Acoustic-visual synthesis technique using bimodal unit-selection

Slim Ouni, Vincent Colotte, Utpala Musti, Asterios Toutios, Brigitte Wrobel-Dautcourt, Marie-Odile Berger, Caroline Lavecchia

EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Acoustic-visual synthesis technique using bimodal unit-selection

Slim Ouni[1*], Vincent Colotte[1], Utpala Musti[2], Asterios Toutios[3], Brigitte Wrobel-Dautcourt[1], Marie-Odile Berger[2] and Caroline Lavecchia[1]

Abstract

This paper presents a bimodal acoustic-visual synthesis technique that concurrently generates the acoustic speech signal and a 3D animation of the speaker's outer face. This is done by concatenating bimodal diphone units that consist of both acoustic and visual information. In the visual domain, we mainly focus on the dynamics of the face rather than on rendering. The proposed technique overcomes the problems of asynchrony and incoherence inherent in classic approaches to audiovisual synthesis. The different synthesis steps are similar to typical concatenative speech synthesis but are generalized to the acoustic-visual domain. The bimodal synthesis was evaluated using perceptual and subjective evaluations. The overall outcome of the evaluation indicates that the proposed bimodal acoustic-visual synthesis technique provides intelligible speech in both acoustic and visual channels.

Keywords: Audiovisual speech; Acoustic-visual synthesis; Unit-selection

1 Introduction

In several situations speech is considered as a bimodal signal. The first modality is audio, provided by the acoustic speech signal, and the second is visual, provided by the face of the speaker. The speech signal is the acoustic consequence of the deformation of the vocal tract under the effect of the movements of articulators such as the jaw, lips, and tongue.

Since some of the articulators directly correspond to facial features, it is quite reasonable to find out that acoustics and facial movements are correlated [1,2].

Research in audiovisual speech intelligibility has shown the importance of the information provided by the face especially when audio is degraded [3-5]. Moreover, Le Gof et al. [4] have shown that when audio is degraded or missing, the natural face provides two thirds of the missing auditory intelligibility, their synthetic face without the inner mouth (without the tongue) provides half of the missing intelligibility, and the lips restores a third of it. For audiovisual synthesis, this suggests that one should pay careful attention to model the part of the face that participates actively during speech, i.e., mainly the lips and lower part of the face.

In the vast majority of recent works, data-driven audiovisual speech synthesis, i.e., the generation of facial animation together with the corresponding acoustic speech, is still considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the facial animation [6-9]. However, achieving perfect synchronization between these two streams is not straightforward and presents several challenges related to audiovisual intelligibility. In fact, humans are acutely sensitive to any incoherence between audio and visual animation. This may occur as an asynchrony between audio and visual speech [10], or a small phonetic distortion compared to the natural relationship between the acoustic and the visual channels [11-14]. The McGurk effect [15] describes the case when the mismatch is more important: when an auditory stimulus 'ba' is paired with a visual stimulus 'ga', and the perceiver reports that the talker said 'da'. This is called a fusion effect. We can observe a combination effect when pairing an auditory ga with a visual ba, and the perceived result is a combined 'bga'. Some perceptual studies may suggest that the acoustic and visual information is processed as

*Correspondence: Slim.Ouni@loria.fr
[1] Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
Full list of author information is available at the end of the article

a 'whole unit' [11,15]. In the field of audiovisual synthesis, it has been shown that the degree of coherence between the auditory and visual modalities has an influence on the perceived quality of the synthetic visual speech [16].

All these studies suggest the importance of keeping the link between the two highly correlated acoustic and visual channels. To reduce the possible existence of incoherency during audiovisual facial animation, we propose to achieve synthesis with its acoustic and visible components simultaneously. Therefore, we consider audiovisual speech as a bimodal signal with two channels: acoustic and visual. This bimodality is kept during the whole synthesis process. The setup is similar to a typical concatenative acoustic-only speech synthesis, with the difference that here the units to be concatenated consist of visual information alongside acoustic information. The smallest segmental unit adopted in our work is the diphone. The advantage of choosing diphones is that the major part of coarticulation phenomena is captured in the middle of the unit, and the concatenation is made at the boundaries, which are acoustically and visually steadier. This choice is in accordance with current practices in concatenative acoustic speech synthesis [17,18].

Although our long-term research goal is to provide a full talking head system, current focus is the synthesis technique itself: combining both channels during the whole synthesis process. Attempts to use bimodal units have been proposed in the past [16,19-22]. For instance, Tamura et al. [20] proposed a synthesis technique to animate a simple lip model synchronously with acoustic. The technique is based on the parameter generation from HMM with dynamic features, using triphone models. Fagel [22] proposed an audiovisual synthesis approach of a German speaker by concatenating synchronous bimodal polyphone segments. The selection of these segments was based on a combined concatenation cost using a weighted sum of costs of audio and visual features. The pre-selection of possible polyphone segments from the four-minute corpus was exhaustive. The visual join cost calculation was based on the pixel-to-pixel color differences in the boundaries of the segments to be concatenated. Mattheyses et al. [16] presented an audiovisual synthesis technique based on the acoustic unit-selection technique extended to the audiovisual domain. They included an additional cost for visual join discontinuities. There are some similarities in terms of the extracted visual features and process with that of Liu and Ostermann [8].

The works of Fagel [22] and Mattheyses et al. [16] share some common characteristics with ours, since they address the audiovisual synthesis problem as one of concatenating units that combine acoustic and visual information. Nevertheless, our technique is unique due to the major differences in the methods used for 2D versus 3D. The 3D case calls for a novel casting of the unit-selection method.

We believe that an ideal audiovisual speech synthesis system should target the human receiver as its final and principal goal. Therefore, we focus on those aspects of audiovisual speech that make it more intelligible. These involve the dynamics of the lips and the lower part of the face: given that the lips are accurately animated, articulation and coarticulation will reproduce similar behavior as that of the real speaker. To achieve this goal, we are using a straightforward but efficient acquisition technique to acquire and process a large amount of parallel audiovisual data to cover the whole face by 3D markers. As can be seen in Figure 1, a large number of these markers mainly covers the lower face to allow accurate reconstruction of the lips and all the area around them.

At the current stage of our long-term research goal, we do not provide a full talking head with a high rendering resolution. We do provide a bimodal synthesis method that can serve as the core of a larger system which will animate a high-resolution mesh of the face with the inner vocal tract, using our simultaneous bimodal synthesis technique. Hence, our attempts are directed towards synthesizing realistic acoustic-visual dynamics that is coherent and consistent in both domains simultaneously: audio and visual.

We have previously presented a preliminary version of our synthesis technique [23]. In the present paper, we provide the details of the synthesis method and its evaluation. We first present our bimodal data acquisition system, acquired corpus, and the modeling of our visual

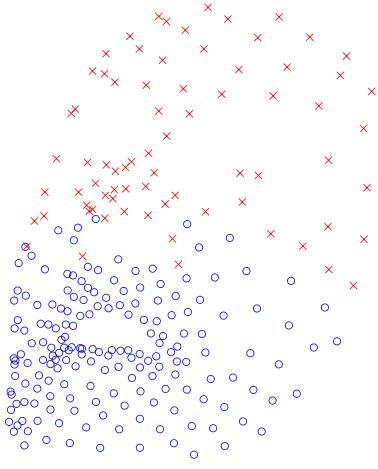

Figure 1 3D positions of the 252 markers. One hundred seventy-eight of these markers (plotted in blue circles) are covering the lower face. The remaining markers (plotted in red crosses) do not reflect explicit speech gestures, in our case.

data using principal component analysis (PCA). Then, we present our method to synthesize audiovisual speech on the principle of bimodal unit-selection. Finally, we present evaluation results that validate the potential benefits of our proposed synthesis method.

2 Data acquisition and modeling

Figure 2 shows an outline of our data acquisition and modeling process. As detailed in the following sections, stereovision data are recorded simultaneously with audio. The acoustic and visual components of the corpus are processed, and the corpus is analyzed linguistically. The final result is stored in a database as diphone entries.

2.1 Acquisition and 3D reconstruction

Visual data acquisition was performed simultaneously with acoustic data recording, using a classical stereovision system we developed few years ago [24].

2.1.1 Setup

During acquisition, the speaker sat in front of a stereo camera pair with a microphone placed at 50 to 60 cm from his mouth. Two synchronized fast monochrome cameras (JAI TM-6740) were used for acquisition (188 fps) thus enabling the tracking of fast movements of the articulators, for instance, release burst of bilabial obstruents. The two cameras were calibrated with the use of a calibration target.

Visual (spatial and temporal) data acquisition requires the same physical points to be tracked over time. As the natural skin is not textured enough, we painted markers on the speaker's face. This method allows control of the size, density, and position of these points of interest.

2.1.2 3D Markers reconstruction and tracking

A preprocessing stage is first used on the images to detect the markers based on their average grayscale, shape and size (white circular points with a radius less than 3 pixels). This low-level process is able to detect most of the markers except some points that are not visible in one image of some stereo image pairs. These points are then matched using the epipolar constraint allowing to retrieve a set of 3D points for every image pair. The majority of markers are reconstructed, but some of them may be missing because they are not detected in some stereo images. This is the case of markers on the temple which disappear when the speaker slightly turns his head. More complex is the case of markers located on the lips, which are occluded during protrusion or mouth closure (Figure 3): markers can disappear or be erroneously matched with the wrong side (lower or upper) of the lip. In addition, the stereovision process may include erroneous points, which have the same photometric features as light reflects on eyes, nose, teeth, or tongue. The use of PCA for modeling the facial dynamics makes it necessary to match physically the 3D points over time, which is a tedious task due to the high speed of lip motion for some sounds. In addition, classical PCA requires the set of points to be determined at each time instant. To cope with these problems, we use a topological mesh which helps us to match temporally the 3D points and to estimate the missing points.

2.1.3 Spatiotemporal mesh reconstruction

The corpus was acquired by sequences of 2 min (around 26,000 frames). For each sequence of stereo pairs, 3D points are built at each time instant. Note that the points located on top of the head are used to compensate for head motion. Then temporal trajectories are built based on the estimated position and velocity.

A topological mesh is then interactively built from the set of points of the time instant for which the largest number of 3D points were reconstructed. The role of the topology is twofold: (a) it defines the neighbors of a point in order to estimate it from its neighbors, when this point is missing in one frame; (b) it prevents the temporal wrong

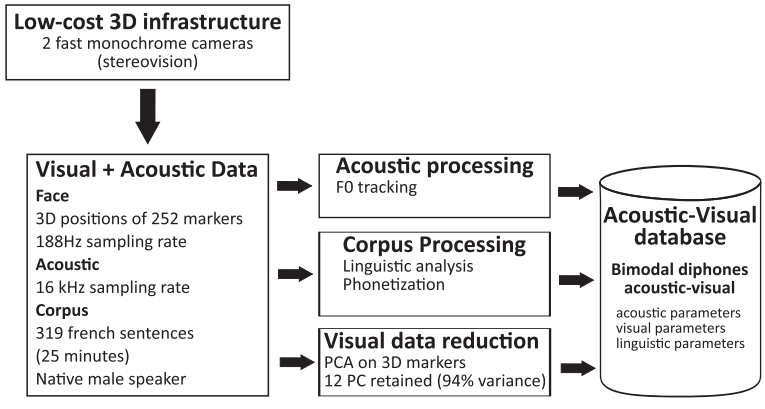

Figure 2 Data acquisition and modeling outline.

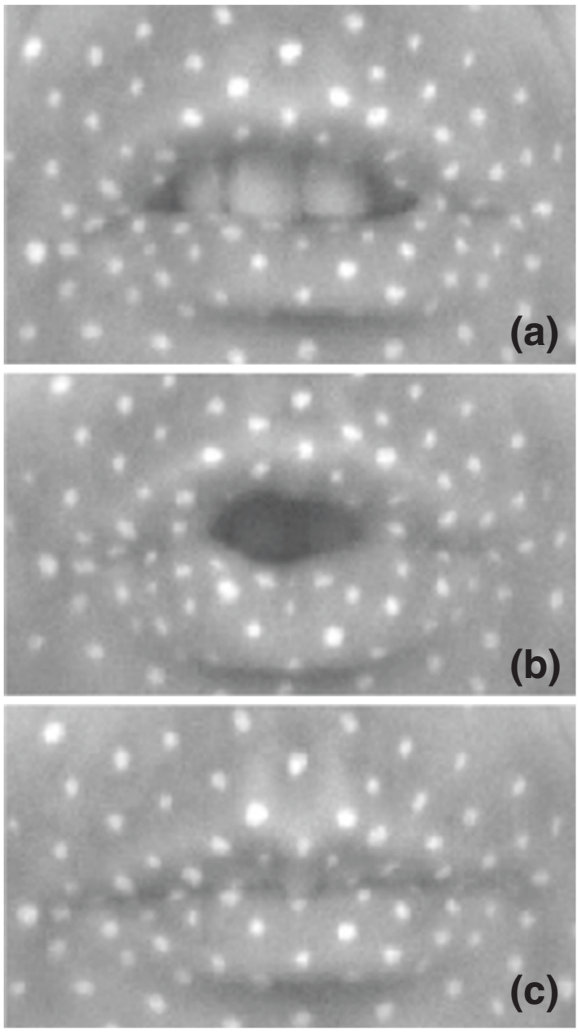

Figure 3 Positions of markers on the lips for (a) sibilant, (b) protruded, and (c) bilabial phonemes.

association of a point of the upper lip with one point of the lower lip.

This 3D mesh then evolves over time based on the temporal trajectories while keeping the same topology, and is used to fill in the gaps: missing points are recovered from the knowledge of their neighbors using a classical interpolation scheme. The topological mesh is also able to eliminate trajectories which link unlikely upper and lower lip markers. Erroneous points are easily eliminated as they do not match any vertex of the mesh.

2.1.4 Visual data acquisition

Recording the full corpus took about 4 h (markers painting, camera calibration, material setup, tests, and recording) giving rise to 25 min of effective speech. We dealt with 32 sequences and provided 3D positions of 252 markers for 585,000 frames. The corpus was made of 319 medium-sized French sentences, covering 25 min of speech, uttered by a male native French speaker. The size of this corpus is large enough compared to other works on audiovisual synthesis, but small compared to works on text-to-speech synthesis. A set of 20 extra sentences was recorded for testing and tuning purposes. The corpus did not cover all diphones possible in French due to the corpus size, but several representations of some diphones were present in different contexts. As in typical concatenative speech synthesis, the corpus was phonetized and analyzed linguistically. A database was then constructed, including acoustic, visual, and linguistic parameters for each bimodal diphone.

2.2 Modeling: principal components

We applied PCA on a subset of markers: in the lower part of the face (jaw, lips, and cheeks - see Figure 1). The movements of markers on the lower part of the face are tightly connected to the speech gestures. As this synthesis technique was designed for neutral speech (affirmative sentences) and not expressive speech, markers on the upper part of the face move very little. We retained the 12 first principal components, which explain about 94% of the variance of the lower part of the face.

These 12 components are shown in Figure 4. The first two components, which explain 79.6% of the lower face variance, both account for combined jaw opening and lip protrusion gestures. For the first component, as the jaw closes, the lips protrude. The effect is reversed for the second component: as the jaw *opens*, the lips protrude. The third component accounts for lip opening, after removal of the jaw contribution. It is in good agreement with the lip opening factor typically described in articulatory models, as in Maeda's model [25], for instance. For the less significant components, it is not entirely clear whether they correspond to secondary speech gestures, or to facial expression features. For instance, components 4 and 5 capture lip spreading; however, due to some asymmetry of our speaker's articulation, lip spreading is divided into two modes: one accounting for spreading toward the left side of the lips and one for spreading toward the right side. Component 6 is a smiling gesture; however, it is not clear whether it is related to speech or pure facial expression. Components 7 to 12 seem to account for extremely subtle lip deformations, which we believe are idiosyncratic characteristics of our speaker.

Preliminary experiments indicated that retaining as few as three components could lead to an animation which would be acceptable, in the sense that it would capture the basic speech gestures and would filter out almost all the speaker-specific gestures. However, such an animation would lack some naturalness, which is mostly captured by

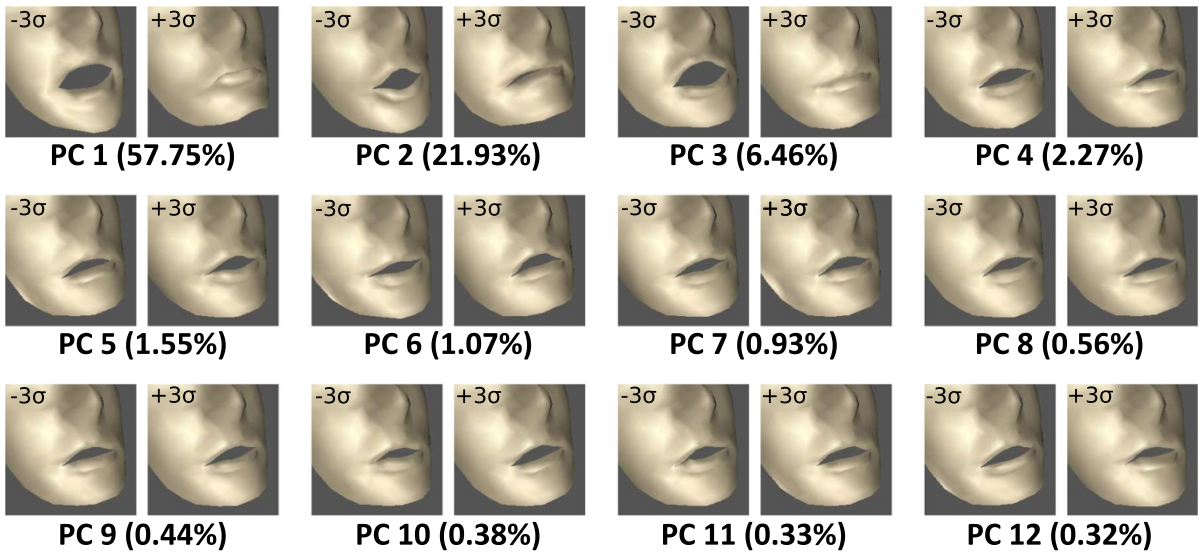

Figure 4 The 12 first principal components of the facial data and their percentage of variance. Each pair of images shows the deformation of the face when the corresponding component assumes a value of −3 (left) or +3 (right) standard deviations.

secondary components. Besides, we are in favor of keeping the specificity of the speaker gestures. Retaining the 12 components leads to animations that are natural enough for all these purposes.

One of the goals of our proposed method is to synthesize trajectories corresponding to the PCA-reduced visual information, for these 12 components, alongside the synthesized speech signal. The visual information of the lower face can be reconstructed using these 12 trajectories. The mean values of the positions of the markers at the upper part of the face may then be added to complete the facial visualization.

3 Bimodal text-to-speech synthesis

Figure 5 shows the overall bimodal synthesis process. The different steps of the text-to-speech synthesis (TTS) are similar to those in typical concatenative acoustic TTS [26]. The engine of our bimodal acoustic-visual synthesis relies on the acoustic-TTS system [27], especially, for the necessary text analysis step. In this section, we present the different steps and show how they are generalized to deal with both acoustic and visual components. First, we present the target specification and how the units are selected using a weighted sum. Then we explain the concatenation of bimodal units.

3.1 Target specification

At execution time, a text to be synthesized is first automatically phonetized and partitioned into diphones. For each diphone required for synthesis, all possible candidates from the database having the same phonetic label are looked up. A special algorithm is available to handle cases where there are no instances of the same diphone in the database. The target specification, used to search for and to rank the diphone candidates, consists of linguistic and phonetic features. It specifies the phonemes being looked up and their linguistic and phonetic content and the context which affects their realization.

It is noteworthy that there is no prosodic model. The prosody is implicitly determined from the linguistic features that cover local and global context. We estimate that the information comprised in these features should be sufficient to provide neutral or 'in reading' prosody, as

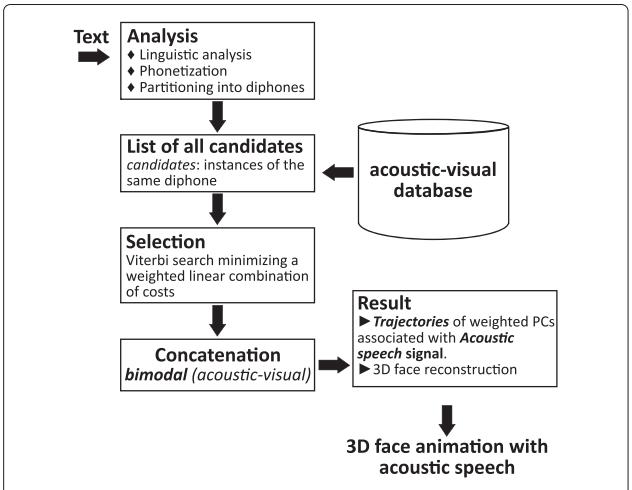

Figure 5 Bimodal text-to-speech synthesis processing.

that of the recorded corpus [28]. We extend this idea to the visual domain, where coarticulation is implicitly integrated similarly by means of linguistic analysis and the use of diphones as a canonical unit of representation.

The target specifications are composed of 50 linguistic and phonetic features (word, syllable position, rhythmic group boundaries, kind of syllable, voicing of context, etc.). The target cost of each of the phonemes is a weighted summation of the difference between the features of the candidate and those of the target. These specifications were introduced in the acoustic-TTS system [27]. To compute the weight of each feature for each phoneme separately, we developed a method based on acoustic clustering and entropy calculation (gain ratio) to determine the relevance automatically and thus the value of weight of the features, in a similar way as in our previous work on acoustic-only synthesis [28]. However, as there is no prosodic model to guide the selection, the quality might get degraded particularly for French due to a chosen unit with an unsuitable duration. For unit selection, we added duration constraints that rely on positions of the units in the sentence (before a short pause, full pause, end of rhythmic group, etc.) and the mean values calculated based on the values met in the recorded corpus. The method is advantageous, as it proposes an implicit duration model that is adapted to the speaker contrary to a generic model.

In the previous set of features, in particular, the phonetic context was reflected as binary values in the target cost. Each of the contextual phonemes was classified as belonging to an articulatory category of phonemes (protruded, spread, etc.). For instance, the phoneme /u/ belongs to the set of protruded phonemes for French. This kind of discrete classification is based on classical phonetic knowledge. We have shown that it is possible to modify the classification of a given phoneme to take into account its main characteristics as observed in audiovisual corpus well [29]. We conducted a statistical analysis on the phonetic articulatory features. The set of articulatory features included lip protrusion, lip opening, lip spreading, and jaw opening. These features were computed from the visual domain using four markers from the lips (for protrusion, opening and spreading) and one marker on the chin (for the jaw) [30]. The results showed that overall the phonetic categories were respected; nevertheless, few phonemes needed to be reconsidered and we modified their categories. For instance, for the two phonemes /ʃ/ and /ʒ/, the articulatory feature representing lip protrusion has the value 0, i.e., phoneme is not protruded. Based on the statistics calculated on the corpus of our speaker, these two phonemes are protruded, and thus their category was modified. The updated phonetic categories have been used during synthesis. Thus, a candidate with a different articulatory context from that of the target phoneme will be penalized in the target cost. In this way, the phonetic features take into account the intrinsic visual/articulatory characteristic of the speaker. We also introduced continuous visual target cost, where real values in the range [0,1] were used rather than binary values [29]. The continuous target costs were calculated based on the articulatory feature statistics.

In our work, the target cost of a diphone candidate is the summation of target costs of the two phonemes composing this diphone. The target cost of each of the phonemes is a weighted summation of the difference between the features of the candidate and those of the target. The considered features rely mainly on linguistic information that have been extended to phonetic characteristics extracted from visual domain.

3.2 Unit selection

The selection among the set of selected candidates is classically operated by resolution of the lattice of possibilities using the Viterbi algorithm. The result of the selection is the path in the lattice of candidates which minimizes a weighted linear combination of four costs, i.e.,

$$C = w_{\mathrm{tc}}\mathrm{TC} + w_{\mathrm{jc}}\mathrm{JC} + w_{\mathrm{vc}}\mathrm{VC} + w_{\mathrm{dvc}}\mathrm{DVC}, \tag{1}$$

where TC is the target cost, as already described. JC is the *acoustic join cost*, defined as the acoustic distance between the units to be concatenated and is calculated using acoustic features at the boundaries of the units to be concatenated: fundamental frequency, spectrum, energy, and duration specification. VC is the *visual join cost* calculated using the values of the PC trajectories at the boundaries of the units to be concatenated, i.e.,

$$\mathrm{VC} = \sum_{i=1}^{12} w_i \left(P_{i,1} - P_{i,2}\right)^2 \tag{2}$$

where $P_{i,1}$ and $P_{i,2}$ are the values of the projection on principal component i at the boundary between the two diphones (see Figure 6). The weights w_i should reflect the relative importance of the components, and we choose them to be proportional to the eigenvalues of PCA analysis, in accordance with [8]. Therefore, as shown in Figure 4, the weights put a lot of emphasis on the first few components. Finally, the *derivative join cost* DVC is calculated in the same manner as VC using the derivatives of the PC trajectories. Derivatives were calculated using a five-point stencil approximation.

The weights w_{tc}, w_{jc}, w_{vc}, w_{dvc} are fine-tuned using an optimization method which involves a series of simple metrics that compare a synthesized utterance to a set of test utterances. These metrics take into account the continuity of the visual trajectory and its first derivative and

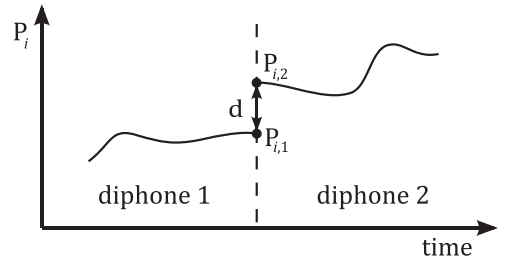

Figure 6 Illustration of the visual cost calculation. The purpose is to minimize the distance d between the points $P_{i,1}$ and $P_{i,2}$ at the boundary of the two concatenated diphones.

of the fundamental frequency (F0), and the correctness of the rhythm structure of the synthesized utterance. They are then merged into a single metric which is minimized over the set of 20 test utterances using a nonlinear optimization technique. See Toutios et al. [31] for the details of this optimization method and the description of the metrics.

3.3 Concatenation

In the acoustic domain, the concatenation of the selected diphone sequence is based on the classical TD-PSOLA-like technique [32]. We use several anchors around the boundaries to carry out the most adapted concatenation and improve the joins of diphones. Firstly, we mark the pitch on important peaks of the signal using F0 detection algorithm and dynamic programming [33]. For each voiced part, we propose two pitchmarks (on minimal and on maximal peaks). Secondly, during concatenation, we choose (by a correlation criterion) the best peak (minimal or maximal) to anchor the pitch period and avoid a dephasing between the pitch periods of the first diphone and the second one. Therefore, we perform a light smoothing around the selected pitchmarks to concatenate the diphones.

Nevertheless, as can be seen in Figure 7d,e, the visual trajectory shows some irregularities in the join boundaries. We apply an adaptive local smoothing around joins which present discontinuities. If the first (Δ) or second ($\Delta\Delta$) derivatives at a given sample of a synthesized visual trajectory lie out of the range defined by ± 3 standard deviations (measured across the whole corpus), then this sample is judged as problematic. We traverse a visual trajectory x_i and check Δ and $\Delta\Delta$ at each sample i. If one of them is out of the desired range, we replace samples x_{i-k} to x_{i+k} by their three-point averaged counterparts, using incremental values for k, until Δ and $\Delta\Delta$ at sample i are within the desired range. This technique reduces the irregularities at the boundaries based on the observed articulatory behavior of our speaker.

3.4 Synthesis examples

Figure 7 shows the trajectories of the first principal component for a synthesized utterance, in several synthesis scenarios. The first example Figure 7a shows the case where only the acoustic cost is minimized. Several discontinuities are visible that result in visible jerks during the animation of the face. On the contrary, in the visual-only Figure 7b and bimodal Figure 7c cases, the resulting visual trajectories are sufficiently continuous. The synthesized acoustic speech of the visual-only result, while still intelligible, has several problems related to duration of diphones, intonation and some audible discontinuities at boundaries between diphones. The three cases in Figure 7a,b,c are using non-optimized weights. The result using optimized weights [31] is presented in Figure 7d. When using a different set of weights, several selected diphones are different, which is reflected in both acoustic and visual channels. The adaptive visual smoothing method presented in Section 3.3 produced smoother animation Figure 7e.

Figure 7f shows a comparison of the synthesized trajectory with recorded trajectory. All the half-phones (the two half-phones of a diphone) of the synthesized sentence and the recorded sentence were resampled individually to make the number of visual samples equal. It is worth noticing that the synthesized trajectories is following the same trends as the recorded trajectory. Additional examples of reconstruction and synthesis are presented in the Additional files 1, 2, 3, 4, 5, and 6.

4 Perceptual and subjective evaluations

Evaluating an audiovisual synthesis technique is always subtle and needs careful attention to draw the correct conclusion. As in our work where we are manipulating both channels, acoustic and visual, the problem is twofold. Both audiovisual speech (animation) and acoustic speech need to be evaluated. It is probably possible to provide some conclusion on the quality of the visual synthesis based on the obtained visual trajectories shown in Figure 7, for instance. The trajectories are smooth and are similar to some test utterances. We used a cross-validation technique to evaluate the synthesis by comparing the synthesized sentences with the original ones [29]. We used root mean square error (RMSE) and correlation coefficients for the evaluation. The results showed high correlation coefficients and the RMSE was very low.

However, we consider that the main evaluation criterion should be the intelligibility and the ability of the synthesis to send an intelligible message to the human receiver.

The audiovisual speech intelligibility focuses mainly on how well both audio and visual channels are integrated and how any mismatch or asynchrony influences human perception. If the acoustic or visual channel

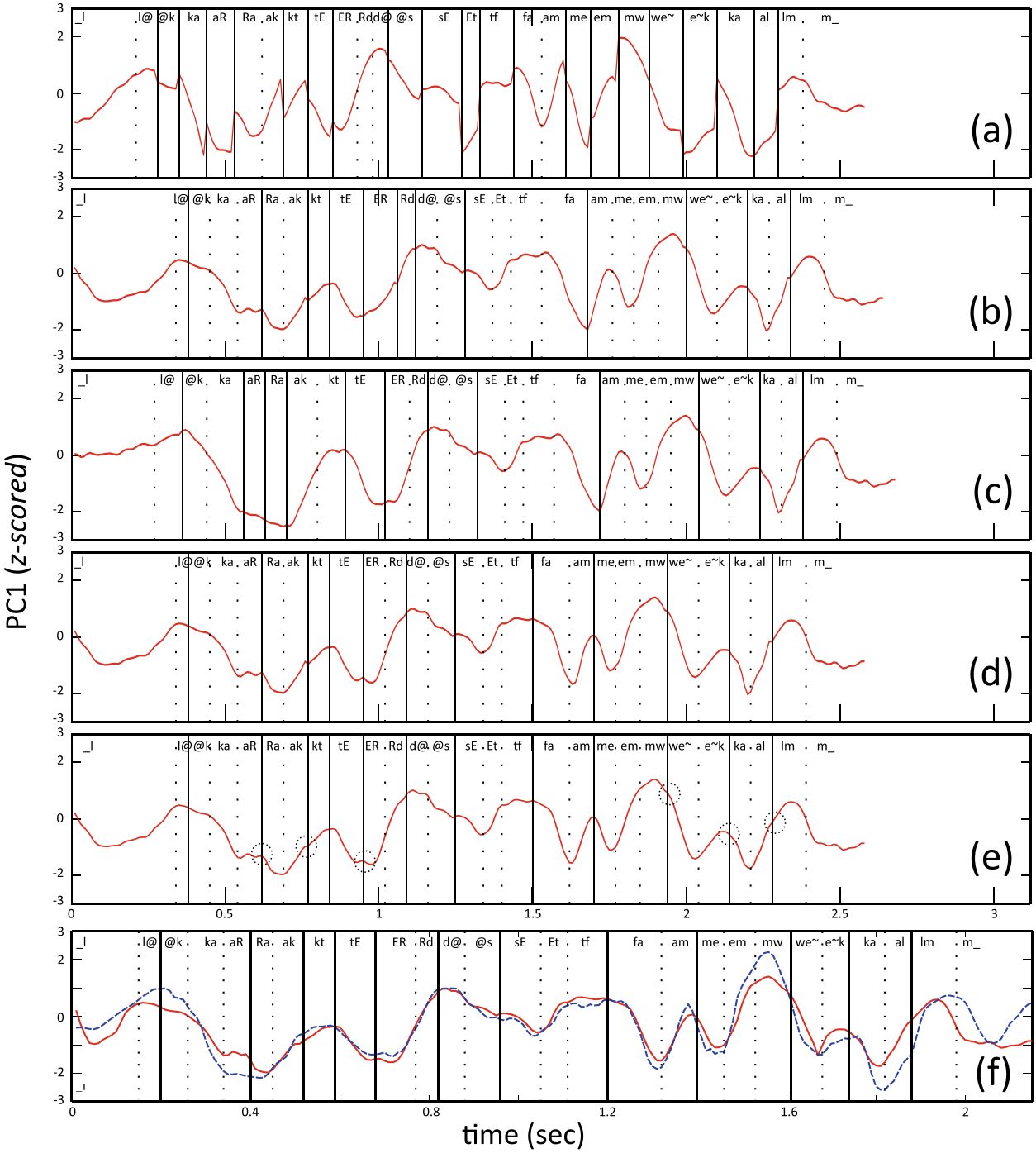

Figure 7 Visual trajectories. First visual principal component (in *z*-scored units) for the sentence 'Le caractère de cette femme est moins calme' when only acoustic join costs is minimized **(a)**, only visual cost minimized **(b)**; both acoustic and visual costs minimized using non-optimized weights **(c)**; then using optimized weights without processing at the visual joins **(d)** and when synthesized using the optimized weights, after processing visual joins **(e)**. Note the corrected details are marked with circles. **(f)** Original recorded trajectory (dashed) compared to the synthesized trajectory (solid) in (e). In (f), the duration of the diphones were adjusted to be able to make such comparison. Horizontal axes denote time in seconds. The boundaries between diphones are marked. Dashed lines indicate that the combination of the two diphones exists consecutively in the corpus and is extracted 'as is' from it, solid lines otherwise. SAMPA labels for diphones are shown.

does not have a good quality, both acoustic and visual channels together might provide an overall result with higher intelligibility compared to taking each channel separately. When dealing with acoustic speech intelligibility, the focus is not just how comprehensible speech is (the degree to which speech can be understood), but also how natural and how pleasant the acoustic speech sounds.

It is not easy to conceive a method to evaluate both channels simultaneously. For this reason, we designed a perceptual experiment to evaluate the intelligibility of synthesized visual speech, and a subjective experiment to evaluate the intelligibility of synthesized acoustic speech. Even though both experiments seem to be independent, they are implicitly highly correlated. The synthesis quality of one channel is related to the synthesis quality of the other channel due to the synthesis design. Therefore, the perceptual experiment also provides hints on how good the acoustic speech is, and the subjective experiment will also provide insights on how good the visual speech is.

4.1 Methods

We carried out two experiments: (1) a human audiovisual speech recognition experiment and (2) a subjective mean opinion score (MOS) experiment. For the first experiment, the two presentation conditions were (a) unimodal auditory speech and (b) bimodal audiovisual speech. In the unimodal auditory presentation, participants can hear only the audio of the synthesized words. In the bimodal audiovisual presentation, participants can see and hear the synthesized face pronouncing the different words.

4.1.1 Participants

Thirty-nine native French speakers, 15 females and 24 males, aged 19 to 65 (average of 30.5 years, SD = 10.97), participated in both experiments. They all reported normal hearing and normal seeing abilities, and the purpose of the experiment was not disclosed to them beforehand.

4.1.2 Test stimuli

The stimuli were either words or sentences. They were synthesized using our acoustic-visual synthesis method. The visual output is the 3D reconstruction of the face using the principal components. Figure 8c shows an example of the presented face. Black eye-glasses have been added for a more pleasant face compared the one without eyes. We made a video for each acoustic-visual synthesis result.

Perceptual evaluation For the perceptual evaluation, we used 50 everyday French words. They were either one or two syllable words. Examples of such words are the following: *anneau* (ring), *bien* (good), *chance* (luck), *pince* (clip), *laine* (wool), and *cuisine* (kitchen). In this experiment, participants were asked to watch (when the face was available), listen to the talking head pronouncing the different words, and type in what they heard. Among the 50 words, we chose 11 *in-corpus* words that were in the corpus used by the acoustic-visual synthesis. They corresponded exactly to what the speaker pronounced during the recording session. These *in-corpus* words give an insight on the intelligibility of the synthesis speech compared to that of the original speaker. Obviously, to be in the same synthesis conditions, we did not use the real speaker videos, but a 3D reconstruction of the face based on the recorded data.

For all the stimuli, the acoustic output was paired with two different white noise signals where the average values of the speech-to-noise ratio (SNR) were either 6 or 10 dB. The noise was added to the stimuli to make it difficult, to some extent, to recognize the words based

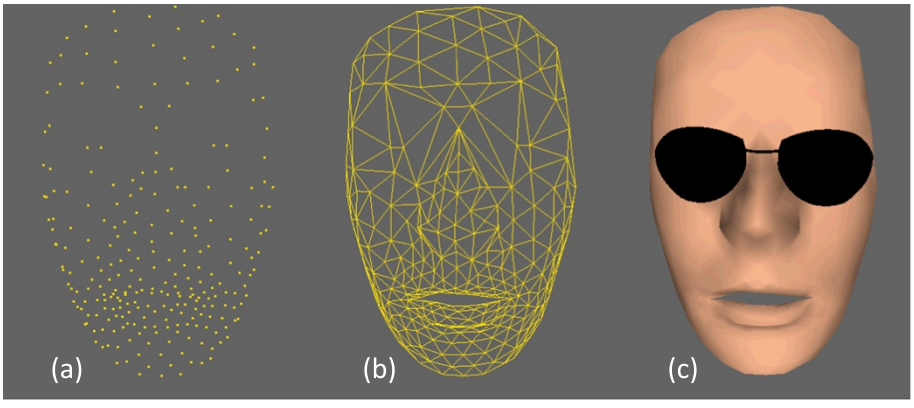

Figure 8 Rendering examples of the face. (a) the 3D vertexes, **(b)** the triangulated mesh, and **(c)** smoothed mesh: the final result. The visual output of the synthesis process is the 3D vertexes that are then rendered as a smoothed mesh with skin-like color.

on audio only. This was done with the intention to push participants to focus on the face, not only on the audio channel. The choice of these two SNR values was done after performing several testing experiments. In fact, our purpose was not to degrade the audio too much because we intended to evaluate also audio besides video. We had to find a compromise to be able to evaluate both channels.

Subjective evaluation We used the MOS test to *subjectively* measure the quality of the synthesis as perceived by participants. Twenty synthesized acoustic-visual sentences were presented (without any added noise), and participants were asked to rate each sentence by answering five questions. The rating scheme ranges from 5 (excellent) to 1 (bad). Table 1 presents the translation of the five questions and the rating scheme.

Similar to the first experiment, among the 20 sentences, we used seven *in-corpus* sentences that were in the corpus used by the acoustic-visual synthesis. These sentences corresponded exactly to what the speaker pronounced during the recording session. These *in-corpus* sentences give an insight on the intelligibility of the synthesis speech compared to the original speaker production. As explained in the first experiment, we did not use the real speaker videos, but a 3D reconstruction of the face based on the recorded data.

4.1.3 Apparatus

We designed a web application to run the experiment. The potential advantage of such an application is that it is accessible to a wider number of participants. It has been shown that web-based evaluation of talking heads provides comparable results as the experiments performed in labs [34]. Nevertheless, several technical aspects were handled carefully to control the experimental conditions as much as possible. Therefore, the application decides whether or not the experiment can be run in a given environment, based on the operating system, browser, and screen resolution, and it adapts the content if possible. It also computes the response time and removes any participant scores where there is suspicious behavior (long absence, for instance). In these experiments, participants were asked to run the experiment in a quiet environment, using a headphone. After instructions and a configuration wizard, the application presents stimuli and collects the responses. Before running the experiments, the application asked participants to configure their system by adjusting overall volume after showing a video of the talking head with noisy audio. This step could be repeated until a particular satisfactory audio volume was reached. The subjective experiment was launched first, followed by the perceptual experiment. The order of the stimuli presentation was randomized from one participant to the other.

Table 1 The five MOS questions and the rating scheme

Question	Rating
Q1 - Does the lip movement match the pronounced audio speech?	(5) Always - (1) Never
Q2 - Is this sentence an affirmation (neutral reading)?	(5) Totally agree - (1) Not at all
Q3 - Does the voice sound natural?	(5) Very natural - (1) Not natural
Q4 - Does the face-only look natural?	(5) Very natural - (1) Not natural
Q5 - Is the pronunciation of this sentence by the talking head pleasant?	(5) Very pleasant - (1) Not at all

4.2 Perceptual evaluation results

Table 2 presents the overall scores across all the 39 participants under the two noise conditions and the two presentations (unimodal audio and bimodal audiovisual). An answer was considered correct when the word was totally recognized by the participant. Across the two noise conditions, the performance of the audiovisual presentation improved compared with unimodal audio presentation, and the difference was significant [low noise level (audio $M = 0.47, \mathrm{SD} = 0.08$; audiovisual $M = 0.51, \mathrm{SD} = 0.09$), $t(76) = -2.25, p = .03$; high noise level (audio $M = 0.4, \mathrm{SD} = 0.09$; audiovisual $M = 0.46, \mathrm{SD} = 0.1$), $t(76) = -2.79, p = .007$)]. Although this was the minimum that one can expect from such a technique, this suggests that visual synthesis presents good coherence with audio regardless of the size of the corpus.

To refine the analysis, we also provide the results of *in-corpus* (data as recorded from the original speaker) and *out-of-corpus* (the result of the synthesis) sets, which are presented in Table 3. The results should be seen just as an indication on the intelligibility performance and not as a deep analysis since the number of items in *in-corpus* set is smaller than that of *out-of-corpus*. The purpose of introducing these two sets is to be able to compare the performance of the acoustic-visual synthesis compared with the face of the speaker used to record the corpus. It should be noted that in this evaluation, we are not using the video of the real face of our speaker, but a 3D reconstruction of the 252-vertex-face based on the recorded data. Thus, in our case, we replace the real face by the dynamics or the articulation of the speaker. For this reason, we are interested in comparing the synthetic face to the speaker's articulation. We continue to denote the reconstructed face from the original data as *the natural face*.

To estimate the quality of the synthetic face, we used the metric proposed by Sumby et al. [3] to quantify the

Table 2 Overall scores across all the 39 participants under each condition

	Audio		Audiovisual	
Participant	**High N**	**Low N**	**High N**	**Low N**
1	0.28	0.26	0.32	0.36
2	0.46	0.48	0.46	0.56
3	0.32	0.46	0.52	0.44
4	0.44	0.56	0.54	0.52
5	0.30	0.44	0.44	0.56
6	0.52	0.54	0.44	0.54
7	0.26	0.42	0.36	0.44
8	0.42	0.50	0.50	0.52
9	0.24	0.38	0.38	0.44
10	0.28	0.36	0.32	0.44
11	0.44	0.52	0.46	0.58
12	0.44	0.46	0.42	0.50
13	0.30	0.52	0.42	0.54
14	0.26	0.34	0.24	0.40
15	0.42	0.50	0.42	0.46
16	0.28	0.40	0.40	0.48
17	0.46	0.54	0.58	0.60
18	0.46	0.48	0.50	0.54
19	0.50	0.52	0.56	0.58
20	0.42	0.46	0.52	0.56
21	0.42	0.40	0.38	0.42
22	0.44	0.44	0.50	0.54
23	0.42	0.52	0.54	0.58
24	0.62	0.68	0.70	0.76
25	0.40	0.56	0.64	0.72
26	0.32	0.56	0.50	0.48
27	0.54	0.58	0.56	0.62
28	0.34	0.40	0.46	0.42
29	0.40	0.44	0.44	0.52
30	0.40	0.50	0.46	0.56
31	0.30	0.36	0.32	0.42
32	0.42	0.48	0.42	0.46
33	0.34	0.48	0.46	0.46
34	0.36	0.40	0.36	0.42
35	0.36	0.40	0.28	0.40
36	0.40	0.44	0.42	0.44
37	0.38	0.38	0.50	0.50
38	0.44	0.38	0.40	0.46
39	0.62	0.62	0.64	0.60
Mean	***0.40***	***0.47***	***0.46***	***0.51***
Standard deviation	***0.09***	***0.08***	***0.10***	***0.09***

Hi N, high noise; Lo N, low noise.

Table 3 Mean scores under each condition, split into two set of stimuli: out-of-corpus and in-corpus words

	Audio		Audiovisual	
	Hi N	**Lo N**	**Hi N**	**Lo N**
Out-of-corpus	0.34	0.40	0.40	0.45
In-corpus	0.59	0.69	0.65	0.72

Hi N, high noise; Lo N, low noise; out-of-corpus words, 39 words; in-corpus words, 11 words.

visual contribution to intelligibility. The metric is based on the difference between the scores of the bimodal and unimodal auditory conditions and measures the visual contribution C_v in given noise condition, which is

$$C_v = \frac{C_{AV} - C_A}{1 - C_A}, \tag{3}$$

where C_{AV} and C_A are the bimodal audiovisual and unimodal auditory intelligibility scores. This metric has been used by several researchers for evaluation purpose [4,35]. We propose to use this metric not to compare synthetic face against natural face, but, for each kind of face, we compute its visual contribution to intelligibility.

For the natural face, C_v = 0.146 in high noise level, and C_v = 0.097 in low noise level. For the synthetic face, C_v = 0.091 in high noise level, and C_v = 0.083 in low noise level. This suggests that the visual contribution to intelligibility of the synthetic face is very close to that of the natural face in the same condition. This is actually influenced by the quality of the audio.

Table 3 shows the improvement made by the synthetic face compared to that of using only the natural audio. The difference in performance between synthetic and natural audios shows that the acoustic synthesis has a scope for improvement to reach natural audio performance. In all cases, the perceptual experiment clearly shows that visual animation is not conflicting with audio, and there is no doubt of its intelligibility.

4.3 Subjective evaluation results

Table 4 shows the MOS results for each of the five questions. The first row presents the mean ratings over all the 20 sentences. The overall result shows that the audiovisual synthesis is positively perceived by participants.

Table 4 Mean MOS scores across the five questions

	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Overall	***3.88***	***3.93***	***3.04***	***2.92***	***3.02***
Out-of-corpus	3.76	3.78	2.57	2.80	2.65
In-corpus	4.80	4.91	4.56	3.67	4.32

The presented scores are overall mean scores, out-of-corpus mean scores, and in-corpus mean scores.

The rating of question Q1 shows that our technique does not introduce any mismatch or asynchrony between the audio and visual channels. The acoustic prosody seems to be acceptable (question Q2). We recall that the prosody is implicitly generated without using an explicit prosody model. Our synthesis is supposed to provide a natural prosody of an affirmative utterance. The rating of questions *Q*3 and *Q*5, related to the naturalness of the voice, is low. This can be explained by the size of the corpus where some diphones have a small number of candidates to propose during the selection step. We were expecting low rating for question Q4, as the vertexes of the face are not those of a high-resolution face, and the face has no teeth or tongue. However, it seems that having good dynamics can overcome the sparseness of the vertexes. This can also be explained by the fact that humans are tolerant when we are not very close to the uncanny valley [36].

To refine this analysis, we split the overall MOS results into two sets: (1) in-corpus and (2) out-of-corpus. Although the number of the *in-corpus* sentences is small (7 of 20 sentences), the goal is to have an idea about the performance upper bound of the natural face compared to the synthesized one. In fact, we assume that it is extremely difficult for this synthesis technique to perform better than the real speaker (unless the latter's articulation is not intelligible). Therefore, the upper limit should be seen as the performance of the real speaker, not the total score [5]. For questions Q1 and Q2, the scores are high for in-corpus sentences, but the natural talker is still not rated as 'perfect' neither. What one can say though is that, for some questions, the performance of the bimodal synthesis reached 56% to 78% of the performance of natural speaker.

5 Conclusions

We have presented a bimodal unit-selection synthesis technique that performs text-to-speech synthesis with acoustic and visible components simultaneously. It is based on the concatenation of bimodal diphones, units that consist of both acoustic and visual components. During all the steps, both components were used together. The good coverage of the lower face by an important number of markers allowed good dynamics of the lips. We should point out that no coarticulation model has been explicitly used during the whole process of the synthesis. Coarticulation has been integrated implicitly by means of linguistic analysis and the use of diphones as a canonical unit of representation.

We also presented a perceptual and subjective evaluations of the bimodal acoustic-visual synthesis. The results showed that audiovisual speech provided by this synthesis technique is intelligible and acceptable as an effective tool of communication. The use of bimodal units to synthesize audiovisual speech seems to be a very promising technique that should probably be generalized in future projects as an effective audiovisual speech synthesis technique. Regarding the acoustic synthesis quality, the bimodal speech synthesis quality is still not as good as that of the state-of-the-art acoustic synthesis systems. In fact, the latter is usually trained on 3 h or more of acoustic speech, much larger than the 25-minute corpus used in the presented work. To reach equivalent quality, bimodal corpus should obviously be at equivalent size compared to that of the corpora typically used in acoustic speech synthesis. This means that an effort should be made in improving the acquisition technique to be able to acquire larger bimodal corpus. Regarding the visual synthesis, it is worth noticing that we are not yet presenting a complete talking head as we are for now just synthesizing the face and the lips concurrently with the acoustic speech. We are currently focusing on synthesizing the dynamics of the face, to assess that it is possible in practice to provide a synthesis technique where both acoustic and visual channels are considered as one unique bimodal signal.

Additional file

Additional file 1: Video: dots-real-Griffon.mpg. Reconstruction (3d vertexes) using PCA of the utterance 'Le Griffon leva ses deux pattes pour manifester sa surprise' which is extracted from the corpus (real data). Note that the generated face is only the 3d synthesis of sparse mesh of the markers on the face.

Additional file 2: Video: mesh-real-Griffon.mpg. Reconstruction (triangulated mesh) using PCA of the utterance 'Le Griffon leva ses deux pattes pour manifester sa surprise' which is extracted from the corpus (real data). Note that the generated face is only the 3d synthesis of sparse mesh of the markers on the face.

Additional file 3: Video: face-real-Griffon.mpg. Reconstruction (smoothed mesh) using PCA of the utterance 'Le Griffon leva ses deux pattes pour manifester sa surprise' which is extracted from the corpus (real data). Note that the generated face is only the 3d synthesis of sparse mesh of the markers on the face.

Additional file 4: Video: dots-synth-MaPartition.mpg. The following sentence 'Ma partition est sous ce pupitre' is un example of bimodal acoustic-visual 3d synthesis (the utterance does not exist in the bimodal corpus). The final visual output of the synthesis process is presented as the 3d vertexes.

Additional file 5: Video: mesh-synth-MaPartition.mpg. The following sentence 'Ma partition est sous ce pupitre' is un example of bimodal acoustic-visual 3d synthesis (the utterance does not exist in the bimodal corpus). The final visual output of the synthesis process is presented as a triangulated mesh.

Additional file 6: Video: face-synth-MaPartition.mpg. The following sentence 'Ma partition est sous ce pupitre' is un example of bimodal acoustic-visual 3d synthesis (the utterance does not exist in the bimodal corpus). The final visual output of the synthesis process is presented as a smoothed mesh.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by the French National Research Agency (ANR-ViSAC project number ANR-08-JCJC-0080-01).

Author details
[1]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France. [2]INRIA, Villers-lès-Nancy, F-54600, France. [3]Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3740 McClintock Ave., Los Angeles, CA 90089, USA.

Received: 18 February 2013 Accepted: 11 June 2013
Published: 27 June 2013

References

1. J Barker, F Berthommier, in *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS '99)*. Evidence of correlation between acoustic and visual features of speech (International Phonetic Association (IPA), San Francisco, CA, 1–7 August 1999)
2. H Yehia, P Rubin, E Vatikiotis-Bateson, Quantitative association of vocal-tract and facial behavior. Speech Commun. **26**(1–2), 23–43 (1998)
3. W Sumby, I Pollack, Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. **26**, 212 (1954)
4. Le Goff, T Guiard-Marigny, M Cohen, C Benoit, in *2nd International Conference on Speech Synthesis*. Real-time analysis-synthesis and intelligibility of talking faces (ISCA/IEEE, Newark, NY, September 1994), pp. 53–56
5. S Ouni, MM Cohen, H Ishak, DW Massaro, Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP J. Audio Speech Music Process. **2007**, 3–3 (2007). http://dx.doi.org/10.1155/2007/47891
6. G Bailly, M Bérar, F Elisei, M Odisio, Audiovisual speech synthesis. Int. J. Speech Technol. **6**(4), 331–346 (2003)
7. BJ Theobald, in *Proceedings of the International Congress on Phonetic Sciences*. Audiovisual speech synthesis (International Phonetic Association (IPA), Saarbrucken, 6–10 August 2007)
8. K Liu, J Ostermann, Optimization of an image-based talking head system. EURASIP J. Audio Speech Music Process. **2009**, 174192 (2009). doi:10.1155/2009/174192
9. JD Edge, A Hilton, P Jackson, Model-based synthesis of visual speech movements from 3D video. EURASIP J. Audio Speech Music Process. **2009**, 597267 (2009). doi:10.1155/2009/597267
10. NF Dixon, L Spitz, The detection of audiovisual desynchrony. Perception. **9**, 719–721 (1980)
11. KP Green, PK Kuhl, The role of visual information in the processing of place and manner features in speech perception. Percept. Psychophys. **45**, 34–42 (1989)
12. KP Green, PK Kuhl, Integral processing of visual place and auditory voicing information during phonetic perception. J. Exp. Psychol.: Hum. Percept. Perform. **17**, 278–288 (1991)
13. J Jiang, A Alwan, PA Keating, ET Auer, LE Bernstein, On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EURASIP J. Appl. Signal Process. **11**, 1174–1188 (2002)
14. J Jiang, LE Bernstein, T Edward, J Auer, in *Proceedings of AVSP*. Realistic face animation from sparse stereo meshes (AVISA, British Columbia, 24–27 July 2005)
15. H McGurk, J MacDonald, Hearing lips and seeing voices. Nature. **264**, 746–748 (1976)
16. W Mattheyses, L Latacz, W Verhelst, On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EURASIP J. Audio Speech Music Process. **2009**, 169819 (2009). doi:10.1155/2009/169819
17. A Hunt, A Black, in *Proceedings of ICASSP*. Unit selection in a concatenative speech synthesis system using a large speech database (IEEE, Atlanta, 7–10 May 1996)
18. P Taylor, in *Text-to-Speech Synthesis*. (Cambridge Univ. Press, Cambridge, 2009)
19. A Hallgren, B Lyberg, in *Proceedings of the AVSP*. Visual speech synthesis with concatenative speech (AVISA, Terrigal-Sydney, 4–6 December 1998)
20. M Tamura, S Kondo, T Masuko, T Kobayashi, in *Proceedings of the Eurospeech Conference*. Text-to-audio-visual speech synthesis based on parameter generation from HMM (Budapest, 5–9 September 1999), pp. 959–962
21. S Minnis, A Breen, in *Proceedings of the Interspeech 2000*. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis (ISCA, Beijing, 16–20 October 2000)
22. S Fagel, in *Proceedings of the International Conference on Speech and Computer*. Joint audio-visual units selection - the JAVUS speech synthesizer (SPIIRAS, St. Petersburg, June 2006)
23. A Toutios, U Musti, S Ouni, V Colotte, B Wrobel-Dautcourt, MO Berger, in *Interspeech 2010*. Setup for acoustic-visual speech synthesis by concatenating bimodal units (Visac Publications, Makuhari, Japan, 2010)
24. B Wrobel-Dautcourt, M Berger, B Potard, Y Laprie, S Ouni, in *Proceedings of the AVSP*. A low-cost stereovision based system for acquisition of visible articulatory data (AVISA, British Columbia, 2005)
25. S Maeda, in *Speech Production and Speech Modelling*, ed. by WJ Hardcastle, A Marchal. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model (Kluwer Academic, Dordrecht, 1990), pp. 131–149
26. R Clark, K Richmond, S King, Multisyn: Open-domain unit selection for the festival speech synthesis system. Speech Commun. **49**(4), 317–330 (2007)
27. V Colotte, A Lafosse, Soja: french text-to-speech synthesis system. http://soja-tts.loria.fr/. Accessed 21 June 2013
28. V Colotte, R Beaufort, in *Interspeech Proceedings*. Linguistic features weighting for a text-to-speech system without prosody model (ISCA, Lisbon, 4–8 September 2005)
29. U Musti, V Colotte, A Toutios, S Ouni, in *International Conference on Auditory-Visual Speech Processing - AVSP2011*. Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer (Volterra, 31 August to 3 September 2011)
30. V Robert, B Wrobel-Dautcourt, Y Laprie, A Bonneau, in *5th Conference on Auditory-Visual Speech Processing - AVSP 2005*. Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for French (AVISA, Vancouver Island, 24–27 July 2005)
31. A Toutios, U Musti, S Ouni, V Colotte, in *12th Annual Conference of the International Speech Communication Association - Interspeech 2011*, ed. by ISCA. Weight optimization for bimodal unit-selection talking head synthesis (ISCA, Florence, 27–31 August 2011)
32. E Moulines, F Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. **9**(5–6), 453–467 (1990)
33. V Colotte, Y Laprie, in *XI European Signal Processing Conference - EUSIPCO 2002 (2002)*. Higher precision pitch marking for TD-PSOLA (EURASIP, Toulouse, 3–6 September 2002)
34. B Weiss, C Kühnel, I Wechsung, S Möller, S Fagel, in *IVA '09 Proceedings of the 9th International Conference on Intelligent Virtual Agents*. Web-Based Evaluation of talking heads: how valid is it? (Springer, Amsterdam, 14–16 September 2009), pp. 552–553
35. S Ouni, MM Cohen, DW Massaro, Training Baldi to be multilingual: a case study for an Arabic Badr. Speech Commun. **45**, 115–137 (2005)
36. M Mori, The uncanny valley. Energy. **7**, 33–35 (1970)

doi:10.1186/1687-4722-2013-16
Cite this article as: Ouni *et al.*: **Acoustic-visual synthesis technique using bimodal unit-selection.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:16.