# Acquisition of a 3D Audio-Visual Corpus of Affective Speech

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise,
and Luc Van Gool

**Abstract**

Communication between humans deeply relies on our capability of experiencing, expressing, and recognizing feelings. For this reason, research on human-machine interaction needs to focus on the recognition and simulation of emotional states, prerequisite of which is the collection of affective corpora. Currently available datasets still represent a bottleneck because of the difficulties arising during the acquisition and labeling of authentic affective data. In this work, we present a new audio-visual corpus for possibly the two most important modalities used by humans to communicate their emotional states, namely speech and facial expression in the form of dense dynamic 3D face geometries. We also introduce an acquisition setup for labeling the data with very little manual effort. We acquire high-quality data by working in a controlled environment and resort to video clips to induce affective states. In order to obtain the physical prosodic parameters of each utterance, the annotation process includes: transcription of the corpus text into the phonological representation, accurate phone segmentation, fundamental frequency extraction, and signal intensity estimation of the speech signals. We employ a real-time 3D scanner for the recording of dense dynamic facial geometries and track the faces throughout the sequences, achieving full spatial and temporal correspondences. The corpus is not only relevant for affective visual speech synthesis or view-independent facial expression recognition, but also for studying the correlations between audio and facial features in the context of emotional speech.

## I. Introduction

With their increasing abilities, computers are becoming more and more part of our everyday life. Being it computer-driven information kiosks, intelligent automobile systems, games, animated movies, or serving robots for museums or hospitals, the one- or bi-directional communication between humans and machines has already become common practice. In order for computer systems to become fully integrated in people's everyday life, they need to be able to recognize and simulate affective states, which are fundamental capabilities in human-human communications. Interacting with an artificial agent will never be perceived as natural, particularly by people who are less familiar with computers, unless the machine can guess the user's current emotional state and react accordingly.

Although there has been substantial progress in the field of affective computing during the last few years [1], the development of emotion-capable systems greatly depends on the collection of affective corpora. Datasets are needed for the training and evaluation of algorithms for both recognition and synthesis of emotional states. Because authentic data is very hard to collect and label, publicly available datasets are still a bottleneck for research in this field. Affective expressions in humans are rare, often suppressed, highly context-related, and thus difficult to capture. For this reason, many of the studies on affective computing s concentrated so far on acquiring posed data. However, research has shown that spontaneous emotional behavior clearly differs from its deliberate counterpart; Whissell shows this in [2] for what concerns spoken language, while numerous other works studied the differences between posed and natural facial expressions, as in [3] and [4]. For recording spontaneous affective behavior, a good trade-off between the acquisition of natural emotional expressions and data quality needs to be found: corpora collected in controlled environments are by definition unnatural, but moving towards more unconstrained settings increases the amount of noise. When accurate data is required, induction methods represent a good compromise, and the literature is rich of examples where videos [5], still photographs [6], music [7], or manipulated games [8] have been used to elicit emotions. Although these methods are not a replacement of pure naturalism, they are well established and have shown to evoke a range of authentic emotions in a laboratory environment [9].

Another important issue related to the acquisition of an affective corpus is the type and number of modalities which should be captured. In the past few years, the research community has converged towards the idea that automatic affect recognition systems should use cues coming from several modalities, in a way imitating humans. Even though emotional cues can be extracted from physiological measurements (e.g. [10], [11]), invasive methods might influence the subject's emotional state. Humans do not have access to the physiological state of others, yet we are very good at guessing someone's affective state using cues such as facial expressions, voice modulation, and body pose.

In this work, we focus on the voice and facial expressions modalities and present an acquisition setup for building an audio-visual corpus with very little manual effort for labeling. We collect high quality data in a controlled environment and use emotional video clips to induce affective states. Because we are interested not only in facial expressions but also in affective speech, the speakers are asked to repeat short sentences from longer scenes extracted from feature films. While the video clips provide a context in the spirit of film-based induction methods, the repetition of the emotional sentences serves in itself as an eliciting method, following the idea of Velten [12]. We also introduce a consistency check by asking the speakers to evaluate the emotion of the actor uttering the sentence in the video clip.

The evaluation, and thus the annotation of the recorded data, requires some definition of emotion, which is still an unsolved problem in itself. Cowie et al. [13] divide the emotional phenomenons categories in "episodic" and "pervasive". Episodic emotions are described as states of mind where the feeling dominates the person's awareness, generally for a short period of time. A pervasive emotion, on the other hand, is something that is integrated in people's most mental states, influencing the way they experience the world and act. People's everyday life is largely emotional, yet most of the technological research on emotion focuses on the episodic types, which rarely appear in real interpersonal or human-computer communication. In particular, the majority of affective state recognition and synthesis approaches are limited to the six basic emotion categories, based on the cross-cultural studies on facial expressions of Ekman [14] and only include happiness, sadness, fear, anger, disgust, and surprise. These few discrete categories, however, are not suitable to describe the mainly pervasive emotional states relevant for communication. An alternative are continuous representations where the emotions are mapped to a low-dimensional space, e.g., a 2D space based on activation (strength of emotion) and evaluation (positive vs. negative) [9], [15]. Since this representation is not intuitive and difficult to use for inexperienced users, we employ a list of affective adjectives to be weighted according to their perceived strength. Lists are easy to use and allow multiple labels whereas a single label procedure (as it is commonly used for the basic emotion categories) is insufficient to describe emotions as shown in [9]. While the perceptional evaluation of the affective states needs input from humans, the audio and visual data is processed with very little manual effort. In contrast to the Facial Action Coding System (FACS) [16], used by trained experts for manually labeling videos by means of Action Units (facial muscles activations), we capture the temporal deformation of the face in real-time using a 3D scanning system [17]. The dynamic depth images are then transformed into a consistent 3D mesh representation by fitting a generic template to each scan. Such representation is not only suitable for emotional visual speech as in [18], but also for view-independent facial expression recognition, as demonstrated in [19]. The audio stream is simultaneously captured in a sound absorbent environment such that prosodic features, like sentence melody, speech rhythm, loudness, and specific phonemes, can be accurately and automatically extracted.

Motivated by the need of multi-modal corpora for the development of affect-aware systems [1], we propose an acquisition setup (Section III) for building a corpus of affective speech and corresponding 3D face dynamics. Most of the data is processed automatically (Section IV) thus avoiding expensive manual labeling procedures. In order to obtain accurate data, we record the corpus in a controlled laboratory environment and use induction methods to elicit emotions. The induction method is evaluated in Section V together with some preliminary studies on the acquired data. Compared to existing databases (Section II), we acquire the first corpus containing both 3D detailed dynamic face geometries and affective speech. The corpus will be made available for research purposes.

## II. Related Work

### A. Audio-Visual Datasets

Databases for training and evaluation of affective recognition systems can be divided depending on whether the recorded emotions are naturalistic, artificially induced (e.g., in Wizard-of-Oz scenarios), or posed (either by professional actors or not). A comprehensive overview of the existing audio-visual corpora can be obtained from [13], [20], and [1], we refer only to datasets where affective speech is in focus.

The HUMAINE Network of Excellence has been an important step forward in the field of affective computing, producing a collection of databases [20] containing a large number of audio-visual recordings, only partly labeled, divided into naturalistic and elicited.

Among the naturalistic databases, the ‘Vera am Mittag’ dataset [21] consists of 12 hours of recordings from a German TV talk show, containing spontaneous emotional speech coming from authentic discussions between the guests of the talk show. Most of the data was labeled by a large number of human evaluators using a continuous scale for three emotion primitives: valence, activation, and dominance. The Belfast naturalistic database [22] contains TV recordings and interviews judged relatively emotional, annotated using the FEELTRACE [9] system. The Castaway Reality Television Database [20] consists of audio-visual recordings of 10 people competing in activities like touching snakes or lighting outdoor fires on a remote island; post-activity interviews and diary-type extracts are included. The EmoTV corpus [23] contains interactions extracted from french TV interviews, both outdoor and indoor, with a wide range of body postures; only a very coarse visual annotation is provided.

Among the elicited sets, the Sensitive Artificial Listener (SAL) database [20] contains audio-visual recordings of humans conversating with a computer. The SAL interface is designed to let the user work through a range of emotional states. Four personalities are implemented as different virtual characters (happy, gloomy, angry, and pragmatic) and the user decides to which to talk to. Another corpus containing elicited emotions is the SmartKom database [24], comprising recordings of people interacting with a machine in a Wizard-of-Oz scenario. The subjects were asked to solve specific tasks provoking different affective states. In the Activity Data and Spaghetti Data sets [20], volunteers were recorded while respectively engaging in outdoor activities and feeling inside boxes containing various objects like spaghetti or buzzers going off when touched. The subjects recorded the emotions they felt during the activities. In the first set, both positive and negative affective states with a high level of activation were elicited, while in the Spaghetti set, a range of brief, relatively intense emotions were provoked. The Green Persuasive Dataset [20] was recorded in a scenario where a persuader tried to convince several people to adopting more environmentally-friendly lifestyles. Audio-visual recordings of eight dialogs are available, each featuring a different persuadee. The entire dataset is annotated for persuasiveness, i.e., with continuous-valued labels indicating how convincing was the persuader at any given instant, according to a third observer. The eWiz database [25] contains 322 sentences pronounced by the same speaker with varying prosodic attitudes (declarative, question, exclamation, incredulous question, suspicious irony, and obviousness) suggested by reading a text specifying the context in which the sentence should have been uttered. In [26], the EmoTaboo protocol is introduced, consisting in letting pairs of people play the game Taboo while their faces, upper bodies, and voices are recorded. One of the subjects is a confederate, making sure that enough emotional reactions are observed in the other, naïve, person. The authors of [27] recorded children while playing with Sony AIBO robots. The subjects believed of being actually interacting with the robots, which were wirelessly controlled by a human operator. Videos are not available due to privacy restrictions.

Going towards acted corpora, the GEMEP corpus [28], following the method of Banse and Scherer [29], comprises recordings of the voices, faces, and full bodies of professional stage actors while uttering meaningless sentences. The set of displayed emotions is an extension of the six basic ones, and the actors were guided by reading introductory scenarios for each emotion. Professional Italian actors were hired for the DaFEx database [30], resulting in a number of low resolution videos depicting each actor while uttering the same sentence in seven different emotional states (six basic emotions plus neutral) at three levels of intensity. The actors were asked to read a short introductory text before each session. In [31], a database was collected where 100 students had their facial movements and speech recorded by means of a videocamera and a microphone while pronouncing a set of sentences, each representing one of eleven emotional states, an extension of the six basic emotions.

In contrast to video recordings which are difficult to annotate, marker-based motion capture has been used to obtain 3D information. The Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [32] consists of ten actors in dyadic sessions motion-captured with markers on the face, head, and hands. Scripted and spontaneous spoken communication scenarios represent emotions such as happiness, anger, sadness, frustration and neutral state.

Some works aimed at visual speech modeling for synthesis purposes also acquired corpora containing actors engaged in affective speech, usually employing motion-capture techniques as in [33]–[36]. Because placing markers on someone's face is error prone and might even influence the subject's emotional state like other invasive physiological measurements, marker-based motion capture is only suitable when acquiring posed data from actors.

### *B. 3D Face Datasets*

Relevant to our new corpus are also databases containing only 3D geometries of faces. Yin et al. [37] collected the facial 3D shapes and textures of 100 students while posing the six basic emotions at four intensity levels plus neutral. The same authors published a dynamic 3D facial expression database [38] where 101 subjects were recorded by means of a real-time 3D scanner while changing their facial expression from neutral to one of the six basic emotions and back. The Bosphorus Database for 3D face analysis [39] includes 3D scans of 81 people showing the six basic emotions plus a selected subset of Action Units [40], variations in head pose and different kinds of occlusion are also included in the set. The University of York 3D Face Database [41] contains 3D facial scans of different subjects showing different expressions and head poses. The IV2 Multimodal Biometric Database [42] also includes 3D expressive static face geometries of several people. Other popular databases contain only neutral scans, as the Gava Database [43], the FRAV3D Face Database [44], the BJUT-3D Large-Scale Chinese Face Database [45]. The Basel Face Model [46] is a publicly available statistical face model built from the 3d scans of 200 people in neutral pose, but the original training data is not available. To our knowledge, there is no currently available dataset containing both the audio and dense dynamic 3D facial representations of affective speech.

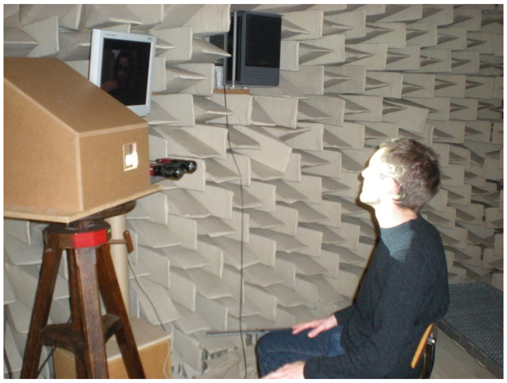

Fig. 1. Recording setup: one speaker sits in front of the 3D scanner in the anechoic room while watching one of the eliciting videos clips.

## III. Acquisition Setup

In order to simultaneously record affective speech and detailed 3D face geometries, we have employed a real-time 3D scanner and a professional microphone. To benefit from automatic speech segmentation, the noise level should be kept as low as possible; we therefore acquired the data in an anechoic room, with walls covered by soundwave-absorbing materials. The authors could operate the system and communicate with the speakers from a separate room. Fig. 1 shows the setup, with a volunteer being scanned while watching one of the eliciting videos on the screen.

### *A. Recording Setup*

For acquiring detailed dynamic face deformation data, we employed the 3D scanner described in [17]. The system combines stereo and active illumination based on phase-shift for robust and accurate 3D scene reconstruction. Stereo overcomes the traditional phase discontinuity problem and motion compensation is applied in order to remove artifacts in the reconstruction. The system consists of two high-speed monochrome cameras, a color camera, and a DLP projector without the 4-segment color wheel (RGBW), so that it projects three independent monochrome images at 120 Hz (sent as the R, G and B channel). The two monochrome cameras are synchronized and record the three images. The texture camera is also synchronized, but uses a longer exposure to integrate over all three projected images. Thanks to the above scanner, accurate depth maps of the subjects' faces could reliably be captured at 25 fps.

For the audio recordings, we used a studio condenser microphone placed in front of the speaker. The microphone was connected to a computer controlled by the authors in a separate room, running standard open source audio acquisition software. To reduce the level of direct noise coming from the projector's cooling fans, we enclosed the projector in a wooden box, thus making the background noise level on the recorded audio stream low enough for automatic speech segmentation.

### B. *Acquisition Protocol*

Our corpus is comprised of 40 short English sentences extracted from feature films. Having the goal of building a database of affective speech for the English language, we contacted native speakers who volunteered to have their voice and facial movements recorded. We gathered 14 subjects, 8 females and 6 males, aged between 21 and 53 (average 33.5). Each person was required to sit alone in the anechoic room and asked to pronounce each sentence twice, first with a neutral tone and then with the emotional one. For synchronization purposes, the volunteers clapped their hands in front of the cameras before uttering the sentences.

Neutral speech was achieved by asking the speakers to read the sentence from text displayed on a screen placed above the scanner's cameras. At a second stage, the emotional film scene corresponding to the previously read text was shown to the volunteers, followed by a repetition of the part of the video containing the specific sentence. The speakers had the chance to see the clips again and were asked to rate the sentence that they had just heard by means of the emotions they thought it conveyed. After the evaluation was completed, the short clip containing the sentence was shown again and the speakers were asked to repeat it according to their feelings. A detailed description of the evaluation process is presented in Section V.

## IV. Data Processing

### A. *Video Processing*

The real-time 3D scanner is employed to capture detailed 3D geometry and texture of the performances of each speaker, as shown by the first two images in figure 2. Facial expression analysis, however, requires full spatial and temporal correspondences of the 3D data. To achieve this goal, we use the two-step procedure introduced in [47]: First, a generic template mesh is warped to the reconstructed 3D model of the neutral facial expression of a speaker. Second, the resulting personalized template is automatically tracked throughout all facial expression sequences.

*1) Personalized Face Template:* In order to build a person-specific face template, each speaker is asked to turn the head with a neutral expression and as rigidly as possible in front of the real-time 3D scanner. The sequence of 3D scans is registered and integrated into one 3D model using the online modeling algorithm proposed in [48]. Small deformations arising during head motion violate the rigidity assumption, but in practice do not pose problems for the rigid reconstruction. Instead of using the 3D model directly as a personalized face template, a generic face template is warped to fit the reconstructed model. Besides enabling a hole-free reconstruction and a consistent parameterization, using the same generic template has the additional benefit of providing full spatial correspondence between different speakers.

Warping the generic template to the reconstructed 3D model is achieved by means of non-rigid registration, where for each mesh vertex $\mathbf{v}_i$ of the generic template a deformation vector $\mathbf{d}_i$ is determined in addition to a global rigid alignment. This is formulated as an optimization problem, consisting of a smoothness term minimizing bending of the underlying deformation [49], and a set of data constraints minimizing the distance between the warped template and the reconstructed model. As the vertex correspondences between generic template and reconstructed model are unknown, closest point correspondences are used as approximation similarly to standard rigid ICP registration. A set of manually labeled correspondences are used for the initial global alignment and to initialize the warping procedure. The landmarks are mostly concentrated around the eyes and mouth, but a few correspondences are selected on the chin and forehead to match the global shape. The manual labeling needs to be done only once per speaker and takes at most a couple of minutes. The resulting personalized template accurately captures the facial 3D geometry of the corresponding person.

The diffuse texture map of the personalized template is automatically extracted from the rigid registration scans by averaging the input textures. The face is primarily illuminated by the 3D scanner, and we can therefore compensate for lighting variations using the calibrated position of the projection. Surface parts that are likely to be specular are removed based on the half angle. The reconstructed texture map is typically oversmoothed, but sufficient for the tracking stage.

*2) Facial Expression Tracking:* The personalized face template is used to track the facial deformations of each performance. For this purpose, non-rigid registration is employed, in a similar manner as during the template creation. In this case, the distances between the template vertices and the 3D scans are minimized. To ensure temporal continuity, optical flow constraints are also included in the optimization. The motion of each vertex from frame to frame should coincide with the optical flow constraints. During speaking, the mouth region deforms particularly quickly, and non-rigid registration may drift and ultimately fail. This can be compensated for by employing additional face-specific constraints such as explicitly tracking the chin and mouth regions, making the whole process more accurate and robust to fast deformations. Figure 2 shows a personalized model adapted to a specific frame of a sequence.

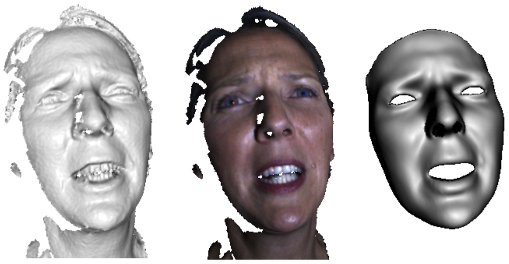

Fig. 2. From left to right, the image shows the 3D reconstruction of a person's face, the corresponding texture mapped on it, and the personalized face templated deformed to fit the specific frame.

### *B. Audio Processing*

Different affective states are manifested in speech by changes in the prosody, see [50] for an overview. As certain prosodic differences are small but still audible, a careful setup of an audio-visual corpus requires accurate extraction of prosodic parameters from the audio signal.

Speech prosody can be described at the perceptual level in terms of pitch, sentence melody, speech rhythm, and loudness. The physically measurable quantities of a speech signal are the following acoustic parameters: fundamental frequency ($F_0$), segment duration, and signal intensity. $F_0$ correlates with pitch and sentence melody, segment duration correlates with speech rhythm, and signal intensity correlates with loudness.

The annotation process necessary for obtaining the physical prosodic parameters of the utterances includes a number of steps: First, the sentence's text is transcribed into the phonological representation of the utterance. Then accurate phone segmentation, fundamental frequency ($F_0$) extraction, and signal intensity estimation are achieved by analyzing the speech data. For the extraction of these physically measurable prosodic quantities, we applied fully automatic procedures provided by SYNVO Ltd. In the following, we give an overview of the extraction procedures for fundamental frequency, signal intensity, and segment duration.

*1) Transcription:* The phonological representation contains the sequence of phones for the sentences in the audio-visual corpus, the stress level of syllables, the position and strength of phrase boundaries, plus the indicators of phrase types. Initial phonological representations of the sentences are obtained applying the transcription component of the SYNVO text-to-speech synthesis system to the text version of the corpora. See [51] for a description of such transcription component. These initial phonological representations contain the standard phonetic transcription, also called canonical phonetic transcription,

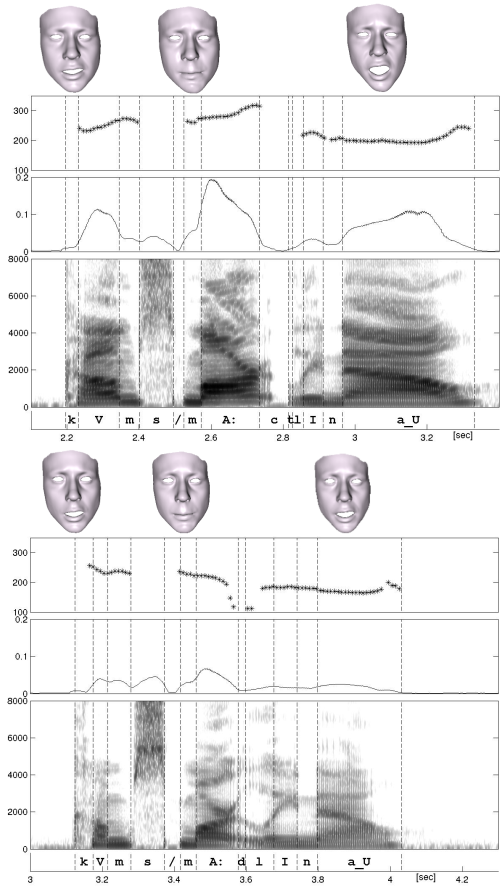

Fig. 3. Comparison of emotional (top) and neutral (bottom) tone of the same phrase, as pronounced by one speaker. For each utterance, spectrogram, signal intensity contour, and fundamental frequency contour of the speech signal, as well as sample faces are shown from bottom to top. The emotional utterance clearly shows higher overall signal intensity and at the second syllable, a high rising fundamental frequency contour in contrast to the low falling one of the neutral utterance. Also, it is clearly visible that syllable nucleus durations are longer in the emotional state, while consonant durations are similar.

of the sentences.

The phonological information (phrase type, phrase boundary, and sentence accentuation) of these automatically generated representations is then adapted to the speech signals. Neural network-based algorithms are employed for automatic phrase type, phrase boundary, and syllable accent identification. Detailed information on this procedure can be found in [52].

*2) Fundamental Frequency Extraction:* $F_0$ values of the natural speech data of the prosody corpus are computed every 10 ms using a pitch detection algorithm based on combined information taken from the cepstrogram, the spectrogram, and the autocorrelation function of the speech signal, cf. [52]. Signal sections judged as unvoiced by the algorithm are assigned no $F_0$ values. Figure 3 shows examples of such fundamental frequency contours.

*3) Signal Intensity Extraction:* Signal intensity values of the natural speech data are computed every 1 ms. The root mean square value of the signal amplitude calculated over a window of 30 ms length is used. Signal intensity contours of two utterances are displayed in figure 3.

*4) Segment Duration Extraction:* An accurate extraction of phone and speech pause durations requires an exact segmentation of the natural speech data of the audio-visual corpus into adjacent, non-overlapping speech or pause segments, and a correct assignment of labels to these segments indicating the segment type. This assignment is commonly termed “labeling”.

Since the phonological representation contains the standard phonetic transcription of an utterance (see Section IV-B1), it is convenient to use this standard transcription for automatic segmentation and labeling. However, a close phonetic transcription, also referred to as matched phonetic transcription, indicating

| | |
|---|---|
| vowels | i: I U u:<br>e<br>@<br>q 3 3: V O:<br>A A: Q |
| diphthongs | @_U a_I a_U e_I E_@ I_@ O_I o_U U_@ |
| consonants | p p_h b t t_h d k k_h g<br>m n N<br>r<br>f v T D s z S Z x h<br>j w<br>l |
| affricates | t_S d_Z |
| pauses | c / |

TABLE I

Segment types of English phones and speech pauses used for transcription of the speech data of the audio-visual corpus.

pronunciation variants made by the speaker, results in a much better segmentation and labeling.

*5) Segment Types:* Segment types correspond to the phone types determined in the transcription. Plosives are additionally segmented into their hold and burst parts, which are labeled separately. While the burst part of a plosive is denoted by the same symbol used for the plosive phone type, a "c" denotes the hold part, also called closure or preplosive pause. Speech pauses corresponding to phrase boundaries are labeled with the symbol "/". For a plosive following a speech pause, no preplosive pause is segmented. Table I lists all segment types used for the transcription of natural speech data.

*6) Automatic Segmentation Procedure:* Manual transcription and segmentation of the speech prosody corpus would take too much time. We apply a segmentation procedure first presented in [53], which simultaneously delivers a highly accurate phonetic segmentation and a close phonetic transcription.

This segmentation procedure relies on iterative Viterbi search for best-matching pronunciation variants and on iterative retraining of phone hidden Markov models (HMMs). In contrast to existing state-of-the-art segmentation systems like [54] and [55], this procedure does not require elaborate features, but only "standard" mel-frequency cepstral coefficients (MFCCs) and voicing information.

The segmentation procedure consists of two stages: First, context-independent three-state left-to-right phone HMMs with 8 Gaussian mixtures per state are trained on the natural speech data of the audio corpus using the standard phonetic transcription of the utterances by applying a so-called "flat start" initialization, cf. [56].

For the second stage, a small set of language- and speaker-dependent pronunciation variation rules is applied to the canonical transcriptions and a recognition network is generated for each utterance. Such a network includes all pronunciations allowed by the rules.

A Viterbi search then determines the most likely path through the networks and thus delivers an adapted phonetic transcription of each utterance. These new transcriptions are used to retrain the HMMs that are in turn used in the next iteration for the Viterbi search. The procedure stops when the number of insertions, deletions, and replacements of phones between the current and the previously adapted transcriptions falls below some predefined threshold. Details on this segmentation procedure can be found in [57].

Since the length of the analysis window restricts the accuracy of boundary detection of certain segments, e.g., preplosive pauses, a post-processing step was added to the second stage, correcting segment boundary placement of specific segment classes based on the speech signal amplitude and voicing information.

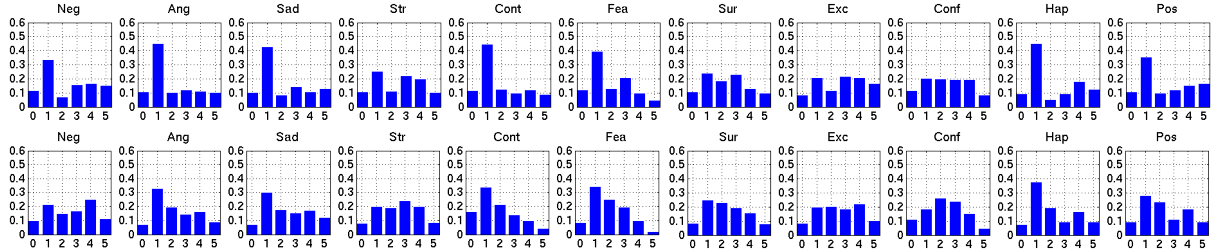

Fig. 4. Histograms showing the evaluation of all movie clips as expressed by the speakers (top row) and by the users of the online survey (bottom row). For each emotional label, the bars indicate how many times (in percentage) that particular label was given the corresponding grade shown on the x axis. The ratings of these two groups are very consistent.

## V. Corpus

In total, we recorded 14 native English speakers uttering 1109 sequences, 4.67 seconds long on average. Some of the extracted audio-visual features are shown in Figure 3. For labeling the affective states, we have carried out three evaluations. First, we have evaluated the eliciting video clips by asking the speakers, followed by an online survey to check the consistency. The third online evaluation rates the affective content of the actually recorded and processed data.

For the first evaluation, the speakers were asked to rate each video containing the sentence just before pronouncing it in the recording room. The volunteers filled out a paper form, giving grades between 0 and 5 to a set of 11 suggested emotional adjectives (“Negative”, “Anger”, “Sadness”, “Stress”, “Contempt”, “Fear”, “Surprise”, “Excitement”, “Confidence”, “Happiness”, and “Positive”), where 0 means “I don’t know”, 1 corresponds to “Not at all”, and 5 to “Very”. An additional field was provided in order to allow the suggestion of labels which might have been considered appropriate for the clip. On a later stage, the eliciting movie clips were presented to a second, larger group of human observers, by means of an online survey, presenting the same structure as the form presented to the speakers. The clips where shown in a randomized order, allowing the user to quit the evaluation at any time. In total, 122 people took part in the survey (64.75% males and 35.25% females; 20.5% native English speakers), grading over 1000 video clips.

Figure 4 shows the results of both evaluations: the top row corresponds to the volunteers’ judgment, the bottom row to the online survey. The histograms show how many times (in percentages of all movie clips) each label was given the corresponding grade on the x-axis. When comparing the results, one sees that the distributions of the grades are very similar between the two groups although the volunteers tended

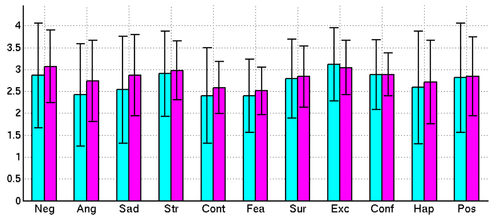

Fig. 5. For each emotional label, the mean and standard deviation of the grades are given respectively by the volunteers (cyan) and by the anonymous users of the online survey (magenta). In general, the average grade is lower for the speakers, who evaluated the eliciting videos as part of the acquisition setup in the anechoic room. Moreover, the standard deviations for the speakers are larger since they tend to give the grades 1 and 5 more often as shown in Figure 4.

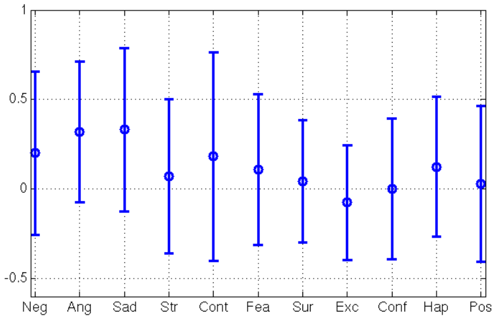

Fig. 6. Mean and standard deviation of the difference between the grades for each sentence given to each emotional label by the speakers and by the people who took the online survey. The plot supports the observation of Figure 5, where the grades for the online survey are slightly higher in average. Note that the standard deviation for all affective labels is less than 1, which is the distance between two grades.

to give the grade “Not at all” more often than people from the online survey. Figure 5 shows mean and standard deviation of the grades given to each emotional label over all sentences. The cyan bars on the left represent the grades given by the speakers in the recording room, while the magenta bars on the right represent the corresponding evaluations given by the online survey. Apart from a decrease in the standard deviations, we notice that the average generally slightly increases in the results yielded by the online survey. Figure 6 shows the mean and standard deviation computed over the differences in the grades for each sentence given to all videos in the two groups of observers for each suggested affective adjective. The plot supports the observation of Figure 5, where the grades for the online survey are slightly higher in average. Note that the standard deviation for all affective labels except contempt is less than 1, which

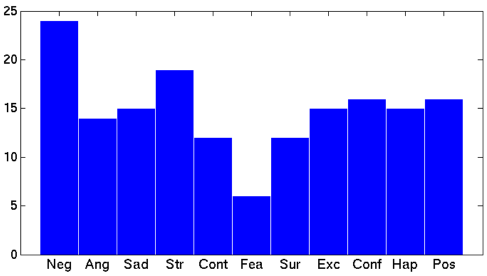

Fig. 7. For each emotion, the number of sentences evaluated as such is shown. There is a clear predominance of negative emotions, while fear is the least perceived affective state.

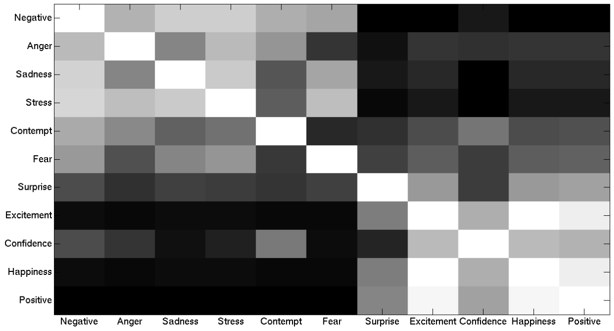

Fig. 8. Correlations between the affective adjectives. There is a high correlation (bright fields) among positive and negative emotions, in particular "Excitement" was very often mentioned together with "Happiness". Note that there is also a significant correlation between some of the basic emotions, like "Sadness" and "Fear" or "Surprise" and "Happiness".

is the distance between two grades. The tendency that the speakers give the grades 1 and 5 more often than the group of the online survey might be explained by the difference in the environment. While the speakers evaluated the video clips in the recording room without any distraction, the online survey could have been evaluated in any environment such that the affective states seem to be perceived as less intense. Nevertheless, Figure 6 shows that the evaluation obtained by the two groups is very consistent, which indicates that the laboratory environment had only a minor impact on the perception of affective states.

Figure 7 shows for each label the number of sentences which were given an average grade greater than 3 during the online survey, thus giving an idea of the emotional content of the eliciting videos. There is a clear predominance of negative emotions, while fear is the least perceived affective state. Some of the labels are naturally depending on each others, as can be seen in Figure 8, where the brighter upper and lower corners indicate a high correlation among positive and negative states. We notice that there

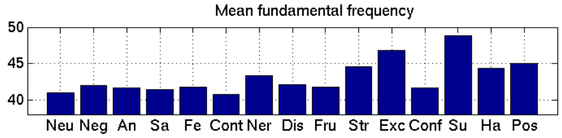

Fig. 9. Fundamental frequency averaged over each proposed affective adjectives on the x-axis, i.e., over the corpus sentences which were given a mean grade $> 3$ for that particular label.

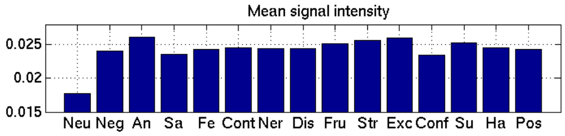

Fig. 10. Average audio signal intensity depending on the emotional labels on the x-axis.

is also a significant correlation between some of the basic emotions ("Sadness", "Fear", "Surprise", and "Happiness"). This indicates that a single label procedure based on the basic emotions is insufficient to describe emotions.

In order to assess the quality of the acquired data, videos were created containing renderings of the tracked 3D faces and the original audio signals. A third survey was thus designed, where the number of suggested emotional labels was enriched by the three additional adjectives most commonly proposed during the first two evaluation steps ("Nervousness", "Disappointment", "Frustration"), and by the additional label "Emotional". Note that each sentence was recorded twice for each person, one after reading from text, and one after showing the corresponding induction video.

Being the number of sequences to be judged very large, it has not yet been possible to achieve a sensible number of evaluations for each of them[1]. For a preliminary study, only sentences which were rated at least 3 times have been considered. We proceeded by selecting as neutral the utterances with an average grade less than 3 for the label "Emotional", and the ones with "Emotional" mean grade greater than 3 as the remaining affective states. The plots in Figures 9 and 10 show the relations of the affective adjectives and simple audio features, averaged over all sequences labeled according to the above rule. In particular, Figure 9 refers to the fundamental frequency $F0$ and Figure 10 to the intensity of the audio signal. Already such simple features show a pattern, especially for the intensity features: emotional

[1] The online survey is available at
http://www.vision.ee.ethz.ch/~gfanelli/emo_survey/survey.cgi

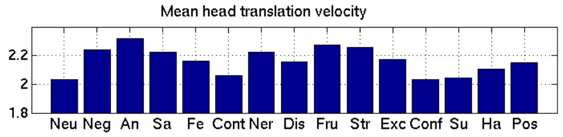

Fig. 11. Mean head translation velocity computed from sequences labeled as each of the adjectives on the x-axis.

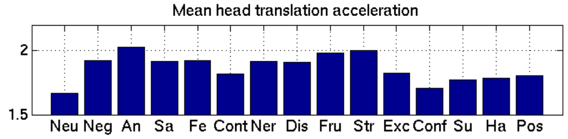

Fig. 12. Average head translation acceleration for the sequences labeled as the adjectives on the x-axis.

sequences present on average higher values. The fundamental frequency $F0$ shows a higher variation for positive emotions than for negative emotions. Simple features extracted from the 3D face data are shown in Figures 11 and 12, respectively describing the mean over the first and second derivatives computed over the magnitude of the rigid translation movements of the heads, i.e., mean velocity and acceleration. Again, sequences judged as non-emotional show on average lower values. These plots indicate that a single feature is not enough to recognize the affective state but that already several low-level audio-visual cues can give some evidence for the affective state of the user.

Using the accurate phoneme segmentation technique described in Section IV-B, we can easily arrange the 3D face scans into groups corresponding to a particular phoneme. That is, groups where all observed facial shapes corresponding to a particular phoneme are represented. Figure 13 shows the result of applying Principal Component Analysis (PCA) to the set of scans corresponding to the phoneme “I”, as uttered by the same subject. It is possible to apply PCA in this case only thanks to the non rigid 3D face tracking technique discussed in Section IV-A, which gives us spatial and temporal correspondences among all facial scans. As training set, we picked one facial scan for each phoneme, right at the middle of the interval of occurrence returned by the automatic segmentation procedure for that phoneme; in this way we try to limit the variations due to coarticulation, i.e., the dependency of the current mouth shape from the preceding and following phonemes. The three rows in Figure 13 show the three main modes of variation observed in the data, with the average in the middle and the faces generated by setting the corresponding weights to $-3$ std. on the left and $+3$ std. on the right, respectively. It appears that much of the variation in the facial configuration depends on the emotion with which the phoneme was uttered. While this plot only gives a simple example, it suggests that the acquired data could be used to build

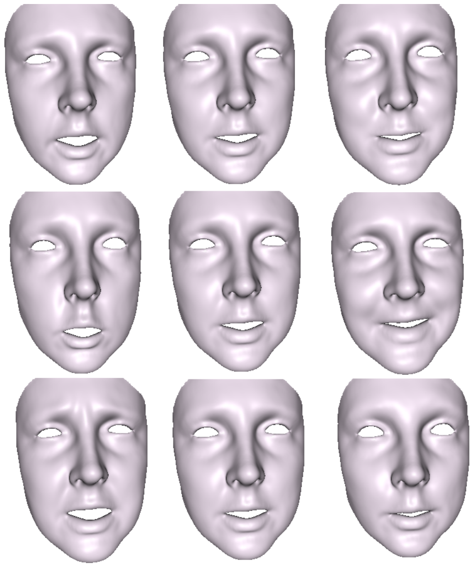

Fig. 13. First three modes of the PCA model of the phoneme 'I'. The middle column shows the average face, while the left and right columns represent the result of setting the mode's weight to $-3$ std. and $+3$ std., respectively.

advanced models for audio-visual animation or recognition purposes.

## VI. Discussion

In this work, we have presented a system for the acquisition of an audio-visual corpus comprised of affective speech and corresponding dense dynamic 3D face geometries. The system was designed to ease the data acquisition process by reducing the required manual labeling effort. Current time-consuming steps of our setup are the recording of the raw data, which takes about 1.5 hours for 80 short sentences spoken by one person, and the evaluation of the affective states in the processed data. While the recording process cannot be speeded up, the evaluation can be widely spread using a web-based survey.

The corpus currently comprises 1109 sentences uttered by 14 native English speakers; it includes the audio synchronized with the faces' depth data. For the speech signal, a phonological representation of the utterances, accurate phone segmentation, fundamental frequency extraction, and signal intensity are provided. The depth signal is converted into a sequence of 3D meshes, which provide full spatial and temporal correspondences across all sequences and speakers. This is an important requirement for generating advanced statistical models which can be used for animation or recognition applications. One of the future work is to increase the size of the corpus and to extend the setup in order to extract additional audio-visual features like phrase type, phrase boundary, and sentence accentuation. We also emphasize that our system is not limited to English and could be directly applied to other languages like German, French, Italian, Spanish, or Portuguese.

The described affective corpus can be used to model coarticulation for the synthesis of emotional visual speech. Other applications include view-independent facial expression recognition, emotion-independent lip reading, or audio-visual emotion recognition. Our experiments indicate that the corpus might also be useful for studying the correlations between audio and facial features in the context of emotional speech. To this end, the list of affective adjectives might be mapped to a low-dimensional space or the online survey might be adapted to obtain additional emotional information. Because we intend to provide both the raw and the processed data, the corpus could also be used as a benchmark dataset. For example, a comparison of the results of different 3D face trackers could be achieved by measuring the loss of information between the original and the processed data in terms of the perceived emotional content.

## REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[2] C. M. Whissell, *The dictionary of affect in language*. Reading, MA: R. Plutchik and H. Kellerman, 1972.

[3] J. Cohn, “Automated analysis of the configuration and timing of facial expression,” in *What the face reveals (2nd edition): Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, P. Ekman and E. Rosenberg, Eds., 2005, pp. 388 – 392.

[4] M. Valstar, H. Gunes, and M. Pantic, “How to distinguish posed from spontaneous smiles using geometric features,” in *Proceedings of ACM Int’l Conf. Multimodal Interfaces (ICMI’07)*. New York, NY, USA: ACM, November 2007, pp. 38–45.

[5] J. Gross and R. Levenson, “Emotion elicitation using films,” *Cognition and Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[6] M. Bradley, B. Cuthbert, and P. Lang, “Picture media and emotion: Effects of a sustained affective context,” *Psychophysiology*, vol. 33, no. 6, pp. 662–670, 1996.

[7] D. Clark, “On the induction of depressed mood in the laboratory: Evaluation and comparison of the velten and musical procedures,” *Advances in Behaviour Research and Therapy*, vol. 5, no. 1, pp. 27–49, 1983.

[8] K. Scherer, T. Johnstone, and T. Bänziger, “Automatic verification of emotionally stressed speakers: The problem of individual differences,” 1998.

[9] R. Cowie and R. R. Cornelius, “Describing the emotional states that are expressed in speech,” *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[10] R. W. Picard, “Toward computers that recognize and respond to user emotion,” *IBM Systems Journal*, vol. 39, no. 3-4, pp. 705–719, 2000.

[11] R. Davidson, D. Jackson, and N. Kalin, “Emotion, plasticity, context and regulation: Perspectives from affective neuroscience,” *Psychological Bulletin*, vol. 126, pp. 890–906, 2000.

[12] E. Velten, “A laboratory task for induction of mood states,” *Behaviour research and therapy*, vol. 6, pp. 473–482, 1968.

[13] R. Cowie, E. Douglas-Cowie, and C. Cox, “Beyond emotion archetypes: Databases for emotion modelling using neural networks,” *Neural Networks*, vol. 18, no. 4, pp. 371–388, 2005.

[14] P. Ekman, “Constants across cultures in the face and emotion,” *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[15] R. Craggs and M. M. Wood, “A two dimensional annotation scheme for emotion in dialogue,” in *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.

[16] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[17] T. Weise, B. Leibe, and L. V. Gool, “Fast 3d scanning with automatic motion compensation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[18] D. Vlasic, M. Brand, H. Pfister, and J. Popović, “Face transfer with multilinear models,” *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 426–433, 2005.

[19] J. Wang, L. Yin, X. Wei, and Y. Sun, “3d facial expression recognition based on primitive surface feature distribution,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1399–1406, 2006.

[20] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, “The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data,” in *Proc. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal.*, 2007, pp. 488–500.

[21] M. Grimm, K. Kroschel, and S. Narayanan, “The vera am mittag german audio-visual emotional speech database,” in *Multimedia and Expo, 2008 IEEE International Conference on*, 23 2008-April 26 2008, pp. 865–868.

[22] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of databases,” *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[23] J. C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, “Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of tv interviews,” *Personal Ubiquitous Comput.*, vol. 13, no. 1, pp. 69–76, 2009.

[24] U. Türk, “The technical processing in smartkom data collection: a case study,” LMU Munich, Tech. Rep., July 2001.

[25] Y. Morlec, G. Bailly, and V. Aubergé, “Generating prosodic attitudes in french: Data, model and evaluation,” *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[26] A. Zara, V. Maffiolo, J.-C. Martin, and L. Devillers, “Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics,” in *ACII*, 2007, pp. 464–475.

[27] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D’Arcy, M. Russell, and M. Wong, “”You stupid tin box” - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus.” in *Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004*, ELRA, Ed., 2004. [Online]. Available: http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2004/Batliner04-YST.pdf

[28] T. Bänziger and K. R. Scherer, “Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus,” in *ACII*, 2007, pp. 476–487.

[29] R. Banse and K. R. Scherer, “Acoustic profiles in vocal emotion expression,” *Journal of Personality and Social Psychology*, vol. 70, pp. 614–636, 1996.

[30] A. Battocchi, F. Pianesi, and D. Goren-Bar, “Dafex: Database of facial expressions.” in *INTETAIN*, ser. Lecture Notes in Computer Science, M. T. Maybury, O. Stock, and W. Wahlster, Eds., vol. 3814. Springer, 2005, pp. 303–306. [Online]. Available: http://dblp.uni-trier.de/db/conf/intetain/intetain2005.html#BattocchiPG05

[31] L. S.-H. Chen, “Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction,” Ph.D. dissertation, Champaign, IL, USA, 2000, adviser-Huang, Thomas S.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, “Iemocap: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*.

[33] Y. Cao, W. C. Tien, P. Faloutsos, and F. H. Pighin, “Expressive speech-driven facial animation,” *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1283–1302, 2005.

[34] Z. Deng, U. Neumann, J. P. Lewis, T.-Y. Kim, M. Bulut, and S. Narayanan, “Expressive facial animation synthesis by learning speech coarticulation and expression spaces,” *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 6, pp. 1523–1534, 2006.

[35] K. Wampler, D. Sasaki, L. Zhang, and Z. Popovic, “Dynamic, expressive speech animation from a single mesh,” in *Symposium on Computer Animation*, 2007, pp. 53–62.

[36] N. Mana, P. Cosi, G. Tisato, F. Cavicchio, E. Magno, and F.Pianesi, “An italian database of emotional speech and facial expressions,” in *Proceedings of Workshop on Emotion: Corpora for Research on Emotion and Affect in association with 5th International Conference on Language, Resources and Evaluation (LREC2006), Genoa, Italy*, May 2006.

[37] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3d facial expression database for facial behavior research,” in *FGR ’06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 211–216. [Online]. Available: http://dx.doi.org/10.1109/FGR.2006.6

[38] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, “A high-resolution 3d dynamic facial expression database,” in *FG*, 2008, pp. 1–6.

[39] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, “Bosphorus

database for 3d face analysis," in *BIOID*, 2008, pp. 47–56.

[40] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto: Consulting Psychologists Press, 1978.

[41] "University of york 3d face database," http://www-users.cs.york.ac.uk/ nep/research/3Dface/tomh/3DFaceDatabase.html.

[42] D. Petrovska-Delacrtaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, E. Krichen, M. Anouar-Mellakh, A. Chaari, S. Guerfi, M. Ardabilian, J. D Hose, and B. Ben Amor, "The IV2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic and Talking Face Data) and the IV2-2007 Evaluation Campaign," in *IEEE Second International Conference on Biometrics: Theory, Applications and Systems (BTAS 08)*, IEEE, Ed., Sep. 2008, pp. 1–7. [Online]. Available: http://liris.cnrs.fr/publis/?id=4186

[43] A. B. Moreno and A. Sánchez, "Gavabdb: a 3d face database," in *Workshop on Biometrics on the Internet*, Vigo, March 2004, pp. 77–85.

[44] C. Conde, A. Serrano, L. J. Rodriguez-Aragon, and E. Cabello, "An automatic 2d, 2.5d & 3d score-based fusion face verification system," in *ASAP '07: Proceedings of the 2007 IEEE International Conference on Application-Specific Systems, Architectures and Processors*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 214–219.

[45] "The bjut-3d large-scale chinese face database," Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory Beijing University of Technology, Tech. Rep., August 2005.

[46] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*. Genova, Italy: IEEE, 2009.

[47] T. Weise, H. Li, L. V. Gool, , and M. Pauly, "Face/off: Live facial puppetry," in *Symposium on Computer Animation*, 2009.

[48] T. Weise, T. Wismer, B. Leibe, , and L. V. Gool, "In-hand scanning with online loop closure," in *IEEE International Workshop on 3-D Digital Imaging and Modeling*, October 2009.

[49] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 213–230, 2008.

[50] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," in *Affective Information Processing*, J. Tao and T. Tan, Eds. Springer, 2008.

[51] H. Romsdorfer and B. Pfister, "Text analysis and language identification for polyglot text-to-speech synthesis," *Speech Communication*, vol. 49, no. 9, pp. 697–724, September 2007.

[52] H. Romsdorfer, "Polyglot text-to-speech synthesis. Text analysis and prosody control," Ph.D. dissertation, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January 2009.

[53] ——, "An approach to an improved segmentation of speech signals for the training of statistical prosody models," Tech. Report Nr. 2, CTI-Project Nr. 6233.1 SUS-ET. Institut TIK, ETH Zürich, Tech. Rep., May 2004.

[54] J. P. H. van Santen and R. Sproat, "High-accuracy automatic segmentation," in *Proceedings of Eurospeech'99*,

Budapest, Hungary, 1999, pp. 2809–2812.

[55] J.-P. Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information," Ph.D. dissertation, Oregon Graduate Institute of Science and Technology, May 2000.

[56] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Cambridge University Engineering Departement, 2002.

[57] H. Romsdorfer and B. Pfister, "Phonetic labeling and segmentation of mixed-lingual prosody databases," in *Proceedings of Interspeech 2005*, Lisbon, Portugal, 2005, pp. 3281–3284.