Joint Analysis of the Emotional Fingerprint in the Face and Speech: A single subject study

Carlos Busso and Shrikanth S. Narayanan
Speech Analysis and Interpretation Laboratory (SAIL)
Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089
Email: busso@usc.edu, shri@sipi.usc.edu

Abstract— **In daily human interaction, speech and gestures are used to express an intended message, enriched with verbal and non-verbal information. Although many communicative goals are simultaneously encoded using the same modalities such as the face or the voice, listeners are generally good at decoding each aspect of the message. This encoding process includes an underlying interplay between communicative goals and channels, which is yet not well understood. In this direction, this paper explores the interplay between linguistic and affective goals in speech and facial expression. We hypothesize that when one modality is constrained by the articulatory speech process, other channels with more degrees of freedom are used to convey the emotions. The results presented here support this hypothesis, since it is observed that facial expression and prosodic speech tend to have a stronger emotional modulation when the vocal tract is physically constrained by the articulation to convey other linguistic communicative goals.**

I. Introduction

Communicative channels, such as gestures, facial expressions and speech, are jointly invoked to convey and express an intended message, which includes not only the spoken content (verbal channel), but also implicit cues (non-verbal channel) that enrich human communication. Notable among these cues are the emotional states, which play a crucial role in daily interaction.

Under the Brunswik's lens model, adapted by Scherer to study emotions [1], different communicative modalities are used to encode the affective state, such as speech [2], facial expression [3], head motion [4] and even body posture [5]. The nature of this encoding is non-trivial, since the same modalities are simultaneously used to convey other communicative goals. However, listeners are particularly good at decoding each aspect of the message, even when the cues are only slightly expressed.

The emotional modulation in the acoustic and articulatory domain in speech has been analyzed at the phoneme level [6], [7]. The results showed that low vowels, such as /a/, with less restricted tongue position presented stronger emotional modulation than high vowels, such as /i/. In [8], we showed that some broad phonetic classes, such as front vowels, have stronger emotional variability in the spectral domain than other phonetic classes, such as nasal sounds. These observations suggest an interaction in the acoustic domain between linguistic and affective goals. Likewise, the interplay between articulation and emotions in facial expressions has been analyzed. In our previous work, we studied the facial areas that are commonly used to convey emotions during expressive utterances [9]. The results showed that the upper face region, such as the forehead, has more degrees of freedom to communicate non-verbal cues than the orofacial area, which is constrained by the articulatory process. Although the encoding of emotions has been individually studied for these modalities, a joint analysis of the emotional fingerprints in different modalities is needed to discover the underlying emotional encoding process. Only by simultaneously studying the emotional fingerprint in different modalities, will we understand the interplay between different channels observed during expressive utterances.

This paper studies the emotional fingerprint in facial expressions and speech during expressive utterances. In particular, our hypothesis is that when some communicative channels are used to fulfill linguistic goals, other modalities with less restrictive articulatory constraints are used to convey affective goals. These ideas resemble the *water-filling* principle in communication, in which the power (bits) is first assigned to the channels with lowest noise (variance) [10]. Here, the *emotional bits* are assigned to the modalities that are not used to convey other communicative goals.

Toward exploring this hypothesis, the emotional interplay observed in speech spectral envelope and prosodic features, as well as in facial expressions is jointly analyzed. In the proposed approach, the phonetic boundaries are used as a basic segmental unit across modalities. Therefore, the instantaneous behavior that is simultaneously conveyed in different communicative channels can be compared. Then, a measure of the emotional content is estimated based on the differences between the features extracted from neutral and emotional utterances. Therefore, the emotional modulation in each modality can be quantified, to analyze how the emotions are displayed. The database used in the analysis was recorded from a single subject, using audio and facial motion capture (marker based), which provide detailed facial information. In addition to neutral speech, the database is comprised of sad, angry and happy utterances. The results reported here support our hypothesis, since it is observed that the pitch and the facial areas tend to have a stronger emotional modulation for the phonemes that are physically constrained by articulation, such as nasal sounds. The analysis proposed is novel and provides thoughtful insights to help design better emotional models, capable of capturing the complex behavior observed in human interaction.

The paper is organized as follows. Section II describes the proposed approach to jointly study the emotional fingerprints in speech and facial expressions. The audio-visual database is also presented. Section III provides the results for the emotional modulation observed in speech and facial expressions. Finally, Section IV concludes the paper with the discussion and future directions of this work.

II. Proposed method

A. Motivation

In our recent work, acoustically neutral models were used to measure the degree of similarity between the input speech and the reference neutral models [8]. The aim of this approach was to segment emotional speech, since this speech would likely be mismatched with the neutral acoustic models. In the approach, the emotionally-neutral TIMIT corpus was used to build standard *Hidden Markov Models* (HMMs), trained with *Mel Frequency Bank* (MFB) outputs, for broad phonetic classes that share similar articulation configuration

TABLE I

BROAD PHONETIC CLASSES

	Description	Phonemes
B	Mid/back vowels	ax ah axh ax-h uw uh ao aa ux
D	Diphthong	ey ay oy aw ow
F	Front vowels	iy ih eh ae ix
L	Liquid and glide	l el r y w er axr
N	Nasal	m n en ng em nx eng
T	Stop	b d dx g p t k pcl tcl kcl qcl bcl dcl gcl q epi
C	Fricatives	ch j jh dh z zh v f th s sh hh hv
S	Silence	sil h# #h pau

(see Table I). After recognition, the likelihood scores provided by the Viterbi algorithm were used as a *fitness measure*.

An interesting result from this work was that some phonetic classes have more degrees of freedom to express emotions. These results are reproduced in Figure 1, for the speech taken from the database presented in Section II-C (the original table presented in [8] was created with another corpus). This figure shows the mean and standard deviation of the likelihood scores for the emotional categories, in terms of the broad phonetic classes. As can be observed, the emotional fingerprint in these spectral features is stronger in phoneme classes such as front vowels. In contrast, the likelihood scores for the nasal sounds are similar across emotional categories, suggesting that for these phonemes this vocal tract feature does not have enough degrees of freedom to convey the affective goals.

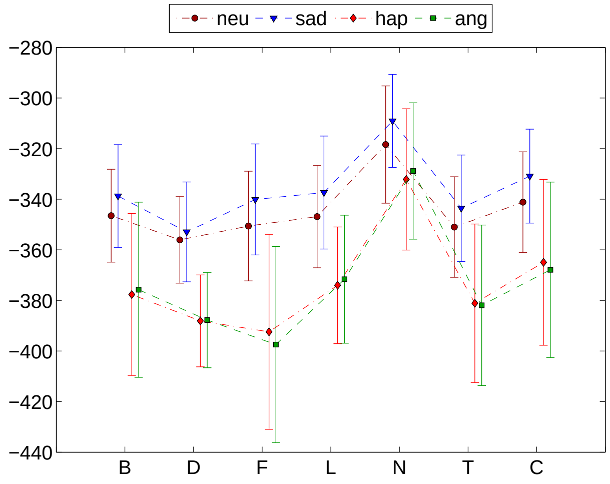


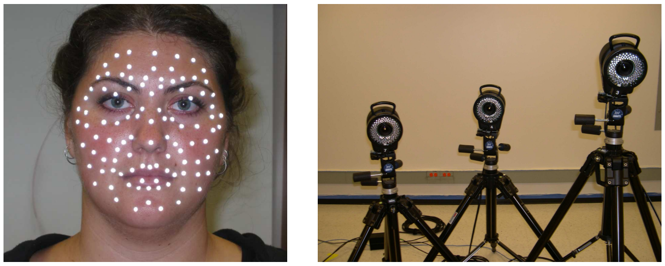
Fig. 1. Mean and standard deviation of the likelihood scores in terms of broad phonetic classes, evaluated with the corpus presented in Section II-C. The neutral models were trained with MFB features.

Motivated by these findings, this paper explores the patterns shown in other modalities to discover whether other communicative channels compensate the articulatory restrictions shown in these spectral features. In this work, we also grouped the phonemes according to the broad phonetic classes described in Table I (based on *manner* of articulation).

B. Methodology

The proposed approach consists of projecting the phonetic segmental boundaries to other communicative channels and analyzing the results in terms of these acoustic units. Specifically, the modalities that we examined were facial expressions and prosodic features such as pitch and energy.

Notice that the modalities considered here have different time scales and are known to be asynchronous. For example, it has been reported that in the orofacial area there may be a phase difference of hundreds of milliseconds, because of the co-articulation process and articulator inertia [11]. However, in this work we are interested in analyzing the instantaneous behavior simultaneously displayed during the duration of the phonemes. Therefore, it is theoretically consistent to project the acoustic boundaries to other modalities, in the context of the proposed method.



(a) (b)

Fig. 2. Audio-visual database collection. (a) The figure shows the facial marker layout, and (b) the figure shows the motion capture system.

For the analysis, the results for the emotional utterances are compared with the results obtained with neutral utterances to quantify the relative changes produced by the emotional fingerprints in different modalities.

C. Audio-visual database and features

The audio-visual database used for the analysis was recorded from one actress with 102 facial markers (Fig. 2-a). She was asked to read a custom-made, phoneme-balanced corpus four times, expressing different emotions at each reading: sadness, anger, happiness, and neutral state. A VICON motion capture system with three cameras was used to track the 3-d spatial position of each marker at a sampling rate of 120 frames per second (Fig. 2-b). Simultaneously, the speech was recorded at 48Khz with a close talking SHURE microphone. The detailed facial information and the size of the corpus (more than 600 sentences) are particularly useful for the kind of analysis presented here.

After the data was collected, the facial markers were translated to make a nose marker the center of the coordinate system. Then, the head motion rotation was compensated by using a technique based on *Singular value Decomposition* (SVD) [12]. The details can be found in [4].

For the analysis, each marker, except the reference nose marker, was used as a facial feature. To display the results in the figures and tables, the markers were aggregated in facial areas. This subdivision is based on the data-driven QR factorization algorithm presented by Lucero *et al.*, in which *independent* markers selected by the technique are used as a basis for the facial kinematics [13]. Here, the QR factorization algorithm was applied to our data. For clustering, the *dependent* markers were associated with the *independent* marker with highest weight in the linear combination. Figure 3-a shows the results of the grouping. This data-driven facial division was modified by including symmetry and muscle activity considerations. Figure 3-b shows the seven facial subdivisions used in this paper: forehead (F1), left eye (F2), right eye (F3), left cheek (left intraorbital triangle) (F4), right cheek (right intraorbital triangle) (F5), nasolabial (F6), and chin (F7).

The acoustic signal was downsampled to 16KHz. The fundamental frequency and the RMS energy were then estimated with the Praat speech processing software [14]. The phoneme transcription was obtained with forced alignment, implemented with the HTK toolkit [15].

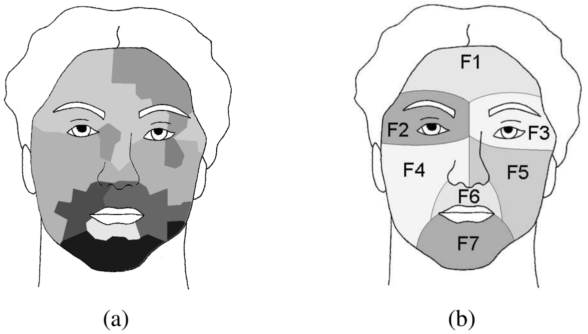


Fig. 3. (a) Data-driven approach to cluster the facial region [13]. (b) Facial subdivition: (F1) Forehead, (F2) Left eye, (F3) Right eye, (F4) Left cheek, (F5) Right cheek, (F6) Nasolabial, and (F7) Chin.

TABLE II
LIKELIHOOD SCORES FOR MFB (*Sad* = SADNESS, *Ang* = ANGER, *Hap* = HAPPINESS, *Neu* = NEUTRAL)

	Average likelihood scores				Ratio emotion/neutral		
	Sad	Ang	Hap	Nen	Sad/Neu	Ang/Neu	Hap/Neu
B	-338.7	-375.8	-377.7	-346.5	0.98	1.08	1.09
D	-352.9	-387.8	-388.1	-356.1	0.99	1.09	1.09
F	-340.1	-397.5	-392.4	-350.6	0.97	1.13	1.12
L	-337.4	-371.6	-374.1	-346.9	0.97	1.07	1.08
N	-309.1	-328.9	-332.2	-318.4	0.97	1.03	1.04
T	-343.6	-381.9	-381.1	-351.0	0.98	1.09	1.09
C	-330.9	-367.9	-365.0	-341.2	0.97	1.08	1.07

III. EXPERIMENTAL RESULTS

This section presents the analysis of the emotional fingerprints observed in speech and facial expressions. Sections III-A and III-B describe the results for acoustic speech features, and Section III-C gives the analysis for facial expressions.

A. Spectral speech features

As mentioned in Section II-A, an alternative approach to quantize the emotional modulation for spectral features is to measure how well the emotional input speech matches the reference spectral-based neutral models. Table II shows the average likelihood scores for the four emotional categories in terms of the phonetic classes (Table I). These values were obtained after recognizing these broad phonetic classes, with the neutral models built with the MFB outputs (trained with the TIMIT database). In addition to the average likelihood scores, Table II shows the ratio of the average values for the emotional categories (sadness, anger and happiness) to the ones for the neutral set. These values are particularly useful to quantify the relative changes between the neutral and emotional speech.

Table II quantifies the results observed in Figure 1. For emotions with a high level of arousal, such as happiness and anger, the magnitude of the average likelihood scores for vowels (F, B) increases approximately 10%, compared with the values observed in the neutral case. This results agree with previous work that have shown that the tongue tip, jaw and lip present more peripheral articulation for vowels during expressive speech when compared to neutral speech [7], [16], [17]. In contrast, the magnitude of the likelihood scores for the nasal sounds (N) in those emotions increases only 3% compared with the neutral case. This result suggests that physical constraints in the articulatory domain restrict the degrees of freedom to simultaneously express other communicative goals, such as emotions.

Another approach to measure the differences of the likelihood scores between the emotional and neutral cases is to compare not only their averages, but also their distributions. For this purpose, the *probability mass functions* (PMFs) were approximated with the histograms of the likelihood scores. The width of the bin was chosen such that the average number of samples per bin was higher than 100. Then, the *Kullback-Leibler Divergence* (KLD) was used to measure the distances between the three emotion distributions (sadness, anger and happiness) and the neutral distribution. Figure 4 shows the results, which complement the aforementioned observations. In the figure, longer bars mean larger KLD, which implies stronger differences between the distributions.

TABLE III
AVERAGE VALUES OF PITCH AND ENERGY DURING BROAD PHONETIC CLASSES (*Sad* = SADNESS, *Ang* = ANGER, *Hap* = HAPPINESS, *Neu* = NEUTRAL)

	Average value				Ratio emotion/neutral		
	Sad	Ang	Hap	Neu	Sad/Neu	Ang/Neu	Hap/Neu
	Pitch						
B	208.2	232.3	247.2	194.2	1.07	1.20	1.27
D	203.7	227.2	248.6	192.0	1.06	1.18	1.30
F	209.3	234.2	240.2	196.1	1.07	1.19	1.22
L	202.8	215.5	235.0	194.7	1.04	1.11	1.21
N	194.8	213.3	236.9	184.8	1.05	1.15	1.28
T	217.1	234.7	276.4	197.4	1.10	1.19	1.40
C	209.5	233.7	254.2	196.1	1.17	1.19	1.30
	Energy						
B	71.81	82.38	82.38	75.32	0.95	1.09	1.09
D	72.99	83.12	83.46	76.79	0.95	1.08	1.09
F	71.72	81.41	81.86	74.79	0.96	1.09	1.09
L	71.00	79.71	80.31	74.21	0.96	1.07	1.08
N	68.65	76.79	77.58	70.60	0.97	1.09	1.10
T	62.21	70.70	72.35	64.18	0.97	1.10	1.13
C	63.22	71.05	72.23	65.17	0.97	1.09	1.11

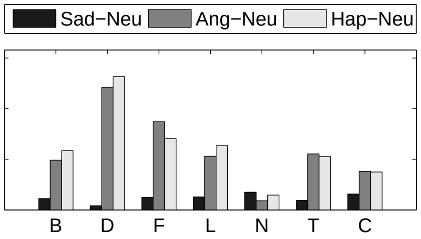


Fig. 4. KLD between the distributions for the neutral and emotional likelihood scores values. Vowels present stronger emotional modulation in the spectral acoustic domain.

B. Prosodic speech features

In this section, we are interested in analyzing the behavior of the pitch and energy during the acoustic boundaries projected from the broad phonetic classes, estimated with forced alignment (Section II-C). These features predominantly describe the source of speech rather than the vocal tract configuration. Notice that pitch values are observed only for voiced phonemes. Therefore, the results reported here for the pitch correspond only to voiced segments.

Table III shows the average values for the pitch and energy in terms of the emotion, for the phonetic classes. Similar to Figure 4, Figure 5 shows the KLD between the emotional and neutral distribution of the pitch and energy. For the pitch, the results reveal a strong emotional modulation for stop and fricatives phonemes, especially for happiness (see also Table III). Notice that in the experiments reported here, those classes do not present strong modulation in the spectral domain (see

Fig. 4). For the energy, Figure 5 shows that the KLD for vowels (classes B, D and F) are higher than other classes.

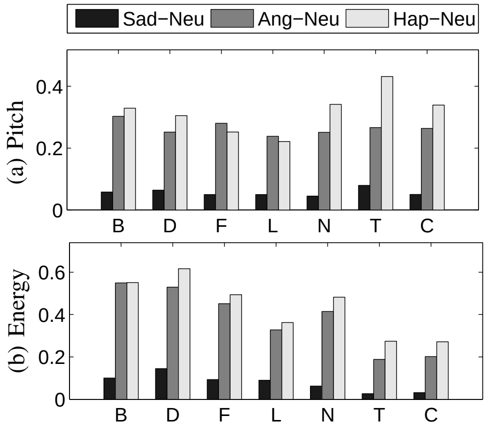


Fig. 5. KLD between the distributions of the neutral and emotional values of the pitch (a) and energy (b).

In summary, while the spectral and energy features tend to have higher emotional modulation for vowels, the pitch, which is relatively independent of the vocal tract configuration, is used to convey emotions in voiced stop and fricative phonemes.

C. Facial expression

In this analysis, we are also interested in studying the emotional modulation in facial expressions that is simultaneously displayed during the phonetic segmental boundaries (also estimated with forced alignment). To quantize the facial activeness, the *displacement coefficient* is defined and calculated for each facial marker. This coefficient is defined as the Euclidean distance between the marker position and the mean vector computed at the sentence-level,

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

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.

After estimating the *displacement coefficient* for the markers, the average value for each facial area was calculated for each phoneme segment. This procedure was applied for each emotional category. For sake of simplicity, only the ration between emotional and neutral results are reported in Table IV. This table shows that the stronger emotional modulation for happiness and anger is observed in nasal sounds. A matched pairs test [18] between the rations reported in Table IV was performed to see whether this result is statistically significant. This statistical test removes the differences observed between facial areas. Therefore, the significance of the emotional effect during the phonetic classes can be measured. The test revealed that the ratios for the class N is significantly higher than the ratios of any other classes ($df = 20$, $p < 0.05$), with the exception of class B.

Similar to Figures 4 and 5, Figure 6 gives the KLD between the distributions of the emotional and neutral *displacement coefficients*. This figure also shows that the nasal sounds are among the phonetic classes with higher emotional modulation.

Figure 6 also shows that the emotional modulation in the upper region of the face (areas F1 to F3) is higher than in the orofacial area (areas F6 and F7). This result agrees with our previous work which showed that the forehead area is less constrained by the articulation process, and therefore, it has more freedom to convey other communicative goals, such as emotions [9].

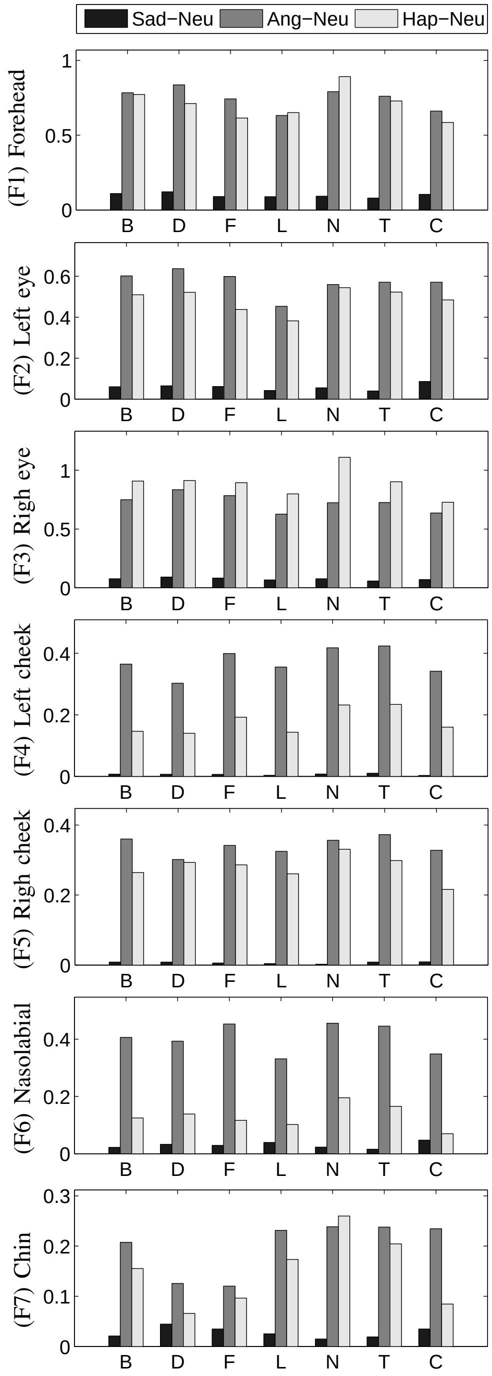


Fig. 6. KLD between the distributions of the neutral and emotional values of the *displacement coefficient* for the facial areas.

Notice that emotional differences are observed in the results presented here. In the acoustic domain, the emotional modulation for happiness tends to be higher than the modulation for anger and sadness (see Fig. 4 and 5). However, in the facial expression domain, anger is the emotion with stronger modulation, especially in the lower face regions (see Fig. 6). These results suggest that different modalities are used to emphasize the affective goals for happiness and anger.

TABLE IV

RATIO BETWEEN THE AVERAGE DISPLACEMENT COEFFICIENT OBSERVED IN EMOTIONAL AND NEUTRAL UTTERANCES. (S = SADNESS, A = ANGER, H = HAPPINESS, N = NEUTRAL)

	(B) Mid/back vowels			(D) Diphthong			(F) Front vowels			(L) Liquid and glide			(N) Nasal			(T) Stop			(C) Fricatives		
	S/N	A/N	H/N	S/N	A/N	H/N	S/N	A/N	H/N	S/N	A/N	H/N	S/N	A/N	H/N	S/N	A/N	H/N	S/N	A/N	H/N
Forehead	1.36	2.40	2.29	1.39	2.41	2.13	1.31	2.38	2.07	1.31	2.21	2.15	1.31	2.46	2.31	1.30	2.33	2.25	1.35	2.24	2.06
Left eye	1.29	2.37	2.04	1.30	2.36	1.95	1.29	2.37	1.91	1.23	2.15	1.83	1.27	2.36	1.98	1.23	2.26	2.03	1.37	2.31	2.00
Right eye	1.33	2.66	2.59	1.38	2.75	2.51	1.33	2.70	2.43	1.30	2.44	2.43	1.31	2.66	2.63	1.27	2.56	2.54	1.31	2.42	2.31
Left cheek	1.05	1.76	1.36	1.00	1.68	1.42	1.03	1.82	1.50	1.04	1.79	1.39	1.10	1.87	1.52	1.11	1.83	1.53	1.04	1.73	1.38
Right cheek	1.05	1.75	1.55	1.01	1.69	1.62	0.99	1.75	1.63	1.02	1.71	1.54	1.03	1.75	1.67	1.08	1.77	1.66	1.03	1.70	1.45
Nasolabial	0.88	1.68	1.40	0.87	1.66	1.41	0.87	1.68	1.34	0.83	1.63	1.36	0.88	1.71	1.47	0.90	1.69	1.43	0.82	1.54	1.20
Chin	0.89	1.55	1.49	0.84	1.41	1.25	0.85	1.39	1.38	0.88	1.54	1.50	0.91	1.60	1.70	0.89	1.56	1.59	0.86	1.52	1.31
Average	1.12	2.03	1.82	1.11	1.99	1.76	1.09	2.01	1.75	1.09	1.92	1.74	1.12	2.06	1.90	1.11	2.00	1.86	1.11	1.92	1.67

IV. DISCUSSION AND CONCLUSIONS

The analysis presented in this paper provides evidence about the emotional encoding process, in which different modalities are used to convey the affective goals. The results support the idea that *emotional bits* are assigned to the modalities that are not constrained by other communicative goals. In particular, it was observed that facial expressions and pitch tend to have stronger emotional modulation when the articulatory configuration does not have enough freedom to express emotions, given the physical constraints in the speech production system. This emotional assignment compensates the temporal limitation observed in some modalities, and plays a crucial role in the interplay between communicative goals.

An interesting question is whether the phonetic classes used here, which are based on the *manner* of articulation, is the best phonetic description to link acoustic and visual modalities. An alternative approach is to subdivide the phonemes according to the *place* of articulation (e.g. bilabial, dental, palatal). Since this classification is connected with the mapping between visemes and phonemes, it may be useful to compare similar analysis with this phonetic subdivision.

A limitation of this work is that the database was recorded from only one subject. To overcome this, we are collecting a similar data with more subjects. This database will be useful to validate and extend the results presented in this paper.

While the results presented here are interesting from a theoretical point of view, many practical applications such as believable human-like animation and cognitive interfaces can be enhanced by understanding how the emotions are jointly encoded in different communicative channels. The analysis presented in this paper provides useful insights to design emotional models that capture the underlying relationships and interplays between modalities. We are currently exploring these challenging areas.

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, conclusions and/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. Thanks also go to Matthew Black and Emily Mower for their thoughful comments.

REFERENCES

[1] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[2] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

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

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

[5] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 117–139, June 2004.

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

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

[8] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *to appear in Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007.

[9] C. Busso and S. Narayanan, "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.

[10] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[11] T. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082–1089, May 2006.

[12] K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, September 1987.

[13] J. Lucero, A. Baigorri, and K. Munhall, "Data-driven facial animation of speech using a QR factorization algorithm," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 135–142.

[14] P. Boersma and D. Weeninck, "Praat, a system for doing phonetics by computer," Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, Technical Report 132, 1996, http://www.praat.org.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England, December 2006.

[16] S. Lee, E. Bresch, and S. Narayanan, "An exploratory study of emotional speech production using functional data analysis techniques," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 525–532.

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

[18] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*. Upper Saddle River, NJ, USA: Prentice-Hall, 2006.