Which Prosodic Features Contribute to the Recognition of Dramatic Attitudes?

Adela Barbulescu[1], *Rémi Ronfard*[1], *Gérard Bailly*[2]

[1]Univ. Grenoble Alpes, Inria, LJK
[2]GIPSA-lab, CNRS & Univ. Grenoble Alpes, Grenoble, France
adela.barbulescu@inria.fr, remi.ronfard@inria.fr, gerard.bailly@gipsa-lab.fr

Abstract

In this work we explore the capability of audiovisual prosodic features (such as fundamental frequency, head motion or facial expressions) to discriminate among different dramatic attitudes. We extract the audiovisual parameters from an acted corpus of attitudes and structure them as frame, syllable and sentence-level features. Using Linear Discriminant Analysis classifiers, we show that prosodic features present a higher discriminating rate at sentence-level. This finding is confirmed by the perceptual evaluation results of audio and/or visual stimuli obtained from the recorded attitudes.
Index Terms: audiovisual expressive speech; affective database; dramatic attitudes; perceptual correlates

1. Introduction

Attitudes refer to the expression of social affects and present acoustic and visual manifestations which are linked to conventions and cultural behaviors [1]. Thus, attitudes differ from basic emotional expressions, which may be seen as more spontaneous and universal expressions [2] [3].

The study of audiovisual parameters which encode the paralinguistic content of speech plays an essential role in improving the recognition and synthesis of expressive audiovisual speech. To this goal, there has been a great amount of work on the analysis and modeling of features which are found to help in the discrimination between expressive styles. Audiovisual features such as voice quality [4], acoustic prosodic features (F0, rhythm, energy) [5][6] [7], head motion [8] and facial expressions [9], have proven to be efficient in discriminating between basic emotions, attitudes or speaker identity.

While recognition of emotion, psycho-physiological state or co-verbal activities (drinking, eating, etc) is largely based on signal-based data mining and deep learning with features collected with a sliding window over multimodal frames, early studies on the expression of verbal attitudes have proposed that speakers use global prosodic patterns to convey an attitude [10][11]. These patterns are supposed to be anchored on the discourse and its linguistic structure, rather than encoded independently on parallel multimodal features. We recently evidenced the relevance of such patterns in facial displays [12].

The main aim of this work is to further explore the effectiveness of using audiovisual features at different structural levels to discriminate among expressive styles. We thus compare below the discrimination between attitudes at different structural levels (frame, syllable and sentence) and with different acoustic and visual features in order to evaluate the importance of the positioning of discriminant audiovisual events within the utterance. To that purpose, we performed a series of Linear Discriminant Analyses (LDA) on an expressive corpus of dramatic attitudes. In line with Iriondo et al [13] who used the results of a subjective test to refine an expressive dataset, we compare our best classification results with perceptual evaluation tests for the set of attitudes which are best discriminated.

The paper is structured as follows: section 2 presents approaches in related studies, section 3 presents our corpus of attitudes and the extraction of audiovisual features. Section 4 presents the experiments we carried out for automatic classification and section 5 presents the perceptual evaluation and comparison techniques, followed by conclusions in section 6.

2. Related work

Although recent years have brought a substantial progress in the field of affective computing [14], the development of emotion-modeling systems strongly depends on available affective corpora. As training and evaluation of algorithms require a great amount of data which is hard to collect, publicly available datasets represent a bottleneck for research in this field. Moreover, the majority of available datasets are limited to the six basic emotion categories proposed by Ekman [15] and include happiness, sadness, fear, anger, disgust, and/or surprise.

Databases containing affective data can be categorized under several criteria: data types used (2D or 3D visual data, speech), spontaneity (naturalistic, artificially induced or posed by professional actors or not), affective state categorization (emotion, attitudes etc). Audiovisual recording is obviously more expensive and time-consuming than audio-only recording. This is proven by the comparative amounts of publicly available audio and audiovisual datasets. For instance, the Interspeech Computational Paralinguistic Challenge [1] provides audio data from a high diversity of speakers and different languages, such as (non-native) English, Spanish, and German.

A comprehensive overview of the existing audiovisual corpora can be obtained from [24][14]. Table 1 presents a set of expressive datasets which are most relevant to our work. The works listed in the table present publicly available data that are used in several research topics: analysis, affective recognition, expressive performance generation, audiovisual conversion etc. IEMOCAP [23] and CAM3D [21] contain motion capture data of the face and upper-body posture from spontaneous performances. Large data variability is presented by Bosphorus [19] and CK+ [20] as they include more than 100 subjects posing over 20 expressions each, in the shape of action units and combinations. Another important work is the Mind Reading dataset [22] which includes video recordings of 412 expressive states classified under 24 main categories. While they serve as valuable references for the expressive taxonomy, these datasets often do not contain audio data. To our knowledge, the only publicly available affective datasets that include 3D data and

[1]http://compare.openaudio.eu/

Table 1: **Datasets for affect recognition systems**.

	# subjects	# samples	Data type	Speech & # sentences	Categories	Spontaneity
BIWI [16]	14	1109	3D face & video	Yes - 40	11 affective labels	acted
4D Cardiff [17]	4	N/A	3D face & video	Yes - N/A	10 expressions	spontaneous
MPI facial expressions [18]	6	N/A	3D face	N/A	55 expressions	acted
Bosphorus [19]	105	4652	3D face (static)	N/A	37 expressions (action units, basic emotions)	acted
CK+ [20]	123	593	video	N/A	23 expressions (action units, combinations)	posed and non-posed
CAM3D [21]	7	108	3D face & torso	Yes - N/A	12 mental states	spontaneous
Mind Reading [22]	6	N/A	video	Yes - N/A	412 emotions in 24 categories	acted
IEMOCAP [23]	10	N/A	3D face & torso	N/A	8 emotions	acted and spontaneous

speech are the BIWI, CAM 3D and 4D Cardiff corpora. Although the expressive categories contained extend the set of basic emotions, only a few present conversational potential (thinking, confused, frustrated, confidence). Most importantly, the sentences used in the datasets do not present a high variability or a systematic variation of syllable lengths.

Using data collected from such databases, a large amount of studies have been conducted for the analysis of audiovisual prosody. Swerts et al [25] present a detailed overview of studies carried out on audiovisual prosody. One aspect of this research relates directly to the communication functions that were traditionally attributed to auditory prosody: prominence, focus, phrasing. Analysis on head nods, eye blinks, eyebrows movement showed that these visual cues present a high influence on word prominence [26] [27][28]. Other areas of study include emotion, attitude and modality. Busso [29] analyzed different combinations of acoustic information and facial expressions to classify a set of 4 basic emotions. The results showed that the acoustic and visual information are complementary. Improved results were also obtained by considering sentence-level features, especially for visual data. In [8] authors performed statistical measures on an audiovisual database, revealing characteristic patterns in emotional head motion sequences. Ouni et al [30] analyze the acoustic (F0, energy, duration) and 3D visual data (facial expressions) captured by an actor performing 6 basic emotions. While no universal feature is found to discriminate between all the emotions, a few observations are noted: anger, joy, fear and surprise have similar speech rates, the facial movements are more important for joy, surprise and anger.

The expression of attitudes is highly dependent on the studied language. The following works focus on the study of attitudes and/or generation of prosody (intonation): [5] (French), [31] (Vietnamese), [32] (Brazilian Portuguese), [33](German). However, the datasets recorded for these studies are not publicly available and do not feature 3D data. Moraes et al [32] conducted a perceptual analysis of audiovisual prosody for Brazilian Portuguese using video data recorded by two speakers. They studied 12 attitudes categorized as social (arrogance, authority, contempt, irritation, politeness and seduction), propositional (doubt, irony, incredulity, obviousness and surprise) and assertion (neutral). An attitude recognition test showed the following: the difference in perception between the two speakers for certain attitudes such as the different strategies developed for irony and seduction, different dominant modality such as one speaker is better recognized in audio while the other in video, better overall recognition rates for audio-video among all modalities, the propositional and social attitudes show different perceptual behaviors. Another work on perception of audiovisual attitudes is focused on the expression of 16 social and/or propositional attitudes in German. Honemann et al [34] perform a set of attitude recognition tests. While the observations are valuable, these studies focus on the perceptual results of attitude recognition tests and do not carry out a complete analysis, including facial features and voice parameters.

To our knowledge, there are no extensive studies of the correlation between acoustic features and nonverbal gestures for the production of a large set of complex attitudes. Except for a few works related to modalities, such as interrogation [35][28][36], there is no qualitative analysis dedicated to the dynamics of visual prosodic contours of attitudes. We designed and recorded an expressive corpus consisting of attitudes performed by two French speakers. The corpus is designed to include sentences of varied sizes to allow our exploration of discriminating audiovisual features at different structural levels. The data gathered consists both in audio, video and 3D motion capture of the recorded performances. The following section describes the recording process.

3. Corpus of dramatic attitudes

We designed and recorded an acted corpus of "pure" social attitudes, i.e. isolated sentences carrying only one attitude over the entire utterance.

Selected text. We extracted the sentences for our database from a French translation of the play "Round dance" by Arthur Schnitzler [37]. The text is represented by 35 sentences (with a distribution of number of syllables spanning between 1 and 21).

Selected attitudes. We selected a subset of 10 attitudes from Baron-Cohen's Mind Reading project [22]. The source taxonomy proposed by Baron-Cohen comprises a total of 412 attitudes grouped under 24 main categories, each comprising several layers of sub-expressions. The attitude choice was made in collaboration with a theater director, such that the attitudes were compatible with the selected text. Table 2 contains the list of attitudes we decided to analyze for this study. Typical facial displays of these attitudes are illustrated in figure 1.

Recordings. The synchronized recording of voice signals and motion was done using the commercial system Faceshift® (http://www.faceshift.com/) with a short-range Kinect camera and a Lavalier microphone. Faceshift enables the creation of a

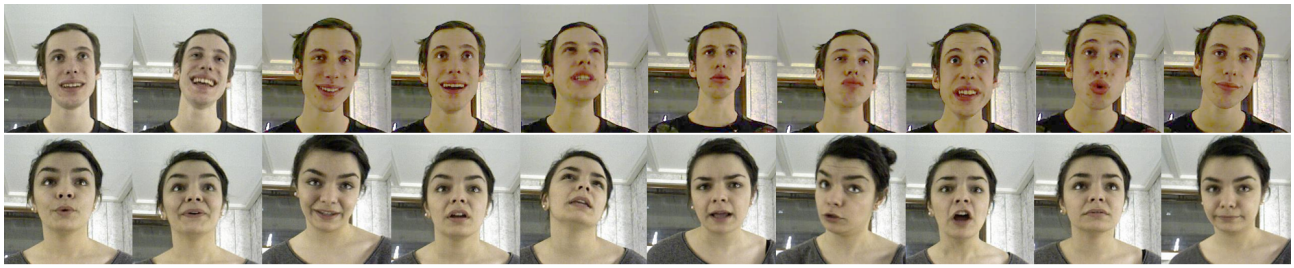

Figure 1: Examples of the 10 attitudes interpreted by the two actors in our data set.

Table 2: Presentation of chosen attitudes: Category, Subgroup and Definition are shown as they appear in Mind Reading [22].

Category	**Subgroup**	**Our labels**	**Abbr.**	**Definition**
Kind	Comforting	Comforting	**CF**	Making people feel less worried, unhappy or insecure
Fond	Liking	Tender	**TE**	Finding something appealing and pleasant; being fond of something
Romantic	Seductive	Seductive	**SE**	Physically attractive
Interested	Fascinated	Fascinated	**FA**	Very curious about and interested in something that you find impressive
Thinking	Thoughtful	Thinking	**TH**	Thinking deeply or seriously about something
Disbelieving	Incredulous	Doubtful	**DO**	Unwilling or unable to believe something
Unfriendly	Sarcastic	Ironic	**IR**	Using words to convey a meaning that is the opposite of its literal meaning
Surprised	Scandalized	Scandalized	**SC**	Shocked or offended by someone else's improper behavior
Hurt	Confronted	Confronted	**CO**	Approached in a critical or threatening way
Sorry	Embarrassed	Embarrassed	**EM**	Worried about what other people will think of you

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, brows up, jaw open etc). Faceshift also outputs estimations of the head motion and gaze direction. The sentences were recorded by two semi-professional actors under the guidance of a theater director. The two actors recorded the 35 sentences uttered first in a neutral, "flat" style, then with each of the selected 10 attitudes. This technique called "exercices in style" is inspired by Queneau [38] who uses this method of retelling the same story in 99 different styles to train comedians.

The recording session began with an intensive training of the actors, which consisted in fully understanding the interpreted attitudes and developing the ability to dissociate the affective state from the meanings of the sentences. Actors were also instructed to maintain a constant voice modulation, specific for each attitude, throughout uttering the entire set of 35 sentences. The actors performed as if they addressed a person standing in front of them. They did not receive any instructions related to gestural behaviors. A perceptual screening was carried out during the recordings by the theater director and an assistant. If needed, certain utterances were repeated.

Annotation. All utterances were automatically aligned with their phonetic transcription obtained by an automatic text-to-speech phonetizer [39]. The linguistic analysis (part-of-speech tagging, syllabation), the phonetic annotation and the automatic estimation of melody were further corrected by hand using a speech analysis software [40]. The manual verification of melodic contours represented an extensive effort due to the large amount of data recorded (a total of 3 hours of speech).

4. Data analysis

This section presents the analysis of the recorded data: feature extraction, stylization and discriminant analysis.

4.1. Feature extraction

There has been a great amount of work on the analysis of features which are found to help in the discrimination between expressive styles. Along with voice pitch (melody), energy, syllable duration and spectrum, we include gestural data: head and eye movements and facial expressions.

Fundamental frequency. As mentioned in the previous section, melody was obtained by automatic phonetic aligning followed by manual verification using Praat [40]. Therefore, we obtained reliable F_0 contours which we further normalized and then converted to semitones.

$$f_0ref = \begin{cases} 210Hz, & \text{if female} \\ 110Hz, & \text{if male} \end{cases} \tag{1}$$

(2)

$$F_0[tone] = \frac{240}{\log 2} \log \frac{F_0[Hz]}{F_{ref}} \tag{3}$$

where F_{ref} represents the speaker's register (210 Hz for the female speaker and 110 Hz for the male speaker). The resulting F_0 contours are comparable across speakers.

Rhythm. For rhythm we used a duration model [41] [42] where 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 corresponding to the deviation of the syllable duration Δ relative to an expected duration Δ':

$$C = \frac{\Delta - \Delta'}{\Delta'} \tag{4}$$

$$\Delta' = (1 - r) \cdot \sum_i \bar{d_{p_i}} + r \cdot D \tag{5}$$

where i is the phoneme index within the syllable, $\bar{d_{p_i}}$ is the average duration of phoneme i, D is the average syllabic duration (=190ms here) and r is a weighting factor for isochronicity (=0.6 here). We note C as the rhythm coefficient which is computed for every syllable in all sentences in the corpus.

Energy. Energy is extracted at phone-level and computed as mean energy (dB):

$$energy[dB] = 10 * log10(\frac{\sum_{i=1}^{|y|} y_i^2}{|y|}) \quad (6)$$

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

Spectrum. The spectrum is extracted using the vocoder STRAIGHT [43] which returns the voice spectra, aperiodicities and fundamental frequency. We use 24 mel-cepstral coefficients, from the 2^{nd} to the 25^{th} (i.e. excluding the energy).

Head and gaze. Head and gaze motion are obtained directly from the processing of the Kinect RGBD data by the Faceshift ®software and processed at 30 frames/s. We consider that an expressive performance is obtained by adding expressive visual prosodic contours to the trajectories of a "neutral" performance. Since motion has a linear representation, we obtain the visual prosody by simply aligning an expressive performance with its neutral counterpart and computing the difference of the vectors.

Facial expressions. Facial expressions are returned by the Faceshift software as blendshape values. We compute the differential blendshape vectors, to which we apply Non-negative Matrix Factorization (NMF). We split these into two main groups: *upper-face expressions* (8 components) and *lower-face expressions* (8 components).

4.2. Feature stylization

By *stylization* we mean the extraction of several values at specific locations from the feature trajectories with the main purpose of simplifying the analysis process while maintaining a constant number of characteristics of the original contour for all structural levels whatever the linguistic content. We propose the following stylization methods:

- *audio*: the audio feature contours are stylized by extracting 3 values: at at 20%, 50% and 80% of the vocalic nucleus of each syllable.
- *visual*: the visual feature contours are stylized by extracting contour values at 20%, 50% and 80% of the length of each syllable.
- *rhythm*: the rhythm is represented by one parameter per syllable: the lengthening/shortening coefficient.

We add *virtual syllables* to account for the pre- and post-phonatory movements for all visual components. As previously observed [44], preparatory movements are discriminant to a certain degree for specific emotion categories. We therefore introduce two virtual syllables with a duration of 250 ms preceding and following each utterance. This duration was chosen because preparatory blinking occurs within 250 ms of utterance beginning in our dataset. Therefore, stylization of motion is also done for the virtual syllables, by extracting motion contour values at 20%, 50% and 80% of the duration of each virtual syllable.

4.3. Blinking interval

Another prosodic feature we mention in this work is the *blinking interval*, defined as the time elapsed between consecutive blinks. We extract blinks by thresholding the blendshapes corresponding to eyelid lowering and then we compute the mean and deviation values of the intervals per attitude. Figure 2 illustrates the blinking strategies presented by the two actors. While this information is useful for attitude characterization, this feature cannot be stylized at syllable level and therefore cannot be used in discriminant analysis at different structural levels.

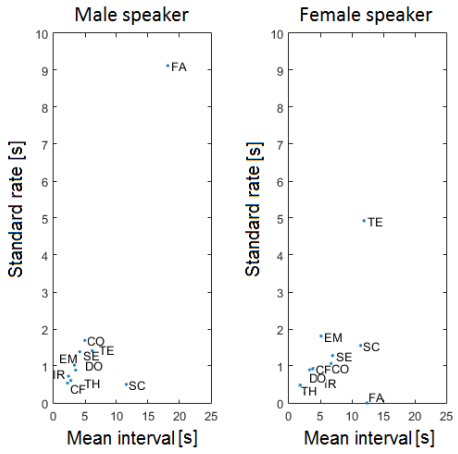

Figure 2: Mean and standard blinking intervals for the two actors. Note that attitudes such as Fascinated, Tender and Scandalized present higher blinking intervals.

4.4. Discriminant analysis

Discriminant analysis between the 10 attitudes is performed using Fisher classification with 10-fold cross-validation. Speaker-dependent and speaker-independent classification of attitudes were performed at three structural levels for each feature separately, for the concatenation of prosodic features and the concatenation of all audiovisual features:

- *frame-level*: a feature represents data extracted from each stylization point
- *syllable-level*: a feature represents the concatenation of the frame-level features at each syllable, including virtual syllables for the visual features
- *sentence-level*: for audio, a feature represents the concatenation of the syllable-level features from the first and last syllables for a given sentence. For visual, we also concatenate the syllable-level features from the two virtual syllables of that sentence. Note that for sentences composed of one syllable, we perform data duplication to obtain the desired feature dimension (see table 3).

Table 3: Dimension and size for all features: F0, rhythm (Rth), energy (Enr), spectrum (Spec), head motion (Head), gaze motion (Gaze), upper-face blendshapes (Up) and lower-face blendshapes (Low).

	F0	**Rth**	**Enr**	**Spec**	**Head**	**Gaze**	**Up**	**Low**
Dimension	1	1	1	24	6	2	8	8
Frame size	1	-	-	1	1	1	1	1
Syllable size	3	1	1	3	3	3	3	3
Sentence size	6	2	2	6	12	12	12	12

Table 4: F1-scores for the automatic classification. LDA classifiers are trained using sentence-level features over 10 attitudes: for the male speaker (a) and female speaker (b). Values in bold are greater than 0.6.

(a) F1-score for the male speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
Comforting	0.24	0.23	0.48	0.33	**0.76**	**0.62**	**0.64**	0.58	**0.90**
Tender	0.56	0.00	0.53	0.46	**0.72**	0.45	**0.60**	**0.75**	**0.92**
Seductive	0.48	0.07	0.24	0.45	**0.73**	0.49	0.58	0.55	**0.89**
Fascinated	0.38	0.00	0.22	**0.65**	**0.67**	**0.68**	**0.73**	**0.69**	**0.94**
Thinking	**0.62**	0.37	0.50	0.27	**0.82**	0.43	**0.71**	0.39	**0.94**
Doubtful	0.46	0.23	0.22	0.39	**0.66**	0.40	**0.60**	**0.69**	**0.96**
Ironic	0.16	0.05	0.21	0.59	**0.83**	0.54	**0.78**	**0.82**	**0.96**
Scandalized	**0.68**	0.22	**0.87**	0.39	**0.81**	0.49	**0.60**	0.43	**0.90**
Confronted	0.47	0.05	0.36	0.27	**0.60**	0.32	0.51	0.24	**0.84**
Embarrassed	0.43	0.21	**0.68**	**0.87**	**0.82**	**0.73**	**0.73**	**0.83**	**0.99**
Mean	0.44	0.15	0.43	0.49	**0.74**	0.53	**0.66**	**0.58**	**0.91**

(b) F1-score for the female speaker

	F0	Rth	Enr	Spec	Head	Gaze	Up	Low	All
Comforting	0.38	0.29	0.23	**0.60**	**0.64**	0.45	0.50	**0.71**	**0.88**
Tender	**0.68**	0.08	0.47	0.49	0.36	0.38	0.36	**0.85**	**0.89**
Seductive	0.47	0.10	0.26	0.31	**0.76**	0.34	**0.64**	**0.69**	**0.89**
Fascinated	0.53	0.00	0.10	0.59	**0.78**	**0.65**	**0.66**	**0.65**	**0.91**
Thinking	0.51	0.26	0.21	0.41	**0.73**	0.35	**0.71**	0.54	**0.93**
Doubtful	**0.70**	0.07	0.19	0.43	**0.64**	0.44	0.50	**0.67**	**0.89**
Ironic	0.30	0.00	0.16	0.39	0.42	0.31	0.32	**0.62**	**0.83**
Scandalized	**0.65**	0.30	**0.92**	**0.78**	0.57	0.26	0.48	**0.60**	**0.97**
Confronted	**0.68**	0.04	0.41	**0.68**	**0.63**	0.50	0.51	**0.90**	**0.99**
Embarrassed	0.45	0.41	0.35	0.40	**0.67**	**0.64**	**0.70**	**0.97**	**0.97**
Mean	0.55	0.16	0.32	0.45	**0.64**	0.45	0.55	**0.71**	**0.90**

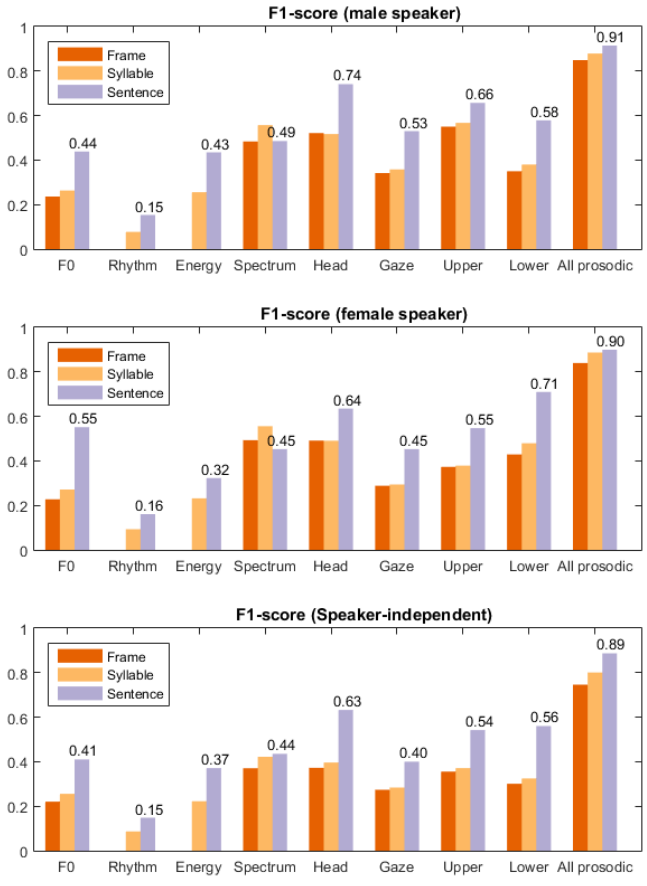

Figure 3: Average F1-scores obtained for F0, rhythm, energy, spectrum, head motion, gaze motion, upper-face expressions, lower-face expressions and concatenated prosodic features at: frame, syllable and sentence-level. The figures are shown for the male speaker (top), the female speaker (middle) and speaker-independent (bottom). Marked values represent mean F1-scores for sentence-level features.

Results. We analyze our data using F1, a balanced measure between precision and recall. For this, we compute the F1-score, which represents the harmonic mean between precision and recall. We observe that in the case of prosodic features (F0, energy, rhythm, head motion, gaze, facial expressions), the discrimination rate increases as feature granularity increases (see figure 3).

Higher scores at sentence-level indicate that order matters: the overall shapes of the features within the sentences have better discrimination power than local feature values. This is especially for F0, head motion and gaze, where the average F1-score is increased by more than 30% of the scores obtained for the frame- and syllable-level. In the case of the concatenated prosodic feature, the gain is smaller. This means that these features already contain enough discriminant information at frame- and syllable-level.

In the case of the spectrum, we observe a decrease in score at sentence-level, showing that the overall shapes at this level does not improve the discrimination of attitudes. The lowest scores for sentence-level features are generally obtained for the speaker-independent classification. F0, head motion, upper and lower-face expressions decrease the most relative to speaker-dependent results, showing that these features manifest different strategies for attitude expression.

Table 4 presents the F1-scores for all features at sentence-level. On average, high scores are obtained for the F0, head motion and facial expressions, while rhythm and energy show lower discrimination scores. However, attitudes present different score ranges showing that the speakers express audiovisual attitudes using different strategies. For example, Comforting, Fascinated and Ironic show higher scores for the visual features, while Scandalized shows higher audio features. In order to assess the perceptual correlates of these features we carried out two perceptual tests on recordings of the two actors.

5. Experiments

5.1. Perceptual tests

We carried out two attitude recognition tests using recorded data from the two speakers. The first test used audio and video recordings of the two actors. For the second test, the stimuli were obtained using an animation system, in which the recorded motion is directly mapped to a realistic 3D model of the speaker

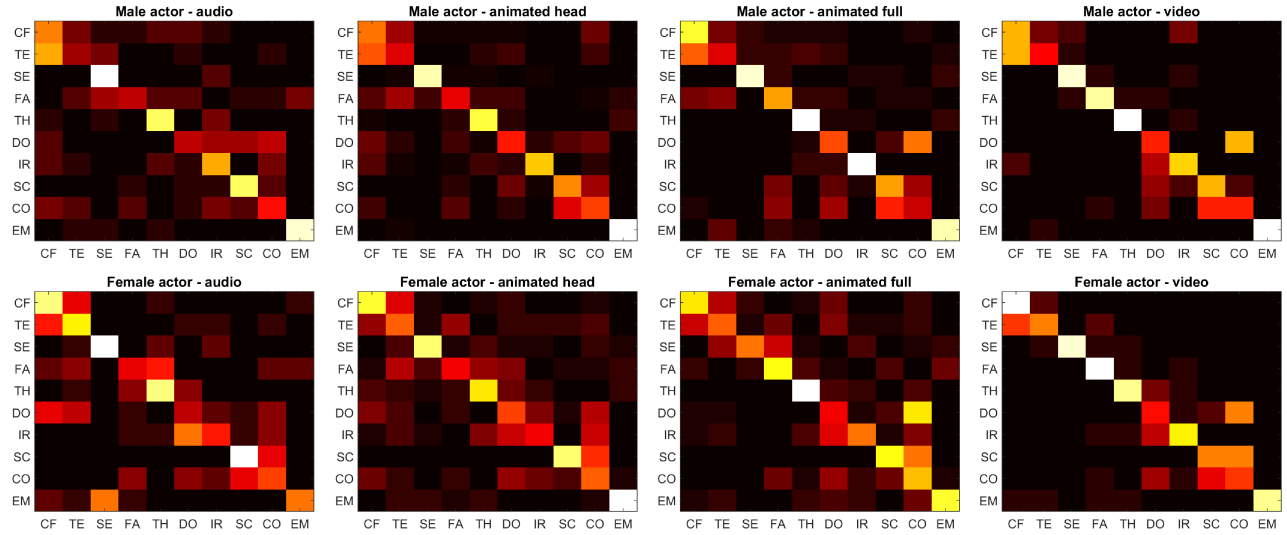

Figure 4: Confusion matrices for the perceptual tests: top images correspond to the male speaker and bottom images correspond to the female speaker. From left to right, confusion matrices obtained on the tests containing : audio, head-only animation, full animation, video. In these figures, rows represent actual attitudes and columns represent the predictions made by the users. Lighter colors indicate higher values.

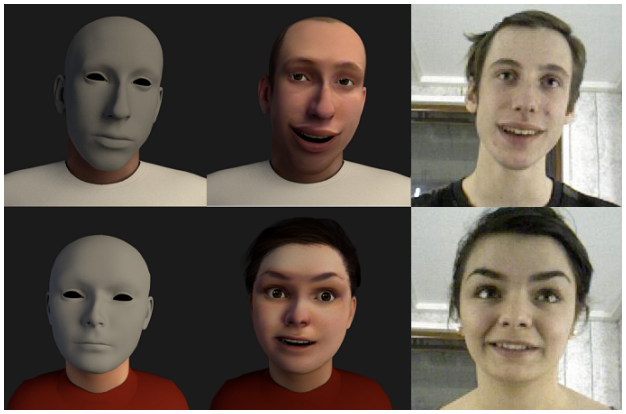

Figure 5: Corresponding frames for the two speakers extracted from performances of the attitude Comforting, in three modalities: head-only animation, full animation and video.

and the audio signal is represented by the original voice recordings. Both tests were closed response set with a single choice. The user would play a video and answer the question: "What is the attitude of the actor in this example?" by checking one option from a list of 10 attitudes.

In the first perceptual test, subjects were asked to recognize the attitudes from a total of 40 stimuli, representing audio and video recordings for the two actors. The video modality also included sound. For each actor, the audio modality was presented first. The sentences presented were randomly picked such that any two consecutive stimuli present the different attitudes and different utterances. A total of 20 French native speakers participated in this experience.

In the second perceptual test, each subject is asked to recognize the attitudes of a set of 40 animations representing the two speakers under two modalities: (1) *full animation*, where all recorded motion is directly retargeted to a 3D model of the actor, (2) *head-only animation*, where only head motion is retargeted, while facial expressions are fixed. The areas of the 3D head model where expressions are fixed are highlighed by replacing the realistic texture with a matte, gray texture. In the head-only animation modality, the eyes are represented as simple, black holes, thus maintaining the appearance of a mask. These modalities are illustrated in figure 5.

The stimuli are presented such that no two consecutive performances contain identical attitudes, sentences or speakers. For one test, random sentences are chosen from a subset of 6 sentences such that each attitude appears twice for each speaker. A total of 36 French native speakers participated in this experience. The first online test can be found at [2] and the second online test can be found at [3].

Results. The confusion matrices obtained for the perceptual tests are illustrated in figure 4 and the F1-scores are presented in table 5.

We observe an overall increase in recognition scores as more information is presented to the subjects. The biggest increases are observed between the *Audio* and *Head-only animation* for the male actor - especially the case for Ironic - and the *Full animation* and *Video* for the female actor - especially the case for the attitudes Comforting, Seductive, Ironic and Embarrassed.

Overall, the best recognized attitudes are Seductive, Thinking, Scandalized and Embarrassed, and lowest are Tender, Doubtful and Confronted, for both speakers. The lowest recognition scores appear because of a high confusion between Tender and Comforting, between Confronted and Scandalized, and an interchangeable confusion between Doubtful and Confronted. This happens because the attitudes in each pair are close in terms of expressive content.

[2] http://www.barbulescu.fr/test_audio_video
[3] http://www.barbulescu.fr/test_attitudes

Table 5: F1-scores obtained in the perceptual tests, per modality and per actor. Values in bold are greater than 0.6.

	CF	TE	SE	FA	TH	DO	IR	SC	CO	EM	Mean
Audio	0.41	0.24	**0.75**	0.34	**0.71**	0.30	0.49	**0.71**	0.36	**0.85**	0.52
Head	0.43	0.32	**0.89**	0.39	**0.75**	0.41	**0.73**	0.56	0.42	**0.92**	0.58
Full	**0.60**	0.33	**0.86**	0.52	**0.82**	0.43	**0.90**	0.52	0.25	**0.83**	**0.61**
Video	0.52	0.45	**0.86**	**0.87**	**0.95**	0.39	**0.65**	**0.60**	0.38	**0.97**	**0.66**

(a) F1-score for the male speaker.

	CF	TE	SE	FA	TH	DO	IR	SC	CO	EM	Mean
Audio	0.54	0.48	**0.70**	0.29	**0.60**	0.21	0.36	**0.71**	0.34	0.52	0.48
Head	0.56	0.35	**0.70**	0.33	0.55	0.34	0.33	**0.72**	0.33	**0.80**	0.50
Full	0.58	0.41	0.54	0.58	**0.75**	0.26	0.48	**0.61**	0.34	**0.68**	**0.52**
Video	**0.77**	0.59	**0.92**	**0.82**	**0.80**	0.36	**0.70**	0.53	0.33	**0.89**	**0.67**

(b) F1-score for the female speaker.

Table 6: F1-scores obtained for LDA scores vs. perceptual tests, per modality and per actor. Values in bold are greater than 0.6.

	CF	TE	SE	FA	TH	DO	IR	SC	CO	EM	Mean
Mod1	0.13	0.32	0.59	0.30	**0.62**	0.30	0.37	**0.71**	0.20	**0.61**	0.41
Mod2	0.43	0.32	**0.70**	0.39	**0.75**	0.41	**0.73**	0.56	0.38	**0.77**	0.54
Mod3	**0.60**	0.33	**0.86**	0.54	**0.82**	0.43	**0.90**	0.52	0.22	**0.83**	**0.61**
Mod4	0.52	0.45	**0.86**	**0.76**	**0.95**	0.39	**0.65**	**0.60**	0.33	**0.97**	**0.65**

(a) F1-score for the male speaker.

	CF	TE	SE	FA	TH	DO	IR	SC	CO	EM	Mean
Mod1	0.31	0.47	0.24	0.17	0.38	0.24	0.30	**0.71**	0.34	0.52	0.37
Mod2	0.52	0.31	**0.67**	0.33	0.29	0.34	0.28	**0.67**	0.40	**0.67**	0.45
Mod3	0.58	0.41	0.54	0.53	**0.71**	0.26	0.48	**0.61**	0.34	**0.68**	0.51
Mod4	**0.77**	0.59	**0.92**	**0.73**	**0.74**	0.36	**0.70**	0.53	0.33	**0.89**	**0.66**

(b) F1-score for the female speaker.

5.2. Comparison between subjective and objective scores

In order to compare the discrimination scores obtained by automatic classification and the perceptual test results, we trained separate LDA classifiers for the two speakers. Data was partitioned into training and testing such that the testing sentences coincide with the ones used in the perceptual test.

After obtaining objective classification results, we are interested in measuring the recognition rates when we consider the classification choices as ground truth data. For this, we compute the confusion matrix where LDA results are predictors and perceptual results are predictions. We test the LDA classifiers on the same sentence that was assessed by the subject of the perceptual test. We are particularly interested in the scores obtained at sentence-level by the concatenation of prosodic features which account for the maximum information that was displayed in the perceptual tests. We therefore define the following feature combinations:

- Mod1 = concatenation of all acoustic prosodic features (F0, energy, rhythm), accounting for the *Audio* modality in the perceptual test
- Mod2 = concatenation of all acoustic prosodic features and head motion, accounting for the *Head-only animation* modality in the perceptual test
- Mod3 = concatenation of all prosodic features, accounting for the *Full animation* modality in the perceptual test
- Mod4 = concatenation of all prosodic features, accounting for the *Video* modality in the perceptual test

Table 6 presents the F1-scores obtained when we compute the confusion matrices between classification scores for these combinations of features and the perceptual scores for their respective modality.

In comparison to the scores presented in table 5, we observe generally smaller scores for the *Audio* and *Head-only animation* modality, and very similar scores for *Full animation* and *Video*. In the case of *Audio*, smaller scores are obtained for attitudes such as Comforting and Seductive, which are performed with specific voice quality ranges. On the other hand Scandalized, which relies on high energy scores, obtains similar scores in all measurements. This shows, that for subtler attitudes, the classification can be improved by fusing a frame-level classifier for segmental features, such as spectrum. The degradation in scores for *Head-only animation*, can also be attributed to the contribution of acoustic information, since the same attitudes are affected.

We observe a certain variability in terms of speaker-dependent strategies and also in attitude-specific strategies. For example, Fascinated obtains low scores in audio modality and higher scores as more visual information is added, while Scandalized scores high in audio modality and lower as visual information is introduced, due to similarities in visual features with Confronted. For this reason, prosodic features bring different contributions to the perception of dramatic attitudes. This contribution depends both on the attitude itself, but also on the individual strategies of performing.

Interrater agreement. For an in-depth look at the relationship between individual features and perceptual results, we evaluate the agreement between the LDA classification and perceptual raters. For this, we compute Cohen's kappa coefficient [45] for the confusion matrix obtained by considering LDA results as predictors and subjective results as predictions. The advantage of using this measure is that we are able to compare attitude scoring between raters, by looking at similarities between correctly or incorrectly classified items. The calculation of the kappa coefficient is based on the difference between how much agreement is actually present compared to how much agreement would be expected by chance alone. A perfect agreement yields the value $k = 1$, while a chance agreement yields the value $k = 0$.

We compute the coefficient on pairs of raters, specifically between each individual rater for the perceptual test and the average LDA rater at frame, syllable and sentence-level. For each pair of raters we test the LDA classifiers on the same sentence that was assessed by the subject of the perceptual test. We perform LDA classification for the following features: F0, concatenated audio, head, gaze, upper-face blendshapes, lower-face blendshapes, concatenated head and audio feature, concatenated audiovisual feature. Figure 6 presents the coefficients obtained per actor, per feature and per modality.

Our results show that the values of the agreement coefficient increase as granularity increases. This demonstrates a better agreement between our objective scores for sentence-level features and the perceptual scores. For the sentence-level

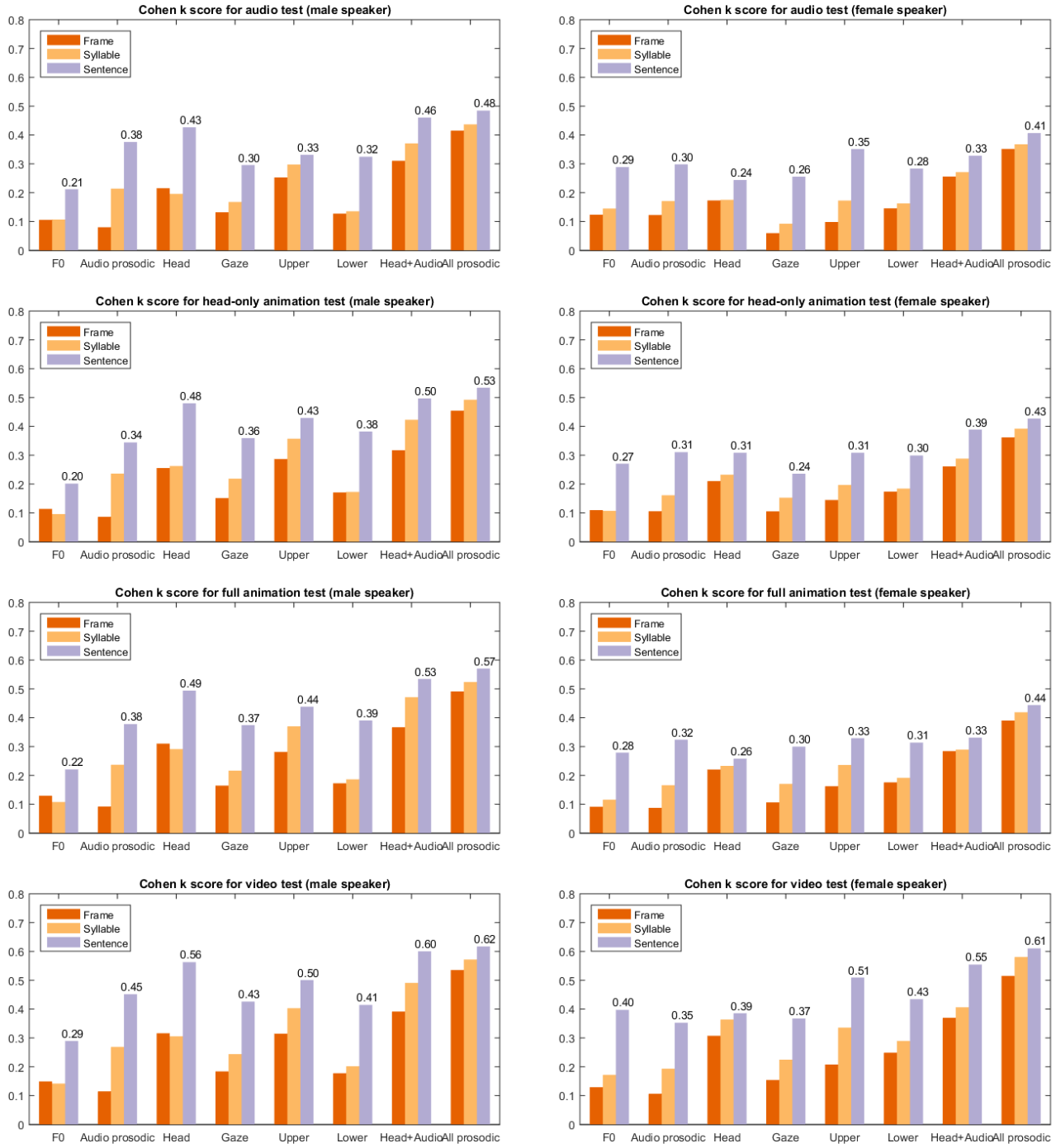

Figure 6: Cohen's kappa agreement values for the male speaker (left) and the female speaker (right). From top to bottom, kappa scores obtained for the modalities: audio, head-only animation, full animation and video.

scores, the interrater agreement values obtained range from fair to substantial ($0.21 \leq k \leq 0.62$), generally with higher agreement for the male actor.

Overall, we observe a higher agreement for head and concatenated head and audio feature for the male actor and higher agreement for F0 for the female actor. Generally, the agreement increases for all features as more information is used in the perceptual tests. A significant increase appears for the female actor for the visual features, as the video modality is used. The lower values for the full animation modality imply that discriminative information - such as subtle expressions and texture - is lost with the usage of 3D animation. However, the usage of *Head-only animation* shows a significant increase for head and concatenated head and audio feature for the female actor also.

6. Conclusion

In this work, we analyzed audiovisual speech utterances with similar content (sentences) in different styles (dramatic attitudes). We found that the expression of dramatic attitudes is speaker-dependent. In a series of experiments, we found that LDA classifiers trained on speaker-dependent data outperform the classifiers trained on data recorded from both speakers.

We also found that LDA classifiers trained on sentence-level features outperform the classifiers trained on either frame-level or syllable-level features. This means that taking temporal context into account improves the attitude classification performance for prosodic features. The improvement is more significant for F0, head motion and gaze. This is also confirmed by the results of perceptual tests, which show a higher agreement with classification scores obtained for sentence-level features.

In these experiments, we noticed that the speakers use different strategies in the expression of attitudes. The objective evaluation tests show that the male speaker presents higher discrimination rates for the energy, head, gaze and upper-face expressions while the female speaker presents higher scores for the F0 and lower-face expressions. Through perceptual tests, we also proved the effective usage of head motion by the male actor.

Our results show that the studied prosodic features contribute differently to the perceptual discrimination of dramatic attitudes. Future work may include a more in-depth study of the relationship between individual prosodic features and perceptual discrimination of attitudes for a higher number of actors. Animated stimuli can still be valuable as they allow control the amount of visual information provided in a perceptual test.

7. Acknowledgements

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.

8. References

[1] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[2] U. Scherer, H. Helfrich, and K. Scherer, "Paralinguistic behaviour: Internal push or external pull?" in *Language: social psychological perspectives: selected papers from the first International Conference on Social Psychology and Language held at the University of Bristol, England*, 1979, p. 279.

[3] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[4] C. Monzo, F. Alías, I. Iriondo, X. Gonzalvo, and S. Planet, "Discriminating expressive speech styles by voice quality parameterization," in *Proc. of ICPhS*, 2007.

[5] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in French: data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[6] I. Iriondo, S. Planet, J.-C. Socoró, and F. Alías, "Objective and subjective evaluation of an expressive speech corpus," in *Advances in Nonlinear Speech Processing*. Springer, 2007, pp. 86–94.

[7] H. Mixdorff, A. Hönemann, and A. Rilliard, "Acoustic-prosodic analysis of attitudinal expressions in German," *Proceedings of Interspeech 2015*, 2015.

[8] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1075–1086, 2007.

[9] C. Davis, J. Kim, V. Aubanel, G. Zelic, and Y. Mahajan, "The stability of mouth movements for multiple talkers over multiple sessions," *Proceedings of the 2015 FAAVSP*, 2015.

[10] I. Fónagy, E. Bérard, and J. Fónagy, "Clichés mélodiques," *Folia linguistica*, vol. 17, no. 1-4, pp. 153–186, 1983.

[11] D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[12] A. Barbulescu, G. Bailly, R. Ronfard, and M. Pouget, "Audiovisual generation of social attitudes from neutral stimuli," in *Facial Analysis, Animation and Auditory-Visual Speech Processing*, 2015.

[13] I. Iriondo, S. Planet, J.-C. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

[14] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[15] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[16] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3d audio-visual corpus of affective communication," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 591–598, 2010.

[17] J. Vandeventer, A. J. Aubrey, P. L. Rosin, and D. Marshall, "4d Cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations," in *International Conference on Auditory-Visual Speech Processing*, 2015.

[18] K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven, "The MPI facial expression database a validated database of emotional and conversational facial expressions," *PloS one*, vol. 7, no. 3, p. e32321, 2012.

[19] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*. Springer, 2008, pp. 47–56.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[21] M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek, "3D corpus of spontaneous complex mental states," in *Affective computing and intelligent interaction*. Springer, 2011, pp. 205–214.

[22] S. Baron-Cohen, *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers, 2003.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

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

[25] M. Swerts and E. Krahmer, "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions," *Journal of Phonetics*, vol. 38, no. 2, pp. 197–206, 2010.

[26] C. Pelachaud, N. I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive science*, vol. 20, no. 1, pp. 1–46, 1996.

[27] B. Granström, D. House, and M. Swerts, "Multimodal feedback cues in human-machine interactions," in *Speech Prosody 2002, International Conference*, 2002.

[28] E. Cvejic, J. Kim, C. Davis, and G. Gibert, "Prosody for the eyes: quantifying visual prosody using guided principal component analysis." in *INTERSPEECH*, 2010, pp. 1433–1436.

[29] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[30] S. Ouni, V. Colotte, S. Dahmani, and S. Azzi, "Acoustic and visual analysis of expressive speech: A case study of French acted speech," in *Interspeech 2016*, vol. 2016, 2016, pp. 580–584.

[31] D.-K. Mac, E. Castelli, and V. Aubergé, "Modeling the prosody of Vietnamese attitudes for expressive speech synthesis." in *SLTU*, 2012, pp. 114–118.

[32] J. A. De Moraes, A. Rilliard, B. A. de Oliveira Mota, and T. Shochi, "Multimodal perception and production of attitudinal meaning in Brazilian Portuguese," *Proc. Speech Prosody, paper*, vol. 340, 2010.

[33] A. Hnemann, H. Mixdorff, and A. Rilliard, "Classification of auditory-visual attitudes in German," in *International Conference on Auditory-Visual Speech Processing*, 2015.

[34] A. Hönemann, H. Mixdorff, and A. Rilliard, "Social attitudes - recordings and evaluation of an audio-visual corpus in German," in *Forum Acusticum*, 2014.

[35] R. J. Srinivasan and D. W. Massaro, "Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English," *Language and Speech*, vol. 46, no. 1, pp. 1–22, 2003.

[36] V. C. Sendra, C. Kaland, M. Swerts, and P. Prieto, "Perceiving incredulity: The role of intonation and facial gestures," *Journal of Pragmatics*, vol. 47, no. 1, pp. 1–13, 2013.

[37] A. Schnitzler, *Round Dance*. Oxford University Press, 2009.

[38] R. Queneau, *Exercises in style*. New Directions Publishing, 2013.

[39] G. Bailly, T. Barbe, and H.-D. Wang, "Automatic labeling of large prosodic databases: Tools, methodology and links with a text-to-speech system," in *The ESCA Workshop on Speech Synthesis*, 1991, pp. 77–86.

[40] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[41] W. N. Campbell, "Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[42] G. Bailly and B. Holm, "SFC: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[43] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[44] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 396–401.

[45] J. Cohen, "A coefficient of agreement for nominal scale," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.