A Generative Audio-Visual Prosodic Model for Virtual Actors

Adela Barbulescu and Rémi Ronfard ■ *Inria*

Gérard Bailly ■ *GIPSA-lab*

An important problem in the animation of virtual characters is the expression of complex mental states using the coordinated prosody of voice, rhythm, facial expressions, and head and gaze motion. The authors propose a method for generating natural speech and facial animation in various attitudes using neutral speech and animation as input.

Natural animation of conversational agents requires a large vocabulary of emotions and attitudes, which are difficult to reproduce because human perception of audio-visual cues is sensitive to matching subtle facial motions with speech content and expressive style. The automation of expressive speech animation is the focus of much research at the intersection of computer graphics, speech processing, and computer vision. A large part of research now tackles this problem using data-driven or machine-learning approaches. Speech signals and visual motions are therefore generated using multimodal behavioral models built from labeled speech and motion capture data.

In the speech community, speaking style is referred to as *prosody*, which represents the set of features that contribute to linguistic functions such as intonation, tone, stress, and rhythm. However, prosody also reflects various features of the speaker (such as physiological or psychological state) or the utterance (such as emphasis, contrast, focus, and the presence of irony or sarcasm). The concept of *audio-visual prosody* refers to the use of multimodal cues for signaling and perceiving linguistic, paralinguistic, and nonlinguistic functions in social communication.

Klaus Scherer and Heiner Ellgring proposed that the affective function in social communication reflects two types of effects: push and pull.[1] The *push effect* underlies psychobiological mechanisms—for example, the arousal that leads to the rise in the fundamental frequency caused by muscle tension. The *pull effect*, on the other hand, is triggered by conventions, norms, and cultural behaviors that pull the voice in certain directions. Such effects include accepted sociocultural speaking styles and vocal and facial display rules. Therefore, affective expression in speech communication happens either involuntarily (expression of emotion) or voluntarily (expression of attitude).

Social attitudes (such as comforting, doubtful, or ironic) are highly conventionalized—that is, entirely part of the language and the speech communication system—and socioculturally built. They trigger specific behaviors of intonation and facial expressions, as Dwight Bolinger notably stated: "Intonation [is] a nonarbitrary, sound-symbolic system with intimate ties to facial expression and bodily gesture, and conveying, underneath it all, emotions and attitudes... [It is] primarily a symptom of how we feel about what we say [attitude], or how we feel when we say it [emotion]."[2]

We are interested in the pull effect and exploring the characteristics of controllable behaviors and the way it triggers speaker-specific prosodic signatures—that is, attitude-specific patterns of audio-visual prosodic trajectories. Our previous work showed that sentence-level audio-visual features lead to higher attitude recognition rates, thus supporting the existence of attitude-specific multimodal contours at the sentence level.[3]

In this article, we propose an end-to-end system for learning generative prosodic models of attitudes from paired examples of neutral and expressive sentences performed by semiprofessional

actors. We apply our generative prosodic model to the task of generating expressive speech animations such that the resulting audio-visual performance preserves the speaker's individuality while encoding the chosen attitudes. Our system is built within an audio-visual conversion paradigm in which we propose the conversion between neutral and expressive performances. The system requires the following input: a neutral version of the audio-visual speech and the label of the desired attitude, which we refer to as *didascalia*, in the context of a dramatic work.

Related Work in Expressive Facial Animation

Previous work in expressive facial animation can be divided into text-driven animation, speech-driven animation, and expressive conversion.

Speech-Driven Facial Animation

The first class of methods implies the use of a prior speech performance to drive realistic face motion. One of the first works to address the problem of generating expressive speech animations was carried out by Erika Chuang and Christoph Bregler, who proposed a bilinear model for facial expressions spanning three emotional styles (happy, angry, and neutral) with the goal of editing existing facial motion.[4]

Stacy Marsella and his colleagues used a hybrid system that combines a speech-driven model-based technique with a rule-based visual text-to-speech (TTS) synthesizer to generate expressive performances for a 3D virtual character using prosody and sentence semantics.[5] The system is able to generate head motion, eye saccades, eye blinks, gazes, and gestures using a complex set of rules derived through a study of video corpora of human behaviors.

These methods impose an implicit constraint because they all require an expressive speech input. Moreover, some co-verbal facial motion during speech cannot be determined from the sole acoustics (such as gaze). Note also that the criteria that are minimized should both ensure that the perceptually significant variance is captured and cope with the highly nonlinear correspondence between acoustic events and smooth facial motions. These dual requirements are difficult to meet, even via latent variables or hidden states.

Text-Driven Facial Animation

The second class of methods is represented by a multimodal TTS system approach, which allows the joint generation of audio-visual speech using only text as input. Prosodic functions such as emotional content are also provided using additional information (prosodic labels, expressive weights, and so on) or via semantics extracted from text.

Irene Albrecht and her colleagues used a rule-based approach to generate facial expressions, head motion, and voluntary blinks from text input.[6] They use a TTS system to obtain the speech and phoneme durations and a lip-sync algorithm to generate speech-related facial expressions. The nonverbal facial expressions and head motion are generated with the use of six types of emoticons: happy, sad, surprised, kidding, angry, and disgusted. Researchers also proposed a similar rule-based method that allows the application of expressive weights for full or parts of sentences.[7]

Expressive Conversion

Our work is closest to the third class of methods, which learn a mapping function between emotion spaces, notably between neutral speech and the desired expressive style. Gaussian mixture models (GMMs) are widely used in voice conversion to modify nonlinguistic information, such as spectrum, while keeping linguistic information unchanged.

Statistical approaches have also been applied to expressive motion capture data in an attempt to jointly synthesize speech and facial expressions.[8] These approaches factorize expressive speech into separate components so that parameterized neutral speech sequences can be modulated with expression parameters.

Rhythm is an important prosodic cue for expressive speech, which has to be considered for both audio and visual. In the case of speech-driven techniques, rhythm is provided by the speech, whereas in the case of expressive conversion, most techniques use the rhythm of the neutral performance. In text-driven synthesis, rhythm is predicted, in most cases by a TTS engine that is generally trained on read speech. In our approach, rhythm is explicitly part of a generative prosody model that is learned from examples of the target attitudes.

Corpus of Dramatic Attitudes

With the goal of synthesizing audio-visual speech for realistic social contexts, we designed and recorded an acted corpus of "pure" social attitudes—that is, isolated sentences carrying only one attitude over the entire utterance.

We selected a subset of 10 attitudes from Simon Baron-Cohen's Mind Reading Project.[9] The source taxonomy proposed by Baron-Cohen consists of

Table 1. Definitions of the 10 attitudes used in our dataset.*

Label	Abbreviation	Definition
Comforting	CF	Making people feel less worried, unhappy, or insecure.
Tender	TE	Finding something or someone appealing and pleasant; being fond of something or someone.
Seductive	SE	Physically attractive.
Fascinated	FA	Very curious about and interested in something that you find attractive or impressive.
Thinking	TH	Thinking deeply or seriously about something.
Doubtful	DO	Unwilling or unable to believe something.
Ironic	IR	Using words to convey a meaning that is the opposite of their literal meaning.
Scandalized	SC	Shocked or offended by someone else's improper behavior.
Confronted	CO	Approached in a critical or threatening way.
Embarrassed	EM	Worried about what other people will think of you.

**The labels, abbreviations, and definitions are a subset of the attitudes from Simon Baron-Cohen's Mind Reading Project.*[2]

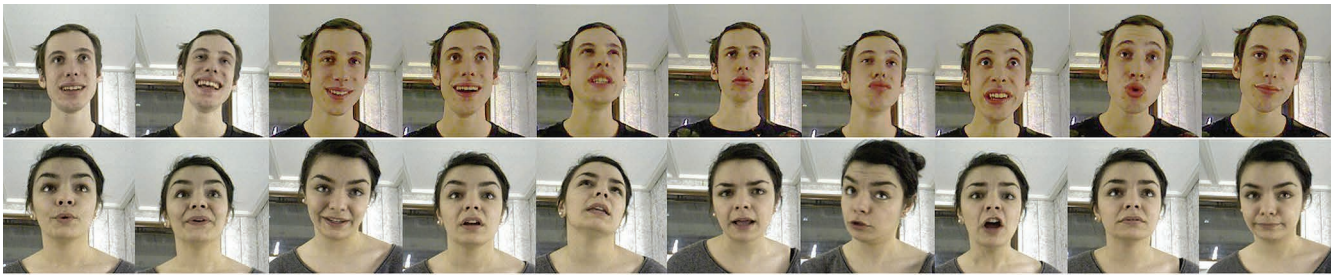


Figure 1. Examples of the 10 attitudes used in our dataset. We recorded two actors interpreting typical facial displays for each attitude.

412 attitudes grouped under 24 main categories, each comprising several layers of subexpressions. Table 1 contains the list of attitudes we analyzed for this study. Figure 1 illustrates typical facial displays of these attitudes.

We extracted the sentences for our database from a French translation of the play "Round Dance" by Arthur Schnitzler. We divided the text into a training set represented by 35 sentences (with a distribution of the number of syllables spanning between 1 and 21) and a testing set represented by a dialogue between a male character with 41 sentences (spanning between 1 and 18 syllables) and a female character with 21 sentences (spanning between 1 and 18 syllables).

The synchronized recording of voice signals and motion was done using the Faceshift software (faceshift.com) with a short-range 3D camera (Primesense Carmine 1.09) and a Lavalier microphone. Faceshift let us create a customized user profile consisting of a 3D face mesh and an expression model characterized by a set of predefined blendshapes that correspond to facial expressions (smile, eye blink, jaw open, and so on). Faceshift also outputs estimations of the head motion and gaze direction. The sentences were recorded by two semiprofessional actors under the guidance of a theater director.

The two actors recorded 35 sentences uttered first in a neutral, flat style and then with each of the selected 10 attitudes. This technique, called "exercises in style," is inspired by Raymond Queneau,[10] who used this method to rewrite the same story in 99 different styles.

The recording session began with an intensive training of the actors, which involved them fully understanding the interpreted attitudes and developing the ability to dissociate the affective state from the sentence meaning. The actors were also instructed to maintain a constant voice modulation, specific for each attitude, while uttering the entire set of 35 sentences. The actors performed as if they were addressing a person standing in front of them. They did not receive any instruction related to co-verbal behaviors.

While recording the dialogue, the actors sat in front of each other across a table. The dialogue sentences were first recorded in a neutral version and then in an expressive version, with a sentence-level didascalia succession that was chosen by the theater director.

All utterances were automatically aligned with their phonetic transcription obtained by an automatic TTS phonetizer. The linguistic analysis (part-of-speech tagging and syllabation), the phonetic annotation, and the automatic estimation of

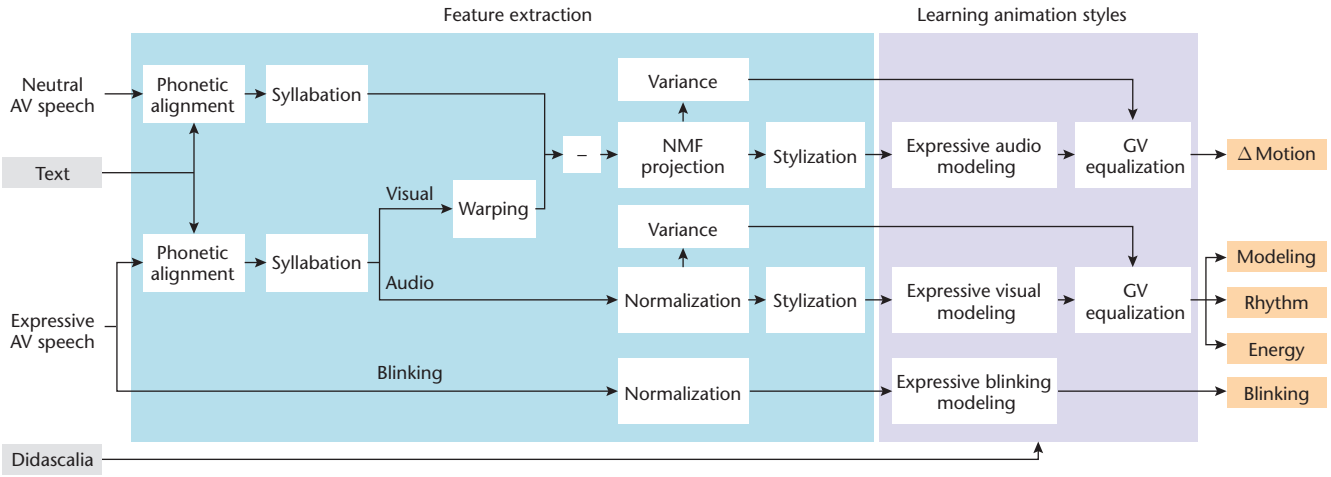


Figure 2. Learning audio-visual speaking styles. We extract audio and visual prosodic features from the training example and learn SFC models and GV equalization parameters for all dramatic attitudes, resulting in a database of audio-visual prosodic contours, including melody, rhythm, and differential motion.

melody were further corrected by hand using the Praat speech-analysis software (www.fon.hum.uva.nl/praat/). The manual verification of melodic contours required extensive effort because of the large amount of data (3 hours of speech).

Learning Audio-Visual Speaking Styles

Our method extends the superposition of functional contours (SFC) model to facial animation.[11] SFC is a comprehensive model of intonation that proposes a method of generating prosodic contours based on the direct link between phonetic forms and prosodic functions (such as attitude, emphasis, segmentation, and dependency relation) acting at different scales (such as utterance, phrase, word, syllable, and phone). SFC proposes that prosodic contours exhibit prototypical shapes that only depend on the size of the carrier linguistic unit (its number of syllables), regardless of the position of the unit within the utterance. Similarly, we hypothesize the existence of visual prosodic signatures as manifestations of the attitudinal functions.

There has been a great deal of work on the analysis and modeling of features that are found to help in the discrimination between expressive styles. Our choice of features was motivated by our earlier research.[3] Along with voice pitch (melody), energy, and syllable duration, we included gestural data: head and eye movements and facial expressions. For the visual component of prosody, we only modeled the difference between the expressive and neutral contours for each attitude once the stimuli have been properly aligned at the phone level. Figure 2 gives an overview of the training process.

Acoustic Prosody Features

After estimating and correcting the voice pitch contours (f_0), we further normalized and converted them to tones using the following equation:

$$f_0[\text{tone}] = \frac{240}{\log 2} \log \frac{f_0[\text{Hz}]}{f_0 \text{ref}}, \quad (1)$$

where f_0 ref represents the speaker's register. The resulting f_0 contours are comparable across speakers.

For rhythm, we use a duration model,[11] where the syllable lengthening/shortening is characterized with a unique z-score model applied to log durations of all constitutive segments of the syllable. We compute a coefficient of lengthening/shortening C equal to the deviation of the syllable duration relative to an expected duration Δ':

$$\Delta' = (1-r)\sum_i \overline{d_{p_i}} + r \cdot D, \quad (2)$$

where i is the phoneme index within the syllable, $\overline{d_{p_i}}$ is the average duration of phoneme i, D is the average duration for a syllable (D = 190 ms), and r is a weighting factor, fixed to r = 0.6. For a syllable with a measured duration Δ, the coefficient is

$$C = \frac{\Delta - \Delta'}{\Delta'}. \quad (3)$$

Note that the coefficient C is computed for every syllable in all the sentences in the corpus and that it incorporates the contingent pause if any as an additional lengthening factor.

We extract the energy values over an audio segment using this equation:

$$\text{energy}[\text{dB}] = 10 * \log 10 \left(\frac{\sum_{i=1}^{n} y_i^2}{|y|} \right), \quad (4)$$

where y is the acoustic signal segment with the length $|y|$.

Visual Prosody Features

We consider that the visual data recorded during speech consists of two main components: verbal and nonverbal motion. *Verbal motion*, such as the mouth opening for pronunciation, depends on the underlying phoneme pronounced at a certain position within the utterance. On the other hand, *nonverbal motion* refers to gestures accompanying speech, where the effects are spread out all over the utterance and have meaning in relation to each other (visual prosody).

In this work, we hypothesize that the verbal and nonverbal components of motion combine linearly. Precisely, we consider that an expressive performance is obtained by adding expressive visual prosodic contours to the trajectories of a "neutral" performance. Because head motion, gaze, and blendshapes have a linear representation, we obtain the visual prosody by simply aligning an expressive performance with its neutral counterpart and computing the difference of the motion trajectory values. Therefore, we view the visual prosodic model as a differential model. A first processing step realigns the expressive contours with the neutral ones. The resulting contour is a stretched version of the neutral contour such that each phoneme duration will match that of the target phoneme duration. Next, the visual prosodic contours are obtained by computing the difference between expressive contours and the aligned neutral contours.

Facial expressions are further processed by splitting them into two main groups: upper face (eyebrow motion, blinking, squinting, and so on) and lower face (smiling, mouth opening, lips protrusion, and so on). We apply nonnegative matrix factorization (NMF) to the two groups separately. This reduces the dimensionality of the extracted features (19 blendshapes for the upper face and 29 blendshapes for the lower face) while preserving significant perceptual changes observed in the reconstructed animations. We keep eight components per group with a reconstruction error of less than 5 percent.

Virtual Syllables

We add virtual syllables to account for the prephonatory and postphonatory movements for all visual components. As previously observed,[12] nonaudible preparatory movements are discriminant to a certain degree for specific emotion categories. We therefore introduce two virtual syllables with a duration of 250 ms (approximately the average syllable duration) preceding and following each utterance.

Stylization

By stylization, we mean the discretization of the prosodic continuum with the main purpose of simplifying the analysis process while maintaining the original contour characteristics.

We propose the following stylization methods for the audio-visual prosodic features:

- *f0*: The log-pitch contour is stylized by extracting three values at 20, 50, and 80 percent of the vocalic part of each syllable, where we know with certainty that voice pitch is defined. The sampling is performed after a polynomial interpolation of the voiced part of the syllable.
- *Motion*: All NMF components obtained are also stylized by extracting contour values at 20, 50, and 80 percent of the length of each syllable. Note that the stylization of motion is also done for the virtual syllables.
- *Rhythm*: Rhythm is stylized by retaining one parameter per syllable, the lengthening/shortening coefficient.
- *Energy*: Energy is also stylized with one parameter per syllable, computed over the speech segment corresponding to the vocalic nucleus.

Expressive Modeling

We learn attitude-specific prosodic contours from our training corpus, given only the set of prosodic functions (attitude) and their scopes (sentence). Therefore, to make a prediction for a given attitude, the only input required is the position and number of syllables of the desired units (phonotactic information).

We choose neural networks for carrying this type of nonlinear mapping between phonotactic information and the stylized contour values. Another important reason is that the model should also be able to extrapolate in the case of new phonotactic information—that is, when we want to generate contours for an utterance with a number of syllables different from the ones seen in the training set. Expressive modeling is carried out separately for each feature (melody, rhythm, energy, and motion) by training a feed-forward neural network with a hidden layer of 17 neurons and a logistic activation function, using the SNNS library (www.ra.cs.uni-tuebingen.de/SNNS/). The neural network's structure was chosen as a result of continuous testing, where the testing error and the generated audio-visual output were inspected.

As input, the neural networks receive a set of linear ramps that give the absolute position (which count the distance toward the beginning and end of the sentence) and relative position (which describe the position of the syllable relative to the end of the sentence) of the current syllable. The output is represented by the prosodic characteristics (stylized contour) for the current syllable. For this reason, we use the term *contour generator* to denominate a neural network trained for a specific attitude and actor. Figure 3 illustrates a contour generator with inputs (ramps) and outputs (stylized contours).

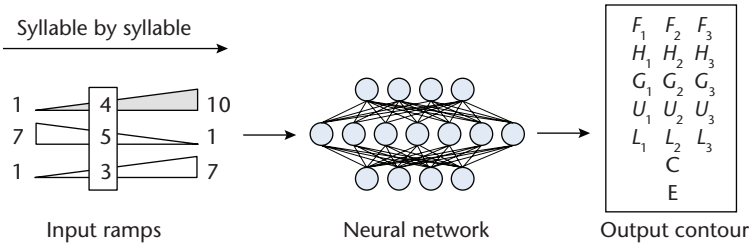


Figure 3. Training of a contour generator for a sentence with seven syllables. The prediction is done syllable by syllable. Here, syllable 3 is being processed with input ramp values (3, 5, 4) and output contour values F (voice pitch), H (head motion), G (gaze motion), U (upper-face expressions), L (lower-face expressions), C (rhythm), and E (energy).

Global Variance Equalization

One downside of using statistical learning of prosodic contours is over-smoothing, which can be alleviated by performing global variance (GV) correction. First, we model the utterance-level variance for all features: voice pitch, rhythm, and differential motion. The utterance-level variance is modeled as a Gaussian distribution. Then, we perform variance scaling at the utterance level, as in earlier work.[13]

Figure 4 presents the stylized contours for predicted and original stimuli and for different audio and visual parameters, before and after global variance correction.

Blinking Modeling

Blinking appears either as a voluntary motion, to express a specific mental state (such as the long duration blinks performed in conjunction with an ironic attitude), or as an involuntary motion, due to physiological mechanisms. Because of its irregular behavior, we analyzed the blinking rate separately. To obtain natural blinking behavior, we learned attitude-specific Gaussian models of blinking rates directly from our training set.

Generating Audio-Visual Speaking Styles

Next, we transformed a neutral speech animation of any given sentence (not in the training set) into an expressive speech animation with any given style (dramatic attitude) using the audio-visual prosodic models just described. Figure 5 presents the overall process of generating an expressive speech animation. We now review each step in more detail.

Duration Reconstruction

Our generative prosodic model for rhythm provides one stretching factor C per syllable. During

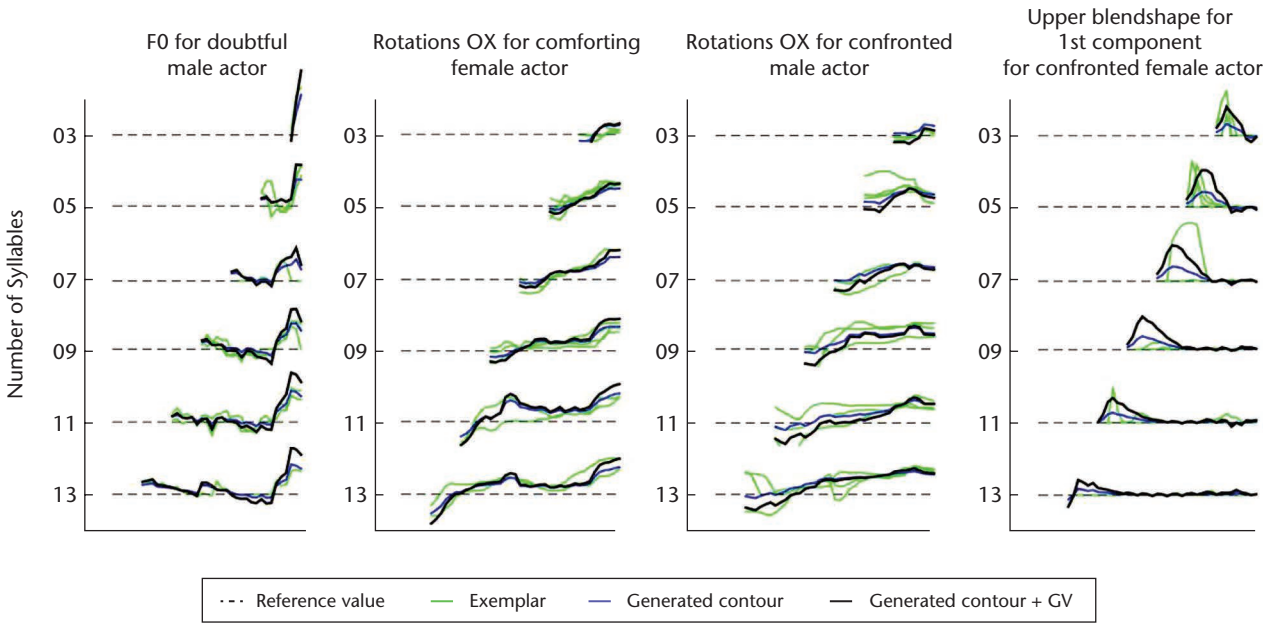


Figure 4. Predicted and recorded stylized contour examples for sentences containing 3, 5, 7, 9, 11, and 13 syllables. The features present specific behaviors for a given attitude, and these behaviors are generalized in the predicted contours.

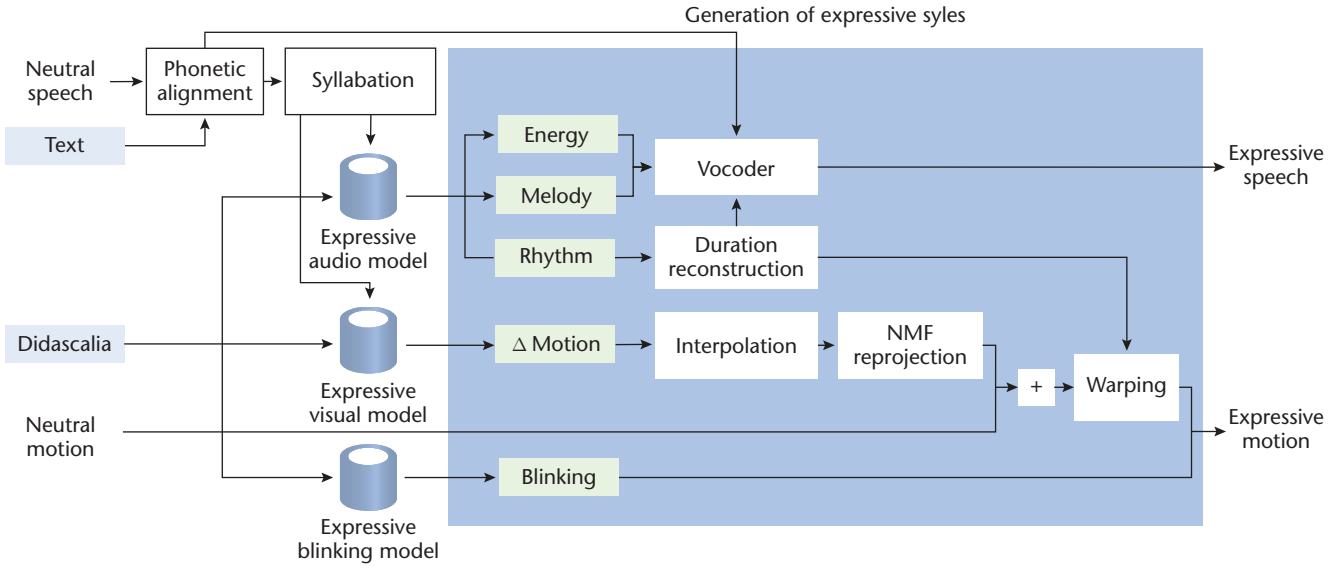


Figure 5. Generating audio-visual speaking styles. Given a neutral sentence, we use the phonotactic information to predict prosodic feature contours. The predicted rhythm is used to compute phoneme durations. The expressive speech is synthesized with a vocoder that uses the neutral utterance, predicted rhythm, energy, and voice pitch, and the facial animation parameters are obtained by adding the warped neutral motion to the reconstructed and warped predicted motion contours.

synthesis, the value of C is used to compute the duration of individual phonemes within the syllable as well as the duration of optional pauses between syllables.

We use the method proposed in earlier work,[11] where all phonemes within one syllable are compressed or elongated in the same fashion. Pauses appear as an emergent phenomenon as a result of excessive lengthening of the syllables.

Motion Reconstruction

The reconstruction of motion is carried out in a succession of steps in inverse order to the steps used in visual feature extraction. That is, the predicted differential motion is first interpolated (reconstructed from stylization by placing them at 20, 50, and 80 percent of the neutral syllable durations and performing cubic spline interpolation) and then added to the neutral motion.

The resulting motion is warped at the phoneme level using the predicted phoneme durations. In the case of facial expressions, we add one step before the interpolation process of the differential motion: NMF reprojection.

Blinking

Blinking rate is predicted by iteratively sampling the Gaussian blinking distribution starting from the beginning of the sentence. The movements of the eyelids are further generated by inserting a prototypical "blink movement" in the respective blendshape contours.

Laura Trutoiu and her colleagues showed a pronounced asymmetry in the time-space domain and presented a fast and full eyelid closing motion, followed by slower eye-lid opening.[14] We propose a generic blink movement with a full closing eyelid and duration equal to the average duration of blinking: 600 ms.

Speech Conversion

Expressive speech is synthesized using the TD-PSOLA technique,[15] which manipulates the prosody of the neutral speech by overlapping and adding speech segments. The approach requires the neutral speech and the new prosodic contours predicted by SFC as input: stylized voice pitch and phoneme durations.

We also modulate the energy of the transformed speech signal according to this equation:

$$y_{\text{enr}} = y * 10^{\frac{\text{enr}_{\text{pred}} - \text{enr}_n}{20}} \quad (5)$$

where y is the acoustic signal synthesized using TD-PSOLA, enr_{pred} is the predicted energy contour, and enr_n is the neutral energy contour measured in decibels.

Experiment

Our corpus contains a testing dataset represented by a dialogue exchange between the two actors in which the didascalia was set by a theater director. We tested our method on reconstructing dramatic dialogue by comparing our results with the expressive version of the recorded dialogue. As input, we used the neutral dialogue and the

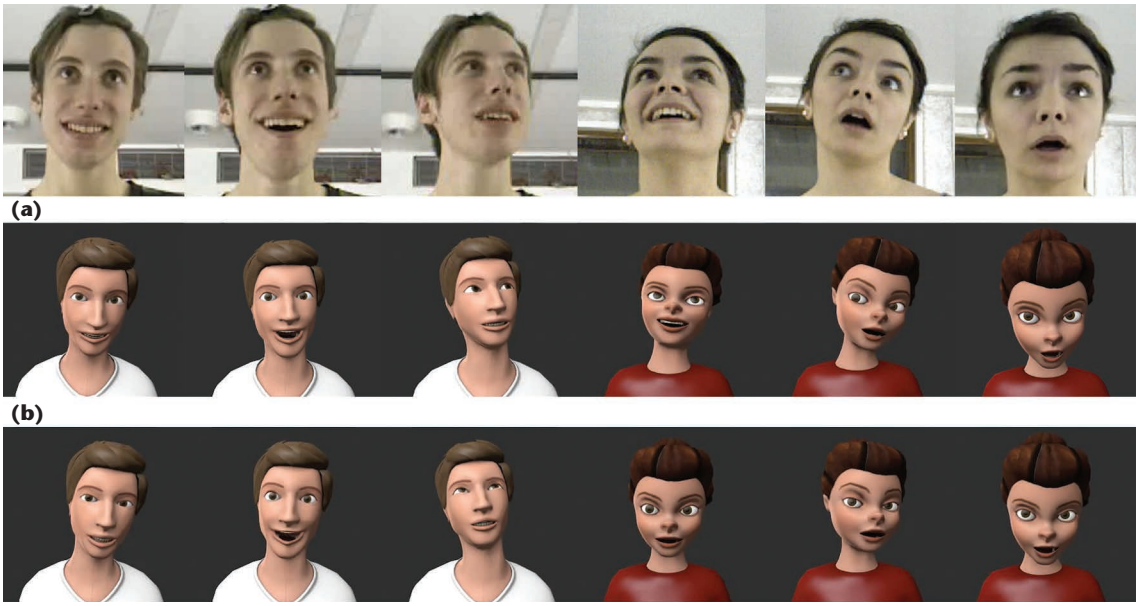


Figure 6. Examples of frames from the three conditions used in the test. The rows present corresponding frames extracted from (a) the video, (b) ground-truth animation, and (c) synthetic animation. From left to right, the images correspond to comforting, fascinated, thinking (male actor), fascinated, ironic, and scandalized (female actor) attitudes.

set of didascalia associated to each utterance as in the expressive version.

An actor may deliver different variations of a sentence for a given attitude, which renders an objective evaluation difficult. Because the aim of this work is to generate expressive animations, we focus on the perceptual evaluation of our results.

Perceptual Test

To assess the perceived expressiveness of our results, we performed an online perceptual test, where subjects were asked to recognize the attitudes performed by the actors in a set of videos and animations. The stimuli used in the perceptual test comprise the original video, the ground-truth animation (from motion capture), and the synthetic animation (obtained by our method). For the animations, we use cartoon-style avatars presenting the full 48 blendshape model. This choice was based on a series of perceptual tests in which users found the cartoon-style appearance friendlier and less disturbing than that of realistic avatars.

We represented the stimuli with short dialogue exchanges in which only one actor was shown on-screen. The visible actor is shown listening to the first utterance delivered by the off-screen actor and then delivering his/her utterance in an expressive manner. We use this type of stimulus in order to exploit the social interactive context provided by the semantics of the extracted dialogue.

We were able to extract from the expressive dialogue a total of 26 short exchanges. In these exchanges, each actor expresses six possible attitudes, with a varied number of examples per attitude. For the male actor, we obtained the following distribution of examples per attitude: comforting (one example), fascinated (two examples), thinking (two examples), doubtful (two examples), confronted (two examples), and embarrassed (three examples). For the female actor, we obtained the following: comforting (two examples), tender (one example), seductive (six examples), fascinated (one example), ironic (two examples), and scandalized (two examples).

Each test contains three separate conditions: video, ground-truth animation, and synthesized animation. The conditions were presented in random order. For each condition, we present a pseudo-randomized set of dialogues such that each attitude was represented once for a given actor, no two consecutive dialogues showed the same attitude, and if possible, a different dialogue was chosen for the same attitude in different conditions. Because each test contains only one example per attitude, per condition, and per actor, the users evaluated a total of 36 dialogue exchanges, leading to an average test duration of 20 minutes.

The test also included a short training part, so users were presented the test requirements and a performance sample of the actors for each attitude. Users were instructed to only evaluate the expressiveness of the on-screen actor. The user played a video and then answered the question, "What is the attitude of the actor in this example?" by checking one option from a list of six attitudes.

The test is available at www.barbulescu.fr/test_dialogue/, and all the dialogue exchanges used in the test are available at www.barbulescu.fr/exchanges/. Figure 6 presents examples for the three conditions.

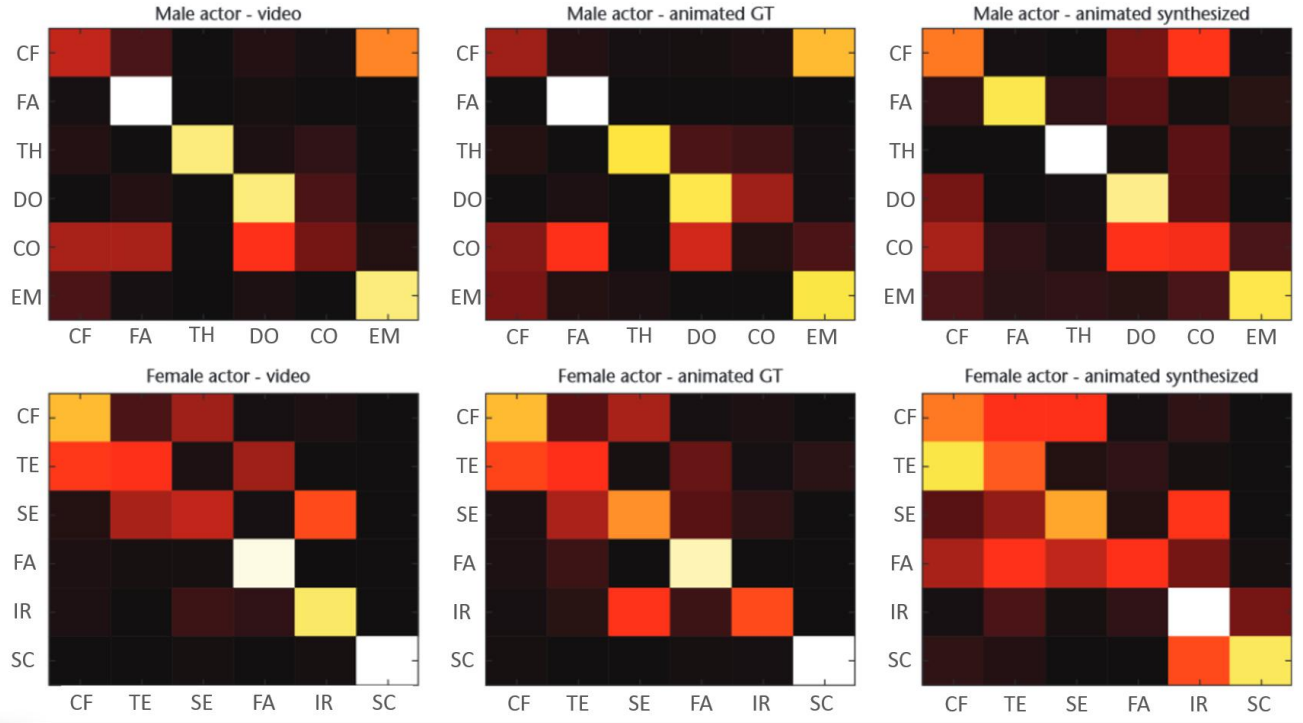


Figure 7. Confusion matrices for the mind-reading experiment represented as hot maps (lighter colors represent higher values): (a) male actor and (b) female actor. From left to right, the images correspond to the video, ground-truth animation, and our method (synthetic animation).

Table 2. F1 scores obtained for the male actor for the video (C1), animated ground truth (C2), and our method (C3).*

Condition	CF	FA	TH	DO	CO	EM	Mean
C1	0.31	0.81	0.90	0.71	0.23	0.68	0.60
C2	0.27	0.79	0.80	0.68	0.08	0.58	0.53
C3	0.48	0.72	0.82	0.61	0.28	0.67	0.59

** See Table 1 for attitude abbreviations.*

Table 3. F1 scores obtained for the female actor for the video (C1), animated ground truth (C2), and our method (C3).*

Condition	CF	TE	SE	FA	IR	SC	Mean
C1	0.58	0.41	0.31	0.82	0.68	0.98	0.63
C2	0.56	0.37	0.51	0.77	0.56	0.94	0.61
C3	0.35	0.33	0.44	0.36	0.56	0.67	0.45

** See Table 1 for attitude abbreviations.*

Results

We obtained results from 51 French subjects who performed the entire test. Figure 7 gives the confusion matrices per condition and actor. Tables 2 and 3 show the F1 scores for the male and female actors, respectively. All F1 scores obtained for the predicted stimuli are above the chance level.

We consider our method successful if the synthetic animations are as good as the ground-truth animations at expressing a given attitude. This means that the users tended to make similar choices in the perceptual tests, whether they are correct or incorrect. For this reason, we tested the hypothesis that there are no significant differences between the results obtained by the two conditions.

We fit two multinomial models per actor to predict an error distribution of the test results. These multinomial models were fit with data representing the two conditions, such that one model was constrained by the condition variable. Then, we conducted a likelihood ratio (LR) chi-squared test to compare distributions of responses predicted by two multinomial models. For the male actor, the test outputs a LR value of 0.62, with $df = 6$ and $p = 0.43$. We fail to reject the null hypothesis, thus showing that there are no statistical differences between the results obtained for the two conditions. However, for the female actor we obtain a LR value of 21.6, with $df = 6$ and $p < 1.e^{3}$. We continued the analysis by removing the data collected for the fascinated and tender attitudes, which are represented by only one example in the test set. For that case, we obtained a LR value of 2.28, with $df = 4$ and $p = 0.32$, which shows that for the other attitudes, there are no statistical differences between conditions. Moreover, removing only the data collected for the fascinated attitude, we obtained a LR value of 6.05, with $df = 5$ and $p = 0.05$, which is situated at the conventionally ac-

cepted significance level of 0.05 where we fail to reject the null hypothesis.

We used the results of this test to compare the perceived expressiveness of the videos and ground-truth animations with our synthetic animations. For this, we computed the correlations between the F1 scores obtained for each condition. For the male actor, we obtained strong correlations between all pairs of conditions (Pearson's correlation $r > 0.9$). For the female actor, we obtained a strong correlation between the video and animated ground truth ($r = 0.9$) and moderate correlation values between the animated ground truth and our method ($r = 0.62$) and video and our method ($r = 0.56$).

For each test, we also gathered user information, such as age, gender, and native language. We performed likelihood ratio tests comparing the combined multinomial model selected attitude (ground-truth attitude + condition + gender + language + age) with the reduced models obtained by eliminating one factor until all remaining factors significantly contributed to the model. Only the ground-truth attitude and condition fulfilled $p < 0.001$. This means that age and gender did not produce a difference in the performance of our attitude-recognition test.

To further illustrate the benefits of our method, we generated different variations of the testing dialogue by assigning all possible attitudes to the two characters over the entire duration of the scene. The video available as a web extra (youtu.be/9UF9xdfO1cs) shows an exchange of six turns acted with a combination of attitudes that is not present in the recorded dataset.

Figure 8 illustrates the predicted poses extracted from the fixed positions within a phrase: the middle of the first virtual syllable, the onset of the first two and last two vowels, and the middle of the second virtual syllable.

Discussion

A notable observation is the difference in the results obtained for the male and female actors. A possible explanation of the impoverished performance of the female actor is that, when delivering certain attitudes in interaction (the dialogue used for the testing dataset), the actors display motions and intonations that may differ from the ones used in isolated sentences (the sentences used for the training dataset). For example, in the case of the fascinated attitude, the female actress is looking toward the camera in the training dataset and looking upward in the testing dataset (see Figure 6, fourth column).

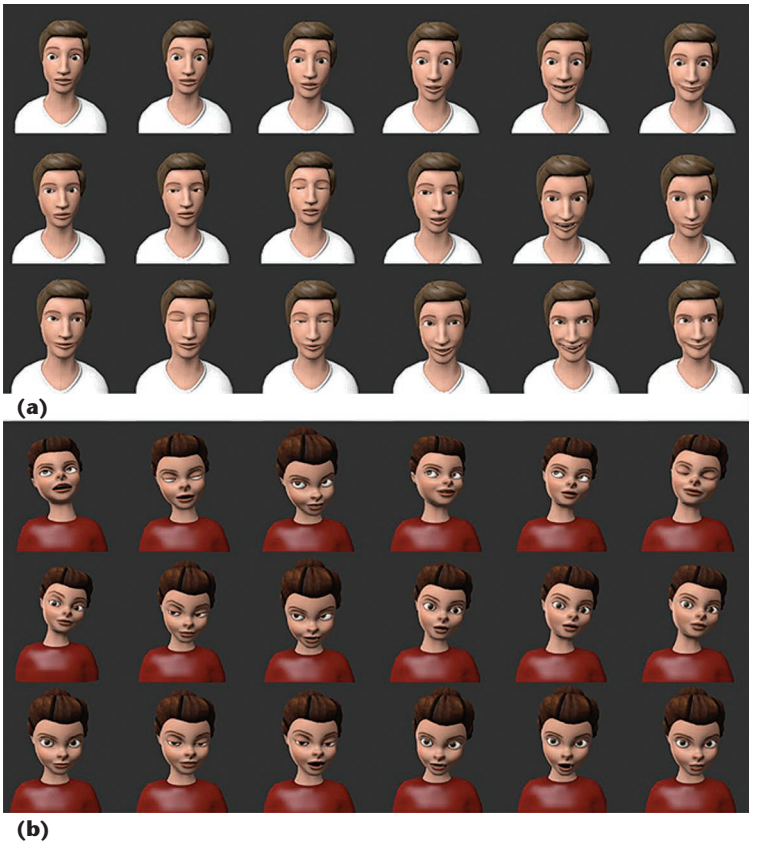


Figure 8. Predicted poses extracted from the fixed positions within a phrase. (a) In the top three rows, the actor performs the sentence "Je vous en prie" [You're welcome] in the fascinated, doubtful, and embarrassed attitudes (from top to bottom, respectively). (b) In the bottom three rows, the actress performs the sentence "Merci de vos jolies fleurs" [Thank you for your lovely flowers] in the thinking, ironic, and scandalized attitudes (from top to bottom, respectively). From left to right, the frames in each line correspond to the middle of the first virtual syllable, the onset of the first two and last two vowels, and the middle of the second virtual syllable.

With the exception of the case where only one example was available in the test set (the fascinated and tender attitudes for the female actor), the expressive animations generated using our method obtained results that show no statistical differences with those of videos and ground-truth animations. This means that our results successfully displayed the expressive styles learned from our expressive corpus.

Limitations

Several limitations need to be emphasized. First, we only consider attitude-specific contours that have no anchor points within the utterance. Some attitudes such as sarcasm might only concern parts of the sentence—for example, a narrow focus on words such as the adjective "really" in the sentence "this guy is really smart." Moreover, the global attitude-specific contours are often modulated by

contours that signal other communicative functions, such as syntactic grouping or emphasis.

Concerning variants, the SFC model captures regularities of attitude-specific prototypical contours. Variants can be considered by splitting the samples according to obvious behavioral alternatives, such as left/right rotations of the head.

Using phonostyles collected via exercises in style for generating convincing interactive dialogues also still requires additional work. This particularly applies to listening models that can be trained using the verbal (typically backchannels) and nonverbal behaviors of the listener.

For this work, we tested our model on 10 dramatic attitudes. The turns performed by two virtual avatars were synthesized and evaluated by asking human raters to identify attitudes in context. This test and the comments we collected via crowd-sourcing demonstrate that our model succeeds in displaying a variety of dramatic attitudes.

Although we consider the use of actors appropriate for our goal, it would be interesting to replicate this work using a spontaneous expressive corpus.

This work paves the route for building comprehensive libraries of dramatic attitudes suitable for expressive animation of conversational agents.

Our study is based on a dramatic text with the goal of generating dialogues with the dramatic styles deployed by two actors. Therefore, we used an acted corpus, as opposed to a spontaneous one. Although we consider the use of actors appropriate for our goal, it would be interesting to replicate this work using a spontaneous expressive corpus.

In future work, we would like to learn a more comprehensive set of attitudes, with more actors, attitudes, and sentences. We also envision creating a balanced testing dataset of dialogue exchanges where each attitude is represented by at least two examples. Additional work and resources are also necessary to extend the approach to include hand gestures and full-body motion.

Acknowledgments

This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the European Research Council advanced grant Expressive (ERC-2011-ADG 20110209). We thank Lucie Carta and Grégoire Gouby for their dramatic performances; Estelle Charleroy, Romain Testylier, and Laura Paiardini for their art work; and Georges Gagneré for his guidance.

References

1. K.R. Scherer and H. Ellgring, "Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?" *Emotion*, vol. 7, no. 1, 2007, pp. 158–171.
2. D. Bolinger, *Intonation and Its Uses: Melody in Grammar and Discourse*, Stanford Univ. Press, 1989.
3. A. Barbulescu, R. Ronfard, and G. Bailly, "Characterization of Dramatic Attitudes," *Proc. 17th Ann. Conf. Int'l Speech Comm. Assoc.* (Interspeech 2016), 2016, pp. 585–589.
4. E. Chuang and C. Bregler, "Mood Swings: Expressive Speech Animation," *ACM Trans. Graphics*, vol. 24, no. 2, 2005, pp. 331–347.
5. S. Marsella et al., "Virtual Character Performance from Speech," *Proc. 12th ACM Siggraph/Eurographics Symp. Computer Animation*, 2013, pp. 25–35.
6. I. Albrecht et al., "'May I Talk to You? :-)' Facial Animation from Text," *Proc. 10th Pacific Conf. Computer Graphics and Applications*, 2002, pp. 77–86.
7. A. Wang, M. Emmi, and P. Faloutsos, "Assembling an Expressive Facial Animation System," *Proc. ACM Siggraph Symp. Video Games*, 2007, pp. 21–26.
8. D. Vlasic et al., "Face Transfer with Multilinear Models," *ACM Trans. Graphics*, vol. 24, 2005, pp. 426–433.
9. S. Baron-Cohen, *Mind Reading: The Interactive Guide to Emotions*, Jessica Kingsley Publishers, 2003.
10. R. Queneau, *Exercises in Style*, New Directions Publishing, 2013.
11. G. Bailly and B. Holm, "SFC: A Trainable Prosodic Model," *Speech Comm.*, vol. 46, no. 3, 2005, pp. 348–364.
12. H.P. Graf et al., "Visual Prosody: Facial Movements Accompanying Speech," *Proc. 5th IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2002, pp. 396–401.
13. H. Silén et al., "Ways to Implement Global Variance in Statistical Speech Synthesis," *Proc. Ann. Conf. Int'l Speech Comm. Assoc.* (Interspeech), 2012, pp. 1436–1439; www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_1436.pdf.
14. L.C. Trutoiu et al., "Modeling and Animating Eye Blinks," *ACM Trans. Applied Perception*, vol. 8, no. 3, 2011, article 17.
15. E. Moulines and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech

Synthesis Using Diphones," *Speech Comm.*, vol. 9, no. 5, 1990, pp. 453–467.

Adela Barbulescu *is a research engineer at GIPSA-lab Grenoble. During this research, she was a postdoctoral fellow at Inria Grenoble in the IMAGINE research team. Her research interests include multimodal interaction with expressive avatars using speech, facial expressions, gaze, gestures, and body pose. Barbulescu has a PhD in mathematics and computer science from the Grenoble Alpes University. Contact her at barbulescu.adela@gmail.com.*

Rémi Ronfard *is a research director at Inria and the scientific leader of the IMAGINE research team at Inria and the University of Grenoble. His research interests include designing novel interfaces between graphic artists and computers. He is a member of ACM and IEEE. Ronfard has a PhD in computer science from Mines Paris Tech. Contact him at remi.ronfard@inria.fr.*

Gérard Bailly *is a senior CNRS research director at GIPSA-Lab Grenoble and head of the Cognitive Robotics, Interactive Systems, & Speech Processing (CRISSP) team. His research interests include multimodal interaction with conversational agents (virtual talking faces and humanoid robots) using speech, hand and head movements, and eye gaze. He is the coeditor of* Audiovisual Speech Processing *(CUP, 2012),* Improvements in Speech Synthesis *(Wiley, 2002), and* Talking Machines: Theories, Models and Designs *(Elsevier, 1992). He is an elected member of the International Speech Communication Association (ISCA) board and a founding member of the ISCA SynSIG and SproSIG special interest groups. Bailly has a PhD in electronics from the Grenoble Institute of Technology. Contact him at gerard .bailly@gipsa-lab.fr.*