7-2015

# Automatic Video Self Modeling for Voice Disorder

Ju Shen
*University of Dayton*, jshen1@udayton.edu

Changpeng Ti
*University of Kentucky*

Anusha Raghunathan
*Intel Corp.*

Sen-ching S. Cheung
*University of Kentucky*

Rita Patel
*Indiana University - Bloomington*

# Automatic Video Self Modeling for Voice Disorder

**Ju Shen · Changpeng Ti · Anusha Raghunathan ·
Sen-ching S. Cheung · Rita Patel**

**Abstract** Video self modeling (VSM) is a behavioral intervention technique in which a learner
models a target behavior by watching a video of him- or herself. In the field of speech language
pathology, the approach of VSM has been successfully used for treatment of language in children
with Autism and in individuals with fluency disorder of stuttering. Technical challenges remain
in creating VSM contents that depict previously unseen behaviors. In this paper, we propose a
novel system that synthesizes new video sequences for VSM treatment of patients with voice dis-
orders. Starting with a video recording of a voice-disorder patient, the proposed system replaces
the coarse speech with a clean, healthier speech that bears resemblance to the patient's origi-
nal voice. The replacement speech is synthesized using either a text-to-speech engine or selecting
from a database of clean speeches based on a voice similarity metric. To realign the replacement
speech with the original video, a novel audiovisual algorithm that combines audio segmentation

Ju Shen, Changpeng Ti, Sen-ching S. Cheung
University of Kentucky
Center for Visualization and Virtual Environments
329 Rose Street
Lexington, KY 40506
Tel: +1-859-218-0299
Fax: +1-859-257-1505
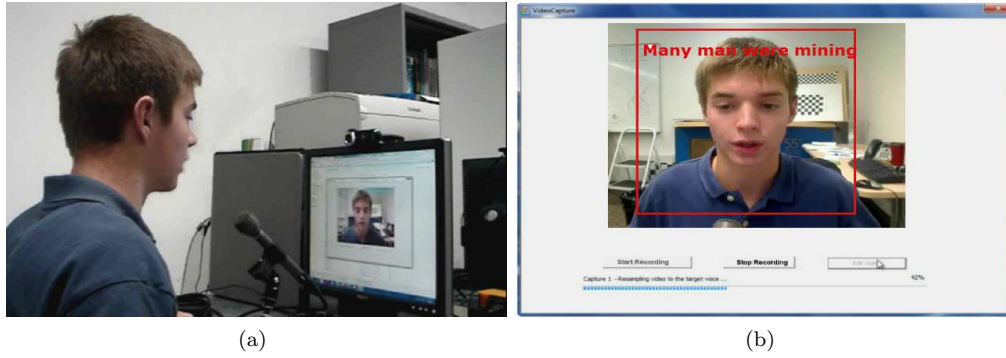E-mail: jushen.tom@uky.edu, changpeng.ti@uky.edu, sccheung@ieee.org

Anusha Raghunathan
Intel corporation
Bowers Avenue,
1900, Prarie City Road, Folsom, CA 95630
Tel: +1-916-673-6625
E-mail: anusha.ragunathan@gmail.com

Rita Patel
Indiana University
Department of Speech and Hearing Science
200 South Jordan Avenue,
Bloomington, IN. 47405
Tel: +1-812-855-3886
E-mail: patelrir@indiana.edu

with lip-state detection is proposed to identify corresponding time markers in the audio and video tracks. Lip synchronization is then accomplished by using an adaptive video re-sampling scheme that minimizes the amount of motion jitter and preserves the spatial sharpness. Results of both objective measurements and subjective evaluations on a dataset with 31 subjects demonstrate the effectiveness of the proposed techniques.

# 1 Introduction



**Fig. 1** (a) System Setup; (b) User Interface

Nowadays, one can learn just about anything by watching a video on the web, on television or from the thousands of DVD/Blue-ray titles available from different sources. Watching a video to learn or model a target positive behavior is in fact a well-studied technique in behavior therapy called Video Modeling (VM) interventions. They are widely used in rehabilitation and education of patients recovering from surgery [28] and cancer [31] as well as job and safety training for hospital staffs [33] and office workers [6]. VM is also effective in a school setting to teach children and young adolescents various skills including social interactions, communication, self-monitoring and emotional regulation [25].

Rather than watching others, some researchers have argued that we can learn even more effectively by watching our own positive behaviors. Such form of self modeling is classically done with a mirror and one of the most prominent examples is the use of the "mirror box" in treating phantom limb pain among amputees [38]. Seeing or visualizing oneself accomplishing the target behavior provides the most ideal form of behavior modeling. Though still in its early development, effectiveness of VSM has been demonstrated for many different types of disabilities and behaviorial problems ranging from stuttering, inappropriate social behaviors, autism, selective mutism to sports training. A summary of this research can be found in [12].

There are two forms of VSM: positive self-review and feedforward [16]. In positive self-review, the portions of the recorded video showing poorly executed routines are removed leaving only the

positive target behaviors. The resulting video will be reviewed to enhance fluency of the skills that have already been acquired by the learner but not yet perfected. On the other hand, the feed-forward VSM focuses on teaching new skills to a learner by showing novel skills that have never been observed but still within the reach of the learner. Evidence shows that the feed-forward approach delivers a more dramatic learning effect than the positive self-review approach [12]. One explanation of these findings comes from the psychological theory of self-efficacy - "I know I can because I have done it before" [9]. There is an inherent difficulty associated with the production of VSM material. Prolonged and persistent video recording is required to capture the rare, if existed at all, snippets that can be used to string together in forming novel video sequences of the target skill. An example of feedforward VSM can be found in [12] – the author records more than six hours of video from a child who can only speak one or two-word utterances to produce a two-minute clip of the same child saying one full sentence.

This problem can be potentially solved by using computational multimedia techniques. From computer generated imagery to speech synthesis, there exist a myriad of multimedia tools that can synthesize realistic video content. The goal is to adapt such tools for the development of feedforward VSM systems that can be used by a learner and his/her therapist in creating VSM contents with minimum amount of training data. For such system to be useful in practice, the synthesis process must be automatic and real-time so that rapid feedback can be provided. The synthetic content should be perceptually indistinguishable from real video footage. Such systems can have a significant impact in reducing the time and effort to achieve the target learning objectives.

In this paper, we propose a novel feedforward VSM content production system for patients suffering from a specific type of vocal disorders called vocal hyperfunction. Vocal hyperfunction refers to the use of excessive muscle force and physical effort in the production of voice, and usually requires a prolonged period of speech therapy. Long-standing untreated voice disorders can detrimentally affect an individual psychosocially and academically and can be a source of substantial economic cost to society in terms of higher health care costs. The goal of our system is to reduce the amount of time on therapy using the principle of video self modeling. Our system records a video of a patient at the clinic reciting a known script, and synthesizes a new video with a "healthier" voice for self-modeling. The purpose of this new video is to encourage patients to continue practising of the skills learned at the clinic in order to accelerate the behavioral changes. The proposed video self modeling approach can be used for voice therapy in individuals with vocal fold nodules, functional dysphonia. It can also be used in speech therapy for post-surgical management of individuals with vocal fold polyps and vocal fold cysts. However, before subjecting the approach of video self modeling to empirical testing with traditional voice therapy approaches it is critical to test the robustness / accuracy of the image processing algorithm. Our system uses a data-driven approach in selecting a replacement speech best resembled that of the patient. A novel joint audio-visual algorithm is developed to synchronize the old speech with the replacement speech. The synchronization process produces a set of time markers which are then used to re-sample the video to achieve perfect lip synchronization. To minimize the amount of motion jitter introduced during the re-sampling process, we introduce a novel adaptive sampling strategy to preserve the motion energy of the original video. Extensive objective and subjective testing have been conducted to demonstrate the effectiveness of the proposed system. Compared with an earlier version of

this work in [42], we have substantially improved the synchronization process and expanded the subjective testing to validate our system.

Figure 1(a) shows the setup of the system. It captures the raw video of the patient through a web camera situated on top of a desktop computer. Figure 1(b) shows the user interface. A red square is shown in the middle of the screen to provide a visual cue to anchor the head position. In order to capture a proper eye gaze, the left-to-right scrolling script is shown near the camera.

The rest of the paper is organized as follows: Section 2 discusses the severity of the vocal hyperfunction problem and motivates the treatment potential of VSM in voice disorders. Section 3 reviews related work in video synthesis and lip synchronization. Section 4 provides an overview of our proposed system, which consists of the analysis and the synthesis portion. The algorithms for our audiovisual analysis of the input signals are presented in Section 5. The synthesis of the speech replacement and video re-sampling are presented in Section 6. Section 7 presents the experimental results. Section 8 concludes the paper with a discussion on possible future work.

## 2 VSM for Voice Disorder

Voice disorders appear to be the most common communication disorder across the lifespan, are disabling and compromise quality of life. Considering that 3-9% of the population has some type of voice disorder at any given point in time [43], is a significant medical problem. According to the National Institute on Deafness and Other Communication 7.5 million people in the United States have trouble using their voices. Voice disorders can have significant personal as well as societal impact. Voice disorders are a source of substantial functional loss for individuals, and a source of substantial economic cost to society in terms of higher health care costs.

Voice therapy is often the primary choice of treatment for voice disorder called vocal hyperfunction. Vocal hyperfunction is one type of voice disorders that is defined as the use of excessive muscle force and physical effort in the production of voice [11]. The traditional model of voice therapy typically involves participation in one or two 40-45 minute sessions per week over the course of eight weeks with the speech language pathologist to facilitate behavior change in production of voice. High dropout rates of 16% to 65% [24,30,40] coupled with reduced long-term success rates of 51% to 68% [41], suggest the need for development of new approaches for delivery of voice therapy to improve treatment success [39] [36].

Access to voice therapy services for management of voice disorders in rural areas and developing countries is particularly lacking, due to difficulties in recruiting and retaining speech language pathologists and the expenses of time and travel for the required voice therapy program necessary to remediate the voice disorder. The use of VSM for voice therapy is a novel application where the pathologist can use the proposed system to create videos of a patient speaking with an improved voice. The patients can either take these videos after their initial visit to the clinic or access them through internet, and continue their behavioral modeling in their home. This new form of treatment has the potential of reducing the length of the treatment program and the number of therapy sessions, thereby reducing health disparities in rural populations.

## 3 Related Work

Our goal of synthesizing new talking head video bears resemblance with the large body of work in facial animation using either real video footage [29] and avatars [15, 37]. The key difference is that we have exploited the requirements from the domain application in developing a fully automated real-time system. For example, we only need to re-sample the video sequence to achieve lip synchronization rather than a complete re-rendering of a new sequence as in [29]. Also it is unimportant for us to preserve emotion as in [15, 37]. On the other hand, there are more stringent audio requirements that we need to overcome in synthesizing a new speech track with a healthy voice that bears strong resemblance to the patient.

The problem of lip synchronization has been extensively studied in literature, which can be grouped into three different categories according to the data source: audio-based, video-based, and joint audiovisual processing. For pure audio-based techniques, Mermelstein's algorithm [34] is an influential rule-based syllable segmentation approach. It locates syllable boundaries by computing the convex hull of the intensity envelope between 500Hz and 4kHz. A modification to Mermelstein's convex hull algorithm based on periodicity and normalized energy was developed in [45] for syllable nuclei detection. In [2], Howitt incorporated Neural Network into an energy-based vowel detector using Mermelstein's algorithm. In [35], a multi-pass automatic speech segmentation algorithm was proposed, which involves a broad segmentation by intensity dips in the filtered speech, followed by further adjustments for syllable nuclei. Despite the sophistication of audio segmentation techniques, consistent segmentation across speakers remains a very challenging problem.
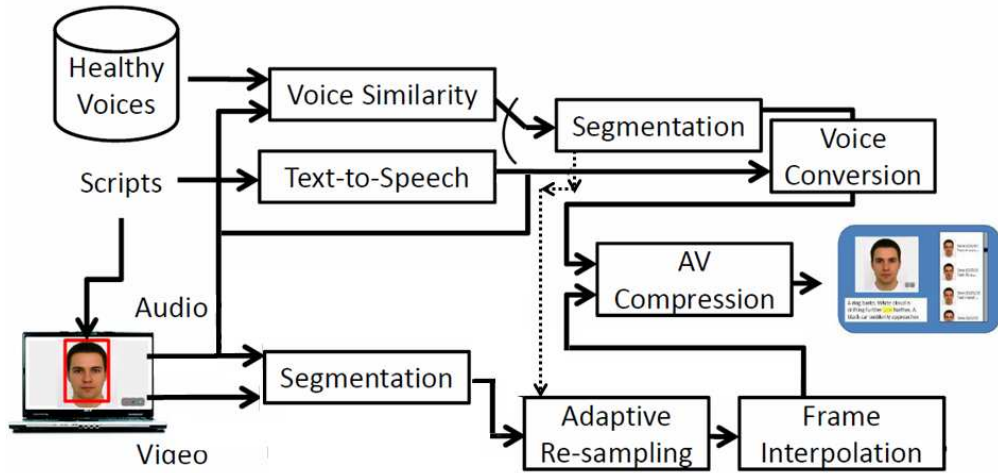
As for video based techniques, a lip-motion recognition method using Hidden Markov Model (HMM) is proposed in [27]. Lip contour boundary was extracted based on the contrast intensity of the image. Similarly, in [18] and [23], a lip segmentation algorithm was developed by contour detection and model fitting. However, these methods require a priori knowledge about lip structure, which makes it difficult to achieve full automation. To avoid the training steps, a geometric deformable model was proposed in [8]. They used spatial fuzzy clustering to create a probability map about the lip image. Then, the lip position was obtained by maximizing the join probability of the lip region and non-lip region. Nevertheless, this method requires heavy computation due to the complex probability model, which makes it hard to achieve realtime or close-to-realtime performance. In [17], the authors provided an efficient implementation through field-programmable gate array (FPGA). In their system, a naive Bayes classifier was used to extract lips features. But this method may be error-prone if the lip color is close to that of the skin. In [14], an improved active contour model was developed to extract the lip shapes by iteratively minimizing the proposed energy functions. Similar to our approach, it is based on the snake algorithm. But ours is simpler since for our problem, we do not need to track the lip motion over time.

Lip-motion and audio analysis are often combined to enhance the accuracy of speech recognition. There is a body of research termed *audio-visual speech recognition*(AVSR) which incorporates lip-motion as additional features in building the speech recognition engines [5, 7, 10, 21, 22, 26]. In our paper, both visual and audio information are also combined in the production of the output video. The difference is that we use them for the alignment of the original speech signal and the replacement speech signal, rather than recognition of specific phones. As such, our focus is on the accuracy in identifying the same set of markers across speeches from different individuals. Once the

alignment is identified, they are used to re-sample the video to establish lip-synchronization with the replacement speech.

## 4 System Overview

Figure 2 shows the audiovisual analysis and the VSM content generation process. After the raw video is captured, the audio track is extracted. The audio is segmented to extract time markers corresponding to the phone boundaries. The system then generates a replacement speech using either perceptually similar pre-recorded healthy voice or text-to-speech synthesis. The merits of both methods will be studied in the experimental section. Time markers for phone boundaries in the replacement speech will also be identified using the same segmentation module. Both sets of markers are needed to align the video track with the replacement speech in order to minimize motion jitter and provide lip-synchronization. Frame interpolation is then applied to re-sample the video track which is then combined with the new speech track. Our audio-visual synchronization process is described in Section 5 while the generation of the replacement speech and video re-sampling are presented in Section 6.



**Fig. 2** VSM Contenet Generation

## 5 Audiovisual Analysis

The goal of the audiovisual analysis is to identify a set of time timers that partition the speech signal at the phone boundaries. Two sets of markers are identified based on the variation of the loudness of the speech signal and the lip openness detected in the video signal. They are then combined to obtain a robust alignment between the original and the replacement speech tracks using dynamic programming. The details are described in the following subsections.

5.1 Audio Analysis

Our audio segmentation module uses a derivative-based approach which consists of three steps: signal envelop calculation, differentiation, and minima identification. An example of this three-step process is shown in Figure 3. First, the signal envelope is computed by applying a low pass filter on the cepstrum of the smoothed input signal. The envelope is obtained by the following equation:

$$E = exp(F^{-1}(W(F(log(y)))))  \qquad (1)$$

where $y$ is the input audio signal, and $W$ is a low-pass window. $F$ is the Fourier transformation. During the recording phase, we use sliding text to control the display speed for the speaker to read the script. This can roughly partition the speech signal into alternative continuous speech and silence periods. Starting with these rough partitions, we use a sliding window over the envelope to compute the short-term signal variance. A significant increase in the variance indicates the beginning of the speech period and a significant drop represents the end. Second, we compute the derivative of the envelope by convolving it with a Gaussian derivative filter. In the final step, the local minima of the envelope are identified based on the increasing zero-crossings of the derivative signal. Each local minimum is treated as the boundary between two phones.

The audio segmentation method does not necessarily recover the true partition of all phones. When the algorithm runs through multiple phones where there is no volume fall-off, it will interpret them as a single phone, resulting in time marker omissions. Such an error is not consistent across different speakers. Discrepancies in time markers between the source and target speech signals can significantly degrade the performance of the subsequent speech alignment. A misalignment due to an omission of a time marker in one signal can propagate to a much longer period before synchronization can be re-established at the end of a continuous speech period. To enhance the correct detection of all time markers, we turn to the video and analyze the lip movement.
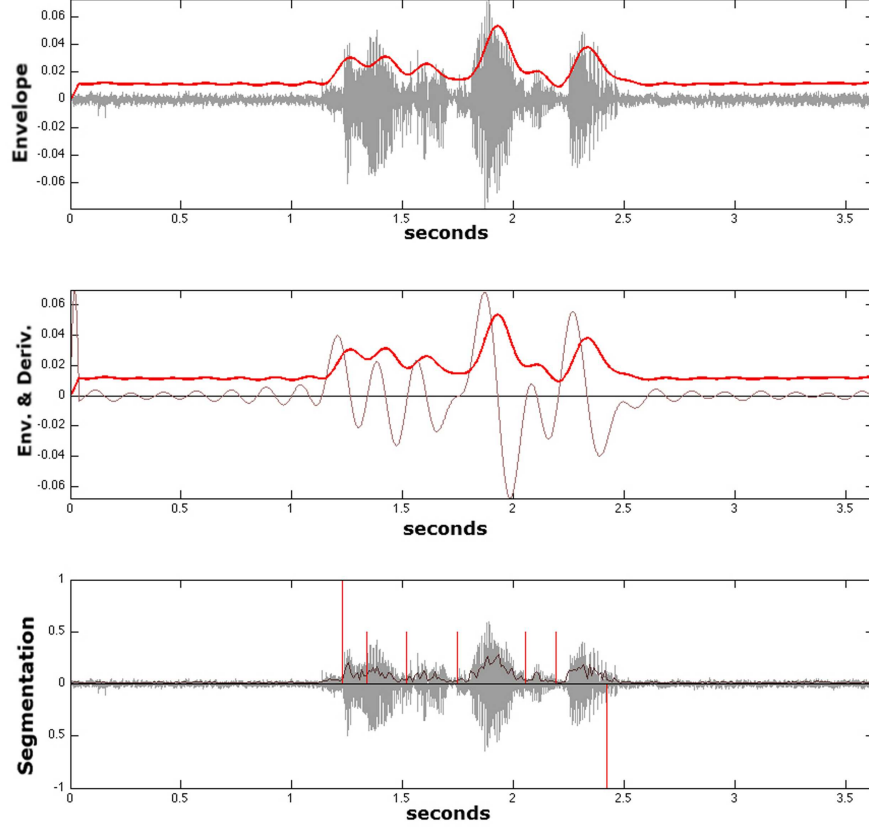
5.2 Video Analysis

Lip state detection is employed to enhance the accuracy and robustness of phone segmentation. Unlike lip-reading techniques, it is not necessary for us to identify the exact lip shape. Instead, we notice a significant change in lip change from open to close or vice versa coincide well with time markers of some phones. As such, we employ the following procedure to detect changes in the shape of the lip.

*5.2.1 Face Detection*

For each frame, the speaker's face is first detected using an Adaboost classifier on Haar-like features [44]. As this classifier is primarily for detecting frontal upright faces, the input image is rotated about the center by a range of small angles and the classifier is applied to each rotated image to ensure proper face detection.
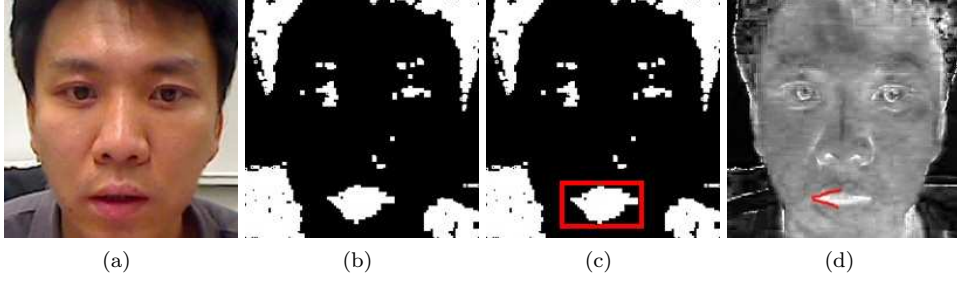
**Fig. 3** Three steps of phone segmentation: envelope calculation (top), differentiation of envelop (middle), minima identification (bottom).

### 5.2.2 Mouth Detection

In this step, we modify the approach described in [19] to detect the mouth region. The face region is first converted to the HSV space. An edge map is then obtained by applying the Sobel edge filter on the difference between the hue and the luminance channels. Connected component clustering is then applied to the edge map. The mouth blob is determined to be the largest blob straddling the vertical centerline of the face region. Sample results from these steps are illustrated in Figure 4(a)-(c).

### 5.2.3 Lip Contour Tracking

After detecting the mouth region, the contours of the lips are extracted in this step. Figure 4(d) shows the saturation channel of a face image. As the lip is more saturated in color than the skin tone, we can take advantage of this observation to track the contour of the inside of the lip and determine if the mouth is open or close. To track the contour of the inside lip, we use the active snake algorithm from [32]. A snake is simply a piece-wise linear contour that is computed based on the optimization of an appropriately chosen objective function. Two snakes, tracking the

**Fig. 4** Processing results on the face image: (a) Original face (b) Sobel Filter on hue-luminance differnce (c) Mouth boundary by blob detection (d) Inside-lip contour

upper and lower lips, are initiated from the left corner of the mouth which is detected using the feature point detection from [20]. The extension of the snakes from their starting point is guided by the gradient vector field of the hue channel. Specifically, the end-points of the segments of a snake, $\{\mathbf{P_0}, \mathbf{P_1}, \mathbf{P_2}, \dots, \mathbf{P_n}\} \in \Re^2$ with a fixed starting point $\mathbf{P_0}$, are recursively computed by maximizing the normalized line integral along each line segment:

$$\mathbf{P_i} \triangleq \arg\max_{\mathbf{P} \in S_i} \int_{\overline{\mathbf{P_{i-1}P}}} \frac{\nabla H(\mathbf{r}) \cdot d\mathbf{r}}{|\overline{\mathbf{P_{i-1}P}}|} \tag{2}$$

where $H(\cdot)$ is the hue image and $\overline{\mathbf{P_{i-1}P}}$ is the line segment between $\mathbf{P_{i-1}}$ and $\mathbf{P_i}$. $S_i$ defines the search region for the endpoint $\mathbf{P_i}$. In our implementation, the integral in Equation (2) is approximated by the corresponding discrete sum, $\nabla H(\mathbf{r})$ by the Sobel edge detector over the hue channel, and $S_i \triangleq \{(x_0 + i\Delta_x, y): y_{i-1} - \Delta_y^1 < y < y_{i-1} + \Delta_y^2)\}$ where $\mathbf{P_i} = (x_i, y_i)$. $\Delta_x$ controls the resolution along the $x$-direction and $\Delta_y^{1,2}$ controls the search range in the $y$-direction. $\Delta_y^1 < \Delta_y^2$ is used for the upper snake and $\Delta_y^1 > \Delta_y^2$ for the lower snake. These parameters are all empirically determined.

### 5.2.4 Lip States Classification

In our system, the lip shape is classified into two states: *open* and *close*. The state is determined by the angle made by the upper and lower snakes. As such, the two snakes are only grown until they reach the centerline of the mouth. Two regression lines are then fit to the upper and lower snakes and the angle $\vartheta$ between them is measured. An example of the two regression lines are shown in Figure 4(d). A temporal median filter is then used to remove noise in the time series of $\vartheta$. The *close* state is assigned to time intervals where $\vartheta$ is close to zero. Time markers are recorded when the lip state changes from *open* to *close* or vice versa. The measurement of lip state provides a visual cue for lip synchronization that cannot be provided by the audio segmentation step. However, not every phone involves closing of the lip and the number of lip-state changes is typically far fewer than the actual number of phones. To provide an accurate set of time markers, the time markers from this section must be intelligently combined with those from Section 5.1 to obtain the final answers.

5.3 Fusion of audio and video time markers for alignment

From Sections 5.1 and 5.2, we find that results from audio and video segmentations could both be used to perform temporal alignment. Yet neither one produces accurate enough markers for alignment. Audio segmentation may miss markers that separate closely spaced phones, a phenomenon that differs from speaker to speaker. Lip-state segmentation is consistent across speakers, but not every phone can be detected based on changes in lip states. Robust alignment is only possible if we combine both sets of time markers together.

The combination is performed after the time markers for both the original and the replacement speech track have been produced. We first normalize all the time markers by the corresponding duration of the speech tracks so that they are between 0 and 1. Denote the two sets of *lip-state time markers* as $M_1 = \{s_1, s_2, \ldots, s_m\}$ and $M_2 = \{t_1, t_2, \ldots, t_n\}$ with $m \leq n$. The alignment $t_{i(k)}$ in $M_2$ for each $s_k$ in $M_2$ is obtained by minimizing the following objective function:

$$S \triangleq \min \sum_{k=1}^{\min(m,n)} |s_k - t_{i(k)}| \tag{3}$$

The optimal alignment with the constraint of a monotonic increasing $i(k)$ can be obtained by applying dynamic programming to minimize $S$. The video alignment produces a coarse but reliable alignment of the two speech tracks. For each pair of corresponding segments $(s_{k-1}, s_k)$ and $(t_{i(k-1)}, t_{i(k)})$, we collect all the audio time markers within and apply exactly the same alignment procedure again to these time markers. This step provides the finer level of alignment of phones between successive video markers.

The fusion of video and audio alignment enhances the matching accuracy and is more robust to handle errors. For example, if there is an absence of audio time markers, it could decrease the overall accuracy of the time marker matching between speech tracks. However, with the fusion of the audio and video, the influence of the wrong or missing time marker can be reduced. As we first use the lip state to divide all the (audio) time marks into several subsets. Only time markers from the corresponding subset are considered for matching. So a wrong or missing audio time maker can only affect the matching accuracy of those time markers from the same subset. In this way, the errors are stopped from propagating down to the whole speech track.

## 6 VSM content synthesis

In this section, we describe the process of generating the replacement speech signal and re-sampling of the original video signal for lip synchronization based on the optimal alignment determined in Section 5.
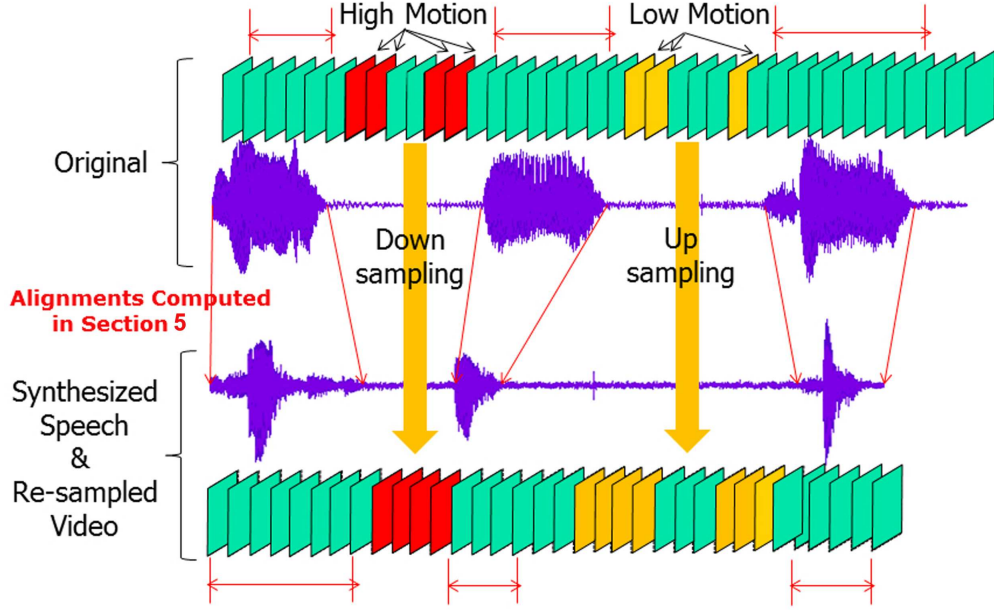
6.1 Replacement Speech Generation

To generate the replacement speech track, we have tested two different approaches - the first one is to use a commercially available text-to-speech synthesizer from Cereproc [13] and the second one is to use a speech corpus of healthy voices. The motivation of using the second approach is due to

the questionable quality of the synthesized speech from the text-to-speech engine. While the text-to-speech engine offers great flexibility in generating arbitrary scripts and produces reasonably sounding speech, it still lacks the naturalness in real human speech. Since the scripts used in a typical therapy session are usually fixed, we collect speech clips from a diverse set of individuals with healthy voice reciting the same script used in the therapy session. Then, we identify among all the speakers in the corpus the one who sounds most similar to the patient's voice. To compute speaker similarity, we use a state-of-the-art text-independent speaker identification system called ALIZE [4]. ALIZE represents individual speaker models using Gaussian Mixture Model (GMM) over linear frequency cepstral coefficient features. We use the data collected from a generic speech corpus to construct a 2048 component world GMM model, which is then adapted to individual speaker models in our voice corpus. In the actual deployment, we use the patients voice as input and find the speaker that produces the maximum likelihood ratio between the respective GMM models among all speakers in the corpus. To make the selected or generated speech signal sound even closer to the patient, we have further experimented with a non-parallel voice conversion process described in [3]. This module modifies the speech based on the vocal tract model constructed using the patients speech. The voice conversion algorithm warps the source speakers spectrum to the target spectrum in time domain using vocal tract model. During the training phase, the warping parameter and the fundamental frequency ratio are computed. During the conversion, the synthetic speech from the text-to-speech engine or the healthy voice speech selected from the corpus is warped using these parameters towards the target spectrum.

6.2 Adaptive Video Re-sampling

The objective of the video processing unit is to re-sample the input video track so that it will be lip-synchronized with the replacement speech track. Due to the differences in the word durations between the original and replacement voice tracks, adaptive re-sampling must be applied to achieve lip synchronization. During the segmentation phase in Section 5, time markers have already been identified for all segments containing phones. The result is a one-to-one mapping between the segments from the original and from the replacement speech tracks. The goal of the re-sampling scheme will be to re-sample each segment of the original video track to match the length of the corresponding segment in the replacement speech track.

The most straightforward approach is to apply uniform re-sampling for each segment independently. Based on our preliminary study, we notice that while the differences in the duration between corresponding word segments from two speech tracks are relatively small, there are large variations among the corresponding silence segments in between. Significant up-sampling or down-sampling creates unevenness in motion or motion jitter, making the resulting video unnatural. While we maintain a uniform re-sampling for all the word segments, we adopt a different approach for the silence segments to preserve the original motion as much as possible. In the case of down-sampling, we would keep more frames at the portions with higher motion to better preserve the movement. In the case of up-sampling, we would add frames or expand the static portions so that we will not slow down or distort the significant object movements. This procedure is illustrated in Figure 5.

**Fig. 5** Up-sampling and Down-sampling

This results in the proposed adaptive re-sampling algorithm for the silence segments shown in Algorithm 1. The re-sampling is treated as a process of creating the set of output frames with a target number of frames $M$ from the input set of $N$ original frames. Step 1 preserves all the original frames in the case of up-sampling. Motion energy is computed between successive frames in step 2. In step 3 and 4, we identify the pairs of original frames that have the highest or the lowest motion energy, depending on whether the goal is to up-sample or down-sample the sequence. For up-sampling, the new frame will be added to the lowest motion energy to stretch the static region. For down-sampling, the new frame is added to give priority to the portion with the highest energy. The routine INTERPOLATE is used to interpolate a video frame between two different frames. The simplest technique is to use bilinear interpolation which can lead to motion blurriness and ghosting. As such, we have also tested bidirectional interpolation based on dense optical flow vectors. The forward and backward optical flow vectors are estimated based on the pyramidal Lucas-Kanade algorithm as implemented in the OpenCV library. The flow vectors are then smoothed by a simple median filter. The temporally-scaled forward and backward vectors are then used in identifying pixels on the input frames that can be combined in creating the intermediate frames. For pixels in the intermediate frame that are not mapped by neither a forward or backward vectors, straightforward bilinear interpolation is applied. In step 6 of Algorithm 1, we deliberately remove portions of the sequence where we have already added new frames - this step prevents clustering of added frames in a small number of low/high motion areas. The parameter $\Delta$ is empirically determined to be two frames.

---

**Algorithm 1** Silence Segment Re-sampling

**Input** Input frames: $I = \{I_1, I_1, \ldots, I_N\}$
**Output** Output frames: $J = \{J_1, J_1, \ldots, J_M\}$

1. For up-sampling (i.e. $M \geq N$), insert all frames of $I$ into $J$.
2. Compute mean-square error between consecutive input frames:
   $e_i = MSE(I_i, I_{i+1})$ with $I_i, I_{i+1} \in I$.
3. For up-sampling, select the pair $(I_i, I_{i+1})$ from I with the minimum $e_i$.
4. For down-sampling ($i.e. M \leq N$), select the pair $(I_i, I_{i+1})$ from I with the maximum $e_i$.
5. Create new frame $J = INTERPOLATE(I_i, I_{i+1})$ and add $J$ into $J$ with the time order preserved.
6. Remove $I_{i-\Delta+1}, I_{i-\Delta+2}, \ldots, I_{i+\Delta}$ from I.
7. Repeat previous step 3-6 until $|J| = M$.

---

## 7 Experiments

To test the performance of our system, we capture video clips from a total of 31 participants. The participants are native English speakers of ages between 18-40. During the recording stage, each participant read the same script commonly used in speech therapy[1], which consists of a series of isolated words and short sentences. The experiment is conducted in a quiet room that only has a researcher and the testing subject involved. To accurately track the speakers face, there is a certain limit in the range of the distance (0.4m to 0.8m) that the subject sits from the camera. The size of the extracted face region ranges from $160 \times 240$ to $210 \times 325$ in pixels.

We use a Logitech QuickCam Pro 9000 to capture the video at a resolution of $640 \times 480$, and an EMU 0404/Electro Voice PL5 combination for the audio recording. The proposed algorithm is implemented in C++ with the OpenCV library that runs on a computer with the hardware setting: Intel Core$^{TM}$ i7-2820QM CPU at 2.3 GHz and 8.0GB of RAM. The video clips are on average 2 minutes and 8 seconds long, and are captured at 30fps (video) and 22.5kHz (audio). For the audio segmentation, it takes about 43 seconds to identify the boundaries between phones. The time cost for the audio-visual matching is about 3 seconds. The frame rate of lip state detection is 17 fps. It takes about 0.244 seconds to synthesize a new video frame from the original sequence.

Out of the 31 participants, 25 of whom are considered to have healthy voice. The speech recordings and the text-to-speech recordings form the candidate dataset for speech replacement. The remaining six participants are voice experts who can imitate the strained voice commonly present in patients with vocal hyerfunction. The speech tracks of their video will be replaced by one of the tracks from the candidate set, followed by voice conversion process. The optimal alignment between the original and the replacement tracks is identified and adaptive video re-sampling is applied to achieve lip-synchronization. In the sequel, we measure the performance of individual components of our proposed system.

### Replacement Speech Generation

We first consider the effect of different audio processing steps in producing a speech sample that best resembles the healthy voice of the subject. We use the following log-likelihood ratio measured

---

[1] The script can be found in http://vis.uky.edu/nsf-autism/speaktome

by ALIZE in gauging the similarity between the candidate speech $S$ and the original speech $L$:

$$LLR(S|L) = \log\left(\frac{l(S|L)}{l(S|W)}\right) \tag{4}$$

where $l(S|L)$ is the likelihood of $S$ based on the adapted GMM model generated using $L$ as the training data and $l(S|W)$ is the likelihood of $S$ based on the world GMM model. The world model is trained based on the entire TIMIT dataset [1]. This dataset contains 6300 utterances from 630 speakers with both male and female from 8 major dialect regions of United States. Table 1 shows the log-likelihood ratios of different replacement speech candidates. It is unsurprising to see

| $S$ | $LLR$ |
|---|---|
| Mimicked Voice | 9.03e-2 |
| Best-human | 2.26e-2 |
| Best-human + Voice Conversion | 0.47e-2 |
| Text-to-speech | 1.39e-2 |
| Text-to-speech + Voice Conversion | -0.63e-2 |

**Table 1** Similarity to Healthy Voice

that the mimicked voice is the one closest to the healthy voice. Among the other candidates, the best human voice is ranked top followed by the text-to-speech version. On the other hand, the application of voice conversion seems to have a detrimental effect. One possible explanation is the proper selection of the warping parameter $\alpha$ and the fundamental frequency ratio $r$. The parameters computed directly by the software produce voices that are non-human like, most likely due to the non-natural hoarseness in the mimicked voice. We have tuned the parameters in such a way that the voice is more human but it adversely affects the overall similarity to the target voice.

**Time Marker Alignment**

Once the replacement speech has been identified, the alignment process is performed between the original and the replacement speech. Time markers from the automatic audiovisual analysis are compared against the ground-truth alignment which is manually obtained by listening to the original and replacement speech tracks. To arrive at an appropriate measurement of alignment, consider a pair of speech tracks $A$ and $B$. Let the ground-truth time markers for $A$ and $B$ be $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$ and $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Note that the number of markers in A and B are identical and the phone (or silence) in $A$ during $[\mathbf{t_{i-1}}, \mathbf{t_i})$ is the same as those in $B$ during $[\mathbf{s_{i-1}}, \mathbf{s_i})$.

After our proposed alignment process, we obtain the correspondences between the automatically-determined time markers from $A$ and $B$. Denote the correspondences as $\mathbf{a}_i \leftrightarrow \mathbf{b}_i$ for $i = 1, 2, \ldots, m$ where $\mathbf{a}_i$ and $\mathbf{b}_i$ are time markers from $A$ and $B$. To compare with the ground-truth, we first identify the ground-truth interval $[\mathbf{t}_{j(i)}, \mathbf{t}_{j(i)+1})$ in $A$ that contains $\mathbf{a}_i$ for each $i$. If the ground-truth is correct, $\mathbf{a}_i$ from $A$ should roughly correspond to $\mathbf{a}'_i$ in $B$ based on linear interpolation:

$$\mathbf{a}'_i = \mathbf{s}_{j(i)} + \frac{(\mathbf{a}_i - \mathbf{t}_{j(i)})(\mathbf{s}_{j(i)+1} - \mathbf{s}_{j(i)})}{\mathbf{t}_{j(i)+1} - \mathbf{t}_{j(i)}}$$

Thus, the absolute error of the correspondence $\mathbf{a}_i \leftrightarrow \mathbf{b}_i$ is $|\mathbf{b}_i - \mathbf{a}_i'|$. Each ground-truth interval may contain zero or more such markers and we want to first measure the matching error over the entire interval. Consider the set of all automatically-determined time markers $M_A(j(i))$ in $[\mathbf{t}_{j(i)}, \mathbf{t}_{j(i)+1})$ of $A$. We can compute the average relative error:

$$E_A(j(i)) = \begin{cases} 1 \text{ if } M_A(j(i)) \text{ is empty} \\ \frac{1}{|M_A(j(i))|} \sum_{\mathbf{a} \in M_A(j(i))} \frac{|\mathbf{b} - \mathbf{a}'|}{|\mathbf{t}_{j(i)+1} - \mathbf{t}_{j(i)}|} \quad \text{otherwise} \end{cases} \tag{5}$$

where $|M_A(j(i))|$ denotes the number of markers in of $M_A(j(i))$. We use relative error so that we do not bias against long intervals. Note that for significantly skewed alignment, i.e. $\mathbf{b}_i \notin [\mathbf{t}_{j(i)}, \mathbf{t}_{j(i)+1})$, the relative error can be bigger than 1.

This error measurement is not symmetric for the case when $\mathbf{b}_i \notin [\mathbf{t}_{j(i)}, \mathbf{t}_{j(i)+1})$. To derive a symmetric metric, we reverse the role of $A$ and $B$: if $\mathbf{b}_i$ is in interval $[\mathbf{s}_{k(i)}, \mathbf{s}_{k(i)+1})$ in $B$, the time marker in $A$ that corresponds to $\mathbf{b}_i$ is

$$\mathbf{b}_i' = \mathbf{t}_{k(i)} + \frac{(\mathbf{b}_i - \mathbf{s}_{k(i)})(\mathbf{t}_{k(i)+1} - \mathbf{t}_{k(i)})}{\mathbf{s}_{k(i)+1} - \mathbf{s}_{k(i)}}$$

The absolute error in this direction would be $|\mathbf{a}_i - \mathbf{b}_i'|$ and the corresponding average relative error within $[\mathbf{s}_{k(i)}, \mathbf{s}_{k(i)+1})$ is analogously defined:

$$E_B(k(i)) = \begin{cases} 1 \text{ if } M_B(k(i)) \text{ is not empty} \\ \frac{1}{|M_B(k(i))|} \sum_{\mathbf{b} \in M_B(k(i))} \frac{|\mathbf{a} - \mathbf{b}'|}{|\mathbf{s}_{k(i)+1} - \mathbf{s}_{k(i)}|} \quad \text{otherwise} \end{cases} \tag{6}$$

Finally, a symmetric average relative error over all the ground-truth intervals can be computed as follows:

$$E = \frac{1}{2n} \sum_{i=1}^{n} (E_A(i) + E_B(i)) \tag{7}$$

Using the best human sample as the replacement speech, we measure the average relative error of the alignment for the six strained speech sample. The results are tabulated in Table 2. For comparison, the performance of using audio only and video only for alignment are also listed. In

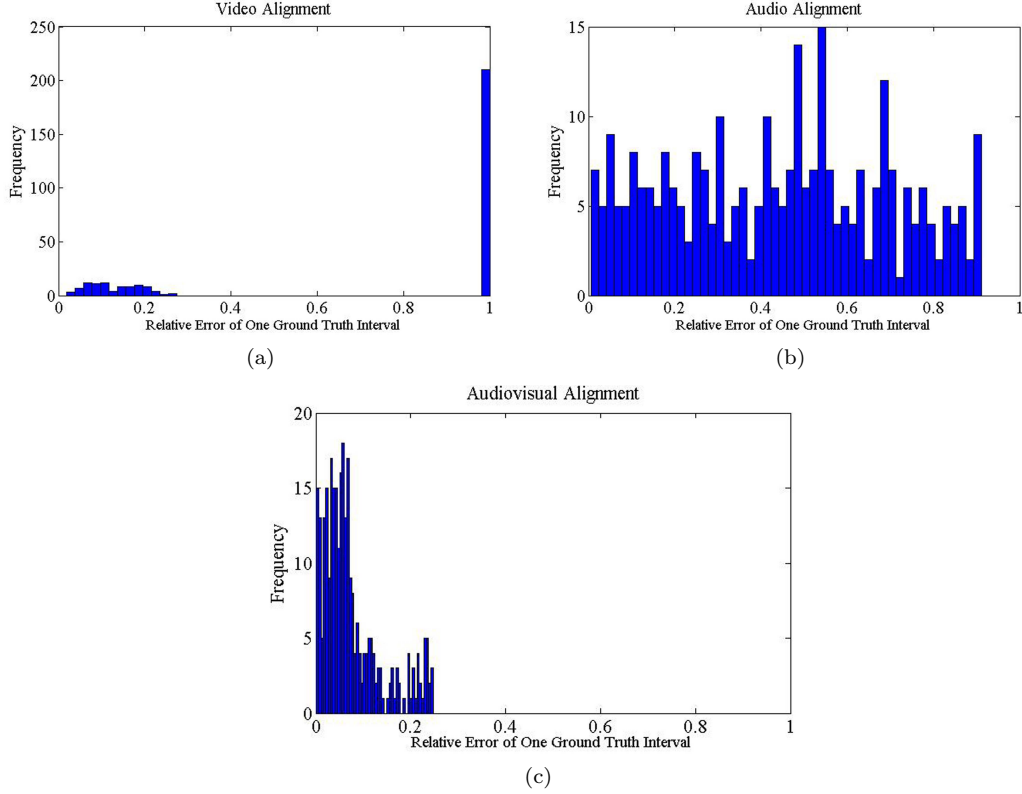| Video Pair Number | Error (video) | Error (audio) | Error (combined) |
|---|---|---|---|
| $< S_1, H_4 >$ | 0.7350 | 0.1854 | 0.0760 |
| $< S_2, H_{10} >$ | 0.7413 | 0.2066 | 0.0859 |
| $< S_3, H_{10} >$ | 0.7472 | 0.2040 | 0.0692 |
| $< S_4, H_{23} >$ | 0.7378 | 0.2125 | 0.0751 |
| $< S_5, H_{24} >$ | 0.7408 | 0.1721 | 0.0704 |
| $< S_6, H_1 >$ | 0.7458 | 0.1714 | 0.0767 |

**Table 2** Results of forced choice tests

the table, the first column $< S_*, H_* >$ are pairs of strained and healthy voices, in which each healthy voice is identified by ALIZE measurement as one bearing maximum resemblance to the corresponding strained voice. The third column shows the average relative error for alignment using lip-state changes only. The error is very high but consistent across different speakers. The fourth column shows the average relative error for alignment using audio segmentation. The results are

better than those using lip-state only but higher variation is observed across different speakers. The last column shows the average relative error of combining both together using our proposed scheme. There is a dramatic reduction in the relative error and the results are accurate across different speakers.
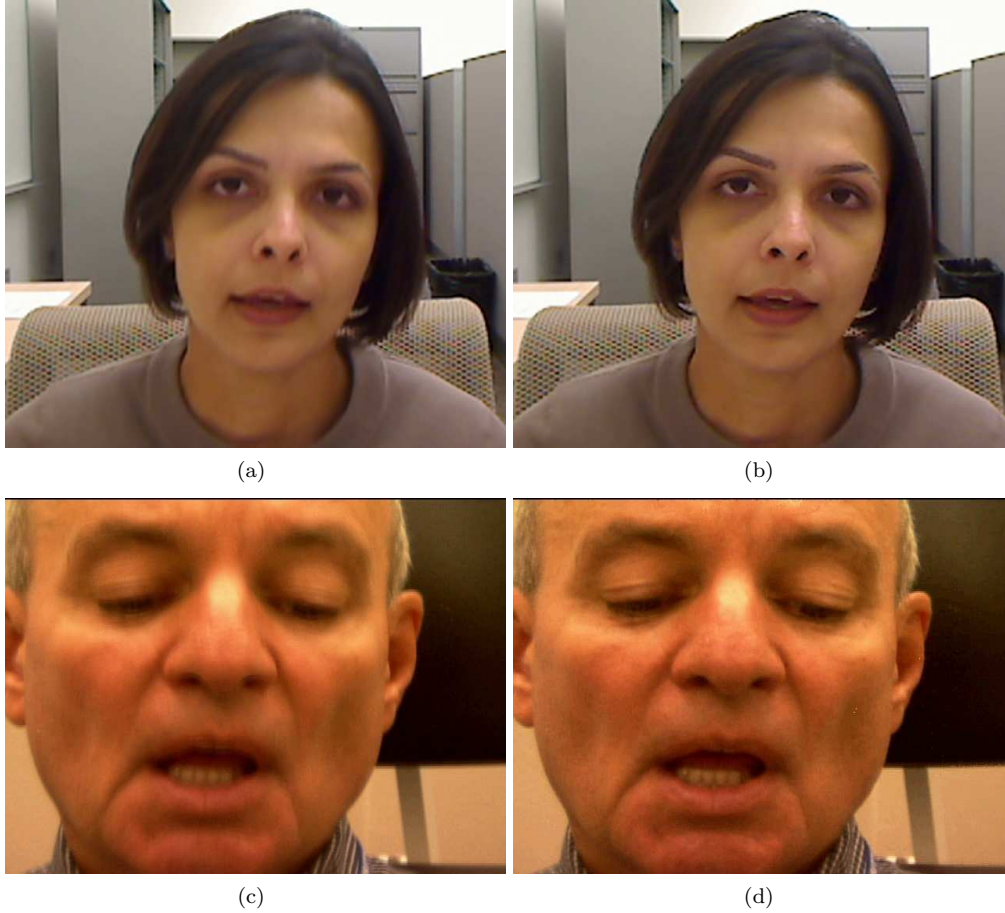
While Table 2 shows the average relative error, more information can be obtained by showing the histogram of the relative error $(E_A(i) + E_B(i))/2$ across all ground-truth intervals for all pairs of matched speech tracks. The three histograms corresponding to video only, audio only, and the combined scheme are shown in Figure 6(a), 6(b), and 6(c) respectively. The bimodal nature of Figure 6(a) indicates that many phones cannot be detected with the changes of lip-states. The high variance in the relative error shown in Figure 6(b) is characteristics of audio segmentation as the accuracy is highly speaker-dependent. The sharp concentration on low average error shown in Figure 6(c) shows the superior performance of our proposed audiovisual scheme over the other two.



**Fig. 6** Histograms of relative error using (a) video only, (b) audio only, and (c) proposed audiovisual approach

## Video Interpolation

Figure 7(a) and 7(c) show two sample frame using bilinear interpolation, while Figure 7(b) and 7(d) the corresponding frame using optical flow interpolation. As expected, optical-flow interpolation produces a much sharper image, especially around high-motion areas such as eyelids and mouth.
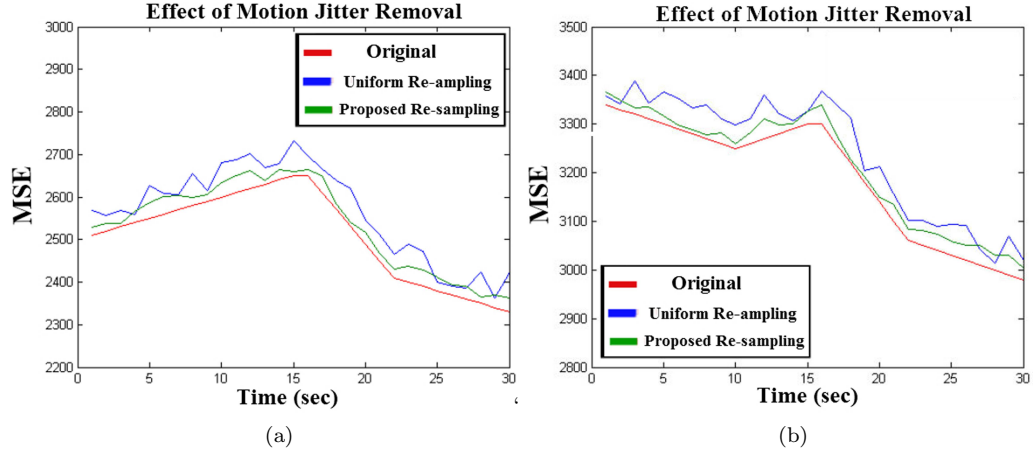


<div align="center">(a)                                                (b)</div>

<div align="center">(c)                                                (d)</div>

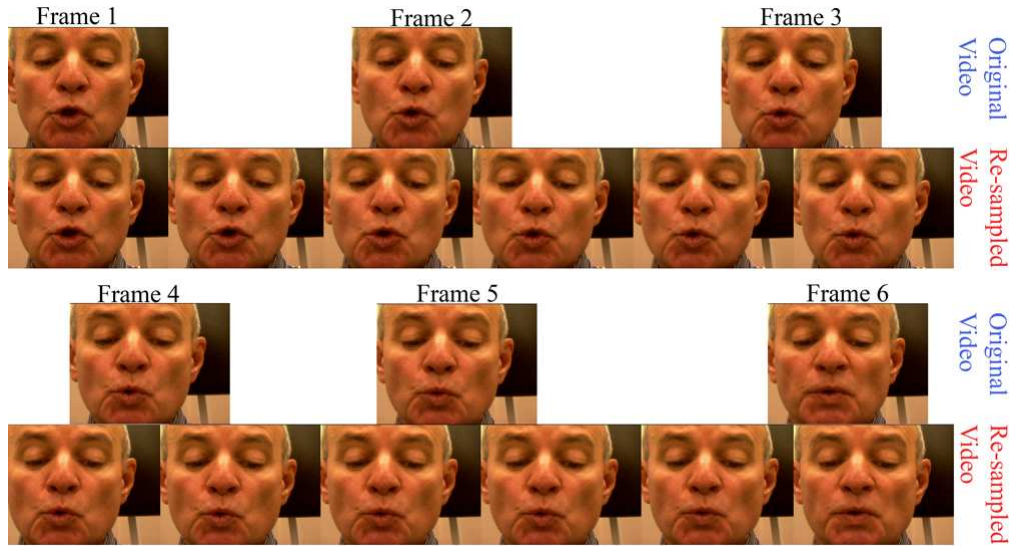**Fig. 7** Interpolation methods comparison: (a),(c) Bilinear Interpolation; (b),(d) Optical Flow Interpolation

## Adaptive Re-sampling

We also study the effect of our adaptive re-sampling of silence segments. In Figure 8(a), we first plot the MSE measurements between successive frames for the original sequence. While keeping all the "word" segments intact, we reduce all the silence segments into one quarter of their original length. Two methods are tested: uniform re-sampling and our proposed adaptive re-sampling. MSE between consecutive frames are then measured and the curves re-sampled to be the same time scale as the original curve. As shown in Figure 8(a), our proposed approach provides a curve that can better

preserve the original temporal energy than the uniform re-sampling approach. Figure 8(b) shows a similar trend when we up-sample all silence segments by a factor of four. Figure 9 demonstrates the up-sampling case. To synchronize the lip motion with the replacing audio, the original video (on the first and third rows) is prolonged by generating multiple intermediate frames.



**Fig. 8** MSE curves by Uniform Re-sampling and Adaptive Re-sampling: (a) down-sampling (b) curve-sampling



**Fig. 9** Video Sequence Re-sampling Example: the original video (row 1, 3) is prolonged in the synthesized video (row 2, 4).

**Subject Testing**

The above objective measurements, however, may not truly reflect the subjective quality of the overall videos. As such, we have also performed a series of subjective evaluation tests. Two sets of questions are administered to 15 test takers who are unaware of the details of our proposed system. In the first test, they were asked to view and compare different video sequences and rate which one is more natural and of higher quality. The key to the acronyms used in describing different schemes are as follows: CA = Combined Alignment, AA = Audio-only Alignment, VC = Voice Conversion, TTS = text-to-speech, BH = best-human voice, BI = binary interpolation, AS = adaptive re-sampling and US = uniform re-sampling. The average results for different tests are tabulated in Table 3. The results for the 15 tests are as expected: testers prefers best-human over TTS, the absence of voice conversion, optical flow over bilinear interpolation, and adaptive re-sampling over uniform re-sampling in the test.

| Test | % favored $1^{st}$ | Common parameters |
|------|------|------|
| CA vs. AA | 100% | BI+AS, BH |
| no VC vs. VC | 100% | BI+AS, TTS |
| no VC vs. VC | 100% | BI+AS, BH |
| BH vs. TTS | 100% | BI+AS, VC |
| BH vs. TTS | 100% | BI+AS |
| OI vs. BI | 100% | AS, BH |
| US vs. AS | 100% | BI, BH |

**Table 3** Results of forced choice tests

In the second test, the testers first view the mimicked voice video. Then, they are asked to view five different videos and rank them based on their likelihood of being the heathy voice after therapy. The results of the subjective evaluation are given in Table 4. While most testers choose the "correct" answer, i.e. the real video with heathy voice, the synthetic video with best human comes close. This result is promising as it demonstrates the possibility of using synthetic video in depicting unseen behavior of an individual, which is precisely the goal of the VSM therapy.

| Test Video | Average Rank |
|------|------|
| Healthy Voice | $1.8 \pm 1.8$ |
| BI+AS, BH | $2.0 \pm 0.7$ |
| BI+AS, BH+VC | $3.4 \pm 1.3$ |
| BI+AS, TTS | $3.2 \pm 0.4$ |
| BI+AS, TTS+VC | $4.6 \pm 0.5$ |

**Table 4** Results of rank test

# 8 Conclusion

In this paper, we have demonstrated the use of computational multimedia techniques in automatically generating video material for video self modeling intervention. The advantage of computational techniques lies in its flexibility in creating unseen behaviors. Our proposed system is designed specifically for voice therapy and produces a video with the patient's coarse voice replaced by a healthy voice. Experimental results have shown that natural human voice selected through speaker similarity provides the best subjective results. No additional benefit has been found by using voice conversion techniques due to the inaccurate target models created with the coarse voice. Optimal alignment between the original and replacement speech has been accomplished through a combination of automatic audio segmentation and lip-state extraction. Based on the alignment, an adaptive re-sampling algorithm has been proposed to preserve the motion energy during the lip-synchronization process. Extensive objective and subjective evaluations have demonstrated the advantages of our design and a clinical test is currently underway to study the effectiveness of our system in a larger scale. While our proposed system is domain specific, we believe that the concept of using multimedia techniques for video self modeling has far-reaching importance in many different areas of health care and behavioral intervention.

# References

1. TIMIT acoustic-phonetic continuous speech corpus.
2. A, H.: Automatic syllable detection for vowel landmarks. PhD Thesis (2000)
3. et al., D.S.: Time domain vocal tract length normalization. In: Signal Processing and Information Technology (2004)
4. et al., J.F.B.: Nist'04 speaker recognition evaluation campaign: new lia speaker detection platform based on alize toolkit. In: Proceedings of NIST Speaker Evaluation (2004)
5. Aleksic, P., Katsaggelos, A.: roduct HMMs for audio-visual continuous speech recognition using facial animation parameters. In: International Conference on Multimedia and Expo (ICME), pp. 481–484 (2003)
6. Alvero, A.M., Austin, J.: The effects of conducting behavioral observations on the behavior of the observer. Journal of Applied Behavior Analysis **37**, 457–468 (2004)
7. Arsic, I., Thiran, J.: Mutual information engenlips for audio-visual speech. In: 14th European Signal Processing Conference (2006)
8. A.W.C Liew S.H. Leung, W.L.: Lip contour extraction using deformalbe model. In: International conference on Image Processing (2000)
9. Bandura, A.: Self-effiicacy: The Exercise of Control. New York: Freeman (1997)
10. Bartels, J.G.A.S.C., Bilmes, J.: Dbn based multi-stream models for audio-visual speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 993–996 (2004)
11. Boone, D.R., McFarlane, S.C.: The Voice and Voice Therapy. Prentice Hall (2006)
12. Buggey, T.: Seeing Is Believing: Video Self Modeling for people with Autism and other developmental disabilities. Wodbine House (2009)
13. CereProc, http://www.cereproc.com: Text to Speech Technology
14. Chen, J., Tiddeman, B., Zhao, G.: Real-time lip contour extraction and tracking using an improved active contour model. In: Lecture Notes in Computer Science, vol. 5359, pp. 236–245 (2008)

15. Deng, Z., Neumann, U.: Expressive speech animation synthesis with phoneme-level controls. Computer Graphics Forum **27**, 2096–2113 (2008)

16. Dowrick, P.W.: Self-modeling. In: Using Video: Psychological and Social Applications. New York: Wiley (1983)

17. Duy, N., David, H.: Real-time face detection and lip feature extraction using field-programmable gate arrays. In: IEEE Trans. Systems Man Cybernet, pp. 902–912 (2006)

18. Eveno, N., Caplier, A., Coulon, P.Y.: Key points based segmentation of lips. In: IEEE International Conference on Multimedia and Expo, 2002. (2002)

19. Eveno, N., Caplier, A., Coulon, P.Y.: Accurate and quasi-automatic lip tracking. Circuits and Systems for Video Technology, IEEE Transactions on **14**, 706 – 715 (2004)

20. feature point extraction, A., tracking in image sequences for arbitrary camera motion: Qineen zheng and rama chellappa. International journal of computer vision **15**, 31 – 76 (1995)

21. Garg, G.P.C.N.G.G.A., Senior, A.: Recent advances in the automatic recognition of audio-visual speech. In: Proc. IEEE, vol. 91, pp. 1306–1326 (2003)

22. Gurban, M., Thiran, J.: Audio-visual speech recognition with a hybrid svm-hmm system. In: 13th European Signal Processing Conference (2005)

23. HAMMAL, Z., N.EVENO, A.CAPLIER, COULON, P.: Parametric models for facial features segmentation. In: IEEE Journal in Signal Processing (2005)

24. Hapner E Portone-Maira C, J.M.: A study of voice therapy dropout. j voice. Journal of Voice **23**, 337–40 (2009)

25. Hitchcock, C.H., Dowrick, P.W., Prater, M.A.: Video self-modeling intervention in school-based settings: A review. Remedial and Special Education **24**(1), 36–45 (2003)

26. Kaucic, R., Dalton, B., Blake, A.: Real-time lip tracking for audio-visual speech recognition applications. In: Lecture Notes in Computer Science, vol. 1065, pp. 376–387 (1996)

27. Kazuhiro, N., Noriaki, M., Kazuyoshi, T., Naofumi, T.: A real-time lip reading lsi for word recognition. In: Proc. IEEE Conf. ASIC, pp. 303–306 (2002)

28. Krouse, H.J.: Video modeling to educate patients. Journal of Advanced Nursing **33**, 748–757 (2001)

29. Ma, J., Cole, R., Pellom, B., Ward, W., Wise, B.: Accurate visible speech synthesis based on concatenating variable length motion capture data. IEEE Transactions on Visualization and Computer Graphics **15**, 485–500 (2005)

30. MacKenzie, K., Millar, A., Wilson, J.A., Sellars, C., Deary, I.J.: Is voice therapy an effective treatment for dysphonia? A randomized controlled trial. British Medical Journal **323**, 658–661 (2001)

31. McDaniel, R.W., v. A. Rhodes: Development of a preparatory sensor information videotape for women receiving chemotherapy for breast cancer. Cancer Nursing **21**, 143–148 (1998)

32. contour models, S.A.: Michael kass and andrew witkin and demetri terzopoulos. INTERNATIONAL JOURNAL OF COMPUTER VISION **14**, 321 – 331 (1988)

33. Nielsen, D., Sigurdsson, S.O., Austin, J.: Preventing back injuries in hospital settings: the effects of video modeling on safe patient lifting by nurses. Journal of Applied Behavioral Analysis **42**(3), 551–561 (2009)

34. P, M.: Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. **58**, 880–883 (1975)

35. P, M.: Automatic segmentation of speech into syllables. ECST pp. 2009–2013 (1987)

36. Patel, R., Bless, D., Thibeault, S.: A novel intensive approach to voice therapy. Journal of Voice **25**, 562–569 (2011)

37. Queiroz, R., Cohen, M., Musse, S.R.: An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behavior. ACM Computers in Entertainment **7**(4), 58:1 – 58:20 (2009)

38. Ramachandran, V.S., Rogers-Ramachandra, D.C., Cobb, S.: Touching the phantom. Nature **377**, 489–490 (1995)

39. Ramig, L.O., Verdolini, K.: Treatment efficacy: Voice disorders. Journal of Speech, Language and Hearing Research **41**, 101–116 (1998)

40. Roy, N., Weinrich, B., Gray, S., Tanner, K., Stemple, J.C., Sapienza, C.M.: Three treatments for 2 teachers with voice disorders: a randomized clinical trial. J Speech Lang Hear Res **46**, 670–688 (2003)

41. Roy N Bless DM, H., NC, F.: Manual circumlaryngeal therapy for functional dysphonia: an evaluation of short- and long-term treatment outcomes. Journal of Voice **11**, 321–331 (1993)

42. Shen, J., Raghunathan, A., Cheung, S.C., Patel, R.: Automatic content generation for video self modeling. In: Proceedings of IEEE International Conference on Multimedia Expo (ICME 2011) (2011)

43. Verdolini, K., Ramig, L.O.: Review: occupational risks for voice problems. Logopedics, Phoniatrics, Vocology **26**(1), 37–46 (2001)

44. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, pp. 511–518 (2001)

45. Xie Z, N.P.: Robust acoustic-based syllable detection. In: INTERSPEECH'06 (2006)