesized mesh $\mathcal{M}_{\mathcal{S}}^t$. Here, $i(t) \in \{1,\dots,f\}$ stands for the retrieved frame index in the target sequence that corresponds to the current index t in the synthesized sequence. To this end, we look for similarities in the blending weights that drive the mouth motion of the mesh sequences $\mathcal{M}_{\mathcal{T}}^t$ and $\mathcal{M}_{\mathcal{S}}^t$.

Let β_j, $1 \leq j \leq m$, denote the set of blending weights that are responsible for the mouth motion, as identified in Sec. 4.2. Then we can represent the synthesized mouth expression at a frame t by the blending weight vector $\boldsymbol{\beta}_{\mathcal{S}}^t = \left(\beta_{\mathcal{S},1}^t,\dots,\beta_{\mathcal{S},m}^t\right)^\top$ and the synthesized sequence of mouth expressions by $\boldsymbol{B}_{\mathcal{S}} = (\boldsymbol{\beta}_{\mathcal{S}}^1,\dots,\boldsymbol{\beta}_{\mathcal{S}}^f)$. Our retrieval problem aims at finding an optimal rearrangement of target indices $(i(1),\dots,i(f))$, such that the corresponding sequence of captured expressions $\hat{\boldsymbol{B}}_{\mathcal{T}} = (\boldsymbol{\beta}_{\mathcal{T}}^{i(1)},\dots,\boldsymbol{\beta}_{\mathcal{T}}^{i(f)})$ is as close as possible to $\boldsymbol{B}_{\mathcal{S}}$. We can write this optimization problem as:

$$\min_{(i(1),\dots,i(f))} E\left(\hat{\boldsymbol{B}}_{\mathcal{T}}, \boldsymbol{B}_{\mathcal{S}}\right) \quad , \tag{4}$$

where E denotes a multi-objective function that measures the similarity of blending weights along with their change over time, and the adjacency of frames, described as follows.

5.1.1. Blending Weight Distance

The similarity between a target and a synthesized mouth expression is computed as the L_2-norm of their difference. The index $i(t)$ of the target mesh, that is closest to the current synthetic mesh at frame t, has to minimize

$$d_{\mathrm{b}}(\boldsymbol{\beta}_{\mathcal{T}}^{i(t)}, \boldsymbol{\beta}_{\mathcal{S}}^t) = \left\|\boldsymbol{\beta}_{\mathcal{T}}^{i(t)} - \boldsymbol{\beta}_{\mathcal{S}}^t\right\|_2 \quad . \tag{5}$$

This distance measure is based on the assumption that, for a given person, face meshes with similar expression, and thus underlying blending weights, have similar skin detail.

5.1.2. Motion Distance

To regularize the retrieval, we consider the change in expression over time, i.e., the difference between consecutive blending weights $\boldsymbol{\beta}^{t-1}$ and $\boldsymbol{\beta}^t$. Given the expression change from $t-1$ to t in the synthesized sequence, we enforce that the currently retrieved blending weights $\boldsymbol{\beta}_{\mathcal{T}}^{i(t)}$ must undergo a similar change w.r.t. the previously retrieved weights $\boldsymbol{\beta}_{\mathcal{T}}^{i(t-1)}$. In other words, $i(t)$ and $i(t-1)$ have to minimize

$$d_{\mathrm{m}}(\boldsymbol{\beta}_{\mathcal{T}}^{i(t-1)}, \boldsymbol{\beta}_{\mathcal{T}}^{i(t)}, \boldsymbol{\beta}_{\mathcal{S}}^{t-1}, \boldsymbol{\beta}_{\mathcal{S}}^t) = \left\|(\boldsymbol{\beta}_{\mathcal{T}}^{i(t-1)} - \boldsymbol{\beta}_{\mathcal{T}}^{i(t)}) - (\boldsymbol{\beta}_{\mathcal{S}}^{t-1} - \boldsymbol{\beta}_{\mathcal{S}}^t)\right\|_2 \quad . \tag{6}$$

This measure assumes that similar changes in expression induce similar changes in skin detail. We remark that the retrieved indices $i(t-1)$ and $i(t)$ do not have to be consecutive in the original target sequence, since the search is global.

5.1.3. Frame Distance

Strong transitions in the retrieved detail are more likely if $i(t-1)$ and $i(t)$ lie far apart in the original target sequence. To enforce smoothly varying detail, we penalize the temporal distance of the retrieved neighboring indices, as follows:

$$d_{\mathrm{f}}(i(t-1), i(t)) = 1 - \exp\left(-|i(t-1) - i(t)|\right) \quad . \tag{7}$$

This measure assumes that the captured detail of close-by frames is more similar than that of distant frames.

5.1.4. Global Energy Minimization

The optimal rearrangement of target indices is then found by minimizing the energy in Eq. 4, which is the weighted sum of the three distances over all frames:

$$\begin{aligned} E\left(\hat{\boldsymbol{B}}_{\mathcal{T}}, \boldsymbol{B}_{\mathcal{S}}\right) = \quad & w_{\mathrm{b}} \sum_{t=1}^f d_{\mathrm{b}}(\boldsymbol{\beta}_{\mathcal{T}}^{i(t)}, \boldsymbol{\beta}_{\mathcal{S}}^t) \\ & + w_{\mathrm{m}} \sum_{t=1}^f d_{\mathrm{m}}(\boldsymbol{\beta}_{\mathcal{T}}^{i(t-1)}, \boldsymbol{\beta}_{\mathcal{T}}^{i(t)}, \boldsymbol{\beta}_{\mathcal{S}}^{t-1}, \boldsymbol{\beta}_{\mathcal{S}}^t) \\ & + w_{\mathrm{f}} \sum_{t=1}^f d_{\mathrm{f}}(i(t-1), i(t)) \quad , \end{aligned} \tag{8}$$

where w_{b}, w_{m}, and w_{f} control the influence of each term.

A greedy approach could find the unknown indices sequentially by progressively retrieving the currently nearest one. A better solution that solves for the complete sequence $(i(1),\dots,i(f))$ at once could be obtained by finding the shortest path in a weighted directed graph where each node represents a target index and each edge is weighted by the distances described above (see Fig. 5). A solution can be found using Dijkstra's algorithm, but since the starting node is unknown, its complexity is $O(f^3)$ in the number of frames, which prohibits its use for long sequences. Instead, we resort to methods based on hyper-heuristics [BHK*10] to arrive at an approximate solution that lies provably close to the global optimum. Hyper-heuristics are automated methods for selecting or generating local search operators to solve a hard combinatorial problem [BGH*13]. In our particular implementation, we define three local operators which independently minimize the three terms in Eq. 8, as well as a fourth operator that randomly disrupts the local optimum at a random index location. The latter ensures that the algorithm can explore new solutions, avoiding stagnation in local minima. To guide the search for the optimal solution, we define a hyper-heuristic approach that adaptively selects these four operators by reinforcement learning, as in [GC12].

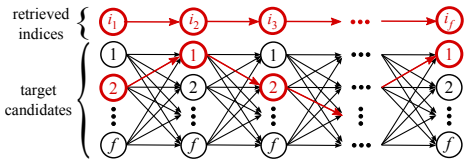

Figure 5: *Shortest path in a graph of candidate indices.*

Blend shape models can be overcomplete and multiple blend shape combinations may produce the same expression. We observed that different actors can activate distinct blend shapes when uttering the same words. As a consequence, facial expressions can not be compared reliably using a distance between blending weights. We overcome this by performing Principal Component Analysis (PCA) on our blend shape model and replacing the blending weights in Eq. 8 by the set of PCA weights that explains 99% of the mouth motion. Note that PCA does not change the face model; it only removes redundancy to make the frame retrieval more accurate. In the motion transfer step in Sec. 4, however, a blend shape representation is still preferred since the dimensions are spatially localized and easier to interpret [LAR*14]. This provides an extra level of control to the user who can globally scale the blend shape curves to modify the expressiveness (see the supplementary video).

5.2. Detail Transfer

Once a sequence of target indices has been retrieved, we transfer the skin detail of the retrieved target mesh $\mathcal{M}_{\mathcal{T}}^{i_t}$ to the current synthesized mesh $\mathcal{M}_{\mathcal{S}}^{t}$. The detail is added as a per-vertex displacement expressed in the local vertex coordinate frame (see Sec. 4.1). We only transfer new detail in the influence region of the mouth, given by the mask of Fig. 3. Outside this region we preserve the original detail of the captured mesh $\mathcal{M}_{\mathcal{T}}^{t}$. At the mask boundary, we ensure a smooth transition between both detail layers using alpha blending.

Despite temporal regularization, the retrieved indices may still introduce slight jumps in the transferred detail (only the original ordering of target indices produces smooth detail over time, but does not resemble the dubbing performance). Thus, we temporally smooth out the transferred detail layer by filtering the displacements in a sliding Gaussian window of 5 frames. The detail transfer is illustrated in Fig. 4.

6. Speech Alignment and Rendering

We improve the synchronization of the lip motion and the dubbed audio by modifying the blending weights to enforce lip closures where needed. The synthesized meshes are then rendered into the target camera using the estimated scene lighting and a per-vertex estimate of the skin reflectance. In a last step, the mouth cavity and the teeth are rendered and combined to produce the final composite.

6.1. Audio Alignment

To determine the precise time instances of visually salient speech gestures, we analyze the audio of the dubbing sequence independently of the video stream. Since the content of the utterances spoken by the dubber is known, we segment the audio into phonetic units using an automatic speech recognizer in forced-alignment mode [YEG*06]. In the resulting phonetic segmentation, lip closure events are aligned by analyzing all instances of bilabial consonants /p/, /b/, and /m/. In many cases, the automatically determined segment boundaries are sufficient, but where reverberation or background noise in the recording affects the reliability of the automatic segmentation, we manually correct the lip closure intervals using visual and acoustic cues in the phonetic analysis software Praat [BW01]. The output is a sequence of time intervals associated with all speech-related lip closure events in the video sequences, at a precision far higher than can be achieved when analyzing only the dubber video footage.

The detected intervals are used to improve the timing of bilabial consonants in the synthesized video by forcing the blending weights responsible for lip closure to zero. To avoid jerky motion, enforcement is done in a small Gaussian window centered around the detected intervals (see Fig. 6).

6.2. Rendering the Synthesized Geometry

Although complex light transport mechanisms, such as subsurface scattering, influence the perceived skin color, we assume pure Lambertian skin reflectance, which is sufficient under most conditions. To this end, we use the following formulation of the rendering equation:

$$B(\boldsymbol{x}) = \rho(\boldsymbol{x}) \int_{\Omega} L(\omega) V(\boldsymbol{x}, \omega) \max(\omega \cdot \boldsymbol{N}(\boldsymbol{x}), 0)\, \mathrm{d}\omega \ , \quad (9)$$

where B is the irradiance, ρ the skin albedo, $\boldsymbol{N}$ the normal at vertex $\boldsymbol{x}$, L the incident lighting from direction ω sampled on the hemisphere Ω, and V the visibility.

The facial performance capture method of [GVWT13] estimates the lighting L in the target scene and a coarse, piece-wise constant estimate of the actor's skin albedo (see Sec. 4.1). However, this coarse albedo is insufficient for a convincing rendering of the actor and we require a per-vertex albedo $\rho(\boldsymbol{x})$ instead. We estimate the dense skin albedo by projecting each vertex $\boldsymbol{x}$ of the captured mesh $\mathcal{M}_{\mathcal{T}}^{t}$ into the target frame $I_{\mathcal{T}}^{t}$ and assigning the intensity to $B(\boldsymbol{x})$ in Eq. 9. Dividing the irradiance by the integral on the right then gives us an estimate of $\rho(\boldsymbol{x})$. We then render the synthetic mesh by solving the rendering equation for each vertex of $\mathcal{M}_{\mathcal{S}}^{t}$.

If the dense albedo is estimated for each frame independently, it may suffer from small imprecisions in the captured face geometry and lead to undesirable intensity changes in the rendered images. To avoid this, we assume that the albedo is constant over time and estimate a single value in each vertex via a least squares fit over all captured meshes.

(a) (b) (c) (d) (e) (f)

Figure 4: *Motion transfer and detail synthesis for the example of Fig. 1. The facial performances of the actor (a) and the dubber (b) are captured and the estimated mouth-related blending weights are transferred from the dubber to the actor, in this case opening the actor's mouth (c). Fine-scale facial detail from the captured mesh in the current frame (d) and detail from the captured mesh in the retrieved frame (e) are combined to produce a detailed synthetic mesh (f).*

To improve spatial sampling, albedo computation and rendering are performed on upsampled versions of the face meshes ($n = 200000$).

6.3. Teeth, Inner Mouth and Final Composite

The rendered face lacks teeth and a mouth cavity. For the upper and lower teeth, we create a 3D proxy consisting of two billboards that are attached to the blend shape model and move in accordance with the face under the control of the blending weights (see Fig. 6). They are colored with the static texture of a target frame in which the teeth are visible. The inner mouth is created by warping a single image of the mouth cavity using the facial landmarks obtained from the synthesized facial performance. We uniformly adjust the brightness of the teeth and inner mouth according to the degree of mouth opening to create a realistic shading effect.

The rendered face, rendered teeth and warped inner mouth layers are blended in with the target image by feathering around the boundaries to assure a smooth transition (see Fig. 6). We only blend the synthesized face inside the projection of the mask of Fig. 3, while preserving the original face elsewhere. The result is the synthesized sequence I^t_S.

7. Experiments

We applied our method to three target sequences of German-speaking actors recorded under constant, unknown illumination. A dubbing studio † translated our scripts and recorded a new English language track for each sequence using a professional dubber. We filmed the dubber in the studio with the set-up of Fig. 1. The central camera is used for performance capture, while the two satellite cameras are only used for the 3D reconstruction needed for the blend shape creation. All videos where shot with an SLR camera at 25 fps in HD quality. The German audio was recorded with a USB microphone and the English audio with the dubbing studio equipment.

† SPEEECH Audiolingual Labs, www.speeech.de

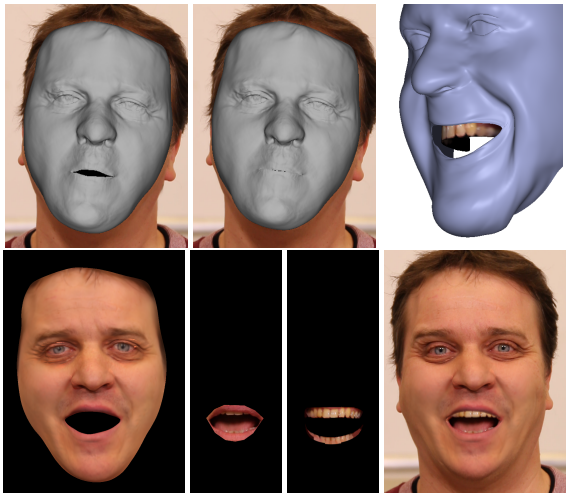

Figure 6: *Speech alignment and rendering. Lip closure is enforced to improve audio-visual quality (upper row left and center). The textured 3D tooth proxy attached to the blend shape model (upper row right). The rendered face, the inner mouth and tooth layers and the final composite (bottom row).*

7.1. Dubbing Results

Fig. 7 presents our result for a target actor reciting a dialog of a movie in German. This sequence is 1.5 min long and the actor remained mostly still while speaking. The upper row in the figure shows example frames from the target sequence, while the middle row shows the corresponding frames from the English dubber sequence. These are assumed to be correctly aligned in time such that the English and German sentences overlap. As most professional dubbing studios record single sentences in separate takes, we had to perform this alignment manually (only at the beginning of the takes). The bottom row of the figure shows the corresponding synthesized results. The mouth motion, the actor appearance, and the mouth interior are plausibly synthesized. The supplementary video further demonstrates that the synthesized mouth motion matches the dubbed audio track well.

Figure 7: *Target (top), dubber, and synthesized (bottom).*

Figure 8: *Target (top), dubber, and synthesized (bottom).*

Another result for a different target actor performing a scene of a passion play can be seen in Fig. 8. This sequence is challenging due to the fast head motion and the expressive facial gestures. As can be seen in the figure and the supplementary video, the new mouth motion and appearance are plausibly generated and much of the emotional content is preserved, which demonstrates that out method is capable of dealing with fast and expressive motion. Finally, Fig. 9 shows the same actor answering interviewer's questions. This video attempts to simulate a television interview where the spoken lines are spontaneous and not scripted beforehand. Also for this result, the expressions of the actor, including laughter and pondering gestures, are preserved well.

Figure 9: *Target (top), dubber, and synthesized (bottom).*

7.2. Validation

Target Frame Retrieval To quantify the influence of the energy terms in Eq. 8, we compared several retrieval results obtained with different values for the weights w_b, w_m, and w_f. To this end, we recorded a control sequence for the experiment of Fig. 7, in which the target actor is reading the English dubbing transcript under target conditions. The target and control sequences thus depict the same actor reciting the same dialog, both in German and in English. Based on the English audio, we selected the corresponding words in the dubbing sequence and control sequence that had a comparable timing, and identified 142 frames in which the visual utterance of the actor was identical to that of the dubber.

These 142 control frames were compared to the frames that were retrieved by our method from the German target sequence. If our frame retrieval is successful, the control frame and the retrieved target frame should depict the same utterance and look very similar. As a similarity measure, we used the average PSNR over all 142 frames. Small differences in the actor's pose were accounted for by warping the faces to a common reference pose. Retrieving the closest frames in time ($w_b = w_m = 0$) was least successful with an average PSNR of 22 dB. Retrieval purely based on the similarity of the PCA weights ($w_m = w_f = 0$) was more successful (28.0 dB), while adding the motion distance (28.2 dB) and the frame distance (28.6 dB) increased the similarity further. We attained the best results by using the combination $w_b = 1$, $w_m = 10$ and $w_f = 1000$, which was utilized in all of our experiments. Note that we did not compare the control frames directly to our final synthesized images, since the rendering and the compositing can affect the PSNR adversely.

User Study We conducted a user study in which we asked users with an understanding knowledge of English to com-

pare our results of Figs. 7–9 with those of traditional dubbing. Traditionally dubbed results were provided by the studio that recorded the dubbing actor and consist of the original German target video overlaid with the English language track, which was further modified by an expert for a better audio-visual alignment (included in the additional video).

45 participants, from countries where dubbing is both a common practice (Germany, France) and not (US, UK, Chile), rated the results on a Likert scale from 0 (very bad) to 5 (very good) based on their audio-visual experience, including viewing discomfort and how natural the video-audio combination was perceived. Our modified videos and the traditionally dubbed videos were displayed side-by-side in a web form. Over all three sequences, traditional dubbing received an average score of 3.2, while our system received a score of 2.7. Overall, 35% of the respondents said they felt more comfortable watching our modified video. These scores seem low at first, but actually indicate a big step ahead in solving this extremely difficult problem. The human eye is tuned to the slightest visual artifact in a rendered face and it is very hard for an automatic system to produce visually plausible results that do not fall in the uncanny valley, especially in a side-by-side comparison against real video.

Despite the professional quality, traditional dubbing was not favored by everyone. In fact, our result of Fig. 7 was preferred by 47% of the users and we believe this shows considerable progress towards a system that can replace facial performances in video. The same result received an absolute score of 2.7, which is only slightly less than the 2.9 score of traditional dubbing. In overall, our result shown in Fig. 9 received the highest score of 3.0. All p-values were lower than 0.01, except for Fig. 7, whose preference was close to 50%.

Comparison to Image-based Methods We compared our 3D model-based approach with an image-based approach that rearranges the input target frames and only applies 2D face warping to produce the final composite. To this end, we modified VDub to use the image reordering and non-rigid warping strategy of [GVR*14] to generate a temporally smooth target performance that is close to the dubber's performance. Such a method is similar to a purely image-based technique, like Video Rewrite [BCS97], but with better image warping. We refer to the additional document for details on how we turned VDub into an image-based approach.

The additional video shows the image-based result for the sequence of Fig. 8. Although the performance matches the dubbed audio, the synthesized animation is much less expressive and suffers from temporal ghosting artifacts and inner mouth instabilities. It also struggles with strong head motion and the compositing introduces stretching and shrinking of the face. Our result has a higher temporal resolution, and the use of a full 3D face model and geometric face detail combined with detailed albedo enables us to better merge synthesized and original performance under larger head pose variations, as well as appearance changes.

Figure 10: *Rendering with (left) and without detail (right). Without added detail the face looks too smooth.*

Rendering Fig. 10 and the supplementary video demonstrate the importance of facial detail synthesis for photo-realistic rendering by comparing our result with a system that renders the face using a blend shape model without fine-scale detail. This corresponds to facial replacement techniques that use a coarse 3D parametric model without a detail layer [DSJ*11]. In contrast to our method, important skin features, such as laugh lines, are hardly visible and realistic shading effects on the chin and upper lip are missing. The supplementary video also compares alternative strategies to create the inner mouth, showing that ours is the best.

7.3. Discussion

Our work takes a notable step ahead over previous facial expression transfer or facial video modification approaches. Unlike video rewrite or model-based replacement methods that mix identities [DSJ*11], we can synthesize results when target and dubbing actor are *different*, which is essential for any practical application. The use of an accurate parametric face model, along with detailed lighting and albedo information enables photo-realistic synthesis of face appearance, even on long videos with moderate out-of-plane head motion. As shown in the experiments, our 3D model-based resynthesis approach bears several advantages over purely or model-assisted image-based methods, which often exhibit ghosting artifacts or temporal aliasing, merely show results without compositing, and can only handle marginal out-of-plane head motion [BCS97, LO11, ASWC13].

We also take a step towards easing and streamlining the workflow of traditional dubbing: we no longer require a translation of the original text that matches the visual utterances in the target video on a viseme level. Since we resynthesize the mouth motion entirely, the translation can be more free. Furthermore, our method relies on very little manual preprocessing, most notably the creation of the blend shape model and teeth proxy; otherwise, it is fully automatic and can be integrated into an industrial pipeline.

7.4. Limitations

Our approach is a step towards a challenging goal and thus has limitations. First, a static 3D face reconstruction to build the blend shape model may not be available for every actor, e.g., in vintage movies. Model reconstruction from video is

an interesting problem for future work. Our approach currently transfers idiosyncrasies of the dubbing actor to the target actor and our renderings may thus reflect the characteristics of the dubber rather than of the original actor. For instance, in our test data we measured an asymmetry in the blending weights of the dubber as part of his natural way of speaking and this asymmetry was reproduced in the actor (refer to the result of Fig. 7 in the supplementary video). We believe that more sophisticated expression transfer methods (e.g. [SLS*12]) would also transfer dubber characteristics and think that this is rather a question of user control. Certain aspects and weights can be manually amplified to control expressiveness (see supplementary video), or differences between the two actors could be learned in order to achieve a certain style; all of this is feasible in our representation.

We compute an average albedo, which can be blurred if correspondences over time are not accurate. Our lighting model may be challenged in scenes with strong and sudden light changes, and the current monocular tracking may fail in extreme facial poses, e.g., completely lateral views. We only detect lip closure and opening events from the audio track. As audio cues provide a precision far higher than video only, more complex information, such as triphones could be extracted and used [BCS97, Bra99], e.g. to train a hidden Markov model. This may further improve results.

We only replace the mouth region and not the full facial expression, which may not convey all of the visual information in the speech. For plausible results, we assume that the dubber is capable of playing the same routine as the actor with similar emotional content. This is a reasonable assumption and mostly fulfilled in practice. However, there may be facial regions where a match between the new mouth motion and the original video is challenged, e.g., the larynx in our results does not move according to the dubbed audio.

The blending weight transfer between two different actors is made possible through the use of a personalized blend shape model that is produced from a static scan. This implicitly assumes that all actors share the same blend shape displacements for the basis expressions and that a true rest pose can be reliably identified. In practice, this assumption does not fully hold and its violation can introduce biases in the results. We propose a correction strategy (Sec. 4.3), but more elaborate face models, with more consistent motion dimensions across actors, will be addressed as future work.

8. Conclusion

We have presented one of the first automatic solutions for transferring expressions between two different real-life actors and rendering photo-realistic, plausible mouth motion in an existing video that visually correlates to a dubbed audio track in a different language. The approach is based on highly detailed monocular 3D face reconstruction, as well as lighting and albedo estimation. New 3D mouth performances are synthesized by using a new motion parameter transfer step between dubbing and target actor, and a space-time retrieval method that synthesizes plausible high-frequency shape detail. The synthesized results, including the interior of the mouth, are photo-realistically rendered and attention is paid to a proper synchronization of the mouth motion with salient events in the audio track. Resynthesizing facial motion at video quality is extremely challenging as our perception is attuned to the slightest inaccuracies. Through qualitative comparison and a user study we showed that we can create plausible results and that we have taken an important step towards solving this challenging problem.

Acknowledgments

This work was partially supported by the ERC Starting Grant CapReal and by Technicolor.

References

[ASWC13] ANDERSON R., STENGER B., WAN V., CIPOLLA R.: Expressive visual text-to-speech using active appearance models. In *Proc. CVPR* (2013), pp. 3382–3389. 3, 10

[BBVP03] BLANZ V., BASSO C., VETTER T., POGGIO T.: Reanimating faces in images and video. *CGF (Proc. EUROGRAPHICS) 22*, 3 (2003), 641–650. 3

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video Rewrite: Driving visual speech with audio. In *ACM TOG (Proc. SIGGRAPH)* (1997), pp. 353–360. 3, 10, 11

[BGH*13] BURKE E. K., GENDREAU M., HYDE M., KENDALL G., OCHOA G., OZCAN E., QU R.: Hyper-heuristics: a survey of the state of the art. *Journal of the Operational Research Society 64*, 12 (2013), 1695–1724. 6

[BHB*11] BEELER T., HAHN F., BRADLEY D., BICKEL B., BEARDSLEY P., GOTSMAN C., SUMNER R. W., GROSS M.: High-quality passive facial performance capture using anchor frames. In *ACM TOG (Proc. SIGGRAPH)* (2011), vol. 30, pp. 75:1–75:10. 3

[BHK*10] BURKE E., HYDE M., KENDALL G., OCHOA G., OZCAN E., WOODWARD J. R.: *A Classification of Hyper-heuristic Approaches*. Kluwer, 2010. 6

[Bra99] BRAND M.: Voice puppetry. In *ACM TOG (Proc. SIGGRAPH)* (1999), pp. 21–28. 2, 11

[BW01] BOERSMA P., WEENINK D.: Praat, a system for doing phonetics by computer. *Glot International 5*, 9/10 (2001), 341–345. 7

[BWP13] BOUAZIZ S., WANG Y., PAULY M.: Online modeling for realtime facial animation. In *ACM TOG (Proc. SIGGRAPH)* (2013), vol. 32, pp. 40:1–40:10. 3

[CE05] CHANG Y., EZZAT T.: Transferable video-realistic speech animation. In *Proc. SCA* (2005), pp. 29–31. 3

[CFKP04] CAO Y., FALOUTSOS P., KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Proc. SCA* (2004), pp. 347–355. 3

[CHZ14] CAO C., HOU Q., ZHOU K.: Displaced dynamic expression regression for real-time facial tracking and animation. In *ACM TOG (Proc. SIGGRAPH)* (2014), vol. 33, p. 43. 3

[DN06] DENG Z., NEUMANN U.: eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. In *Proc. SCA* (2006), pp. 251–260. 3

[DSJ*11] DALE K., SUNKAVALLI K., JOHNSON M. K., VLASIC D., MATUSIK W., PFISTER H.: Video face replacement. In *ACM TOG (Proc. SIGGRAPH Asia)* (2011), vol. 30, pp. 130:1–130:10. 3, 10

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *ACM TOG (Proc. SIGGRAPH)* (2002), pp. 388–398. 3

[GC12] GARRIDO P., CASTRO C.: A flexible and adaptive hyper-heuristic approach for (dynamic) capacitated vehicle routing problems. *Fundamenta Informaticae 119*, 1 (2012), 29–60. 6

[GVR*14] GARRIDO P., VALGAERTS L., REHMSEN O., THORMAEHLEN T., PEREZ P., THEOBALT C.: Automatic face reenactment. In *Proc. CVPR* (2014). 3, 10

[GVWT13] GARRIDO P., VALGAERTS L., WU C., THEOBALT C.: Reconstructing detailed dynamic face geometry from monocular video. In *ACM TOG (Proc. SIGGRAPH Asia)* (2013), vol. 32, pp. 158:1–158:10. 3, 4, 5, 7

[Kil93] KILBORN R.: Speak my language: Current attitudes to television subtitling and dubbing. *Media Culture Society 15*, 4 (1993), 641–660. 2

[KIM*14] KAWAI M., IWAO T., MIMA D., MAEJIMA A., MORISHIMA S.: Data-driven speech animation synthesis focusing on realistic inside of the mouth. *Journal of Information Processing 22*, 2 (2014), 401–409. 3

[KM03] KSHIRSAGAR S., MAGNENAT-THALMANN N.: Visyllable based speech animation. In *CGF (Proc. EUROGRAPHICS)* (2003), vol. 22, pp. 632–640. 3

[KSSS10] KEMELMACHER-SHLIZERMAN I., SANKAR A., SHECHTMAN E., SEITZ S. M.: Being John Malkovich. In *Proc. ECCV* (2010), pp. 341–353. 3

[LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F., DENG Z.: Practice and theory of blendshape facial models. In *EUROGRAPHICS STAR report* (2014), pp. 199–218. 7

[LGMCB94] LE GOFF B., GUIARD-MARIGNY T., COHEN M., BENOÎT C.: Real-time analysis-synthesis and intelligibility of talking faces. In *ESCA/IEEE Workshop on Speech Synthesis* (1994), pp. 53–56. 2

[LK81] LESNER S. A., KRICOS P. B.: Visual vowel and diphthong perception across speakers. *Journal of the Academy of Rehabilitative Audiology 14* (1981), 252–258. 2

[LO11] LIU K., OSTERMANN J.: Realistic facial expression synthesis for an image-based talking head. In *Proc. ICME* (July 2011), pp. 1–6. 3, 10

[LTK09] LEVINE S., THEOBALT C., KOLTUN V.: Real-time prosody-driven synthesis of body language. In *ACM TOG (Proc. SIGGRAPH Asia)* (2009), vol. 28, pp. 172:1–172:10. 2

[LXW*12] LI K., XU F., WANG J., DAI Q., LIU Y.: A data-driven approach for facial expression synthesis in video. In *Proc. CVPR* (2012), pp. 57–64. 3

[MCP*06] MA J., COLE R. A., PELLOM B. L., WARD W., WISE B.: Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE TVCG 12*, 2 (2006), 266–276. 3

[MM76] MCGURK H., MACDONALD J.: Hearing lips and seeing voices. *Nature 264*, 5588 (1976), 746–748. 2

[NN01] NOH J., NEUMANN U.: Expression cloning. In *ACM TOG (Proc. SIGGRAPH)* (2001), pp. 277–288. 3

[OB86] OWENS E., BLAZEK B.: Visemes observed by the hearing-impaired and normal-hearing adult viewers. *Journal of Speech, Language, and Hearing Research 28* (1986), 381–393. 2

[PSS99] PIGHIN F., SZELISKI R., SALESIN D.: Resynthesizing facial animation through 3D model-based tracking. In *Proc. CVPR* (1999), pp. 143–150. 3

[SC00] SLANEY M., COVELL M.: FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Proc. NIPS* (2000), pp. 814–820. 2

[SDT*07] SONG M., DONG Z., THEOBALT C., WANG H., LIU Z., SEIDEL H. P.: A generic framework for efficient 2-D and 3-D facial expression analogy. *IEEE Trans. Multimedia 9*, 7 (2007), 1384–1395. 3

[SLC11] SARAGIH J. M., LUCEY S., COHN J. F.: Deformable model fitting by regularized landmark mean-shift. *IJCV 91*, 2 (2011), 200–215. 3

[SLS*12] SEOL Y., LEWIS J. P., SEO J., CHOI B., ANJYO K., NOH J.: Spacetime expression cloning for blendshapes. *ACM TOG 31*, 2 (2012), 14. 11

[SP54] SUMBY W., POLLACK I.: Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America 26*, 2 (1954), 212–215. 2

[SSRMF06] SIFAKIS E., SELLE A., ROBINSON-MOSHER A. L., FEDKIW R.: Simulating speech with a physics-based facial muscle model. In *Proc. SCA* (2006), pp. 261–270. 3

[SSSE00] SCHÖDL A., SZELISKI R., SALESIN D., ESSA I. A.: Video textures. In *ACM TOG (Proc. SIGGRAPH)* (2000), pp. 489–498. 3

[Sum92] SUMMERFIELD Q.: Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society Series B: Biological Sciences 335*, 1273 (1992), 71–78. 2

[TMCB07] THEOBALD B.-J., MATTHEWS I. A., COHN J. F., BOKER S. M.: Real-time expression cloning using appearance models. In *Proc. ICMI* (2007), pp. 134–139. 3

[TMTM12] TAYLOR S. L., MAHLER M., THEOBALD B.-J., MATTHEWS I.: Dynamic units of visual speech. In *Proc. SCA* (2012), pp. 275–284. 2, 3

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. In *ACM TOG (Proc. SIGGRAPH)* (2005), vol. 24, pp. 426–433. 3

[WBLP11] WEISE T., BOUAZIZ S., LI H., PAULY M.: Real-time performance-based facial animation. In *ACM TOG (Proc. SIGGRAPH)* (2011), vol. 30, pp. 77:1–77:10. 3

[Wil90] WILLIAMS L.: Performance driven facial animation. In *ACM TOG (Proc. SIGGRAPH)* (1990), vol. 24, pp. 235–242. 3

[XCXH03] XIANG-CHAI J., XIAO J., HODGINS J.: Vision based control of 3D facial animation. In *Proc. SCA* (2003), pp. 193–206. 3

[XLS*11] XU F., LIU Y., STOLL C., TOMPKIN J., BHARAJ G., DAI Q., SEIDEL H.-P., KAUTZ J., THEOBALT C.: Video-based characters: Creating new human performances from a multi-view video database. In *ACM TOG (Proc. SIGGRAPH)* (2011), vol. 30, pp. 32:1–32:10. 3

[YEG*06] YOUNG S., EVERMANN G., GALES M., HAIN T., KERSHAW D., LIU X. A., MOORE G., ODELL J., OLLASON D., POVEY D., VALTCHEV V., WOODLAND P.: *The HTK Book*. Cambridge University Engineering Department, 2006. 7

[YRVB98] YEHIA H., RUBIN P., VATIKIOTIS-BATESON E.: Quantitative association of vocal-tract and facial behavior. *Speech Communication 26*, 1–2 (1998), 23–43. 2, 3