Realistic Transformation of Facial and Vocal Smiles in Real-Time Audiovisual Streams

Pablo Arias, Catherine Soladie, Oussema Bouafif, Axel Roebel, Renaud Seguier, Jean-Julien Aucouturier

Realistic transformation of facial and vocal smiles in real-time audiovisual streams

Pablo Arias, Catherine Soladié, Oussema Bouafif, Axel Roebel, Renaud Séguier, and Jean-Julien Aucouturier

Abstract—Research in affective computing and cognitive science has shown the importance of emotional facial and vocal expressions during human-computer and human-human interactions. But, while models exist to control the display and interactive dynamics of emotional expressions, such as smiles, in embodied agents, these techniques can not be applied to video interactions between humans. In this work, we propose an audiovisual smile transformation algorithm able to manipulate an incoming video stream in real-time to parametrically control the amount of smile seen on the user's face and heard in their voice, while preserving other characteristics such as the user's identity or the timing and content of the interaction. The transformation is composed of separate audio and visual pipelines, both based on a warping technique informed by real-time detection of audio and visual landmarks. Taken together, these two parts constitute a unique audiovisual algorithm which, in addition to providing simultaneous real-time transformations of a real person's face and voice, allows to investigate the integration of both modalities of smiles in real-world social interactions.

Index Terms—Smiling, Facial expressions, Vocal emotions, Audiovisual, Real-time, Video and audio signal processing

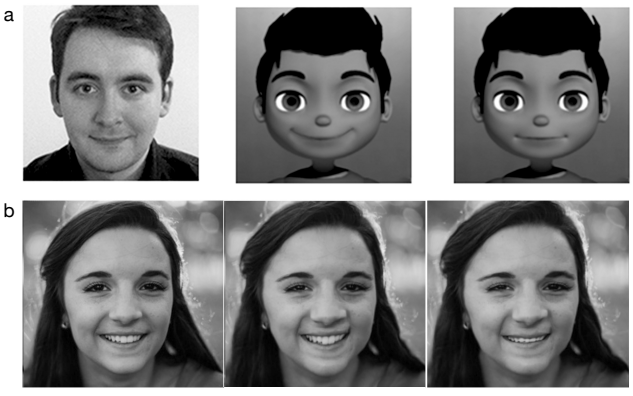
1 Introduction

Smiles are a fundamental element of the human expressive repertoire [1]. We recognize them regardless of age, gender or culture [2]–[4]; they influence what we think of a person (e.g., making them more attractive [5]), how we behave towards them (e.g., with greater empathy [6]), and even provoke unconscious reactions [7].

It is therefore not surprising that smiling constitutes a much-researched part of the behavioral repertoire of embodied agents designed for human-computer interactions [8], [9]. Avatars with smiling faces are judged more attractive and positive [10] and, like smiling humans, trigger unconscious physiological reactions in human observers [11], [12]. More than a feature that can be turned on and off, avatar smiles can be synthesized gradually [13] and with temporal dynamics [14], allowing to experiment with how and when an avatar should smile to improve the quality of a virtual interaction. Avatar smiles were found to have a positive impact on the ongoing interaction [15] and on its later outcomes, including better learning [16]–[18] and problem solving [11].

However, because existing techniques mostly allow to synthesize and manipulate the expression of embodied agents, but not to transform the audiovisual expression of real users, e.g. in a live stream, their scope is mostly limited to human-computer interactions. In the visual domain, techniques allowing to control the morphological parameters of a synthetic face (e.g. cheek-raising, mouth opening, symmetry, lip press [9], [19]) can work in real-time, but can only apply to human-human interactions if they are mediated by a virtual avatar, be it photorealistic [10] (Figure 1-a). Conversely, recent deep-learning techniques able to learn expressive transformations from a corpus of paired images [20] allow realistic facial transformations of arbitrary users, but have yet to operate in real-time (Figure 1-b). Similarly, in the audio domain, hidden-markov models [21]–[23] and formant resynthesis [24] techniques can reproduce realistic characteristics of speech pronounced while smiling or laughing, but only in non-real time applications. In the same way, in the audiovisual domain, techniques using deep neural networks are also proposed to create expressive audiovisual speech synthesis, but are not build with real-time constraints [25]. Given how sensitive humans are to small deviations of interactive synchrony [19] and face realism [15], much more could be achieved if one could realistically control audiovisual smiles in real-time streams between real users, rather than in virtual interactions.

Fig. 1. State-of-art in smile synthesis in the visual domain. (a) Common face synthesis systems allow to generate and control smiling expressions in real-time, but only via the mediation of avatars (adapted from [10]; left: original, middle: avatar with enhanced smile, right: suppressed smile). (b) Deep-learning techniques can learn photorealistic transformations of facial expressions on arbitrary photographs, but these techniques do not typically work in real-time (adapted from [20]; left: original picture, middle: manipulated picture with enhanced smile, right: suppressed smile). Our proposal aims to generate real-time transformations of a user's input, as in A, based on their normal, non-synthetic video stream, as in B.



To this aim, we propose an audiovisual smile trans-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

formation technique able to manipulate an incoming audiovisual stream in real-time to parametrically control the amount of "smiliness" seen on the face and heard on speech, while preserving other characteristics such as the user's identity or the interaction's timing and content. The transformation is composed of separated audio and visual pipelines, both based on a warping technique informed by the real-time detection of visual and audio landmarks.

The visual part of the algorithm tracks morphological features of the face, such as the eyes and lip corners, stretches its position using a predefined parametric model, and resynthesizes pixel grey-levels to map the modified shape of the face. This algorithm significantly extends what constitute, to our knowledge, the only other example to date of real-time smile transformation [26], by making it adaptive to the position of the user (more precisely, to camera-user distance and head pose), allowing users to speak during the transformation (an important limitation of previous work, allowing the simultaneous manipulation of smiled speech), as well as adding the possibility to use specific smile warpings that can be learned from a given user. We describe this algorithm, and how it relates to the existing literature, in Section 2.

Using a similar processing pipeline, the audio algorithm described in Section 3 tracks the frequency positions of the vocal formants (bumps and valleys of the vocal spectral envelope), shifts their positions and amplitude using a predefined parametric model, and reconstructs the audio signal with the new modified spectral cues. As far as we know, this part of the algorithm is the first published technique able to transform running speech in real-time to give it the characteristics of smiling (see [27] for a similar aim with more general emotional expressions). To develop it, we collected and analyzed a corpus of smiled and non-smiled vocalizations and derived a parametric model of how smiling affects the spectral properties of sound, all of which is described in this article.

Taken together, these two parts constitute a unique audiovisual algorithm which, in addition to providing simultaneous smiled transformations in both face and voice, also provides an experimental tool to investigate how humans integrate visual and auditory smiles both in their productive behaviour and in social perception. The last section of the article (Section 4) reports on a perceptual study that both validates each part of the algorithm separately and explores some of these audiovisual interactions.

2 Visual Smile Transformation

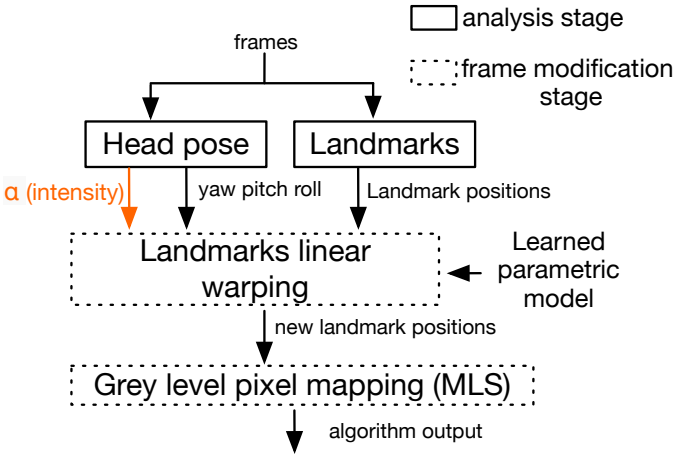
2.1 Transformation algorithm

Smiling involves the activity of several muscles that raise the corners of the mouth and cheek, and lift the lower eyelids [28]. To recreate these distortions in real time on any face, we designed a two-stage image processing algorithm, which stretches morphological features of the face around the lips and the eyes using a pre-learned parametric model, and resynthesizes pixel grey levels to correspond to the modified shape of the face. Figure 2 illustrates the global process.

2.1.1 Landmarks linear warping

The algorithm works in real time and applies a pre-learned smile deformation on a frame by frame basis. For each frame, we first detect 84 landmarks on the face, as well as the head pose (roll, pitch, yaw) using a framework from a generic face tracking SDK provided by Dynamixyz [29] - see figure 3-a.

Fig. 2. Overview of the visual smile transformation. The first stage of the algorithm (solid line) extracts feature from the video frames: head pose and 84 landmarks, from which the system notably computes the distance between the subject's eyes. The second stage (dotted line) operate image manipulation: first, positions of 12 of the landmarks are modified using a learned linear model, then the grey-level pixel intensities of the image are changed using a Moving Least Square algorithm.



Instead of heuristically designing a fixed warping function to simulate the expression of a smile, we made the choice to learn the pattern of landmark distortion on one actor's expression, and then apply this pre-learned pattern to all subsequent input videos. The reasons for this choice are the following: first, while the smile expressions as defined e.g. in the Facial Action Coding System (FACS) [28] make it possible to describe or detect such deformations, we did not find them sufficient to synthesize them with precision. Second, in an adaptive system, it appears interesting to learn the smile deformations that may be specific to a given person or attitude (e.g., genuine vs fake smiles [30]).

Learning is based on two images of the same subject, a neutral face and a slightly smiling face (with mouth shut, no visible teeth). After aligning the two faces, we calculated a linear deformation model for 12 landmarks to model the changes in the Zygomaticus Major (AU 12) and the Orbicularis Oculi (AU 6) muscles involved in smiling [28] (2 landmarks on the lower eyelid for each eye, 2 landmarks on the corners of the lips, three landmarks on the upper lip and three others on the lower lip - see figure 3-b).

In more details, if i is one landmark ($i = 1..12$), X_i^n its 2D coordinates from the neutral face and X_i^s its 2D coordinates from the smiling face, the linear model can be described as

$$X_i^s = (X_i^n + Q_r * \Delta_{xy}) * scale * \alpha_{video} \quad (1)$$

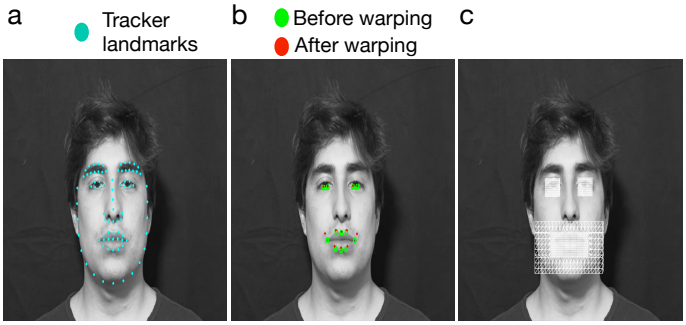
where Δ_{xy} is the learned parameters of the model and α_{video} is the intensity of the smile distortion. To adapt to face-camera distance and head pose, $scale$ is computed as the distance between the two eyes multiplied by cosine of the angle yaw, and Q_r is the rotation matrix corresponding to the roll. Figure 3-b shows an example of original and modified landmarks for a frontal face with $\alpha_{video} = 2.5$.

2.1.2 Pixel grey levels mapping

The second step of the algorithm computes the impact of landmark warping on the pixel grey-level intensity. As in [26], we use the rigid Moving Least Squares (MLS) method [31]. MLS optimizes the deformation made on an image when the position of some landmarks is modified, while maintaining the spatial coherence of the overall shape.

We made two approximations to the standard MLS procedure in order to allow real-time performance. First, we apply MLS only to areas of the image around the mouth and the eyes. Second, we do not apply the algorithm to every pixel of these areas but first approximate the areas with grids (with smaller meshes close to the eyes and mouth) and apply the deformation function to each vertex in the grid. We then fill the resulting triangles using affine warping. Figure 3-c shows an example of the grids after the MLS algorithm.

Fig. 3. Illustration of the tracking, warping and mapping steps in the visual smile transformation. (a) Tracking: 84 landmarks (turquoise dots) are automatically detected on the face. (b) Warping: the positions of 12 of the 84 landmarks (green dots) are transformed using a pre-learned linear model (red dots). (c) Mapping: we create a grid around the mouth and eyes, apply Moving Least Square deformation to each vertex of the grid and interpolate inside each resulting triangle using affine warping.



2.2 Video results

Figure 4 shows some examples of original ($\alpha_{video} = 0$) and transformed video frames with various positive and negative intensities ($\alpha_{video} = -1..1.5$). The second row shows the absolute difference between the grey-level intensities in the original vs transformed, confirming that modified pixels are found inside the grids around the mouth and eye areas.

A detailed quantitative evaluation can be found in Section 4. On a qualitative level, transformations appear plausible even in contexts where the subject is speaking (subject 2) or already smiling (subject 3).

One limitation of the warping model is that, while it is by construction adaptive to the frame-by-frame position of the mouth, it isn't to the qualitative nature of pronounced phonemes during speech. For instance, during real speech production, protruded/round vowels such as [y] may be incongruent with a large smile, whereas smiles on unround vowels such as [i] can be amplified without breaking the acoustic characteristics of the sound. A possible extension of the algorithm would be to learn a separate deformation pattern for different types of phonemes, and apply them adaptively, but this is beyond the scope of the current work.

One limitation of the MLS algorithm is that it cannot create textures that are not present in the original image, such as wrinkles. In particular, there are time-varying features (or "discontinuities") in the mouth and eyes areas (e.g., teeth which appear or disappear behind opening lip tissue, white sclera revealed by opening eyelids), which the algorithm cannot "add" to a frame if not originally present. Finally, at large intensities, the MLS algorithm may stretch geometric shapes, resulting e.g. in unrealistically oval rather than round iris shape, although the effect is not observable at the intensities investigated here.

Finally, we measured the latency of the overall visual algorithm including the 3 processing stages. The mean time (over 1000 iterations) to process a frame depends on the processing power of the machine. Our tests resulted in a mean 61ms processing time for a single frame (45ms for landmark tracking, 7ms for warping and 9ms for MLS) which is suitable for real time applications, for instance at 15 fps. Anyway, the latency can be further diminished either by reducing the number of landmarks in the tracking stage—the most time consuming stage—, or by improving the machine processing power, specially, the CPU speed. You can find examples of the algorithm at https://archive.org/download/StimuliExample, where speaking and head poses/orientations variations are presented.

3 Audio Smile Transformation

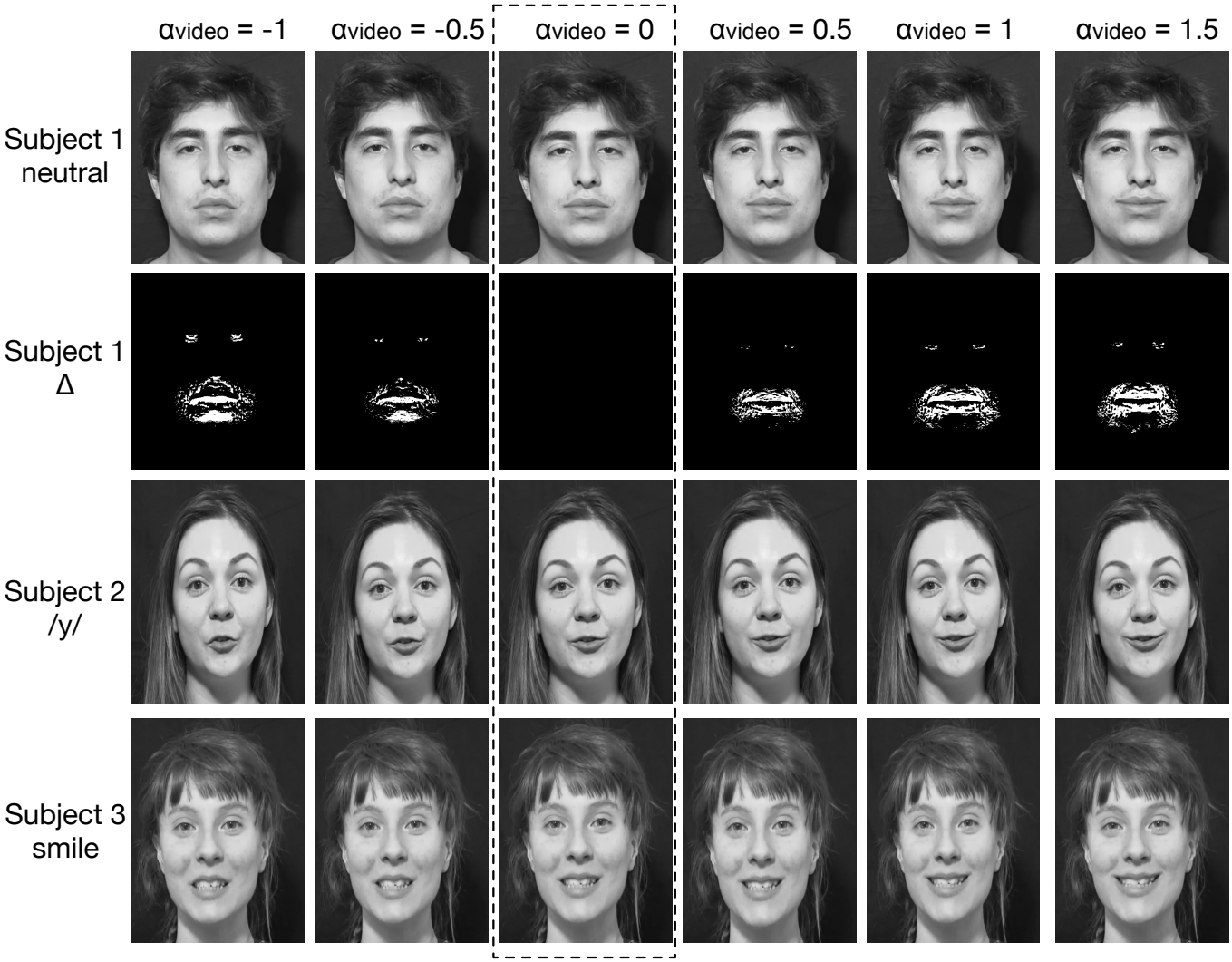
3.1 The acoustics of smiled speech

Although the visual features of a smile have been widely studied, it is still an open secret that smiles can also be heard in speech, even in the absence of visual cues [32], [33]. In a source-filter perspective, stretching lips while speaking changes the shape of the vocal resonator, possibly reducing vocal tract length, and thus transmitting filtered frequency content from the glottal impulses compared to normal speech. However, despite years of research on the acoustics of smiled speech, considerable debate still exists in the phonological community as to what features of speech necessarily result from—or rather simply co-occur with—smiling. Initially, smiled speech was thought to involve prosody similar to that of expressive speech, with high mean pitch and high intensity [24], [34], [35]. However, because smiles can also be perceived in whispered, non-pitched voices [36], pitch and prosody do not appear to be necessary components of smiled speech, which may more primarily affect sound spectrum. Accordingly, in [34], smiling was found associated with an increase of the second formant (F_2) for words with the round vowel /o:/, of intensity as well as F_0. In [37], smiling was associated with an increase of F_2, in [21], [38] with an increase of formants and F_0. Higher F1 and F2 dispersion are also reported [39]. Complementing these results, in [40] the mental representations of a smiled 'a' phoneme was found to have higher F_1 and F_2 as well as increased high frequency content. However, as the acoustic consequences of smiling do not seem to be similar across different phonemes, recent work has not converged to a common parametric model of how smiling affects the sound spectrum.

To clarify this situation, we recorded a dataset of smiled and non-smiled French phonemes and conducted an acoustical analysis of the recordings. We asked N = 8 (male: 6)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

Fig. 4. Examples of original and modified images with various positive and negative intensities ($\alpha_{video} = -1..1.5$). The original image is either neutral (subject 1) or speaking (subject 2) or already smiling (subject 3). Subject 1 Δ presents the difference between the non-modified and the modified image for subject 1, the black areas show where the image is unchanged, the white areas where the image is transformed



participants to pronounce 9 types of phonemes (5 voiced: *a, e, i, o, u* [a,ə,i,ɔ,y] and 4 unvoiced: *s, h, j, f* [s, ʃ, f, ʒ]), with and without stretched lips. Phonemes were pronounced three times each, and at 3 different pitches. The dataset was recorded at sampling rate 44.1kHz, in a sound-proof booth using a high quality microphone (DPA 4088 F). In the following, we analyse the recordings with phonological analysis software to measure the impact of smiling on three aspects of sound spectrum: formants, spectral envelope, and spectral centroid.

3.1.1 Consequences of smiling in formants

We analysed formant frequencies for all the smiled and non-smiled voiced phonemes using the Praat software [41]. Statistical analysis showed a significant increase of mean F_1 between the non smiled and the smile condition (a 5% increase from M=483 Hz to M=507 Hz; paired t-test t(7)=3.5, p=.008), and a marginally significant increase of F_2 (4% from 1572 Hz to 1634 Hz; paired t-test t(7)=1.9, p=.09), see Figure 5-a.

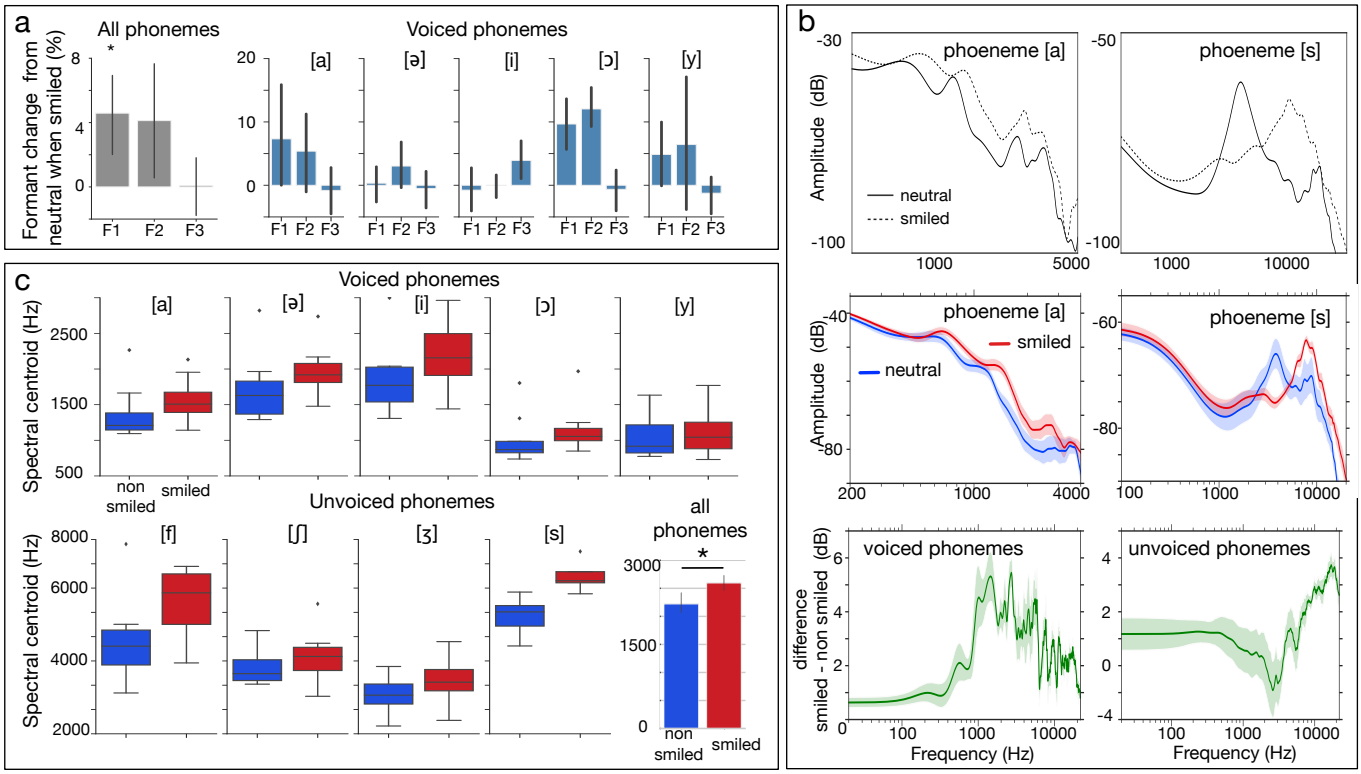
3.1.2 Consequences of smiling on the spectral envelope

We analysed the spectral envelope of the recordings using the adaptive true envelope technique [42], [43]. Spectral envelopes produced when smiling have more energy in the high-frequency regions, both for voiced and unvoiced phonemes (Figure 5-b). For voiced phonemes, the main difference between the smiled and non-smiled envelopes is found between 700 and 4000 Hz, corresponding to a shift and boost of the region around F1-F3. For unvoiced phonemes smiling affects higher frequencies, creating both resonances and antiresonances in the spectral envelope.

3.1.3 Consequences of smiling on the spectral centroid

Finally, we analysed the spectral centroid (where the "center of mass" of the spectrum is, a measure related to perceived brightness) for all the phonemes of the database (Figure 5-c) and found that the mean spectral centroid increases for every phoneme of the database when smiled, regardless of whether the phoneme is voiced, unvoiced, opened or closed. The overall effect is statistically significant (paired t-test t(7) = 6.2, p=.0004).

Fig. 5. Smiled speech corpus analysis. (a) Consequences of smiling on formants: Mean frequency shift of the first three formants, expressed in percentage of the non-smile utterance, averaged for all phonemes (left) and for each voiced phoneme (right) in the corpus. (b) Consequences of smiling on the spectral envelope. Top: Time-averaged spectral envelope of a single utterance of a French phonemes *'a'* and *'s'*, pronounced with and without smile. Middle: Averaged spectral envelope for all *'a'*s and *'a'*s of the corpus in smile and non-smiled conditions. Error bars represent standard errors. Bottom: Mean spectral envelope difference (smile minus non-smile) for all voiced and unvoiced phonemes of the corpus. (c) Consequences of smiling on spectral centroid. Mean spectral centroid for voiced (top), unvoiced (bottom left) and all (bottom right) phonemes in the corpus. Error bars represent 95% confidence intervals on the mean.



In sum, the mean acoustic consequence of smiling on sound spectrum is to shift both formants F1 and F2 by 5-10% (Figure 5-a) and to boost the high frequency energy (Figure 5-b).

3.2 Transformation algorithm

To simulate these changes on arbitrary spoken input, we designed a two-stage signal processing algorithm, which warps the vocal spectral envelope, then filters the reconstructed signal adaptively. Both stages are informed by a prior detection stage which tracks the positions of the formants. Figure 6 shows a general view of the algorithm.

This approach is different from the literature in several ways. First, compared to [21], [23], [24], [34], we implement here a transformation (i.e., operating on real speech input, and preserving its identity, prosody and content) rather than a synthesis technique (i.e., which generates speech from scratch). Second, by operating only on the spectral envelope and preserving the harmonic partials of the original voice, we avoid artifacts caused by the synthetic glottal impulses found with other formant re-synthesis approaches. Finally, like for the visual part of the algorithm, the frame-by-frame architecture of the system makes it suitable for real-time processing.

3.2.1 Piecewise linear frequency warping

In order to model the transformation of the whole vocal tract filter due to smiling, we use a spectral envelope manipulation technique, frequency warping, which does not only transform the local peak resonances (formants) but also the acoustic details besides these local peaks, e.g anti-resonances. Frequency warping was introduced to normalize vocal tract differences across speakers in order to improve the performance of recognition and categorization algorithms [44]. More recently, it has been applied to do speaker de-identification [45] and voice and gender conversion [46], [47]. Here, we use frequency warping to shift the spectral envelope (with its formants) either high or down with the aim of reinforcing or reducing the smile impression of a voice. The algorithm operates on a frame-per-frame basis. For each frame, it estimates the vocal spectral envelope (f_{in}), using the 'true envelope' technique [42], [43], and manipulates it using a non-linear change, or warping, of the frequency dimension (f_{out}). The intensity and direction of the warping are controlled by the parameter α_{audio}, such as $f_{out} = \Phi(f_{in}, \alpha_{audio})$.

The transformation function Φ, illustrated in Figure 7, was heuristically designed to shift the voice's formants by stretching and warping parts of the spectral envelope, to generate similar formant distributions as the ones seen in the voice recordings (Figure 5). Namely, to increase F1 and F2 frequencies. Φ is piece-wise linear with cut-frequencies defined as a function of the input signal's formant frequencies F_i: the output spectral envelope is untransformed below $F_1/2$ and above F_5; the segment between $F_1/2$ and F_2 is warped so that the spectral envelope at F_2 is mapped to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

Fig. 6. Overview of the audio smile transformation. The first stage of the algorithm is a transformation of the audio frames to the frequency domain, followed by both spectral envelope and f0 analysis. Spectral envelope analysis allows to compute speech's formants and F0 analysis to extract its harmonicity, and to categorize it either as a voiced or unvoiced frame. The two dotted blocks are the sound transformation stage, informed by the formant frequencies and harmonicity parameters extracted in the first stage.

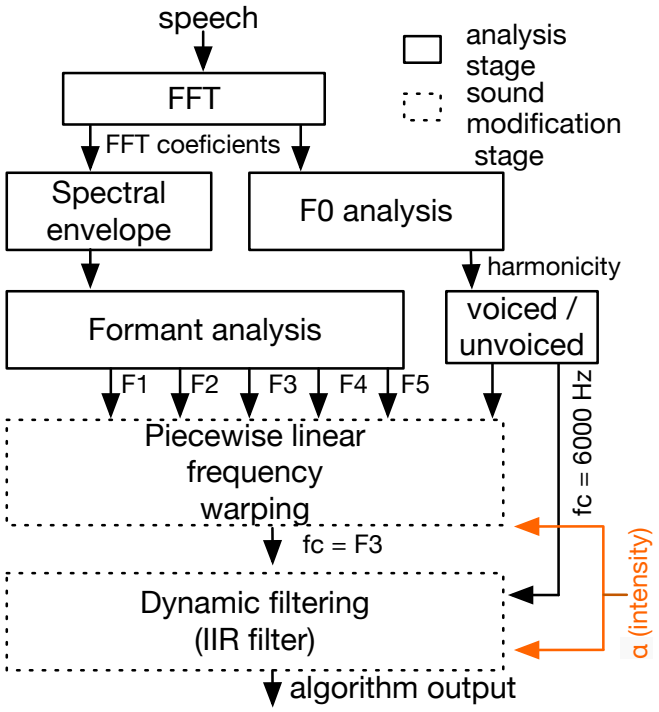
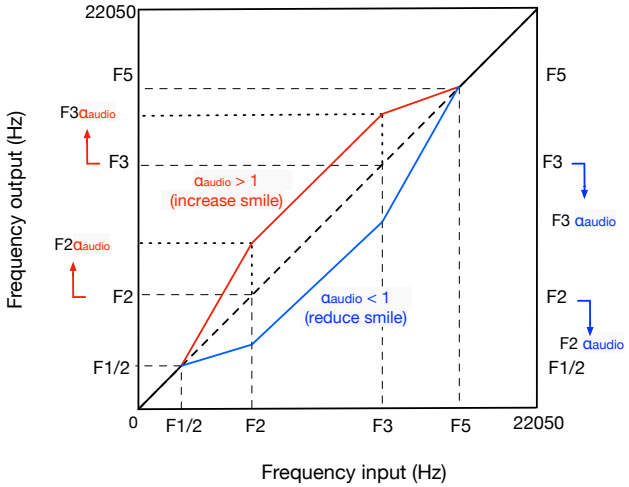


Fig. 7. Piecewise linear warping function mapping the frequency axis of the input envelope to the frequency axis of the output envelope. This function defines how the segments of the input spectral envelope are warped to the segments of the output spectral envelope. For instance, the segment [F1/2 , F2] will be warped to the segment [F1/2, F2α_{audio}], which will shift F2 either towards the high frequencies if $\alpha_{audio} > 1$ or towards the low frequencies when $\alpha_{audio} < 1$. The same logic applies to all the segments of the picewise linear function.



$\alpha_{audio}.F_2$ and F_3 to $\alpha_{audio}.F_3$; eventually, the last segment between $\alpha_{audio}.F_3$ and F_5 is warped to return to identity after F_5. Finally, we reapply the warped spectral envelope to the harmonic information and resynthesize the signal using the phase vocoder technique [48] [49]. Note that, if $\alpha_{audio} = 1$ then $f_{out} = f_{in}$; if $\alpha_{audio} > 1$, the algorithm shifts the envelope towards the high frequencies, and the higher α_{audio}, the higher the shift, which should increase the smile impression in a voice; Conversely, if $\alpha_{audio} < 1$, the acoustic effect is opposite and the envelope is shifted towards the low frequencies, which should reduce the smile impression.

As for the warping of face landmark positions in the visual part of the algorithm, the output of the frequency warping stage is also adaptive to the input signal, since frequency breakpoints follow formant frequencies in the signal. This adaptability can be used at different time scales: at low adaptation rates, if mean formant frequencies are computed for a range of sentences by a given speaker, the algorithm will adapt to speaker characteristics such as sex or body size (e.g., males have lower, more dispersed formants [50]); at faster rates, if formant frequencies are computed for each frame, the mapping will change phoneme per phoneme. In the current implementation and its validation in Section 4, mean formant frequencies were computed for each 1-second sentence in the validation set by averaging the formants over all the harmonic parts of the signal. Formant frequencies are estimated by taking the peaks of the 45-coefficient LPC envelope at a window size of 512 samples and hop size 8 samples (2ms), using the superVP software [49] - a non real-time alternative would be to use the formant estimation algorithm from the Praat software [41].

3.2.2 *Optimal alpha range*

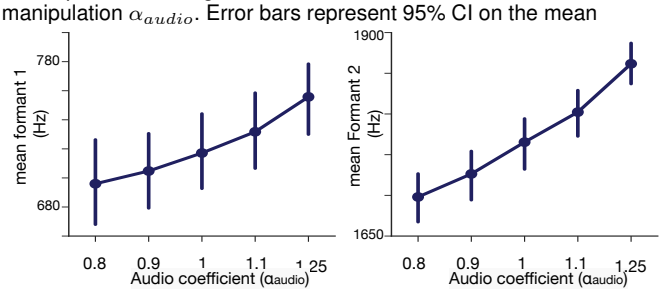
Here, we present an acoustical evaluation of the algorithm performance, to choose the optimal α_{audio} values, and to test weather changes of α_{audio} do produce formant movements comparable to those observed in the corpus analysis.

We analyze the formant frequencies of a set of 15 French speech sentences (mean duration = 2.3s, $F_s = 44100$), for five manipulation intensities (0.8, 0.9, 1, 1.1, 1.25) for which we compute the statistical effect on F1 and F2. The analysis was done with two one-way, within-sound-files, repeated-measures analysis of variance (RM-ANOVA). Data were analyzed using R (R Development Core Team, 2016), effect sizes are reported as generalized η^2 (Eta-Squared), Greenhouse-Geisser adjustment for sphericity corrections was applied when needed, and corrected p-values are reported along with uncorrected degrees of freedom.

The analysis revealed a significant main effect of the audio coefficient α_{audio} on F1 (F(4,56)=61.5, p=7.7e-10, η^2 =0.14) and F2 (F(4,56)=137.1, p=4.8e-13, η^2 = 0.5), as illustrated in Figure 8, showing that the manipulation does indeed shift formant frequencies. The optimal α_{audio} value to recreate the formant movements caused by smiling as observed in the natural recordings in Section 3.1 (5% for F1 and 4% for F2) is $\alpha_{audio} = 1.25$, which increased F_1 of 4.8% (from 717 Hz to 756 Hz) and F2 of 3.9% (from 1765 Hz to 1698 Hz). Conversely, for $\alpha_{audio} < 1$, we observe the opposite acoustic effect—a decrease of formant frequencies—for both F1 and F2. For instance, for $\alpha_{audio} = 0.8$, F1 and F2 decreased 2.9% and 3.8% respectively (from 717 Hz to 696 Hz for F_1; from 1765 to 1698 for F2). Thus, the range

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

Fig. 8. Formant changes as a function of $alpha_{audio}$. F1 frequency and F2 frequencies averaged over 15 validation sentences for intensities of manipulation α_{audio}. Error bars represent 95% CI on the mean



[0.8, 1.25] for α_{audio} seems to recreate the range formant variation seen in the corpus recordings.

3.2.3 Dynamic filtering

In addition to warping the signal's spectral envelope with the consequence of shifting the first formant frequencies, smiling also increases spectral energy in the higher-mids of the signal, between 1 and 4 kHz (Figure 5-b, Bottom left panel) for harmonic signals, a frequency area typically associated with F3. To simulate this element of smiled speech, in a second stage of the algorithm, we filter the reconstructed audio signal with an adaptive bell IIR filter which cut-frequency follows the third formant frequency. The filter gain is computed as $g = 20(\alpha_{audio} - 1)$ dB, which for α_{audio} in the range $[0.75, 1.25]$ varies from -5dB to 5dB, which is in line with the changes observed on real smiled utterances in Section 3.1. The cut-frequency refresh rate for the filter was chosen heuristically at 15ms, thus low-pass averaging the formant frequencies extracted at a rate of 2ms in the previous stage.

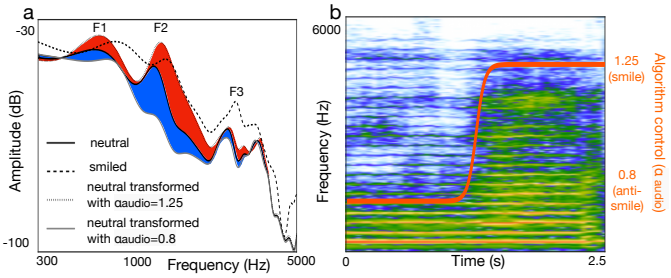
3.2.4 Special case of non-harmonic frames

Unvoiced phonemes, such as *s*, don't have clearly defined formants like voiced phonemes, and when they do, not in the same frequency region. To avoid formant estimation errors, we measure the signal harmonicity frame by frame, using the confidence of the pitch estimation algorithm of superVP. Upon reaching a low-harmonicity frame, neither the frequency warping stage nor the filtering stage update their parameters to the estimated formants of the frame; rather, they continue using the formant frequencies of the last-seen harmonic frame (until a new incoming harmonic frame is process, at which point continuous adaptation resumes with new formant frequencies). In addition, in order to recreate the type of resonance seen in Figure 5-b, non-harmonic frames are processed with a static filter centered at 6000 Hz with a Q of 1.5 and gain $g = 20(\alpha_{audio} - 1)$ dB.

3.2.5 Latency

As all time-frequency based digital audio effects, the overall latency of the algorithm depends on the window size of the FFT. An accurate time-frequency analysis is essential for high quality transformations as it is used to extract both the spectral envelope and the formants in the analysis-resynthesis stage. Here, for a sampling rate of 44100 and for a window size of 1024 samples, which is suitable for human voice signals, the latency of the algorithm is 75ms.

Fig. 9. Examples of the audio transformation. (a) Spectral envelopes of recorded and transformed phonemes [a]: solid bold: original version, pronounced with a neutral tone; dotted bold: original version, pronounced with stretched lips (smiled); dotted light: original version transformed with $\alpha_{audio} = 1.25$; solid light: original neutral transformed with $\alpha_{audio} = 0.8$. Red area represents spectral energy added to the neutral spectral envelope when $\alpha_{audio} = 1.25$; blue area represents energy taken out from the neutral envelope when $\alpha_{audio} = 0.8$. (b) Spectrogram of a single phoneme [a] transformed with the audio algorithm with a time-varying α_{audio} (a sigmoid going from 0.8 to 1.25; orange)



This is satisfactory for real time human-human interactions, but not for sensorimotor feedback [51]. As an example of the overall transformation, figures 9-a and 9-b present the transformed spectral envelope and spectrogram of an utterance of phoneme [a] for different α_{audio} values. You can find examples at https://archive.org/download/StimuliExample.

4 Perceptual Audiovisual Validation

In this section, we present a quantitative validation of how videos processed with the visual and audio smile algorithms are perceived by human observers (manipulation examples can be downloaded from https://archive.org/download/StimuliExample). In a first experiment, we validate the two modalities separately, by presenting our experimental participants with video-only and audio-only stimuli. In a second experiment, we examine how the two modalities of smile transformation interact when the video and audio channels of an audiovisual stream are processed simultaneously.

4.1 Validation of each modality

Ten participants (M=23, SD=3.21, 4 female, 6 men) took part in an experiment measuring the perceptual consequences of both the audio and the visual algorithm. Participants were naive to the fact that stimuli may be algorithmically manipulated, gave informed consent and were compensated for their participation.

Audio and video channels were separated from audiovisual recordings of 15 sentences with neutral content (6 Males, 9 female speakers, same audio as in Section 3.2.2). The 15 audio channels were transformed with the audio smile algorithm at 5 levels of intensity α_{audio} (0.8, 0.9, 1.0, 1.1, 1.25), for a total of 75 audio stimuli. The video channels were transformed with the visual smile algorithm at 6 levels of intensity α_{video} (-1,-0.5, 0, 0.5, 1.0, 1.5), for a total of 90 video stimuli without audio.

The task was composed of two blocks. In the first block, participants heard each of the 75 audio stimuli and had to answer the question "to what extent was this sentence

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

pronounced with a smile?" in a unipolar continuous scale anchored with "not smiling at all" and "with a lot of smile" (a mid-point was also included, which was - perhaps confusingly - labeled as "neutral"). In the second block, participants had to answer the same question, using the same scale, for each of the 90 video stimuli. Both in the audio and in the visual block, the presentation distance between manipulated variants of the same utterance was maximized. Stimulus order was pseudo-randomized following this constraint.

Participants' ratings are presented in figure 10 (where they are z-score normalized for visualisation purpose). The effect of smile intensity parameter α_{audio} on participants' ratings in the first block, and α_{video} on ratings in the second block, was analysed with two separate one-way within-subjects RM-ANOVA (statistics are reported as in Section 3.2.2).

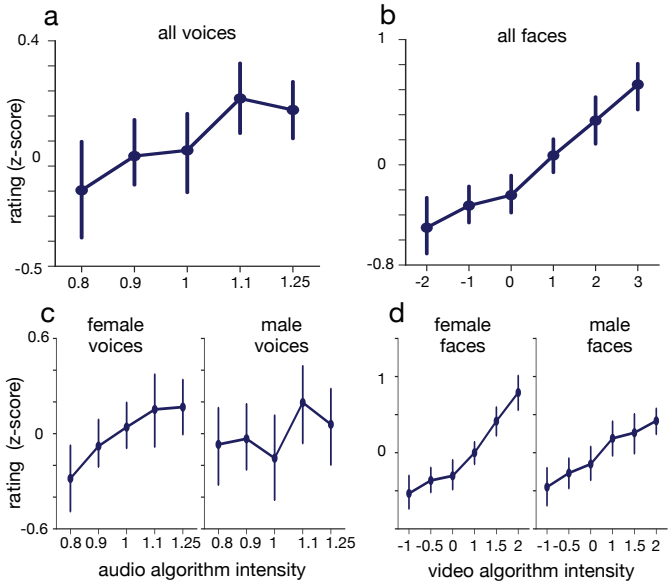
In the audio block, intensity of audio smile (α_{audio}) had a significant main effect on participant ratings of smiliness ($F(4,36)=4.8$, $p=0.004$, $\eta^2=0.07$), with increasing manipulation intensities (i.e. larger formant shifts and energy boost) perceived as increasingly smiling (Figure 10-a). Similarly, in the video block, we found a main effect of the video coefficient ($F(5,45)=11.0$, $p=0.001$, $\eta^2=0.37$), with increasing manipulation intensities (i.e. larger deformation of the mouth corners and eye regions) perceived as increasingly smiling (Figure 10-b). In sum, both the audio and the video manipulations had a significant influence on participants' ratings of smiliness. The effect of the video smile intensity parameter ($\eta^2=0.37$) was about 5 times larger as that of the audio smile intensity ($\eta^2=0.07$).

A breakdown of the effect on male and female videos can be seen in Figures 10-c-d. The visual smile transformation appears relatively stable across speaker gender, but at high visual smile intensity (α_{video}=1.5, 2) female speakers received higher ratings of smiliness than male speakers. Similarly, the audio smile transformation appears to work better on female than male voices. Indeed, although for both male and females an $\alpha_{audio} > 1$ does change positively the impression of a smile, $\alpha_{audio} < 1$ seem to reduce the perception of smiles only for female speakers.

In both modalities, such disparities across speaker gender may result from algorithmic limitations with certain physical features of the input stimuli. For instance, the audio disparities might come from difficulties shifting down the formants in low-pitched male voices. The visual disparities might be due to face size differences between genders—females have generally smaller face size than men [52]. Because the deformation model is linear, the visual transformation may transform more face area in small faces, giving higher ratings to females. Another possibility is that these disparities are in fact perceptual asymmetries. Female speakers may be perceived as more emotional than males at similar levels of expressive intensity.

Finally, it should be noted that we did not evaluate here the perceived naturalness of the transformations, i.e. whether transformed stimuli are readily accepted by listeners as authentic, plausible human expressions. Naturalness is an important consideration for audio/visual transformations. It is often found to be negatively affected by the effect intensity [27]. Future work should investigate this aspect of the transformations, but also how these interact with speaker identity ("sounds/looks human, but not like this speaker"), temporal dynamics ("nobody would smile continuously for such a long time") or semantics ("nobody would smile while saying that").

Fig. 10. Validation of the audio (A-C) and video (B-D) smile transformations in separate modalities. (A) Mean participant ratings of smiliness (z-score) in transformed audio recordings as a function of audio algorithm intensity (male and female stimuli). (B) Mean participant ratings of smiliness (z-score) in transformed video recordings (w/o sound) as a function of visual algorithm intensity (male and female stimuli) (C) Breakdown of audio results by speaker gender. (D) Breakdown of video results by speaker gender. Error bars represent 95% CI on the mean



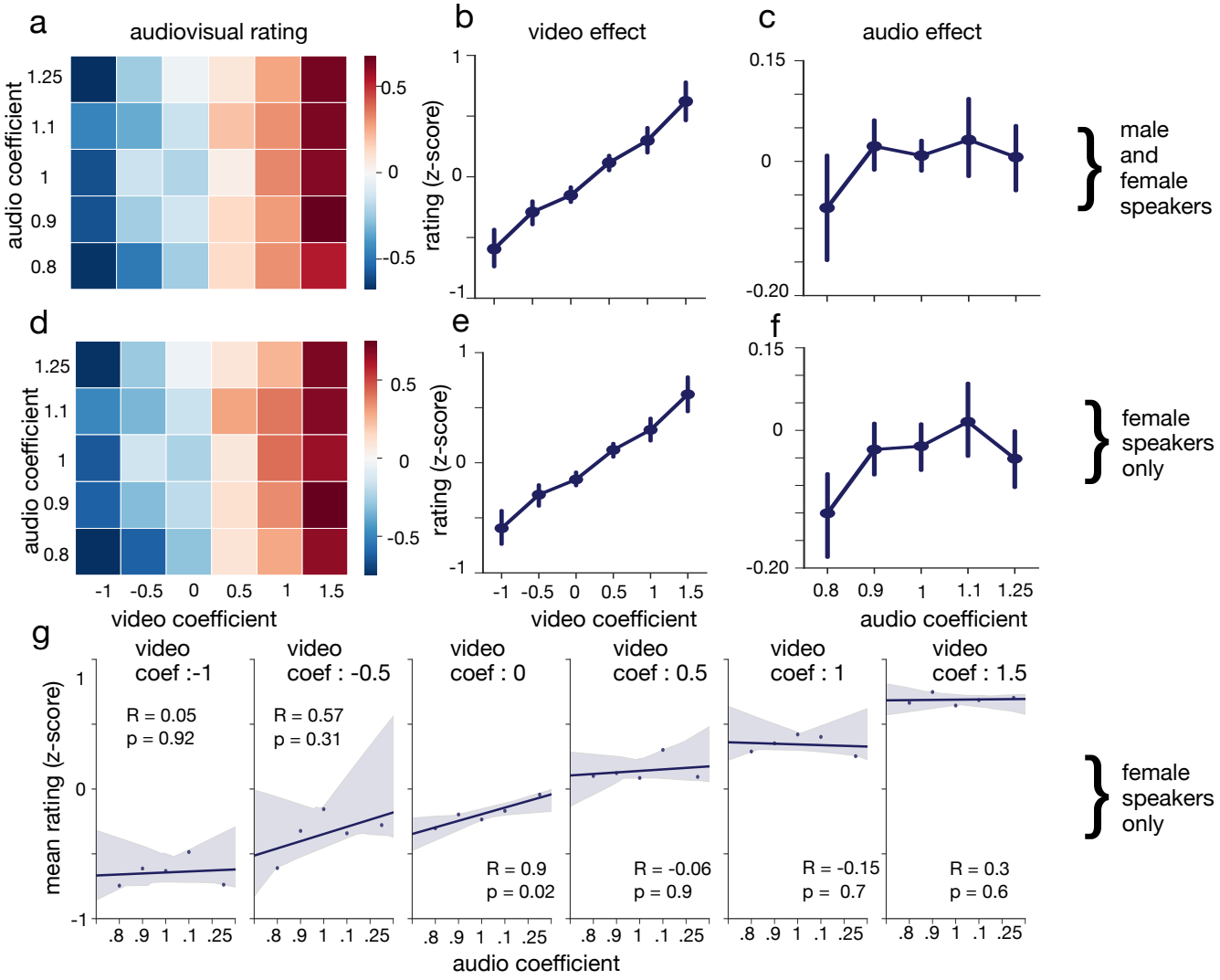
4.2 Audiovisual interaction

In a second experiment, we examine how the two modalities of smile transformation interact when the video and audio channels of an audiovisual stream are processed simultaneously. A separate group of N=15 participants (M=22, SD=3.6, 8 female, 7 men) took part in the study, in similar circumstances as above.

A subset of 12 videos (3 males, 9 females) taken from the first experiment was manipulated using the same five α_{audio} and six α_{video} levels as before, only this time conjointly. For each original audiovisual recording, we thus created 30 (6*5) manipulated videos with all the pairs of possible audiovisual manipulations, for a total of 360 rated videos, in which both congruent and incongruent audiovisual smiles are present. In a single experimental block, participants were presented all 360 stimuli, for each of which they were asked to rate their answer to the question "What is the emotional state of this person?" on a unipolar continuous scale ranging from "negative" to "positive". Note that, in this task, we used a more holistic question about emotional valence as a proxy to smiling, in order to force participants to use both audio and visual cues, as we found in pilot experiments that participants asked to evaluate "smilingness" in multimodal stimuli interpreted the task as the purely visual question whether the speaker's face showed the visual features of a smile, regardless of audio content. Participants ratings are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2018.2811465, IEEE Transactions on Affective Computing

Fig. 11. (A) Mean participant rating of speaker emotionality (z-score) in transformed audio-visual recordings as a function of audio and video algorithm intensity (male and female stimuli). (B) Mean participant ratings of speaker emotionality (z-score) in transformed audio-visual recordings as a function of visual algorithm intensity only (male and female stimuli) (C) Mean participant ratings of speaker emotionality (z-score) in transformed audio-visual recordings as a function of audio algorithm intensity only (male and female stimuli) (D-E-F): same data as A-B-C, restricted to female stimuli. (G): scatter plot and linear fit between the audio coefficient and participant ratings, broken down by level of video transformation intensities (female stimuli only). Error bars represent 95% CI on the mean



presented in figure 11. Figure 11-a presents mean participant rating (z-scored) for each pair of audio and video intensity levels. As can be seen, there was a clear horizontal gradient of emotional ratings from left to right, following the video intensity parameter, but no obvious vertical gradient of ratings following the audio smile intensity parameter. Figures 11-b-c slice through the same data, grouping by separate values of visual smile intensity (a), and audio smile intensity (c). A repeated measures-anova (RM-anova), with two within factors (audio coefficient : 5 levels, and video coefficient : 6 levels), confirmed a significant main effect of the video coefficient (F(5,70)=25.9, p=2.2e-5, η^2 = 0.24) and a non-significant effect of the audio coefficient (F(4,56)=1.6, p=0.19, η^2 = 0.003) on participant rating of the emotional state displayed in these stimuli. Because the audio smile manipulation was found in Section 4.1 to operate more strongly on female than male speakers, we analysed the subset of the current audiovisual data restricted to female stimuli (Figure 11d-f). This time, an RM-ANOVA revealed a significant main effect of both the audio (F(4,56)=X, p=4.7e-2, η^2 = 0.006) and the video (F(5,70)=23.5, p=8.4e-5, η^2 = 0.3) coefficients on participant ratings of emotion, as well a significant interaction between the audio and the video coefficients (F(20,280)=2.04, p=3.5e-2, η^2 = 0.015). Even restricted to female speakers, the size of the effect of the video transformation (η^2 = 0.3) remained 50 times larger than that of the audio effect (η^2 = 0.006), a ten-times increase of the difference in effect size seen in Section 4.1.

In sum, while the audio smile transformation is effective in an audio-only presentation, its effect is largely overridden by that of the video smile transformation in an audiovisual context. It appears cognitively plausible that visual cues are considered more reliable and salient in judging a given speakers emotion, and that in some cases, audio cues are only useful when visual cues are ambiguous or otherwise unavailable. The present data supports this interpretation:

Figure11-g breaks down the relation between the audio coefficient and participant ratings for the different levels of video transformation intensities for female stimuli. At extreme positive and negative video transformation intensities, the relation between α_{audio} and participant ratings is a flat horizontal line. However, for intermediate α_{video} values (i.e. when positive or negative cues are not, or not as much, available in the visual modality), the correlation between the strength of the audio transformation and the ratings becomes positive and statistically significant (R=0.9, p=.02 for $\alpha_{video} = 0$).

5 Conclusion and Perspectives

We created an audiovisual smile transformation algorithm able to manipulate an incoming video stream in real-time to parametrically control the amount of smile seen and heard on the users' face and voice. To simulate visual smiles, we use face recognition and automatic landmark positioning, followed by a warping and a mapping stage. For the audio transformation, we measured the acoustic consequences of phonation with stretched lips, and implemented an algorithm that simulates these acoustic cues on speech using adaptive frequency warping and dynamic filtering, with the consequence of significantly increasing formant frequency in running speech.

We validated the transformation using audio-only, video-only, and audiovisual stimuli processed with these algorithms. In separate modalities, both the audio and video smile transformations were associated with increased participant evaluations of speaker's smiliness, with some gender differences (stronger effects for female speakers). However, in audiovisual contexts, the strength of the audio transformation was largely overridden by that of the video smile transformation, which average effect was quantified as fifty times larger than the audio effect even when restricted to female stimuli. Further analysis revealed that the audio smile transformation was only significantly associated with participants' ratings of speaker emotionality when smile visual cues were weak or ambiguous, which suggests that human observers use a hierarchy of perceptive processes, with higher priority/saliency to visual than audio cues, when judging audiovisual smiles.

This does not entail that audio smile transformations are not useful. First, they find a natural application in audio-only interactions. With more than 60% of all customer experience interactions still happening over the telephone [53], it would be particularly interesting to test the effect of audio smile enhancement on variables like customer satisfaction or retention rates in naturalistic contexts like a call-center [33]. Second, the fact that auditory smile cues may be secondary to visual cues does not diminish their usefulness in contexts where it is inappropriate or otherwise impossible to manipulate visual cues, e.g. when manipulations need to remain undetectable.

Several algorithmic improvements can be pursued for the techniques reported here. In the audio modality, it should be clarified whether the difference in perceived smile intensity between male and female stimuli result from algorithmic limitations when processing male rather than female voices. One notable possibility is that lower-pitched male voices suffer from spectral estimation artifacts at the time resolution used in the algorithm, thus hurting the precision of pitch and formant frequency estimation in the analysis stage, or the precision of phase-vocoder reconstruction in the synthesis stage (similar problems were discussed e.g. in [27]). It remains an intriguing possibility, however, that these differences are explained by a cognitive, rather than an algorithmic, asymmetry, in which observers judge objectively-identical levels of transformation differently depending on speaker gender [54].

In the visual algorithm, the transformation described here is based on a parametric warping model learned from a single subject. While this approximation proved reasonable here, as shown e.g. by quantitative evaluations in video and audio/video contexts, the method could be extended to simulate different kinds of smiles (e.g. genuine vs fake smiles, the discrimination of which may depend on processing the eye region and temporal dynamics [55]) or to convey different types of smile-expressed affect as amusement, joy, or shame. The tool would also readily lend itself to modeling specific smile transformations for different users, potentially allowing more precise or realistic expressions across e.g. gender or individual differences. This may be particularly needed when manipulating facial features of well-known or familiar speakers, for which observers may have more stringent representations of what's real and what's not.

Finally, independently from work in each modality, further improvements to the system may consider its integration with wider information about the context of the interaction. First, algorithms in both modalities can be controlled in real-time using their intensity parameter α, which opens the question of modeling appropriate temporal dynamics for both effects in the course of an interaction, such as e.g., detecting phrase boundaries to smile as a back-channel at the end of a turn [56]. Second, the effects could be integrated in a wider audiovisual emotion recognition system, and thus make the transformed smiles adaptive to a speaker's emotional expression. Finally, an intriguing possibility would also be to use our audiovisual system to generate large amounts of parametrically-varied training examples for face and voice classification algorithms to learn from.

Beyond affective computing and human-computer interaction, we anticipate that this technology will find wide-ranging applications as an experimental method in the behavioral sciences, because it enables a high level of control over the acoustical, visual and emotional content of experimental stimuli in a variety of laboratory situations, including video-mediated real-time social situations. Possible applications include e.g. creating stimuli with systematically-varying degrees of audio and visual smiles to study their audiovisual integration in observers [57], modeling the process of facial mimicry and emotional contagion in observers depending on the intensity of the facial and auditory cues that are presented to them [58] or studying the impact of emotional processes on group performance by manipulating smile expressive cues in a group of participants while they are interacting to solve a problem [59]. Efforts will be made to make the tool available in a user-friendly format to support this type of applications.

ACKNOWLEDGMENTS

Work partially funded by ERC StG CREAM 335536 (to J-J.A.) and ANR REFLETS (to C.S.). Experimental data was collected at the Centre Multidisciplinaire des Sciences Comportementales Sorbonne Universités-Institut Européen dAdministration des Affaires (INSEAD). The authors thank Louise Vasa for her help with the data collection. All data reported in the paper are available on request.

REFERENCES

[1] J. Martin, M. Rychlowska, A. Wood, and P. Niedenthal, "Smiles as multipurpose social signals," *Trends in cognitive sciences*, 2017.

[2] A. N. Meltzoff, M. K. Moore *et al.*, "Imitation of facial and manual gestures by human neonates," *Science*, vol. 198, no. 4312, pp. 75–78, 1977.

[3] P. Ekman, E. R. Sorenson, W. V. Friesen *et al.*, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[4] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological bulletin*, vol. 128, no. 2, p. 203, 2002.

[5] J. ODoherty, J. Winston, H. Critchley, D. Perrett, D. M. Burt, and R. J. Dolan, "Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness," *Neuropsychologia*, vol. 41, no. 2, pp. 147–155, 2003.

[6] V. Surakka and J. K. Hietanen, "Facial and emotional reactions to duchenne and non-duchenne smiles," *International Journal of Psychophysiology*, vol. 29, no. 1, pp. 23–33, 1998.

[7] P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess, "The simulation of smiles (sims) model: Embodied simulation and the meaning of facial expression," *Behavioral and brain sciences*, vol. 33, no. 06, pp. 417–433, 2010.

[8] K. El Haddad, H. Cakmak, S. Dupont, and T. Dutoit, "Laughter and smile processing for human-computer interactions," *Just talking-casual talk among humans and machines, Portoroz, Slovenia*, pp. 23–28, 2016.

[9] H. Yu, O. G. Garrod, and P. G. Schyns, "Perception-driven facial expression synthesis," *Computers & Graphics*, vol. 36, no. 3, pp. 152–162, 2012.

[10] S. Oh, J. Bailenson, N. Krämer, and B. Li, "Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments," *PLoS ONE*, vol. 11, no. 9, p. e0161794, 2016.

[11] T. Partala and V. Surakka, "The effects of affective interventions in human–computer interaction," *Interacting with computers*, vol. 16, no. 2, pp. 295–309, 2004.

[12] N. KräMer, S. Kopp, C. Becker-Asano, and N. Sommer, "Smile and the world will smile with youthe effects of a virtual agents smile on users evaluation and behavior," *International Journal of Human-Computer Studies*, vol. 71, no. 3, pp. 335–349, 2013.

[13] J. Ku, H. J. Jang, K. U. Kim, J. H. Kim, S. H. Park, J. H. Lee, J. J. Kim, I. Y. Kim, and S. I. Kim, "Experimental results of affective valence and arousal to avatar's facial expressions," *CyberPsychology & Behavior*, vol. 8, no. 5, pp. 493–503, 2005.

[14] M. Ochs, C. Pelachaud, and G. Mckeown, "A user perception-based approach to create smiling embodied conversational agents," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 1, p. 4, 2017.

[15] N. Yee, J. N. Bailenson, and K. Rickertsen, "A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 1–10.

[16] H. Meij, J. Meij, and R. Harmsen, "Animated pedagogical agents effects on enhancing student motivation and learning in a science inquiry learning environment," *Educational technology research and development*, vol. 63, no. 3, pp. 381–403, 2015.

[17] H. Maldonado, J.-E. R. Lee, S. Brave, C. Nass, H. Nakajima, R. Yamada, K. Iwamura, and Y. Morishima, "We learn better together: enhancing elearning with emotional characters," in *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!* International Society of the Learning Sciences, 2005, pp. 408–417.

[18] Y. Kim, J. Thayne, and Q. Wei, "An embodied agent helps anxious students in mathematics learning," *Educational Technology Research and Development*, pp. 1–17, 2016.

[19] C. Pelachaud, "Greta: an interactive expressive embodied conversational agent," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 5–5.

[20] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," *arXiv preprint arXiv:1611.09961*, 2016.

[21] K. El Haddad, S. Dupont, N. d'Alessandro, and T. Dutoit, "An hmm-based speech-smile synthesis system: An approach for amusement synthesis," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 5. IEEE, 2015, pp. 1–6.

[22] K. El Haddad, H. Cakmak, A. Moinet, S. Dupont, and T. Dutoit, "An hmm approach for synthesizing amused speech with a controllable intensity of smile," in *Signal Processing and Information Technology (ISSPIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 7–11.

[23] K. El Haddad, H. Cakmak, S. Dupont, and T. Dutoit, "Towards a speech synthesis system with controllable amusement levels," in *4th Interdisciplinary Workshop on Laughter and Other Non-Verbal Vocalisations in Speech, Enschede, Netherlands*, 2015, pp. 14–15.

[24] H. Quené, G. R. Semin, and F. Foroni, "Audible smiles and frowns affect speech comprehension," *Speech Communication*, vol. 54, no. 7, pp. 917–922, 2012.

[25] P. P. Filntisis, A. Katsamanis, P. Tsiakoulis, and P. Maragos, "Video-realistic expressive audio-visual speech synthesis for the greek language," *Speech Communication*, vol. 95, pp. 137–152, 2017.

[26] S. Yoshida, T. Tanikawa, S. Sakurai, M. Hirose, and T. Narumi, "Manipulation of an emotional experience by real-time deformed facial feedback," in *Proceedings of the 4th Augmented Human International Conference*. ACM, 2013, pp. 35–42.

[27] L. Rachman, M. Liuni, P. Arias, A. Lind, P. Johansson, L. Hall, D. Richardson, K. Watanabe, S. Dubal, and J. Aucouturier, "David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech," *Behavior Research Methods (in press)*, 2017.

[28] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[29] Dynamixyz, "Dynamixyz generic face tracking," 2017. [Online]. Available: www.Dynamixyz.com

[30] S. D. Gunnery and M. A. Ruben, "Perceptions of duchenne and non-duchenne smiles: A meta-analysis," *Cognition and Emotion*, vol. 30, no. 3, pp. 501–515, 2016.

[31] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 533–540.

[32] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Attention, Perception, & Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[33] F. Basso and O. Oullier, "smile down the phone: Extending the effects of smiles to vocal social interactions," *Behavioral and Brain Sciences*, vol. 33, no. 06, pp. 435–436, 2010.

[34] H. Barthel and H. Quené, "Acoustic-phonetic properties of smiling revised–measurements on a natural video corpus," in *Proceedings of the 18th International Congress of Phonetic Sciences.–Glasgow, UK: The University of Glasgow*, 2015.

[35] E. Lasarcyk and J. Trouvain, "Spread lips+ raised larynx+ higher f0= smiled speech?-an articulatory synthesis approach," *Proceedings of ISSP*, pp. 43–48, 2008.

[36] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, 1994.

[37] R. J. Podesva, P. Callier, R. Voigt, and D. Jurafsky, "The connection between smiling and goat fronting: Embodied affect in sociophonetic variation," in *Proceedings of the International Congress of Phonetic Sciences*, vol. 18, 2015.

[38] K. El Haddad, I. Torre, E. Gilmartin, H. Çakmak, S. Dupont, T. Dutoit, and N. Campbell, "Introducing amus: The amused speech database," in *International Conference on Statistical Language and Speech Processing*. Springer, 2017, pp. 229–240.

[39] A. Drahota, A. Costall, and V. Reddy, "The vocal communication of different kinds of smile," *Speech Communication*, vol. 50, no. 4, pp. 278–287, 2008.

[40] E. Ponsot, P. Arias, and J.-J. Aucouturier, "Uncovering mental representations of smiled speech using reverse correlation," *JASA-EL*, 2017.

[41] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.
[42] F. Villavicencio, A. Robel, and X. Rodet, "Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.
[43] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *International Conference on Digital Audio Effects*, 2005, pp. 30–35.
[44] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 353–356.
[45] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E. R. Banga, C. Garcia-Mateo, and D. Erro, "Piecewise linear definition of transformation functions for speaker de-identification," in *Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016 First International Workshop on*. IEEE, 2016, pp. 1–5.
[46] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*. IEEE, 2014, pp. 211–215.
[47] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.
[48] A. Roebel, "Shape-invariant speech transformation with the phase vocoder," in *InterSpeech*, 2010, pp. 2146–2149.
[49] M. Liuni and R. Axel, "Phase vocoder and beyond," *Musica, Tecnologia*, vol. 7, pp. 73–120, 2013.
[50] S. Evans, N. Neave, and D. Wakelin, "Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice," *Biological psychology*, vol. 72, no. 2, pp. 160–163, 2006.
[51] J.-J. Aucouturier, P. Johansson, L. Hall, R. Segnini, L. Mercadié, and K. Watanabe, "Covert digital manipulation of vocal emotion alter speakers emotional states in a congruent direction," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 948–953, 2016.
[52] A. V. Badyaev, "Growing apart: an ontogenetic perspective on the evolution of sexual size dimorphism," *Trends in Ecology & Evolution*, vol. 17, no. 8, pp. 369–378, 2002.
[53] I. F. d'Opinion Publique, "Les franais et les services clients," 2015.
[54] S. Whittle, M. Yücel, M. B. Yap, and N. B. Allen, "Sex differences in the neural correlates of emotion: evidence from neuroimaging," *Biological psychology*, vol. 87, no. 3, pp. 319–333, 2011.
[55] E. Krumhuber and A. Kappas, "Moving smiles: The role of dynamic components for the perception of the genuineness of smiles," *Journal of Nonverbal Behavior*, vol. 29, no. 1, pp. 3–24, 2005.
[56] T. Stivers, "Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation," *Research on language and social interaction*, vol. 41, no. 1, pp. 31–57, 2008.
[57] B. De Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cognition & Emotion*, vol. 14, no. 3, pp. 289–311, 2000.
[58] P. Arias, P. Belin, and J.-J. Aucouturier, "Auditory smiles trigger unconscious facial imitations," *submitted*, 2017.
[59] E. A. Van Doorn, M. W. Heerdink, and G. A. Van Kleef, "Emotion and the construal of social situations: Inferences of cooperation versus competition from expressions of anger, happiness, and disappointment," *Cognition & emotion*, vol. 26, no. 3, pp. 442–461, 2012.

Pablo Arias Pablo Arias is a PhD candidate in CNRS/IRCAM. He holds a M.Eng. from Polytech Nantes and a M.Sci. in acoustics, signal processing and computer science applied to sound and music from UPMC/IRCAM. In addition, he is a musician and a music producer. His research is focused in using audio processing techniques to understand how sound induces emotions in humans.

Catherine Soladié Catherine Soladié graduated from the engineering school Supelec in France in 2002. She worked for 8 years at Capgemini as a developer and then project manager. In 2010, she completed a PhD in the field of facial expression analysis. In 2013, she joined the FAST (Facial Analysis, Synthesis and Tracking) lab of Supelec as an Assistant Professor in Image Processing. Her current research focuses on face analysis and synthesis, with a focus on emotional expressions.

Oussema Bouafif Oussema Bouafif is an Internship student in Centrale Supelec/IETR at the Team FAST, and a Master student in Signal and Image processing at the University of Toulouse III. He received his Bachelor's degree, in 2015 from the Faculty of Mathematical, Physical and Natural Sciences of Tunis. His current research interests are related to facial expressions analysis and synthesis.

Axel Roebel Axel Roebel received the Diploma in electrical engineering from Hanover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. He is currently the head of the Analysis/Synthesis Team at IRCAM. His current research interests are related to music and speech signal analysis and transformation.

Renaud Séguier Renaud Séguier received the PhD degrees in Signal Processing, Image, Radar from the University of Rennes I in 1995 and the HDR (Habilitation Diriger des Recherches) in 2012. He worked one year in Philips R&D department on numerical TV and Mpeg2 transport-stream. He joined SCEE (Communication and Electronic Embedded Systems) lab of Supélec in 1997 as Assistance Professeur. In 2015 he created FAST (Facial Analysis, Synthesis & Tracking) a new research team dedicated to emotion analysis and synthesis for medical applications; and reached the rank of Professor and team leader.

Jean-Julien Aucouturier JJ Aucouturier is a CNRS researcher in IRCAM in Paris. He was trained in Computer Science (Ecole Supérieure d'Elecricité, 2001; PhD University of Paris 6, 2006), and held several postdoctoral positions in Cognitive Neuroscience in RIKEN Brain Science Institute in Tokyo, Japan and University of Burgundy, France. He is now heading the CREAM music neuroscience lab in IRCAM, and is interested in using audio signal processing technologies to understand how sound and music create emotions.