7-2011

Automatic Content Generation for Video Self Modeling

Ju Shen
University of Dayton, jshen1@udayton.edu

Anusha Raghunathan
Intel Corporation

Sen-ching S. Cheung
University of Kentucky

Ravi R. Patel
University of Kentucky

AUTOMATIC CONTENT GENERATION FOR VIDEO SELF MODELING

*Ju Shen, Anusha Raghunathan, Sen-ching S. Cheung and Rita Patel**

University of Kentucky
{ju.shen, anusha.raghunathan, sen-ching.cheung, rita.patel}@uky.edu

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. Its effectiveness in rehabilitation and education has been repeatedly demonstrated but technical challenges remain in creating video contents that depict previously unseen behaviors. In this paper, we propose a novel system that re-renders new talking-head sequences suitable to be used for VSM treatment of patients with voice disorder. After the raw footage is captured, a new speech track is either synthesized using text-to-speech or selected based on voice similarity from a database of clean speeches. Voice conversion is then applied to match the new speech to the original voice. Time markers extracted from the original and new speech track are used to re-sample the video track for lip synchronization. We use an adaptive re-sampling strategy to minimize motion jitter, and apply bilinear and optical-flow based interpolation to ensure the image quality. Both objective measurements and subjective evaluations demonstrate the effectiveness of the proposed techniques.

Index Terms— video self modeling, positive feedforward, voice disorder, computational multimedia, frame interpolation, voice imitation

1. INTRODUCTION

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 recovered from surgery [1] and cancer [2] as well as job and safety training for hospital staffs [3] and office workers [4]. 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 [5].

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 [6]. 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 studied for many different types of disabilities and behaviorial problems ranging from stuttering, inappropriate social behaviors, autism, selective mutism to sports training. A summary of these research can be found in [7].

There are two forms of VSM: positive self-review and feedforward [8]. 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 feedforward VSM shows novel skills that have never been observed but still within the reach of the learner. The goal is to teach new skills to a learner. It is the feedforward approach that shows more dramatic learning effect than the positive self-review approach as there is more room for improvements during the initial stages of learning. Creating feedforward video, however, is difficult. An example of feedforward VSM can be found in [7] where the author splices short clips to form a long video of a full sentence from a child who can only speak one or two-word utterances. 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.

In this paper, we consider the use of computational multimedia techniques in creating feedforward VSM contents. From computer generated imagery to speech synthesis, there exists a myriad of multimedia tools that can synthesize realistic video contents. Our goal is to develop feedforward VSM systems that can be used by a learner and his/her therapist in creating VSM contents with minimum amount of training data. The synthesis process should be automatic and in real-time so that rapid feedback can be provided. The synthetic content should be perceptually indistinguishable from real video footage. In order to have our design evaluated by

*The first three authors are with the Department of Electrical and Computer Engineering and the fourth author is with the Department of Health Sciences - Rehabilitation Science.

978-1-61284-350-6/11/$26.00 ©2011 IEEE

domain experts, we focus on the development on a novel feed-forward VSM system for patients suffered from vocal hyper-function.

Using a raw talking-head video footage of a patient reciting a known script, our system re-samples the image frames with minimal motion jitter to lip-synchronize with a new speech track from a healthy voice. Our goals bear some resemblance with the large body of work in facial animation using either real video footage [9] and avatars [10, 11]. 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 [9]. Also it is unimportant for us to preserve emotion as in [10, 11]. 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 rest of the paper is organized as follows: in Section 2, we describe the types of voice disorders we are targeting and motivate the treatment potential of VSM. The proposed system is described in Section 3 and experimental results are presented in Section 4. We conclude the paper with future work in Section 5.

2. VSM FOR VOICE DISORDER

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 [12]. Generally, vocal hyperfunction can be effectively treated with behavioral voice treatment or voice therapy. Participation in voice therapy requires regular, at least once a week, voice therapy sessions with the speech-language pathologist over a period of at least two months to facilitate behavioral change in production [13]. Considering that 3-9% of the adult population under 65 years in U.S. present with voice disorders, this is a significant medical problem [14]. 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 visit to the clinic or access them through internet, and continue their behavioral modeling 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. A clinical study is currently underway at the our clinical voice center to test the effectiveness of the proposed VSM therapy.

3. PROPOSED SYSTEM

3.1. Overview

Figure 1 shows the user interface of the system. The left figure depicts the interface for capturing the raw video from the patient through a web camera situated on top of the laptop computer. 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 in a translucent box near the camera. The right figure shows the user interface where the patients can review different rendered sequences with the accompanying scripts.

Fig. 1. User Interface

Figure 2 shows 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 word 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 word 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. Details of the algorithms used for video and audio processing are provided in the next two sub-sections.

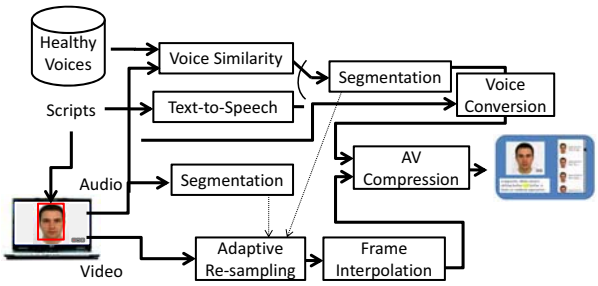

Fig. 2. VSM Content Generation

3.2. Audio Processing

The audio segmentation module is used to identify the time markers that correspond roughly to the boundaries of words. The module has an initial training period of a few seconds before the subject starts speaking. The module determines the ambient noise amplitude level during this training period and uses the average loudness as the threshold level for the segmentation. For the remaining portion of the captured audio, the module identify the beginning of words at the instances when the short-term average loudness level increases above the threshold, and the end of the words at those instances when the loudness level deceases below the threshold for a sufficiently long period of time.

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 [15] 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 [16]. 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 patient's 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 a non-parallel voice conversion process described in [17]. This module modifies the speech based on the vocal tract model constructed using the patient's speech. The voice conversion algorithm warps the source speaker's spectrum to the target spectrum in time domain using vocal tract model. During the training phase, the warping parameter α and the fundamental frequency ratio r 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.

3.3. Video Processing

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 audio segmentation, time markers have already been identified for all segments containing spoken words. There is a one-to-one mapping between the segments (word and silence) from the original and from the replacement voice 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 an 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 results in the proposed adaptive re-sampling algorithm for the silence segments shown in Algorithm 1. 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 Δ is empirically determined to be two frames.

In step 5 of Algorithm 1, 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.

Algorithm 1 Silence Segment Re-sampling

Input Input frames: $\mathbf{I} = \{I_1, I_2, \ldots, I_N\}$
Output Output frames: $\mathbf{J} = \{J_1, J_2, \ldots, J_M\}$

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

4. EXPERIMENTAL RESULTS

In our preliminary experiment, we use two video sequences of a voice expert who is familiar with the voice characteristics of vocal hyperfunction. He prepared two sequences – one using his natural voice and the other mimicking that of a patient suffered from vocal hyperfunction. The two videos are on average 43 seconds long and are captured at a video frame rate of 30 fps and audio sampling rate at 16 khz. The script recited consists of isolated words and a couple of sentences with pauses between each. Despite the fact that we have the normal voice of our voice expert, this normal-voice clip is never used in any of the training of our speaker identification system for identifying similar voices or the vocal tract modeling used in the voice conversion module. In all cases, we use the mimicked voice as the target as it would have been the only data available if he was a true patient. The normal-voice is only used for comparison.

For the CereProc text-to-speech synthesizer, we manually identify one character with southern English accent named "William" to be the closest match to our voice expert. As for the real-voice dataset, we have identified three individuals in the same gender, age and race group as our voice expert and have recorded their voices reciting the same script. None of these three individuals claim to have any voice disorder.

First we compare the effect of image interpolation. Figure 3 shows a sample frame using bilinear interpolation while Figure 4 shows 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. We also study the effect of our adaptive re-sampling of silence segments. In Figure 5, 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 are then re-sampled to be the same time scale as the original curve. As shown in Figure 5, our proposed approach provides a curve that can better preserve the original temporal energy than the uniform re-sampling approach. Figure 6 shows a similar trend when we up-sample all silence segments by a factor of four.

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

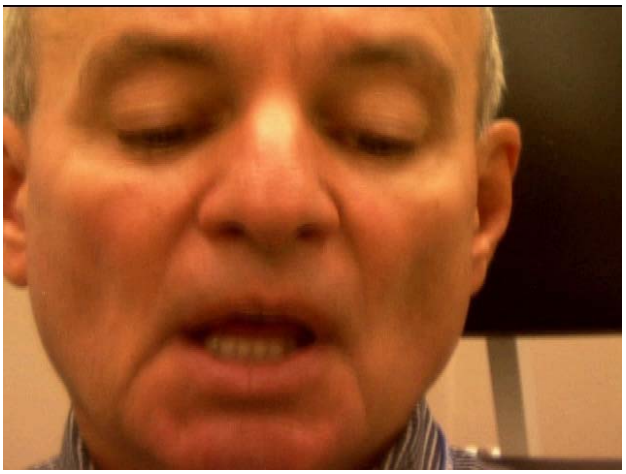

Fig. 3. Bilinear Interpolation

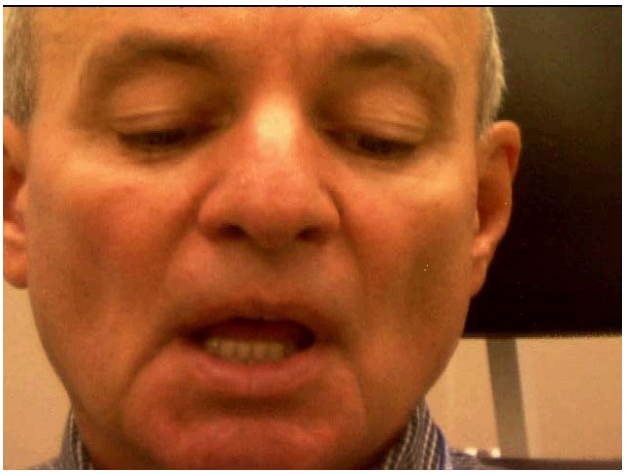

Fig. 4. Optical-flow Interpolation

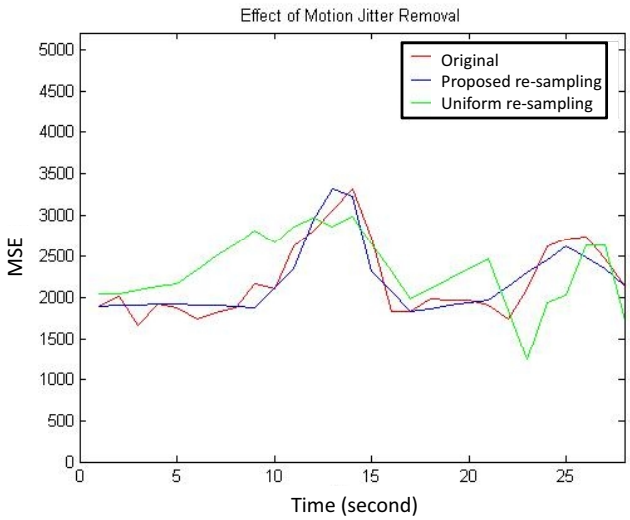

Fig. 5. MSE curves for down-sampling

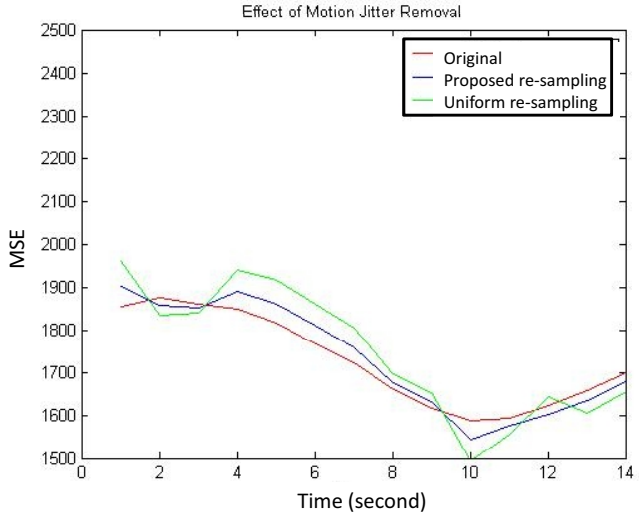

Fig. 6. MSE curves for up-sampling

measured by ALIZE in gauging the similarity between the synthesized voice S and the original voice L:

$$LLR(S|L) = \log\left(\frac{l(S|L)}{l(S|W)}\right) \quad (1)$$

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. Table 1 shows the log-likelihood ratios of different synthesized voices. It is unsurprising to see that the mimicked voice is the one closest to the healthy voice. Among the synthetic ones, 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 α and the fundamental frequency ratio r. The parameters computed directly by the software produce voices that are non-human like, mostly 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.

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

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 five 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 keys to the acronyms used in describing different schemes are as follows: 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 2. The results for the first five tests are as expected: testers prefers best-human over TTS, the absence of voice conversion, and optical flow over bilinear interpolation. On the other hand, it is surprising that 60% testers prefer uniform re-sampling over adaptive re-sampling in the last test. One possible reason is that as we have conducted this test over the web, the network jitter may negate the benefit brought forth by adaptive re-sampling. We are currently conducting more experiments with a larger user group in a more controlled setting.

Test	% favored 1^{st}	Common parameters
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	60%	BI, BH

Table 2. 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 3. 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 ± 1.8
BI+AS, BH	2.0 ± 0.7
BI+AS, BH+VC	3.4 ± 1.3
BI+AS, TTS	3.2 ± 0.4
BI+AS, TTS+VC	4.6 ± 0.5

Table 3. Results of rank test

5. CONCLUSIONS

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 is its flexibility in creating unseen behaviors that can rival the true reality. The proposed system is designed specifically for voice therapy. Replacement track generated with natural human voice best resembled the input represents the current best option for audio. Optical flow interpolation with adaptive re-sampling is used to lip-synchronize the original video with the replacement audio track. Preliminary 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.

6. REFERENCES

[1] H. J. Krouse, "Video modeling to educate patients," *Journal of Advanced Nursing*, vol. 33, pp. 748–757, 2001.

[2] R. W. McDaniel and v. A. Rhodes, "Development of a preparatory sensor information videotape for women receiving chemotherapy for breast cancer," *Cancer Nursing*, vol. 21, pp. 143–148, 1998.

[3] D. Nielsen, S. O. Sigurdsson, and J. Austin, "Preventing back injuries in hospital settings: the effects of video modeling on safe patient lifting by nurses," *Journal of Applied Behavioral Analysis*, vol. 42, no. 3, pp. 551–561, 2009.

[4] A. M. Alvero and J. Austin, "The effects of conducting behavioral observations on the behavior of the observer," *Journal of Applied Behavior Analysis*, vol. 37, pp. 457–468, 2004.

[5] C. H. Hitchcock, P. W. Dowrick, and M. A. Prater, "Video self-modeling intervention in school-based settings: A review," *Remedial and Special Education*, vol. 24, no. 1, pp. 36–45, January/Februrary 2003.

[6] V. S. Ramachandran, D. C. Rogers-Ramachandra, and S. Cobb, "Touching the phantom," *Nature*, vol. 377, pp. 489–490, 1995.

[7] T. Buggey, *Seeing Is Believing: Video Self Modeling for people with Autism and other developmental disabilities*, Wodbine House, 2009.

[8] P. W. Dowrick, "Self-modeling," in *Using Video: Psychological and Social Applications*. New York: Wiley, 1983.

[9] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate visible speech synthesis based on concatenating variable length motion capture data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 485–500, 2005.

[10] R. Queiroz, M. Cohen, and S. R. Musse, "An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behavior," *ACM Computers in Entertainment*, vol. 7, no. 4, pp. 58:1 – 58:20, December 2009.

[11] Z. Deng and U. Neumann, "Expressive speech animation synthesis with phoneme-level controls," *Computer Graphics Forum*, vol. 27, pp. 2096–2113, 2008.

[12] D. R. Boone and S. C. McFarlane, *The Voice and Voice Therapy*, Prentice Hall, 2006.

[13] L. O. Ramig and K. Verdolini, "Treatment efficacy: Voice disorders," *Journal of Speech, Language and Hearing Research*, vol. 41, pp. 101–116, 1998.

[14] K. Verdolini and L. O. Ramig, "Review: occupational risks for voice problems," *Logopedics, Phoniatrics, Vocology*, vol. 26, no. 1, pp. 37–46, 2001.

[15] CereProc, http://www.cereproc.com, *Text to Speech Technology*.

[16] J. F. Bonastre et al., "Nist'04 speaker recognition evaluation campaign: new lia speaker detection platform based on alize toolkit," in *Proceedings of NIST Speaker Evaluation*, 2004.

[17] D. Sundermann et al., "Time domain vocal tract length normalization," in *Signal Processing and Information Technology*, 2004.