AN ACOUSTIC ANALYSIS OF BERLIN DATABASE OF EMOTIONAL SPEECH BASED ON BIO-INFORMATIONAL DIMENSIONS

Zhengyang Cai, Yi Xu

University College London
caizhengy2022@163.com, yi.xu@ucl.ac.uk

ABSTRACT

We performed a systematic acoustic analysis of the Berlin database of emotional speech (Emo-DB), with the aim to explore a) relation of acoustic measurements and individual emotions in light of the bio-informational dimensions (BID) theory, and b) inter-correlations among the acoustic measurements. Results show consistent spectral slope measurements with the size projection hypothesis for happiness and anger, but complex patterns for other emotions. The inter-correlation analysis made three surprise findings. The first was high correlations between median pitch and spectral slope measurements, which was likely related to the Lombard effect. The second was a high correlation between mean intensity and median pitch, which is likely due to data collection procedures in developing the corpus. The third finding was a negative correlation between median pitch and jitter, which was mainly attributable to creaky voice in sadness. These findings may have significant implications for further research on the phonetics of emotional speech.

Keywords: acoustic parameters, bio-informational dimensions theory, correlation analysis, emotional expressions

1. INTRODUCTION

The identification of acoustic correlates of emotions in speech has been difficult, despite continued interest. A major source of difficulty is a lack of proper theorical grounding [17, 18]. Most investigations tend to be data-driven and descriptive, and this is true of both studies that examine acoustic measurements of emotional speech [13, 18, 22, 23, 24] and those that use emotion perception to evaluate synthetic manipulations of acoustic parameters [4, 12]. Also, studies that do try to apply popular psychological models of emotion, such as the dimensional theories, have been unable to identify acoustic correlates of the most theory-essential components such as valence (pleasant/unpleasant) [8, 10]. In the face of this difficulty, an alternative approach to the phonetics of emotion has been gaining ground, namely, the Morton-Ohala hypothesis. This approach posits functional connections between the acoustic properties of emotional vocalizations and the emotional categories and dimensions that conventional approaches have been unable to identify.

1.1. The Morton-Ohala hypothesis

The hypothesis consists of Morton's motivation-structural rules [11] and Ohala's frequency code [15]. It posits that emotional expressions are evolutionally adapted to influence the listener for the benefit of the vocalizer. According to Morton, animals' aggressive calls are for intimidating the hearer by exaggerating the body size of the caller with low pitch and harsh vocal quality, and submissive and fearful calls for making appeasement with raised pitch and pure-tone like vocal quality [11]. Ohala extended this size projection principle to human speech and added vocal tract length as a further indicator of body-size, proposing, in particular, that the smile is for shortening the effective length of the vocal tract to appease the listener [15].

Body-size projection has received support from studies on animal calls and social attributes of human voice, such as gender, attractiveness, etc. [1, 7, 16], but its application for emotional speech has been slow. An important reason is the conflicting findings about the relation of f_0 with confidence or aggression, [15, 17]. Another reason is the lack of evidence for emotion-specific intonation patterns [15] based on frequency code [2, 19]. More recently, however, direct evidence of size projection is found by treating emotional cues as parallel to (i.e., independent of) the phonetics of linguistic contrasts [6, 14, 27, 28]. Nevertheless, body-size projection alone still does not seem enough. For one thing, the conflicting f_0 patterns for anger and dominance remain unexplained. For another, synthetic manipulations based on size-projection alone do not generate speech that sounds emotionally charged [6, 14, 26], which suggests that other factors may also be involved. This has led to the bio-information dimensions (BID) theory [27].

1.2. Bio-informational dimensions

The bio-informational dimensions (BID) theory adopts the core assumption of the Morton-Ohala hypothesis, namely, non-human animals and humans alike use specific acoustic cues in their vocalizations to influence the listener in ways that may benefit the vocalizer. But BID goes beyond body-size projection by adding three further dimensions not directly related to body size, as follows.

The *dynamicity* dimension controls how vigorous the vocalization sounds, depending on whether it is beneficial for the vocalizer to appear strong or weak.

The *audibility* dimension controls how far a vocalization can be transmitted, depending on whether and how much it benefits the vocalizer to be heard over long distance.

The *association* dimension controls associative use of sounds typically accompanying a non-emotional biological function in circumstances beyond the original ones.

There has been initial evidence for the dynamicity dimension from synthesis-perception paradigm [14, 27, 28], but not yet for audibility and association. And there has been no examination of whether evidence of these dimensions can be observed from acoustic measurements of emotional speech. This study is the first attempt to test BID in a production corpus, with the goal to explore a) relation of acoustic measurements and individual emotions in light of the bio-informational dimensions (BID) theory, and b) inter-correlations among the acoustic measurements.

2. METHOD

2.1. The corpus

The Berlin database of emotional speech (Emo-DB) is a widely used corpus consisting of 10 sentences spoken by 5 male and 5 female German speakers of different ages in seven emotions: anger, happiness, fear, disgust, sadness, boredom and neutral [4]. All the sentences have been shown to be highly recognizable for the intended emotions. The corpus is also fully annotated for syllables and phones. For this study, we used the phone labels and extracted all measurements only from the vowel segments.

2.2. Measurements and analysis

Sixteen BID measurements were taken by ProsodyPro, an interactive Praat script developed for large-scale analysis of speech prosody [25]. We first ran ProsodyPro to create a tier in which only vowel intervals were labeled to allow the output of the measurements. Then, all measurements in each sentence were averaged to get a single set of mean measurements. These data were then analyzed using R.

3. RESULTS

Due to space limitations, only the most significant results are discussed in this paper.

3.1. Emotion-specific analysis

Tables 1 displays mean values of all the spectral slope parameters that reflect voice quality. As can be seen, all these measurements indicate less spectral tilt in anger than in happiness: lower h1-h2, H1-A1, H1-A3, EB1000, EB500, Ham, but higher COG. A gentle spectral slope suggests a tense voice, which is consistent with the harsh voice suggested by Morton for aggressive animal calls and findings of earlier studies [14, 28]. In fact, all these parameters also show that anger has the smallest spectral tilt across all the emotions. Oddly, however, except for h1-h2 relative to disgust, happiness has the second flattest spectral slope among all the emotions. The reason for this will become clear in the next section on the correlation analysis. Relatedly, all these parameters show the steepest spectral slope for sadness.

Table 1. Top rows: Mean values of voice quality measurements of 7 emotions, where h1-h2 stands for Amplitude difference between 1st and 2nd harmonics, H1-A1 and H1-A3 stand for Amplitude difference between 1st harmonic and 1st and 3rd formants, EB500, EB1000 stand for energy below 500 and 1000 Hz, Ham stands for Hammarberg index, and COG stands for center of gravity. Bottom rows: ANOVA results with measurements as dependent variables, emotion as independent factor, and sentence and speaker as random factors.

	h1-h2 (dB)	H1-A1 (dB)	H1-A3 (dB)	EB_ 500	EB_ 1000	Ham (dB)	COG (Hz)
Happiness	3.0	1.9	14.5	0.49	0.73	12.8	866
Anger	1.3	0.9	10.7	0.34	0.6	9.3	1171
Disgust	2.9	4.0	21.4	0.65	0.84	15.9	663
Sadness	8.9	15.1	32.0	0.89	0.95	21.7	302
Fear	7.9	9.8	23.0	0.68	0.84	15.0	643
Boredom	4.0	7.5	27.5	0.78	0.93	20.0	422
Neutral	4.6	7.3	26.4	0.73	0.91	18.7	474
p	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05
F	10.02	28.09	31.06	46.17	27.88	22.5	28.34
df	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62

In Table 2, the first two columns show that both formant dispersion measurements indicate longer vocal tract in anger than in happiness, consistent

with previous findings from synthetic manipulations [6, 14, 27, 28]. The significant effect of jitter seems to be purely due to the high value of sadness, which again will be revisited in the correlation analysis.

For fear, it differs significantly from neutral emotion only in median pitch based on paired t-test (t(df) = 6, 62, $p < 0.05$). This is consistent with previous reports of high pitch in fear [9, 18]. For disgust, only duration is significantly different from neutral emotion (t(df) = 6, 62, $p < 0.05$), while median pitch is not lower than neutral emotion as previously reported [23].

Table 2. Top rows: Mean values of formant dispersion, median_pitch, mean_intensity, duration, jitter, shimmer, Harmonicity of 7 emotions. Bottom rows, ANOVA results with measurements as dependent variables, emotion as independent factor, and sentence and speaker as random factors.

	FD1_3	FD1_5	Median pitch	Mean intensity (dB)	Duration (ms)	Jitter	Shimmer	Harm
Happiness	1090	879	254	75.3	73.7	0.02	0.1	11.0
Anger	1064	907	261	73.6	79.0	0.02	0.12	9.6
Disgust	1115	898	180	76.8	85.8	0.02	0.1	10.3
Sadness	1172	863	136	78.0	67.7	0.06	15.5	26.6
Fear	1140	873	225	77.0	59.0	0.02	0.11	10.8
Boredom	1150	912	149	77.6	82.0	0.03	0.07	13.1
Neutral	1136	881	152	78.1	66.4	0.03	0.1	11
p	<0.05	>0.05	<0.05	<0.05	<0.05	<0.05	>0.05	>0.05
F	4.31	0.19	12.65	10.26	12.95	15.7	1.69	1.59
df	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62	6, 62

3.2. Correlation analysis

We have examined correlations between all pairs of measurements, with either individual utterances or specific emotions as raw data. Three sets of correlations were found to be particularly interesting. The first is the high correlation between median pitch and spectral slope measurements, as shown in Table 3. As can be seen, the only exceptions are h1-h2 and H1-A1, which measure the low frequency portions of the spectrum. An example of the scatter plots is shown in Figure 1 for median pitch over center of gravity (COG), which has the highest R^2 value.

Table 3. Slope and R2 values of correlations between median pitch and spectral slope measurements. Slope is the coefficient of the correlation line.

	h1-h2	H1-A1	H1-A3	EB_500	EB_1000	Ham	COG
Slope	-1.41	-5.61	-5.34	-241	-337	-8.93	-0.15
R^2	0.01	0.19	0.42	0.55	0.53	0.44	0.56

These high correlations mean that the interpretation of spectral slope as an indicator of voice quality needs to take median pitch into consideration. For example, the voice quality of an emotion can be judged as extra breathy or extra tense only if it significantly deviates from the correlation line. In Figure 1b, for instance, the mean COG of happiness is located on the left of the correlation line, whereas that of anger is located on the right of the correlation line. This suggests that the voice of happiness is truly breathy while that of anger is truly tense. The same principle is applicable for other emotions. For example, sadness has the lowest COG among all the emotions, which might indicate a breathy voice, consistent with previous reports [5, 9]. But the fact that it also has the lowest median pitch (Figure 1b) raises the question how much true breathiness is involved in sad voice.

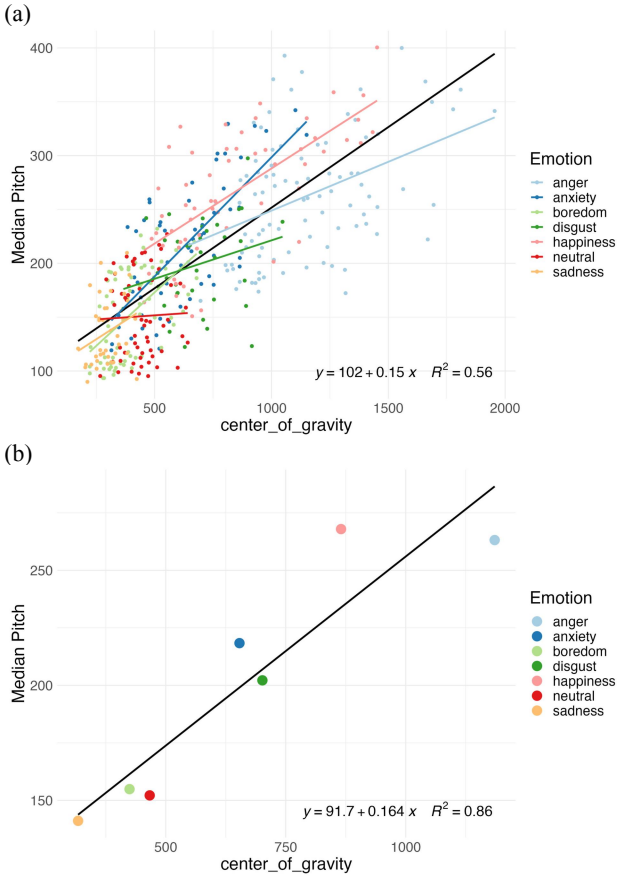

Figure 1. Correlation of COG and median pitch. In (a) each utterance contributes one data point. In (b) each emotion contributes a data point.

Table 4. Slope and R2 values of correlations between mean intensity and spectral slope measurements. Slope is the coefficient of the correlation line.

	h1-h2	H1-A1	H1-A3	EB_500	EB_1000	Ham	COG
Slope	0.19	0.18	0.19	7.19	11.2	0.31	0.005
R^2	0.07	0.13	0.34	0.30	0.36	0.34	0.36

The second interesting sets of correlations are a fairly close relation between mean intensity and spectral slope measurements, as shown in Table 4. Again, an example of the scatter plots is shown in

Figure 2 for mean intensity over center of gravity (COG), which has one of the highest R^2 values. These high correlations are dubious, however, because the emotions with the highest mean intensity are neutral, sadness and boredom, while those with the lowest mean intensity are anger and happiness. Because measured intensity is closely related to recording conditions, we checked the original report [4] on how the corpus was recorded and noticed a number of things. First, the speakers "were instructed not to shout to express anger and to avoid whispering while expressing anxiety," where anxiety = fear. Second, "the recording level had to be adjusted between very loud speech (mostly anger) and very quiet speech (mostly sadness)." Probably as a result, the maximum sound level varies very little across the 7 emotions in the corpus, with the means ranging from 1.008 (boredom) to 1.026 (fear) Pascal, with a standard deviation of 0.005 Pascal. It is likely, therefore, that emotions with relatively flatter spectral tilt have been unproportionally reduced in intensity.

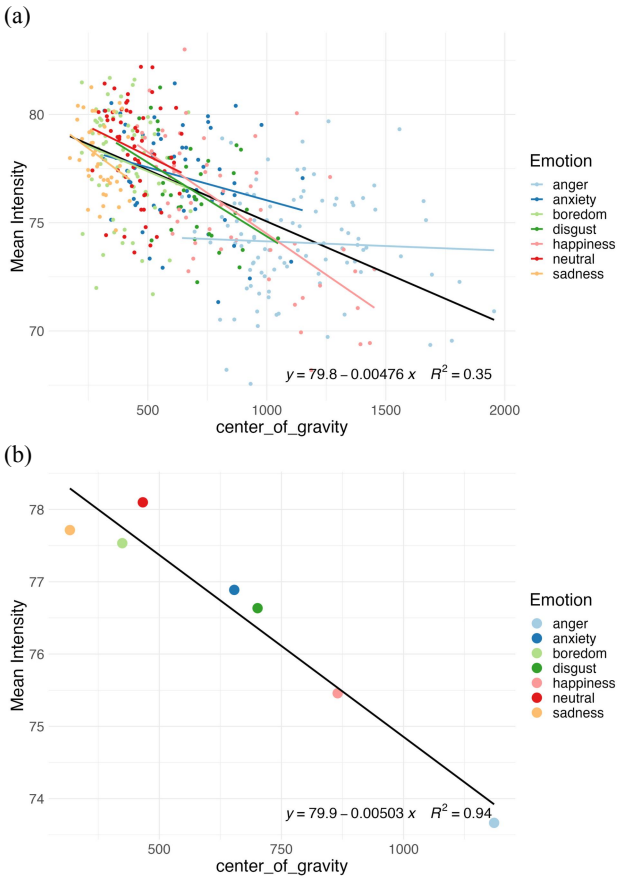

Figure 2. Correlation of center of gravity and mean intensity. In (a) each utterance contributes one data point. In (b) each emotion contributes a data point.

The third interesting correlation is a negative relation between jitter and median pitch, as shown in Figure 3. As can be seen, the negative overall correlation is mainly due to the high jitter values of sadness whose median pitch is very low, which can also be seen in Table 2. One possibility based on our listening impression is that many of the sad utterances have creaky voice, which may have given rise to the high jitter values. In contrast, the jitter values in fear are not high, contrary to some previous reports [9].

4. DISCUSSION AND CONCLUSIONS

A major finding of a detailed acoustic analysis of Emo-DB is the clear contrast between anger and happiness in a) voice quality as shown by all the spectral slope measurements, indicating harsh voice in anger and breathy voice in happiness, and b) vocal tract length as shown by formant dispersion 1-3. These results are consistent with the prediction of body-size projection based on the Morton-Ohala hypothesis [11, 15]. The relatively flat spectral slope and high median pitch in both of these emotions also suggest high orders of dynamicity and audibility, but this is made inclusive given the surprise finding of the correlation analysis.

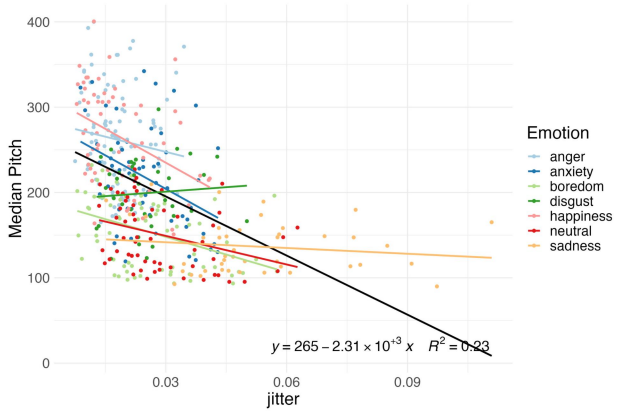

Figure 3. Correlation of jitter and mean median pitch. Each data point from a single utterance.

From the correlation analysis, the first main finding is the close relations between median pitch and spectral slope. A likely source of this relation is related to the Lombard effect, under which pitch and amplitude of a voice are both automatically raised by a noisy environment [3]. Emotions like hot anger and happiness may also involve the same voice raising mechanism, because both presumably require increased vocal effort. So the high pitch often found in anger [9, 15, 18] is likely related to the dynamicity or audibility dimension. Despite the effect, body-size projection is still effectively achieved by spectral slope and formant dispersion, as shown Tables 1 and 2.

The second correlation finding is the close relation between mean intensity and median pitch, which is likely due to the data collection procedures in

developing Emo-DB [4]. The procedure may have neutralized much of the amplitude differences across the emotions. Cautions are therefore needed when evaluating loudness or vocal effort in this and other corpora applying similar recording strategies.

Finally, the negative correlation between median pitch and jitter is likely due to creaky voice in sadness. Equally significant is the lack of high jitter in fear, which is often believed to be associated with a trembling voice [20, 21]. In fact, the general lack of variability of jitter except for sadness questions the usefulness of this measurement for emotional speech.

7. REFERENCES

[1] Anikin, A., Pisanski, K., Massenet, M. and Reby, D. (2021). Harsh is large: nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B: Biological Sciences*.

[2] Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication* **46**: 252-267.

[3] Brumm, H. and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* **148**(11-13): 1173-1198.

[4] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B. (2005). A database of German emotional speech. In *Proceedings of Ninth European Conference on Speech Communication and Technology*

[5] Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In ISCA Tutorial and Research Workshop (ITRW) on speech and emotion.

[6] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2008). Encoding emotions in speech with the size code. *Phonetica, 65*(4)*, 210-230.*

[7] Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M. and Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour* **69**: 561-568.

[8] Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America, 128*(3), 1322-1336.

[9] Juslin, P. N. (2013). Vocal affect expression: problems and promises. *Evolution of emotional communication*: 252-273.

[10] Mauss, I. B. and Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion* **23**(2): 209-237.

[11] Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist, 111*(981), 855-869.

[12] Mozziconacci, S. J. (2001). Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction, 11*(4), 297-326.

[13] Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America, 93*(2), 1097-1108.

[14] Noble, L., & Xu, Y. (2011). Friendly Speech and Happy Speech-Are They the Same?. In *ICPhS* (pp. 1502-1505).

[15] Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*(1), 1-16.

[16] Puts, D. A., Gaulin, S. J. C. and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior* **27**(4): 283-296.

[17] Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin* **99**: 143-165.

[18] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms *Speech communication, 40*(1-2)*, 227-256.*

[19] Scherer, K. R. and Bänziger, T. (2004). Emotional expression in prosody: a review and an agenda for future research. In *Proceedings of Speech Prosody 2004***:** 359-366.

[20] Schuller, B., Wöllmer, M., Eyben, F. and Rigoll, G. (2009). The role of prosody in affective speech. *linguistic insights, studies in language and communication*: 283-305.

[21] Shaver, P., Schwartz, J., Kirson, D. and O'Conner, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* **52**: 1061-1086.

[22] Shami, M., & Verhelst, W. (2007). An evaluation of. the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech communication*, 49(3), 201-212.

[23] Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9), 1162-1181.

[24] Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. The journal of the acoustical society of America, 52(4B), 1238-1250.

[25] Xu, Y. (2013). ProsodyPro—A tool for large scale. systematic prosody analysis. Laboratoire Parole et Langage, France.

[26] Xu, Y. and Kelly, A. (2010). Perception of anger and happiness from resynthesized speech with size-related manipulations. In *Proceedings of Speech Prosody 2010*, Chicago

[27] Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. *Prosody and iconicity*, 33-60.

[28] Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLoS ONE*, *8*(4), e62397.