Interplay between linguistic and affective goals in facial expression during emotional utterances

Carlos Busso , Shrikanth S. Narayanan

[1]Speech Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089

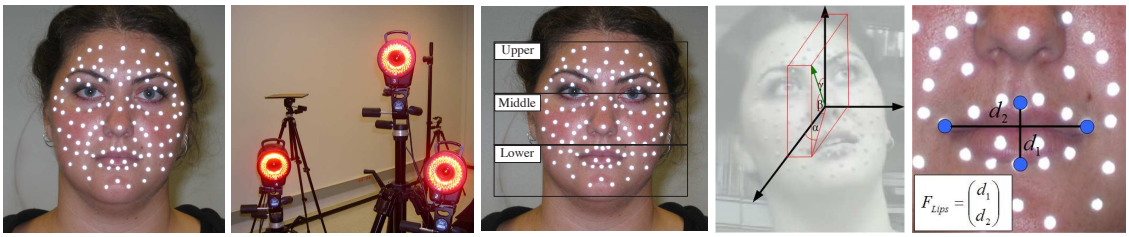
busso@usc.edu, shri@sipi.usc.edu

Abstract. *Communicative goals are simultaneously expressed through gestures and speech to convey messages enriched with valuable verbal and non-verbal clues. This paper analyzes and quantifies how linguistic and affective goals are reflected in facial expressions. Using a database recorded from an actress with markers attached to her face, the facial features during emotional speech were compared with the ones expressed during neutral speech. The results show that the facial activeness is mainly driven by articulatory processes. However, clear spatial-temporal patterns are observed during emotional speech, which indicate that emotional goals enhance and modulate facial expressions. The results also show that the upper face region has more degrees of freedom to convey non-verbal information than the lower face region, which is highly constrained by the underlying articulatory processes. These results are important toward understanding how humans communicate and interact.*

1. Introduction

During human interaction, gestures and speech are simultaneously used to express not only verbal information, but also important communicative clues that enrich, complement and clarify the conversation. Notable among these non-linguistic clues is the emotional state of the speaker, which is manifested through modulation of various communicative channels, including facial expressions (Ekman and Rosenberg, 1997), head motion (Busso et al., 2007), eyebrow movement (Ekman, 1979) and speech (Yildirim et al., 2004; Cowie and Cornelius, 2003; Scherer, 2003). The fact that many of these channels are actively or passively involved during the production of speech (verbal) and facial expressions (non-verbal) indicates that the linguistic and affective goals co-occur during human interaction. Since conflicts may appear between these communicative goals in their realization, some kind of central control system needs to buffer, prioritize and execute them in a coherent manner.

Many studies have shown that acoustic parameters such as the speech rate, speech duration, the fundamental frequency and the RMS energy change during emotional utterances (Yildirim et al., 2004; Cowie and Cornelius, 2003; Scherer, 2003). Articulatory parameters such as the tongue tip, jaw and lip also present more peripheral articulation during emotional speech compared to neutral speech (Lee et al., 2005, 2006). Similar results were reported by Nordstrand et al. (2003) and Caldognetto et al. (2003). In fact, these characteristic patterns during emotional speech have been used in facial animation to generate *viseme* models for expressive virtual agents (Bevacqua and Pelachaud, 2004). In spite of all this spatial-temporal emotional modulation, the linguistic goals are successfully fulfilled, which suggests that the communicative goals are prioritized according to their roles.



(a) (b) (c) (d) (e)

Figure 1: Audio-visual database collection and face parameterization. (a) facial marker layout, (b) motion capture system, (c) facial marker subdivision (upper, middle and lower face regions), (d) head motion features, and (e) lip features.

We believe that affective and linguistic goals interplay interchangeably as primary and secondary control. For instance, during speech, linguistic goals are prioritized over affective goals, which are restricted to enhance and complement the communicative channels, under the constraints imposed by the articulatory processes. However, during acoustic silence, articulatory processes are passive, so affective goals may dominate human gestures. Toward validating this hypothesis, this paper focuses on investigating the interplay between linguistic and affective goals in facial expressions. Although apex poses of expressive faces have been studied before (Ekman and Rosenberg, 1997), it is not clear how these communicative goals affect the face during emotional utterances. The goal of this study is to quantify both the degree of freedom of facial areas to express the underlying emotion, and the articulatory constraints imposed in the face during active speech.

In this study, an actress with markers attached to her face was asked to read semantically neutral sentences in four different emotional states: sadness, happiness, anger and neutral states. The results show that, compared to neutral speech, the activeness of the face, measured as the Euclidean distance between the facial features and their sentence-mean vector, increases for angry and happy sentences, and decreases for sad sentences. After aligning neutral and emotional sentences with similar semantic content with the use of *Dynamic Time Warping* (DTW), the facial features were compared frame-by-frame. The results show that the upper face region has more freedom to convey non-verbal information regardless of the verbal message, which contrasts with the lower face region, which is constrained by the articulatory processes. These results have important implications in areas such as emotion recognition, facial animation and in understanding speech production and perception.

2. Methodology

2.1. Audio-visual database

The database used in this analysis was recorded from an actress who was asked to read a custom-made, phoneme-balanced corpus four times expressing the following emotional states: sadness, happiness, anger and neutral state. The subject had 102 markers attached to her face (Fig. 1-(a)), which were tracked with a VICON motion capture system with three cameras (Fig. 1-(b)). The sampling rate of the system was 120 Hz, but it was downsampled to 60 Hz for the analysis. The audio was simultaneously recorded with a SHURE microphone at a sampling rate of 48 kHz. In total, 404 sentences were used in the analysis. The subject was instructed to express the recorded emotions as naturally as possible.

2.2. Feature Extraction

After the data were captured, the markers were translated by defining a nose marker at the local coordinate center. The head rotation was compensated by estimating a rotational matrix for each frame. Firstly, a neutral pose of the face was used as reference, which was reshaped as a 102 $\times$ 3 dimensional matrix, referred to as M_{ref}^T, with the location of each marker placed in each row. Then, for the frame t, a similar matrix (M_t) was created, following the same order used in the reference matrix. After that, *Singular Value Decomposition* (SVD), UDV^T, of the matrix $M_{ref}^T \cdot M_t$ was calculated (Eq. 1). The rotation matrix R_t was then calculated as VU^T (Eq. 2).

$$M_{ref}^T \cdot M_t = UDV^T \quad (1)$$

$$R_t = VU_T \quad (2)$$

In this paper, each of the facial markers, except the reference nose marker, was used as a facial feature. Three facial areas were defined to summarize the results presented in the tables of Section 3: Upper, middle and lower face regions (See Fig. 1-(c)). The upper face region includes all the markers over the eyes corresponding to the forehead and brow area. The lower face region consists of all the markers below the upper lip. The middle face region includes the markers between the upper and lower face areas corresponding to the cheeks.

In addition to the facial points, head, eyebrow and lip motion features were parameterized and included in the analysis. The head motion rotation was directly derived from the rotational matrix R_t (Fig. 1-(d) and Eq. 2). The eyebrow was parameterized with a two-dimensional feature vector, computed by subtracting the position of two markers chosen in the right eye. After that, the vectors were normalized to be in the range 0 to 1. Likewise, the lip features were estimated with a two-dimensional feature vector, which measures the opening (width and height) of the mouth, as shown in Figure 1-(e).

3. Emotional Modulation

3.1. Temporal emotional modulation

The temporal modulation in the speech of this database was analyzed in Yildirim et al. (2004). The results showed that the mean and variance values of the utterance durations for sadness, anger and happiness were higher than for neutral state. Likewise, the speaking rate had higher average values during emotional speech. With regard to vowel durations, the results showed that the mean values for anger and happiness were significantly higher than for sadness and neutral state.

To further analyze this temporal modulation at the sentence-level, *Dynamic Time Warping* (DTW) was used to estimate the temporal alignment between neutral and emotional speech, for the same sentences. This technique uses dynamic programming algorithms to find the lowest-cost alignment path between two signals. The slope of this path provides valuable information about the overall emotional temporal modulation. The median of the slopes in the alignment path between neutral and emotional speech are 1.14, 1.09 and 1.09 for sad, happy and angry speech, respectively. The results show that emotional utterances are more than 9% longer than the neutral utterances. Interestingly, sad sentences are longer than happy and angry sentences. Since the average phoneme duration for sadness is shorter than in happiness and anger, this result indicates that the inter-word silence to speech ratio was highly modulated during sad speech.

3.2. Spatial emotional modulation

During emotional speech, areas in the face present different levels of movement activity. This section analyzes and quantifies the activeness of facial areas and evaluates whether the activeness

Table 1: Average facial activeness during emotional utterances

Facial Area	N	S	H	A
Head Motion	1.92	4.11	4.65	4.53
Eyebrow	0.05	0.06	0.11	0.12
Lip	4.51	3.55	6.22	7.28
Upper region	0.66	0.79	1.47	1.47
Middle region	0.88	0.88	1.43	1.56
Lower region	3.15	2.46	4.21	4.59

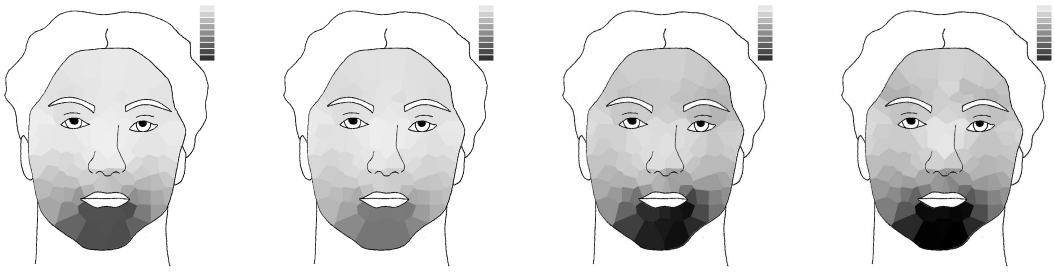
levels are affected by emotional goals. It also includes comparison of the patterns in the facial expressions between neutral and emotional utterances.

To measure the activeness of the face in each sentence, the *motion coefficient* Ψ, given in Equation 3, was calculated. This coefficient is defined as the Euclidean distance between the facial features and the mean vector computed at sentence-level,

$$\Psi_u = \frac{1}{T_u} \sum_{i=1}^{T_u} D_{eq}(\vec{X}_i^u, \vec{\mu}^u) \tag{3}$$

where T_u is the number of frames in sentence u, $\vec{\mu^u}$ is the mean vector, and D_{eq} is the Euclidean distance.

The average results for the *motion coefficient* over the sentences are presented in Table 1 for the parametric features corresponding to head, eyebrow and lip motion, in terms of emotional categories. The table also shows the aggregated results for the markers contained in the upper, lower and middle face regions. Figure 2 shows a visual representations of the *motion coefficient* results. After calculating the activeness of each facial marker, the values were normalized to be in the range between 0 and 1. After that, gray-scale colors were assigned to the markers according to their values (1:black, 0:white). Finally, the Voronoi cells centered in the markers were colored according to the assigned gray-scale values.



(a) Neutral (b) Sad (c) Happy (d) Angry

Figure 2: Facial activeness during speech. The figures show that during speech, the lower face region is the most active region in the face. It also shows important inter-emotional differences (see Section 3.2 for details).

The results shown in Table 1 and Figure 2 reveal that the lower face region has the highest activeness levels. In fact, in neutral speech this area is four times more active than the upper face area. Since this area is directly connected with the production of the speech, this result suggests that the articulatory processes play a crucial role in the movement of the face. This result supports the hypothesis that linguistic goals have priority during active speech.

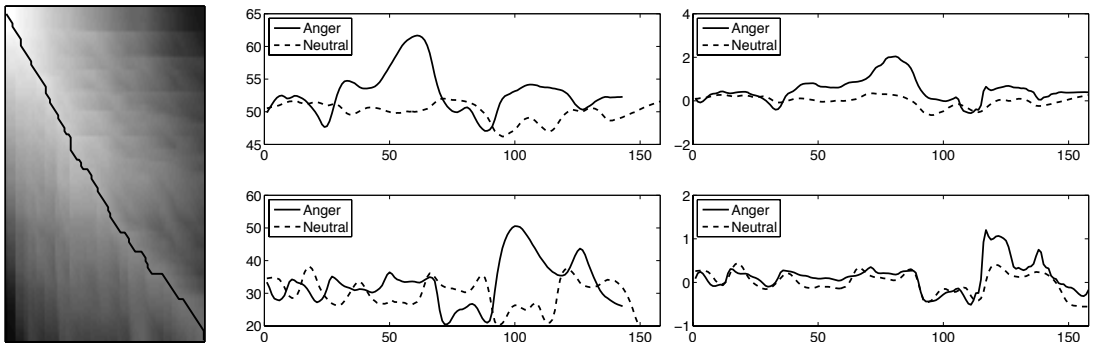


Figure 3: Dynamic time warping to align neutral and emotional facial features. (a) optimum alignment path (b) original lip features (c) normalized and warped lip features.

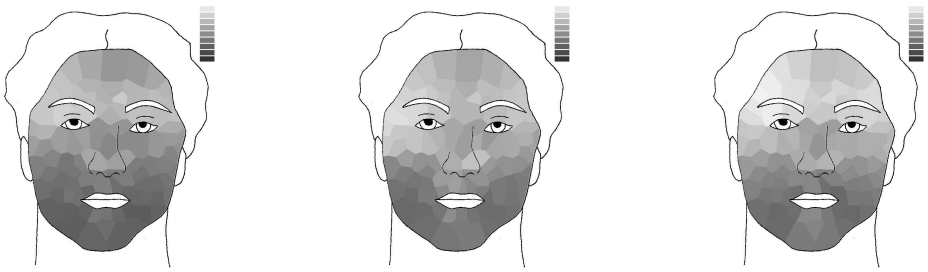
The results also indicate that the activeness levels change during emotional speech. The average *motion coefficient* for happiness and anger is more than 30% higher than in the neutral case. Also, the activeness in the lower face region for sadness decreases 20% compared with the activeness in neutral state. These results reveal that emotional modulation affects the activeness in the face, which agrees with previous work (Lee et al., 2005, 2006; Nordstrand et al., 2003; Caldognetto et al., 2003; Bevacqua and Pelachaud, 2004). Notice that the upper face region presents the highest relative increments for happiness and anger compared to the neutral case (120%). This result suggests that valuable non-verbal information is conveyed in this area.

The *motion coefficient* provides an overall measure of the emotional influence in the face during affective speech. To study this spatial-emotional modulation in more detail, the neutral and emotional facial features for the same sentences were compared frame-by-frame. As discussed in Section 3.1, there are temporal differences between neutral and emotional utterances that need to be taken into account before the analysis. The alignment paths estimated with DTW were used to match the neutral and emotional frames. Figure 3 shows an example with the results between neutral and anger for the lip features extracted during the sentence "*That dress looks like it comes from Asia*". Notice that repetitions of the same sentence will generate differences not only in the gestures, but also in the speech. However, since the emotional content is the most important variable that is changed, the difference can be associated mainly with emotional modulation.

After aligning the utterances, Pearson's correlation was calculated between the neutral and warped versions of the emotional facial features. The goal of this experiment is to quantify how free the facial areas are to convey emotional information. Since the linguistic content is the same, high correlation levels will be associated with facial areas with low degree of freedom to convey non-verbal messages, and vice versa.

The median results of the correlation levels are presented in Table 2 and Figure 4. Figure 4 shows a graphical representation of the results, following the same procedure used for Figure 2. The results clearly indicate that the lower facial region presents the highest correlation levels. Thus, this area is not freely able to convey emotion due to the underlying articulatory constraints. In contrast, the correlation for the upper face region is very low which indicates that this area can communicate non-verbal information regardless of the linguistic content. The same results are observed for head and eyebrow motion, which can be freely modulated to express emotional goals.

In addition to Pearson's correlation, the Euclidean distance between the neutral ($F_t^{(neu)}$)

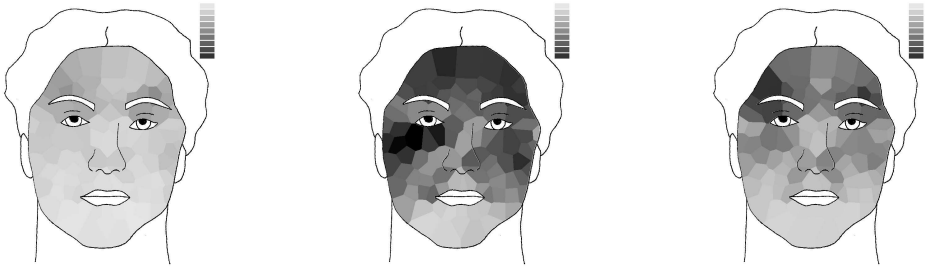


(a) Neutral-Sad (b) Neutral-Happy (c) Neutral-Angry

Figure 4: Graphical representation of correlation levels between neutral and warped version of emotional facial features. Dark areas represent high correlation values, which imply low degree of freedom to convey non-verbal information

Table 2: Frame-by-Frame analysis between emotional and neutral facial features

Facial Area	Pearson's correlation			Euclidean distance		
	Neu-Sad	Neu-Hap	Neu-Ang	Neu-Sad	Neu-Hap	Neu-Ang
Head Motion	0.24	0.25	0.17	4.28	3.83	3.44
Eyebrow	0.25	0.15	0.07	0.69	2.56	1.31
Lip	0.54	0.50	0.53	0.38	1.61	0.82
Upper region	0.27	0.24	0.15	1.08	2.49	2.02
Middle region	0.46	0.38	0.37	0.63	2.12	1.27
Lower region	0.58	0.52	0.53	0.46	0.95	0.71



(a) Neutral-Sad (b) Neutral-Happy (c) Neutral-Angry

Figure 5: Graphical representation of Euclidean distance between the normalized neutral, ($\widehat{F}_t^{(neu)}$) and emotional ($\widehat{F}_t^{(emo)}$) facial features. Dark areas represent large distances, which imply that the areas are not driven by articulatory processes.

and the warped version of the emotional ($\widetilde{F}_t^{(emo)}$) facial features were computed. Since the facial features considered here have different movement ranges, they need to be normalized before they can be directly compared. Firstly, the mean vector of the neutral facial features, $\vec{\mu}_t^{(neu)}$, was removed from both $F_t^{(neu)}$ and $\widetilde{F}_t^{(emo)}$. Then, the amplitude of $F_t^{(neu)}$ was scaled by $\alpha_t^{(neu)}$, such that its range was 1. Finally, the amplitude of $\widetilde{F}_t^{(emo)}$ was also scaled by the same factor $\alpha_t^{(neu)}$. Figure 3-(c) shows an example of the results after this normalization (see y-axis).

$$\widehat{F}_t^{(neu)} = (F_t^{(neu)} - \vec{\mu}_t^{(neu)}) \cdot \alpha_t^{(neu)} \quad (4)$$

$$\widehat{F}_t^{(emo)} = (\widetilde{F}_t^{(emo)} - \vec{\mu}_t^{(neu)}) \cdot \alpha_t^{(neu)} \quad (5)$$

Table 2 shows the median Euclidean distance between the normalized neutral and emo-

tional facial features, in terms of emotional categories. Similar to the Figures 2 and 4, Figure 5 shows a graphical representation of the results. Contrary to the Pearson's correlation results, high values indicate that facial features are more independent of the articulation, and vice versa. These results also show that the upper face region presents the highest differences between the neutral and emotional expression, supporting the fact that emotional goals control this area. Table 2 also quantify the levels of emotional modulation in the face. For this subject, happiness, followed by anger, seems to have the highest indices of spatial-facial emotional modulation.

4. Discussions and conclusions

This paper analyzed the spatial-temporal emotional modulation in the face during active speech. It also presented evidence about the interplay between linguistic and affective goals in facial expression. The results have important implications in areas such as emotion perception, emotion recognition, speech production and facial animation.

The results regarding the activeness of the face showed that facial motion is mainly driven by the articulatory processes. The results also showed that the activeness levels are affected by emotional modulation. Compared to the neutral case, the activeness levels increase for happiness and anger, and decrease for sadness.

The frame-by-frame analysis comparisons between the neutral and warped emotional facial features indicate that the upper face region presents more freedom than the lower face region, which is highly constrained by the underlying articulatory processes, to convey non-verbal information such as the emotional content. These results explain why the upper face region is perceptively the most important facial region to detect visual prominences (Swerts and Krahmer, 2006; Lansing and McConkie, 1999). They also explain why the upper and lower face regions are sufficient to accurately recognize human emotions (Busso et al., 2004). These results suggest that facial emotion recognition systems during active speech should focus primarily in this area, because it is not constrained by the linguistic content. Also, for human-like facial animations, this facial area should be properly modeled and rendered to convey more realistic emotional representations.

In the lower face region, the results reveal that linguistic and affective goals co-occur during active speech. Further analyses need to be conducted to evaluate what happens when conflict between these communicative channels occur. In these cases, areas with more degrees of freedom to convey non-verbal information such as head and eyebrow motion may be used to simultaneously achieve these communicative goals.

In this paper, the facial gestures during active speech were analyzed. An interesting question is how the patterns change during acoustic silence, in which the affective goals can be expressed without articulatory constraints. Another interesting question is how to model this spatial-temporal emotional modulation. These are some of the questions that we are addressing in our ongoing research.

Notice that the facial gestures expressed by one subject were analyzed. We are currently collecting more data from more subjects to generalize and validate these results. Understanding how human beings express communicative goals will help us to design cognitive interfaces more able to recognize and respond to user intentions and goals.

Acknowledgment

This research was supported in part by funds from the NSF (through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army and a MURI award from ONR. Any opinions, findings and conclusions or recommendations expressed in this paper are

those of the authors and do not necessarily reflect the views of the funding agencies. The authors thank colleagues in the emotion research group for their valuable comments.

References

Bevacqua, E. and Pelachaud, C. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 15(3-4):297– 304, July 2004.

Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, In Press, March 2007.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Sixth International Conference on Multimodal Interfaces ICMI 2004*, pages 205–211, State College, PA, 2004. ACM Press.

Caldognetto, E. M., Cosi, P., Drioli, C., Tisato, G., and Cavicchio, F. Coproduction of speech and emotions: Visual and acoustic modifications of some phonetic labial targets. In *Audio Visual Speech Processing (AVSP 03)*, pages 209– 214, S. Jorioz, France, September 2003.

Cowie, R. and Cornelius, R. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, April 2003.

Ekman, P. About brows: emotional and conversational signals. In von Cranach, M., Foppa, K., Lepenies, W., and Ploog, D., editors, *Human ethology: claims and limits of a new discipline*, pages 169– 202. Cambridge University Press, New York, NY, USA, 1979.

Ekman, P. and Rosenberg, E. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York, NY, USA, 1997. ISBN 0-19-510446-3.

Lansing, C. and McConkie, G. Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, 42:526– 539, June 1999.

Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., and Narayanan, S. A study of emotional speech articulation using a fast magnetic resonance imaging technique. In *International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006.

Lee, S., Yildirim, S., Kazemzadeh, A., and Narayanan, S. An articulatory study of emotional speech production. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 497–500, Lisbon, Portugal, September 2005.

Nordstrand, M., Svanfeldt, G., Granström, B., and House, D. Measurements of articulatory variations and communicative signals in expressive speech. In *Audio Visual Speech Processing (AVSP 03)*, pages 233– 237, S. Jorioz, France, September 2003.

Scherer, K. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227– 256, April 2003.

Swerts, M. and Krahmer, E. The importance of different facial areas for signalling visual prominence. In *International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006.

Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. An acoustic study of emotions expressed in speech. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.