Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database

Carlos Busso and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089
busso@usc.edu, shri@sipi.usc.edu

Abstract

Recording high quality data is an important step in the study of emotions. Given the inherent limitations and complexities of the current approaches to capture natural emotions collected in real-life scenarios, the use of professional actors appears to be a viable research methodology. To avoid stereotypical realization of the emotions, better elicitation techniques rooted in theatrical performance are needed. Based on the lessons learned from the collection of the IEMOCAP database, this paper analyzes the advantages and disadvantages of two of the most appealing elicitation approaches: improvisation, and scripted dialogs. These methods are studied in terms of the lexical content, disfluencies, facial activity and emotional content. The results indicate that spontaneous sessions have higher levels of disfluencies and overlapped speech. Also, the emotional content seems to be more intense than in scripted sessions, as revealed in subjective evaluations.

Index Terms: Elicitation techniques, emotional databases, actor, acting styles, portrayal of emotions.

1. Introduction

One of the challenges in the study of emotions is the recording of high quality databases that reflect the expression of emotions observed in real-life human interaction. While collecting natural (non-acted) emotional data has been encouraged by the community, most of the suggested approaches present serious limitations such as ethical issues, restricted domain, or lack of control (i.e., type of sensors, modalities, quality of recording, noise background, lexical and emotional content) [1, 2, 3]. In this context, the use of acting in the study of human emotions appears to be a viable research methodology.

Instead of asking the actors or naïve speakers to read a set of given utterances portraying specific emotional categories, which has been the conventional approach, recent efforts have focused on studying better elicitation techniques to incorporate the strategies of theatrical performance [4, 5, 6]. Based on recent data collection experiences, we suggested some guidelines for recording emotional databases from actors [7]. Some of the important aspects that we proposed relate to emotional contextualization, social setting (interactions rather than monologues), the use of specific acting styles, and the selection of experienced participants. Following these guidelines, we recently recorded the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database at the *University of Southern California* (USC) [4]. This multimodal corpus comprises speech and detailed facial, head and hand motion. Two elicitation approaches were used in the design of this corpus: the use of plays (scripted sessions), and improvisation based on hypothetical scenarios (spontaneous sessions). These techniques, which are rooted in the core of acting training, provide alternative approaches to take advantage of the actor skills to record high quality emotional databases [6].

This paper analyzes the advantages and limitations of scripted and spontaneous techniques to elicit expressive speech.

Figure 1: VICON system with eight cameras. The right pannel shows an actor with the markers on the face and headband.

Based on a thorough analysis of the IEMOCAP database, these approaches are compared in terms of the resultant lexical content, disfluencies, facial activity, and emotional content. The results indicate that spontaneous sessions are characterized as having more disfluencies and overlapped speech (similar to spontaneous human interaction). Furthermore, the vocabulary size is higher and the emotion content is more intense in spontaneous sessions. The scripted sessions contain well-designed transitions between the emotional states of the actors. Also, the semantic context is fixed beforehand. The analysis of the data indicates that the selection of the elicitation technique for future collections should depend on the intended purpose of the analysis. We hope that the results presented here provide useful steps in this selection, toward recording better emotional corpora.

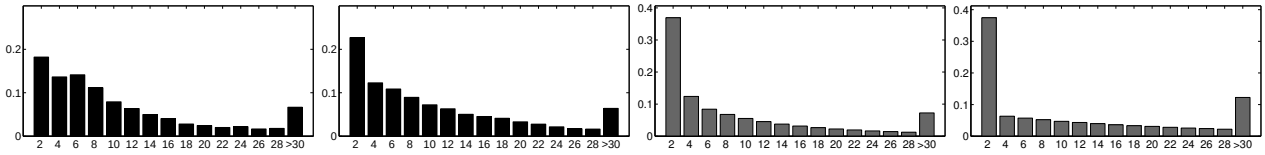
The paper is organized as follows. Section 2 describes the IEMOCAP corpus. Section 3 analyzes the data from spontaneous and scripted sessions. Finally, Section 4 gives the discussion, conclusions and future directions of this work.

2. IEMOCAP database

The IEMOCAP database was designed to study the patterns observed in different modalities during expressive speech communication. Although the target emotional categories were happiness, anger, sadness, frustration and neutral state, the corpus spans a wider emotional spectrum.

Seven professional actors and three senior students in the Drama Department at USC were each recorded in five dyadic sessions (male-female). Sixty-one markers were attached to one participant at a time (2 on the head, 53 on the face, and 6 on the hands). After recording the sessions, the markers were attached to the other actor, and the sessions were recorded again. A VICON motion capture system with eight cameras was used to track the location of the markers (Fig. 1).

As mentioned in Section 1, scripted and spontaneous scenarios were used as elicitation techniques. For the scripted sessions (55% of the corpus), three 10-minute plays with clear emotional content were selected (a theater professional supervised the selection). The actors were asked to memorize and rehearse the scripts. For the spontaneous sessions (45% of the corpus), eight hypothetical scenarios were selected based on the guidelines provided by Scherer *et al.* (e.g., loss of baggage, death of a friend) [8].



(a) Scripted session (b) Spontaneous sessions (c) Fisher (d) Switchboard-I

Figure 2: Histogram with the number of words per turns (in percentage) in the scripted and spontaneous sessions.

The database was transcribed and segmented at the dialog turn level. Six evaluators assessed the emotional content of the database in terms of emotional categories (happiness, sadness, anger, surprise, fear, disgust, frustration, excited, neutral state and other). The evaluators were allowed to assign more than one label. The evaluations were arranged such that three different raters assessed each turn. The inter-evaluator agreement is discussed in Section 3.3. In addition, two different raters are currently evaluating the emotional content of the corpus in terms of *valence* [1-negative, 5-positive], *activation* [1-calm, 5-excited], and *dominance* [1-weak, 5-strong] (VAD). At this point, approximately 85.5% of the data have been evaluated. Details of the corpus are given in [4]. The focus of this paper is to compare the scripted and spontaneous sessions.

3. Spontaneous versus scripted sessions

3.1. Lexical content

Although the scenarios used to record spontaneous sessions are restricted to specific domains, the actors were free to express themselves using their own language. In the scripted sessions, the vocabulary is restricted to the dialog contained in the plays. Although 45% of the IEMOCAP corpus corresponds to spontaneous sessions, the vocabulary size (2864) is two times larger than the vocabulary size of the scripted sessions (1489).

Figure 2 shows the histogram with the number of words per turn for the spontaneous and scripted sessions. For comparison purpose, the figure also shows the statistics for two well-known spontaneous (neutral) speech corpora: Switchboard-1 Telephone Speech, and Fisher English Training speech. Notice that these corpora were recorded from telephone conversations (not face-to-face interaction). Both the scripted and spontaneous elicitations in the IEMOCAP data show similar patterns as in these spontaneous telephony speech corpora. The scripted sessions tend to have longer utterances. In contrast, more than 20% of the spontaneous sessions contain only one word. The most frequently of these words were *yeah*, *okay*, *what*, and *right* which are commonly used as a back channel. Notice that these words are also among the most frequently words in the references Switchboard/Fisher corpora, when the turns contain only one word.

3.2. Disfluencies and overlapped speech

To estimate the disfluencies in the corpus, a naïve approach was implemented to find obvious repetitions in the transcripts. In addition, fillers (*uh*, *um*, *etc.*), discourse markers (*you know*, *well*, *etc.*) and explicit editing terms (*I mean*, *sorry*, *etc.*) were also automatically detected. Other types of disfluencies were not considered since they required more detailed human annotation (e.g., restart, repair). The same code was used to estimate the disfluencies in the Switchboard-I and Fisher corpora. Table 1 presents the percentage of the turns with disfluencies for the different emotion types. Notice that the emotional categories disgust, fear and surprise were not included in the figure since only few turns were labeled with these emotional categories (Fig. 4). Given the complexity of detecting disfluencies automatically, these results need to be considered only as a rough approximation of the true values (notice that even human raters disagree in labeling disfluencies). On average, 44% of the turns in spontaneous sessions present some type of disfluencies (only 30% for scripted sessions). During improvisation, the actors seem to use more disfluencies. Notice that the referenced speech Switchboard/Fisher corpora present even more disfluencies.

Table 1: Percentage of the turns that present some type of disfluencies in the corpus.

	All disf.	Fillers	Discourse marker	Editing term	Repetition
	Scripted sessions				
All	30.05%	7.42%	14.33%	4.43%	8.60%
Neutral	30.15%	4.94%	23.00%	2.39%	3.92%
Anger	30.38%	8.00%	10.09%	2.83%	13.28%
Happiness	31.37%	11.76%	9.80%	5.88%	7.52%
Sadness	23.74%	1.68%	11.55%	8.61%	7.56%
Frustration	31.92%	5.84%	14.30%	4.92%	11.56%
Excited	44.71%	20.63%	16.14%	5.03%	15.08%
	Spontaneous sessions				
All	44.00%	13.40%	20.94%	10.41%	13.84%
Neutral	52.96%	19.78%	28.44%	13.67%	13.95%
Anger	32.29%	4.86%	12.50%	6.94%	13.54%
Happiness	49.29%	21.99%	24.11%	8.87%	14.18%
Sadness	39.21%	5.77%	21.91%	12.36%	12.69%
Frustration	42.05%	6.71%	17.15%	12.71%	17.46%
Excited	43.52%	18.22%	18.52%	6.78%	12.05%
	References				
Fisher	54.43%	30.45%	22.38%	4.13%	15.58%
Switchboard-I	42.79%	28.40%	16.18%	1.90%	12.92%

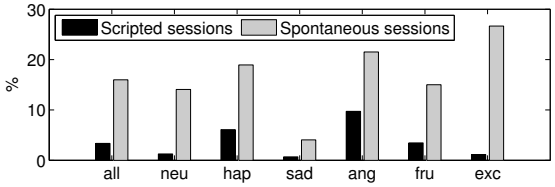


Figure 3: Percentage of overlapped speech in the corpus.

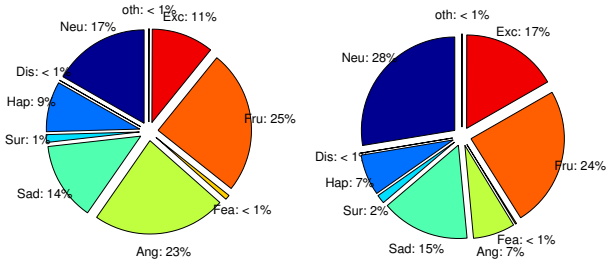
To study the percentage of overlapped speech in the corpus, we use the results from automatic forced alignment. The word boundaries of the segmented turns were projected to the dialog for each actor. Then, the percentage of overlapping speech was estimated for the scripted and spontaneous sessions. Figure 3 presents the results with respect to emotional categories. While the percentage of overlapped speech for the scripted sessions is on average lower than 5%, the value for spontaneous sessions is approximately 15%. This level of overlapping is similar to the level observed in spontaneous small group discussions (meetings) [9]. The figure also shows the strong emotional dependency in the overlapping rate (e.g., sadness versus anger).

3.3. Emotional content

As mentioned in Section 2, the emotional content of the IEMOCAP database has been marked by human subjective assessments. An interesting result from this evaluation is that the inter-evaluator agreement of the emotional categories, measured with the *Kappa* statistic, is higher for the spontaneous sessions ($\kappa = 0.34$) than for scripted sessions ($\kappa = 0.20$). In fact, 83.1% of the spontaneous turns were assigned an emotional label based on majority vote. However, in only 66.9% of the scripted turns the evaluators reached agreement. While the spontaneous sessions were designed to target a specific emotional category, the scripted sessions include progressive

Table 2: Confusion matrices from subjective emotional perceptual assessments. Ground truth defined by majority vote.

	Neu	Hap	Sad	Ang	Fru	Exc	Oth
	Scripted sessions						
Neutral	69.2	2.4	3.8	1.2	17.2	4.8	1.5
Happiness	8.3	69.1	1.6	0.0	1.6	17.7	1.8
Sadness	7.8	2.3	73.4	1.4	10.4	1.2	3.6
Anger	1.3	0.1	0.8	76.7	16.1	0.3	4.7
Frustration	5.8	0.2	3.6	13.6	72.2	0.8	3.9
Excited	5.9	12.9	0.3	0.1	4.1	74.1	2.6
	Spontaneous sessions						
Neutral	76.3	1.8	2.5	1.3	11.4	5.3	1.3
Happiness	9.0	70.4	0.0	0.0	0.3	18.8	1.4
Sadness	7.3	0.3	80.0	1.7	6.6	0.3	3.8
Anger	0.5	0.0	0.4	74.7	21.2	0.2	2.9
Frustration	8.2	0.1	3.5	9.4	75.1	0.5	3.3
Excited	2.9	18.0	0.0	0.0	0.1	76.1	2.9



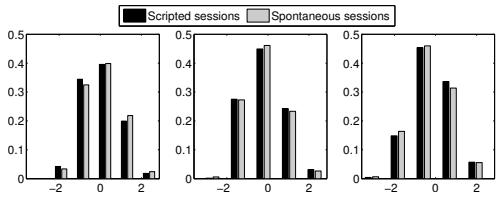
(a) Scripted session (b) Spontaneous sessions

Figure 4: Categorical emotion content of the IEMOCAP corpus.

changes from one emotional state to another (as dictated by the narrative content of the script). As a result, the scripted dialog approach, within a session, typically elicits a wider spectrum of emotional content. Thus, the boundaries between emotional categories become closer, decreasing the inter-evaluator agreement. This is especially clear in the transition turns in which the emotional state of the actors change from one emotion to another. Table 2 supports this observation. This table shows the confusion matrices between emotional categories for the scripted and spontaneous sessions obtained from the subjective evaluation when the emotional labels (majority vote) are considered as ground truth. Although the confusion between a pair of emotions is consistent across these elicitation techniques, the accuracy for spontaneous sessions is higher than in scripted sessions.

Figures 4 and 5 show the emotional content of the corpus. Figure 4 gives the distribution of the emotional categories in which the evaluators reached agreement (majority vote). For the targeted emotions, the results indicate that spontaneous sessions are more balanced than for scripted sessions. In fact, in this corpus, it was significantly simpler to select the spontaneous scenarios than the scripts with a balanced emotional content. Notice that the actors play a key role in the interpretation of the emotions conveyed in the text, so the designer loses some of the direct control (even when a director guides the recording).

Figure 5 gives the distribution of the VAD scores in each dimension. To compensate for inter-evaluator variation, z normalization was used. The results correspond to the portion of the database that has already been evaluated along valence and activation dimensions (85.5% of the sessions). Figure 6 combines the categorical and attribute-based evaluation. In this figure, the mean vector and covariance matrix for the valence and activation scores are estimated for each emotional category. Then, an ellipsoid defining confidence region at 50% is plotted. This figure shows that the emotions for scripted sessions (dashed line) are shifted toward the center of the valence-activation coordinate system (0,0). This result indicates that the emotions displayed during spontaneous sessions are more intense than the ones displayed during scripted sessions, and therefore, they are easier to recognize. While the actors concentrate on remembering the scripts, there is a risk that they may overlook the expression of emotions.



(a) Valence (b) Activation (c) Dominance

Figure 5: Attribute-based emotional evaluation. *Valence* [negative-possitive], *activation* [clam-excited] and *dominance* [weak-strong].

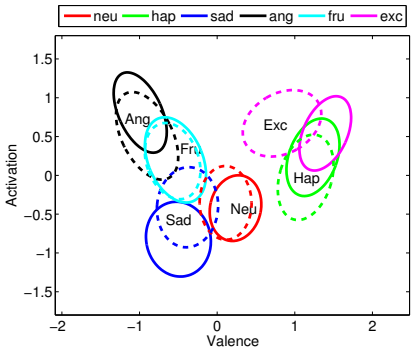


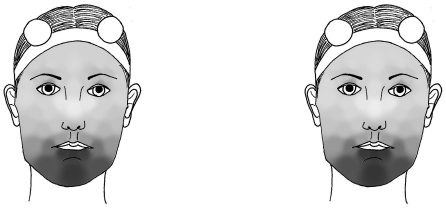
Figure 6: Clustering of emotional categories in terms of the valence and activation attributes. The results are presented for the spontaneous (solid line), and scripted session (dashed line).

From these results, while we are able to compare the two elicitation techniques, we are however not able to conclude which technique induces closer realizations of real-life emotions. Perceptual experiments to assess the naturalness of the emotional displays are needed to illuminate that question.

3.4. Facial expression

The activity of the facial markers was also studied to analyze whether the facial expression presents differences when the two elicitation techniques are used. After reconstructing the markers' trajectories, the facial markers were modified to compensate for the rotation and translation of the head. The resulting trajectories correspond only to facial movements (details are given in [4]).

For each marker, the facial activity was quantified by measuring the *displacement coefficient*. This value is defined as the average Euclidean distance between the marker trajectory, and its mean vector (at sentence-level). Figure 7 presents a graphical representation of the average results across all emotional categories by assigning darker colors to the facial regions closer to the markers with higher displacement coefficients. This figure suggests that the gross facial activity measure is similar for



(a) Scripted sessions (b) Spontaneous sessions

Figure 7: Average facial displacement in the face during speech. Darker areas represent higher facial activity.

Table 3: Average facial displacement (*fo*-Forehead, *le*-Left eye, *re*-Right eye, *lc*-Left cheek, *rc*-Right cheek, *na*-Nasolabial, and *ch*-Chin.)

	Face	fo	le	re	lc	rc	na	ch
	Scripted sessions							
All	1.58	0.95	1.22	1.26	1.09	1.13	1.63	3.82
Neutral	1.28	0.71	0.98	1.00	0.86	0.89	1.25	3.26
Anger	1.90	1.11	1.35	1.42	1.29	1.36	2.04	4.74
Happiness	1.69	0.98	1.26	1.30	1.24	1.30	1.77	4.03
Sadness	1.40	0.83	1.18	1.19	0.93	0.95	1.34	3.37
Frustration	1.53	0.94	1.19	1.25	1.04	1.08	1.52	3.66
Excited	1.97	1.33	1.53	1.64	1.39	1.42	2.10	4.41
	Spontaneous sessions							
All	1.53	0.88	1.11	1.17	1.05	1.09	1.57	3.86
Neutral	1.32	0.70	0.89	0.94	0.90	0.94	1.36	3.51
Anger	1.89	1.10	1.35	1.47	1.23	1.30	2.05	4.82
Happiness	1.61	0.94	1.17	1.16	1.14	1.18	1.69	3.99
Sadness	1.27	0.76	1.00	1.06	0.89	0.91	1.21	3.02
Frustration	1.55	0.88	1.15	1.18	1.02	1.07	1.55	3.95
Excited	1.74	1.01	1.26	1.34	1.21	1.25	1.79	4.34

spontaneous and scripted sessions. However, there are small but significant differences between them. Table 3 presents the average displacement coefficient in terms of emotional categories and facial regions, which are estimated by grouping the facial markers into seven areas: forehead (*fo*), left eye (*le*), right eye (*re*), left cheek (*lc*), right cheek (*rc*), nasolabial (*na*), and chin (*ch*). The motivation behind this particular face subdivision is based on the kinematics and symmetry of the face [10]. The table suggests that spontaneous sessions have higher displacement coefficient than scripted sessions.

A matched pairs test between the means reported in Table 3 was performed to measure whether the average displacement coefficient in each facial area for spontaneous and scripted sessions are statistically significant. The test reveals that the mean displacement for all facial areas excepting the nasolabial (*na*), and chin (*ch*) regions are significantly higher for spontaneous sessions ($df = 6$, $p < 0.05$). This result also supports the claim that spontaneous sessions produce more intense emotions. Notice that the upper and middle face regions, in which the differences are significant, are less constrained by articulatory goals, and therefore, can be simultaneously manipulated by the speaker to express other information such as emotions [11].

4. Discussion and conclusions

The main advantage of using improvisation of hypothetical scenarios to elicit emotions is that the resulting corpus is similar to natural speech in many aspects such as disfluencies, overlapped speech, and turn-taking statistics. Furthermore, the scenarios can be easily designed to achieve a more emotionally balanced corpus with higher vocabulary dimension. However, high levels of overlapped speech and disfluencies directly affect some of the pre-processing steps in the analysis of the corpus (e.g., estimation of speech features). In fact, in recent analyses of spontaneous speech, the turns with overlapped speech have been usually discarded, such as due to difficulties in doing pitch measurements. Another downside of spontaneous sessions is that it requires experienced actors willing to cooperate with each other.

Scripted sessions have the advantage that the lexical content is fixed beforehand. Therefore, the corpus can be analyzed in terms of the (known, defined) underlying lexical structure. Also, the low level of overlapped speech simplifies the post analysis steps. One drawback of this elicitation technique is that remembering the dialogs may affect the emotional quality conveyed by the actors. However, the use of experienced actors, and rehearsal sessions should mitigate this problem.

One important aspect considered in this paper is the emotional content of the corpus produced by spontaneous and scripted sessions. The results suggest that spontaneous sessions seem to elicit more intense emotions. This observation may explain the higher inter-evaluator agreement reached in spontaneous turns. Therefore, listeners can better recognize different emotional categories. On the other hand, the emotional boundaries in scripted sessions are more ambiguous. While this issue is challenging for analysis, it may better represent the emotions observed in real-life scenarios. However, at this point we cannot completely answer which techniques provide closer realizations of the emotions observed in real-life situations. As mentioned in Section 3.3, human perceptual experiments to assess the naturalness of the corpus can be conducted to address this question. Also, a speaker-dependent microanalysis of the facial expression may provide further evidence about the goodness of each elicitation technique. These are topics for future work.

The scripted and spontaneous recordings are only two of the approaches that capitalize, up to some extent, the richness of theatrical performance. We are planning to systematically analyze different acting styles, ranging from fully predetermined (everything is scripted) to fully undetermined (everything is improvised). Our ultimate goal in this direction is to identify better recording methodologies that resemble the emotions observed in real-life scenarios.

5. Acknowledgements

This research was supported in part by funds from the NSF, and the Department of the Army.

6. References

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

[2] K. Scherer and G. Ceschi, "Lost luggage: A field study of emotionantecedent appraisal," *Motivation and Emotion*, vol. 21, no. 3, pp. 211–235, September 1997.

[3] F. Schiel, S. Steininger, and U. Türk, "The SmartKom multimodal corpus at BAS," in *Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, May 2002.

[4] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. In press, 2008.

[5] T. Bänziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," in *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738*, A. Paiva, R. Prada, and R. Picard, Eds. Berlin, Germany: Springer-Verlag Press, September 2007, pp. 476–487.

[6] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actors process," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 6–10.

[7] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[8] K. Scherer, H. Wallbott, and A. Summerfield, *Experiencing emotion: A cross-cultural study*. Cambridge, U.K.: Cambridge University Press, 1986.

[9] C. Busso, P. Georgiou, and S. Narayanan, "Real-time monitoring of participants interaction in a meeting using audio-visual sensors," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, vol. 2, Honolulu, HI, USA, April 2007, pp. 685–688.

[10] C. Busso and S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *International Workshop on Multimedia Signal Processing (MMSP 2007)*, Chania, Crete, Greece, October 2007, pp. 43–47.

[11] ——, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.