Audio Engineering Society

Convention Paper 10671

Presented at the 155th Convention

2023 October 25-27, New York, NY



# Vocal Affects Perceived from Spontaneous and Posed Speech

Eunmi Oh[1] and Jinsun Suhr[1]

[1] *Department of Psychology, Yonsei University, 50 YONSEI-RO, SEODAEMUN-GU, SEOUL, 03722*

Correspondence should be addressed to Author (eunmioh@yonsei.ac.kr)

**ABSTRACT**

This study examines listeners' natural ability to identify an anonymous speaker's emotions from speech samples with broad ranges of emotional intensity. This study aims to compare emotional ratings between posed and spontaneous speech samples and analyzes how basic acoustic parameters are utilized. The spontaneous samples were extracted from the Korean Spontaneous Speech corpus consisting of casual conversations. The posed samples with emotions (happiness, neutrality, anger, sadness) were obtained from the Emotion Classification dataset. Non-native listeners were asked to evaluate seven opposite pairs of affective attributes perceived from the speech samples. Listeners perceived fewer spontaneous samples as having negative valences. The posed samples had higher mean rating scores than those of the spontaneous speeches, only in negative valences. Listeners reacted more sensitively to the posed than spontaneous speeches in negative valence and had difficulty detecting happiness from the posed samples. The spontaneous samples perceived as positive had higher variance in pitch and higher maximum pitch than those perceived as negative. Contrastingly, the posed samples perceived as negative valences were positively correlated with higher values of the pitch parameters. These results can be utilized to assign specific vocal affects to artificial intelligence voice agents or virtual humans, rendering more human-like voices.

## 1 Introduction

This study aims to understand how listeners can perceive and infer speakers' feelings and attitudes from their tone of voice—so-called vocal affect—in everyday life. The focus of our study is on the listener's natural ability to identify a speaker's emotion from voice alone. In daily life, humans effortlessly use their natural skills to detect and recognize a speaker's emotions from the tone of voice, which includes acoustic information [1,2]. If the emotion conveyed by the voice is incongruent with that conveyed semantically, listeners may infer a speaker's attitude mainly by variation in tone [3,4]. We often experience that the vocal affect is quite powerful in interpersonal interaction and communication [5].

Many studies on vocal affects have revealed the relationship between vocal acoustic parameters and emotion. They mostly utilize synthetic speech samples [6–9] and vocal bursts and speech recordings by actors and professionals with posed emotions [10–17]. However, there is a lack of studies on vocal affect with spontaneous speech samples that can reflect casual conversations. One of the issues in evaluating vocal affect is controlling the effect of semantics on emotional rating while maintaining prosodic features

and vocal information. A way to minimize the effect of semantics is to use pseudo-random words [18] and low-pass filtering [19] which would have distorted the acoustic parameters in the analyses. Alternatively, one could use speech in a foreign language. Previous studies have shown that listeners can recognize the emotional states of communicators based on vocal expressions even in foreign languages, although non-native listeners might not fully utilize vocal expressions of emotion available in foreign languages [20–23].

In our previous study, we explored vocal affects with spontaneous speech samples perceived by non-native listeners [24]. The results suggest that even from 2 s speech samples, non-native listeners were able to incorporate some acoustical cues to perceive an anonymous speaker's gender and affective attributes. In the current study, we aim to evaluate non-native listeners' ability to perceive emotional attributes from speech samples with a wider range of emotional intensity and emotions rarely encountered in casual conversations. Thus, the goal of this study is to compare emotional ratings between spontaneous speech samples and posed samples performed by professionals. We further analyze how basic acoustic parameters are utilized when listeners perceive vocal affects from the two types of speech samples.

## 2 Methods

### 2.1 Stimuli

A total of 317 Korean speech samples, including 180 spontaneous and 137 posed samples, were used in this study. The spontaneous speech samples were selected from the Korean Spontaneous Speech corpus [25], which comprises 969 h of open-domain dialogue recorded by 2,000 speakers, that is, casual conversations of 1,000 pairs of native Korean speakers in a quiet environment. The posed speech samples were selected from the Emotion Classification dataset provided by AI Hub [26], which comprises 10,351 videos of 100 Korean actors acting out seven emotions. The acted emotions were happiness, surprise, neutrality, fear, disgust, anger, and sadness. For each emotion, there were 50 different emotion-specific script sentences. The posed samples for four emotion categories (happiness, neutrality, anger, and sadness) were chosen corresponding to the affective attributes that were rated in our experiments.

We then selected a random time point within the spontaneous clip and a point near the beginning of the posed clip and extracted a 2 s portion from each. Some clips consisted of incomplete utterances. All samples were stored in waveform audio format, with their sampling rate at 16 kHz and loudness-matched to one sample's root mean square (RMS) in Adobe Audition. We also applied 150 ms sigmoid fades at both the beginning and end of the samples.

### 2.2 Listeners

Non-native listeners, who did not understand Korean, were recruited on Amazon's Mechanical Turk. All participants provided informed consent and were rewarded $0.5~$1 after the completion of a 10-to-15 min session. In each experiment, we inserted image-recognition catch trials to ensure that participants were concentrating on the tasks. Incorrect responses to the trials led to the elimination of their results from the analysis. After post-screening, the results were obtained from a total of 472 non-native listeners (219 women, 252 men, 1 undisclosed). Their first languages were English (422), Hindi (10), Spanish (1), and other (39).

### 2.3 Procedure

Listeners were required to use earphones or headphones and adjust the volume to a sample speech that had the same RMS loudness as the stimuli under test. At the beginning of each trial, a speech sample was played on a blank screen. The same sample was repeated once while displaying rating scales. The 317 samples were randomly divided into nine batches containing 35 or 36 stimuli per batch. At least 50 listeners were randomly assigned to each batch.

Listeners were asked to evaluate seven opposite pairs of affective attributes of the speech on a 7-point scale with 0 in the center. The seven pairs of affective attributes were *relaxed-stressed*, *content-angry*, *friendly-hostile*, *sad-happy*, *bored-interested*, *intimate-formal*, and *timid-confident* [6,27]. We counterbalanced how the attribute pairs were shown on the screen. For example, for the pair of *relaxed-stressed*, half of the participants had relaxed on the left and stressed on the right of the scale, while the other half had relaxed on the right and stressed on the left of the scale.

All listeners were given a short break (at least 1 min) in the middle of the experiment if they desired. The online experiment was hosted on Pavlovia [28] and written in JavaScript using the jsPsych library [29].

### 2.4 Analyses

Acoustic parameters of each speech sample were obtained using Praat [30]. The parameters included intensity (maximum, minimum), pitch (mean, maximum, minimum), standard deviation (SD) of intensity and pitch reflecting variation, measure of frequency perturbation (jitter), measure of amplitude perturbation (shimmer), number of voice breaks reflecting inter-pulse intervals, and Harmonics-to-Noise Ratio (HNR), which measures the ratio between periodic and non-periodic components of speech segments. As samples were loudness-matched, mean intensity was unanalyzed. For the 2 s speech samples, pitch parameters were measured within a 40 ms analysis window, and intensity parameters were measured within a 42.7 ms analysis window. Pearson correlation analyses were conducted to relate acoustical parameters to listeners' voice perception on emotion.

## 3 Results

### 3.1 Emotion Perception

To obtain a response score for each speech sample and affect, we averaged the results from the listeners. There were seven response scores for each of the 317 speech samples. Figure 1 illustrates the distribution of listeners' ratings for seven affective pairs across the (a) spontaneous and (b) posed samples. Overall, the variation of the rating scores of the posed speech tended to be larger than those of the spontaneous speech, especially for angry, hostile, and stressed emotions. We then divided the speech samples according to their response score for all 14 affects. For example, if the *relaxed-stressed* score for a speech sample was above 0, it would be categorized as a stressed-sounding stimulus; this same sample would also be categorized as content if its *content-angry* affect pair score was below 0. Table 1 shows the number of speech samples categorized into one or the other affect for each affect pair. Note that the sample was not assigned to any affect if the score was equal to 0; the number of 0s is specified in parentheses. Listeners perceived fewer spontaneous samples as having negative valences (*stressed*, *angry*, *hostile*, *bored*, and *timid*). They perceived fewer posed samples categorized into emotions such as boredom and timidness. This might be expected because the intended emotions of our posed samples were happy, sad, and angry.

The absolute values of rating scores were averaged over the samples for each perceived affect specified in Table 1. Figures 2 and 3 illustrate the mean rating scores and the percentage of samples categorized into 14 affect attributes for the spontaneous and posed

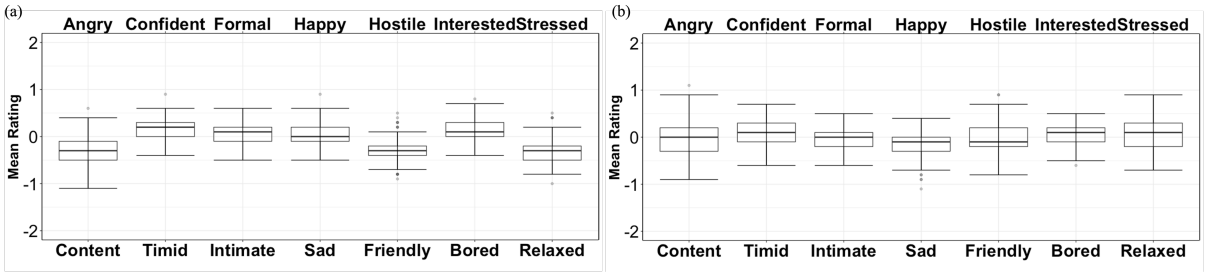


Figure 1. Mean response scores for seven affective pairs across (a) 180 spontaneous speech samples and (b) 137 posed speech samples. Boxplots show interquartile range (box), median (horizontal line in box) with minimum and maximum values (vertical line), and extreme values (dots).

| Perceived Affect / Sample Type | Relaxed /Stressed | Content /Angry | Friendly /Hostile | Sad /Happy | Bored /Interested | Intimate /Formal | Timid /Confident |
|---|---|---|---|---|---|---|---|
| **Spontaneous** | 158/13 (9) | 156/10 (14) | 151/19 (10) | 59/84 (37) | 39/110 (31) | 50/106 (24) | 30/120 (30) |
| **Posed** | 50/72 (15) | 67/57 (13) | 72/50 (15) | 90/34 (13) | 39/77 (21) | 60/49 (28) | 38/83 (16) |

Table 1. Number of speech samples categorized into one of the affects in the affect pair of spontaneous and posed speech samples. Numbers in parentheses indicate the number of speech samples categorized into neither of the affects.

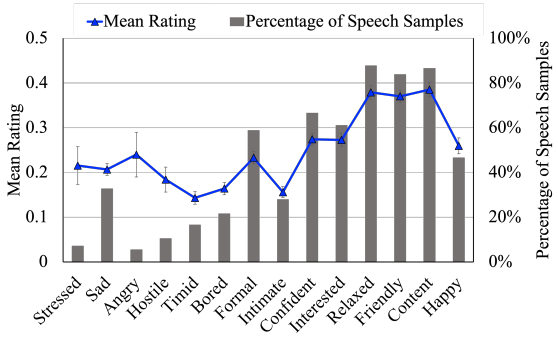


Figure 2. Means rating scores with standard errors and the percentage of spontaneous speech samples for 14 affect attributes from negative (left) to positive (right) valences.

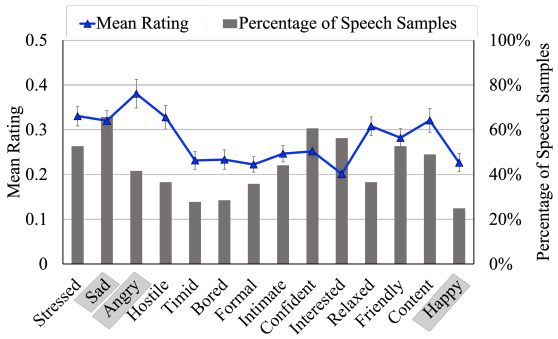


Figure 3. Means rating scores with standard errors and the percentage of posed speech samples for 14 affect attributes. The attributes highlighted in gray indicate the intended emotions for the posed samples.

samples, respectively. Notice that the affective attributes were arranged in order of emotional valence [31,32]. The affective attribute on the far left (*stressed*) is the most negative, while the one on the far right (*happy*) is the most positive. There are much less spontaneous samples perceived as negative valences (*stressed*, *angry*, and *hostile*), however, this skewness is less evident for the posed samples. Figure 2 shows that the mean rating scores of the spontaneous speeches in positive valences were much higher than those in negative valences. Figure 3 illustrates that the mean rating scores of the posed speeches in negative valences were higher than those in positive valences. In comparison with spontaneous and posed samples, the rating scores of the spontaneous speeches were higher than those of the posed speeches in positive valences, while the rating scores of the posed speeches were higher than those of the spontaneous speeches in negative valences.

Table 2 indicates the number of samples perceptually categorized based on the affect pairs of the posed speech samples. The degree to which the intended emotion was expressed could vary among affective attributes. Angry and sad voices were generally perceived by listeners as intended, while happy voices were not. Non-native listeners seemed to have difficulty detecting happiness from the posed samples.

The difference between the posed and spontaneous speech samples were not evident for the affective attributes such as *bored*, *intimate*, and *confident*. This might be expected, since our posed samples only cover the emotions of happiness, neutrality, anger, and sadness.

| **Perceived** / **Posed** | Relaxed /Stressed | Content /**Angry** | Friendly /Hostile | **Sad** /**Happy** | Bored /Interested | Intimate /**Formal** | Timid /Confident |
|---|---|---|---|---|---|---|---|
| **Angry** | 7/**29** (3) | 5/**30** (4) | 10/**25** (4) | **31**/4 (4) | 10/**25** (4) | **21**/10 (8) | 6/**32** (1) |
| **Happy** | 14/**16** (5) | **18**/13 (4) | **21**/12 (2) | **21**/**11** (3) | 11/**19** (5) | **17**/12 (6) | 9/**20** (6) |
| **Sad** | 13/**22** (5) | **25**/11 (4) | **24**/9 (7) | **32**/6 (2) | 16/**17** (7) | **20**/11 (9) | **19**/14 (7) |
| **Neutral** | **16**/5 (2) | **19**/3 (1) | **17**/4 (2) | 6/**13** (4) | 2/**18** (3) | 2/**16** (5) | 4/**17** (2) |

Table 2. Number of speech samples categorized into one of the affects in the affect pair of posed speech samples. Numbers in bold and blue indicate the number of dominantly perceived samples in each affect pair and the number of samples correctly categorized as the intended emotion for the posed samples, respectively.

| Perceived Affects | Types of Stimuli | Acoustic Parameters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min. Intensity | Max. Intensity | S.D. Intensity | Mean Pitch | Min. Pitch | Max. Pitch | S.D. Pitch | Voice Breaks | Jitter | Shimmer | HNR |
| Stressed | Spontaneous | **-0.44*** | 0.07 | 0.15 | 0.09 | 0.16 | -0.02 | 0 | 0 | -0.07 | 0.09 | -0.07 |
| | Posed | -0.09 | -0.14 | -0.06 | **0.33**** | 0.11 | **0.26*** | **0.26*** | 0.04 | -0.18 | -0.09 | -0.09 |
| Sad | Spontaneous | 0.18 | -0.15 | **-0.36**** | 0.06 | -0.1 | -0.06 | -0.14 | 0.04 | -0.08 | -0.1 | 0.08 |
| | Posed | 0.08 | 0.04 | 0.02 | 0.11 | 0.06 | 0.17 | 0.07 | -0.07 | -0.01 | 0.1 | 0.13 |
| Angry | Spontaneous | -0.12 | -0.13 | 0.02 | -0.06 | -0.02 | -0.05 | -0.21 | 0.12 | -0.14 | -0.1 | 0.09 |
| | Posed | **-0.31**** | -0.07 | 0.11 | **0.44**** | 0.13 | **0.37**** | **0.4**** | 0.08 | -0.22 | -0.17 | -0.17 |
| Hostile | Spontaneous | **-0.45*** | 0.29 | **0.5**** | 0.18 | 0.16 | 0.16 | 0.13 | -0.19 | -0.26 | -0.16 | 0.14 |
| | Posed | -0.19 | -0.04 | -0.1 | **0.41**** | 0.23 | **0.41**** | **0.34**** | **0.26*** | -0.18 | -0.09 | -0.16 |
| Timid | Spontaneous | 0.03 | 0.06 | 0.03 | 0.15 | -0.01 | 0.17 | 0.26 | 0.04 | 0.03 | 0.21 | -0.12 |
| | Posed | 0.06 | 0.14 | 0.08 | 0.14 | 0.12 | 0.09 | 0.04 | -0.2 | -0.06 | -0.07 | 0.17 |
| Bored | Spontaneous | 0.13 | -0.18 | -0.19 | **-0.35*** | -0.25 | -0.12 | -0.08 | -0.18 | 0.09 | -0.02 | 0.08 |
| | Posed | 0.3 | -0.02 | -0.22 | -0.08 | -0.1 | -0.02 | -0.13 | -0.24 | -0.15 | 0.08 | 0.28 |
| Formal | Spontaneous | 0.16 | 0.01 | -0.16 | -0.13 | -0.12 | 0.07 | 0.05 | 0.16 | 0.11 | 0.05 | -0.11 |
| | Posed | 0.08 | -0.08 | -0.05 | -0.2 | -0.1 | -0.22 | **-0.25*** | 0.16 | 0.06 | 0.06 | -0.03 |
| Intimate | Spontaneous | 0.06 | 0.01 | -0.12 | 0.15 | 0.1 | 0.03 | 0.06 | -0.11 | -0.08 | **-0.31*** | 0.21 |
| | Posed | 0.05 | -0.19 | -0.25 | 0.1 | 0.11 | 0.15 | 0.09 | 0.04 | 0.16 | -0.04 | 0.02 |
| Confident | Spontaneous | 0.16 | 0.07 | -0.15 | -0.05 | -0.14 | 0.01 | 0.1 | 0.14 | 0.02 | 0.08 | -0.12 |
| | Posed | -0.07 | -0.09 | 0.02 | 0.04 | 0.05 | 0.05 | 0.01 | -0.09 | -0.12 | 0.06 | -0.14 |
| Interested | Spontaneous | -0.12 | 0.05 | 0.04 | 0.16 | 0.02 | **0.18*** | **0.21*** | 0.08 | 0.05 | -0.05 | -0.01 |
| | Posed | -0.17 | 0.15 | 0.17 | -0.08 | -0.07 | 0.05 | 0.06 | -0.04 | -0.01 | 0.03 | -0.06 |
| Relaxed | Spontaneous | -0.05 | 0.02 | 0.07 | -0.06 | -0.04 | -0.01 | -0.01 | -0.06 | -0.1 | -0.09 | 0.08 |
| | Posed | -0.15 | -0.1 | 0.08 | **0.4**** | **0.41**** | **0.48**** | 0.25 | -0.17 | **-0.29*** | -0.19 | 0.22 |
| Friendly | Spontaneous | -0.06 | 0.06 | 0.01 | -0.02 | -0.01 | -0.04 | -0.01 | -0.03 | -0.11 | -0.01 | -0.03 |
| | Posed | 0.02 | 0.03 | 0.09 | -0.01 | 0.07 | -0.03 | -0.12 | -0.15 | -0.02 | -0.04 | -0.08 |
| Content | Spontaneous | -0.09 | 0.01 | 0.01 | -0.07 | -0.08 | -0.01 | -0.01 | 0.01 | -0.03 | 0.02 | -0.05 |
| | Posed | 0.14 | -0.02 | -0.12 | 0.06 | 0.18 | 0.11 | -0.05 | -0.09 | -0.05 | -0.03 | -0.07 |
| Happy | Spontaneous | 0.01 | 0.03 | -0.09 | 0.05 | -0.08 | **0.25**** | **0.29**** | 0.07 | 0.11 | 0.13 | -0.14 |
| | Posed | 0.23 | **-0.34*** | **-0.29*** | **-0.33*** | -0.19 | -0.07 | -0.1 | 0.04 | 0.21 | 0.26 | -0.26 |

*p < .05. **p < .01

Table 3. Correlations between the acoustic parameters and perceived affects of spontaneous and posed samples. Numbers shown in bold indicate statistically significant correlations.

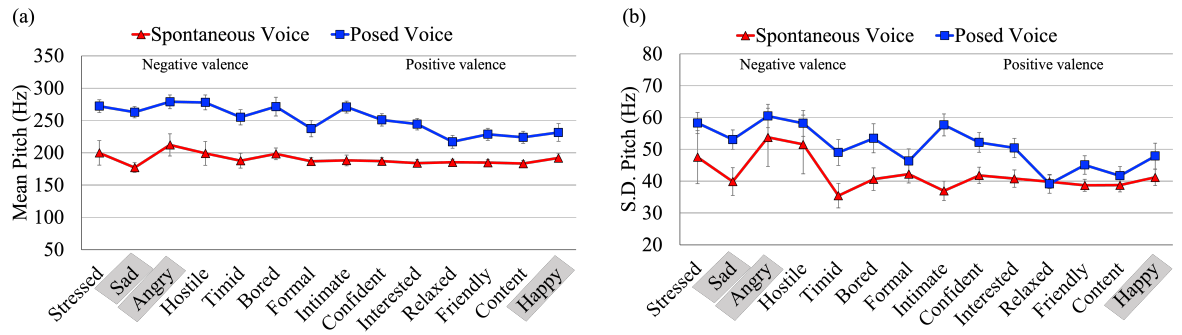


Figure 4. Means and standard errors of (a) mean pitch and (b) variation in pitch of spontaneous and posed speech samples across affective attributes in order of emotional valence.

### 3.2 Acoustic parameters

Table 3 lists the correlations between acoustic parameters and perceived affects for the spontaneous and posed samples. Numbers in bold indicate that the correlations were statistically significant at a level of $p < 0.05$ (*) and $p < 0.01$ (**). Spontaneous speech samples were perceived as *hostile* when the variation of intensity increased and as *sad* when the variation in intensity decreased. Spontaneous samples

perceived as *happy* and *interested* showed positive correlations with the maximum pitch and variation in pitch.

Overall, there were more statistically significant correlations with the posed than spontaneous samples. The posed samples perceived as negative valences (*stressed*, *hostile*, and *angry*) were positively correlated with higher values in pitch parameters (mean, maximum, and SD). For instance, mean pitch and pitch variation are shown in Figure 4 across the affects as a function of emotional valence. The posed samples were perceived as negative valences (*stressed*, *hostile*, and *angry*) when the mean pitch and variation in pitch increased.

## 4 Discussions and Conclusions

Our results indicate several differences between the posed and spontaneous speech samples. First, regarding the number of speech samples categorized for the 14 affective attributes, listeners perceived few spontaneous samples as having negative valences. This may reflect the characteristics of our naturalistic speech stimuli. Although we rarely encounter emotions in negative valences in casual conversations, it is crucial to detect negative valence in everyday life. Therefore, it is necessary to include posed samples in negative valences to understand how listeners can perceive and infer speakers' feelings and attitudes from their tone of voice.

Second, the listeners perceived many more spontaneous samples as having positive valences than negative valences. Furthermore, they perceived higher emotional intensity from the spontaneous samples than the posed samples, only in positive valences. They reacted more sensitively to the posed speeches with negative valences. As specified in Table 2, the results regarding the posed samples suggest that affective attributes in negative valences, such as angry and sad, should be well perceived. However, the participants had difficulty detecting happiness from the posed samples. Our finding regarding the posed samples is in line with previous studies on cross-cultural emotion recognition, which revealed that negative emotions were recognized with higher cross-cultural accuracy than positive emotions [33].

Third, the correlations between basic acoustic parameters and emotional ratings revealed that listeners differed in the utilization of acoustic parameters to detect emotions from the posed and spontaneous samples. The spontaneous samples perceived as positive tended to have higher variance in pitch and higher maximum pitch than those perceived as negative. By contrast, the posed samples perceived as negative valences were positively correlated with higher values of the pitch parameters.

We speculate that the observed differences between the posed and spontaneous samples may derive from two specific factors. One factor could be the smaller sample size for certain affects. There were fewer samples for the posed samples perceived as positive valences and for the spontaneous samples perceived as negative valences. In future research, we will add more spontaneous speech samples with affective attributes whose sample size is small. The other factor could be the cross-cultural and language differences. Cross-language differences exist regarding the range and strength of affective responses as well as the use of acoustic parameters and prosodic features [34,35]. The discrepancy between the posed and perceived emotions by non-native listeners may result from the samples that were not well posed by actors or cross-cultural emotion recognition. In future research, we will explore this discrepancy by conducting experiments with Korean native listeners.

Our experimental results can contribute to forming an appropriate database for training AI algorithms in the computational processing of emotional voice [36]. For more natural interactions with an AI agent, its tone of voice can be adjusted according to the content of speech and facial expressions [37,38]. Furthermore, our listening test results, that is, psychologically annotated data with vocal affects, can be utilized to assign specific vocal affects to voice agents or virtual humans.

## Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020 S1A5B5A16082247, NRF-2022S1A6A4038701)

## References

[1] A. Kappas, U. Hess, and K. R. Scherer, "Voice and Emotion," in R. S. Feldman and B. Rime (Eds.), *Fundamentals of Nonverbal Behavior* (Cambridge University Press, Cambridge, 1991, pp. 200–238).

[2] A. Mehrabian, *Silent Messages: Implicit Communication of Emotions and Attitudes*, 2nd Edition (Wadsworth, Belmont, 1981).

[3] A. Mehrabian, "Decoding of Inconsistent Communication," *J. Pers. Soc. Psychol.*, vol. 6, no. 1, pp. 109–114 (1967).

[4] Y. Lin, H. Ding, and Y. Zhang, "Prosody Dominates over Semantics in Emotion Word Processing: Evidence from Cross-Channel and Cross-Modal Stroop Effects," *J. Speech Lang. Hear. Res.,* vol. 63, no. 3, pp. 896–912 (2020). https://doi.org/10.1044/2020_JSLHR-19-00258.

[5] J. J. Guyer, P. Briñol, T. I. Vaughan-Johnston, L. R. Fabrigar, L. Moreno, and R. E. Petty, "Paralinguistic Features Communicated through Voice Can Affect Appraisals of Confidence and Evaluative Judgments," *J. Nonverbal Behav.*, vol. 45, no. 4, pp. 479–504 (2021). https://doi.org/10.1007/s10919-021-00374-2

[6] C. Gobl and A. N. Chasaide, "The Role of Voice Quality in Communicating Emotion, Mood and Attitude," *Speech Commun.*, vol. 40, no. 1–2, pp. 189–212 (2003 Oct.). https://doi.org/10.1016/S0167-6393(02)00082-1

[7] Y. Li, J. Li, and M. Akagi, "Contributions of the Glottal Source and Vocal Tract Cues to Emotional Vowel Perception in the Valence-Arousal Space," *J. Acoust. Soc. Am.*, vol. 144, no. 2, pp. 908–916 (2018). https://doi.org/10.1121/1.5051323

[8] C. Gobl, and A. N. Chasaide, "Acoustic Characteristics of Voice Quality," *Speech Commun.*, vol. 11, no. 4–5, pp. 481–490 (1992). https://doi.org/10.1016/0167-6393(92)90055-C

[9] M. Cohn, E. Raveh, K. Predeck, I. Gessinger, B. Möbius, and G. Zellou, "Differences in Gradient Emotion Perception: Human vs. Alexa Voices," in *proceedings of the INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, pp. 1818–1822 (Shanghai, China) (2020 Oct.)

[10] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 Emotions Conveyed by Brief Human Vocalization," *Am. Psychol.*, vol. 74, no. 6, p. 698–712 (2018 Dec.). https://doi.org/10.1037%2Famp0000399

[11] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–636 (1996).

[12] K. R. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," *Psychol. Bull.*, vol. 99, no. 2, pp. 143–165 (1986).

[13] S. Shigeno, "The Effects of the Literal Meaning of Emotional Phrases on the Identification of Vocal Emotions," *J. Psycholinguist. Res.*, vol. 47, no. 1, pp. 195–213 (2018 Feb.). https://doi.org/10.1007/s10936-017-9526-7

[14] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping Emotions into Acoustic Space: The Role of Voice Production," *Biol. Psychol.*, vol. 87, no. 1, pp. 93–98 (2011). https://doi.org/10.1177/1754073920949671

[15] J. Sundberg, S. Patel, E. Bjorkner and K. R. Scherer, "Interdependencies among Voice Source Parameters in Emotional Speech," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 162–174 (2011 July–Sept.). https://doi.org/10.1109/T-AFFC.2011.14

[16] T. Bänziger, S. Patel, and K. R. Scherer, "The Role of Perceived Voice and Speech Characteristics in Vocal Emotion Communication," *J. Nonverbal Behav.*, vol. 38, no. 1, pp. 31–52 (2013). https://doi.org/10.1007/s10919-013-0165-x

[17] P. N. Juslin, P. Laukka, and T. Bänziger, "The Mirror to Our Soul? Comparisons of Spontaneous and Posed Vocal Expression of Emotion," *J. Nonverbal Behav.*, vol. 42, no. 1, pp. 1–40 (2018). https://doi.org/10.1007/s10919-017-0268-x

[18] M. D. Pell and S. A. Kotz, "On the Time Course of Vocal Emotion Recognition," *PLoS ONE*, vol. 6, no. 11, e27256 (2011 Nov.). https://doi.org/10.1371/journal.pone.0027256

[19] M. A. Knoll, M. Uther, and A. Costall, "Effects of Low-Pass Filtering on the

Judgment of Vocal Affect in Speech Directed to Infants, Adults and Foreigners," *Speech Commun.*, vol. 51, pp. 210–216 (2009 Mar.). https://doi.org/10.1016/j.specom.2008.08.001

[20] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *J. Nonverbal Behav.*, vol. 33, no. 2, pp. 107–120 (2009 Jun.). https://doi.org/10.1007/s10919-008-0065-7

[21] S. Paulmann and A. K. Uskul, "Cross-Cultural Emotional Prosody Recognition: Evidence from Chinese and British Listeners," *Cogn. Emot.,* vol. 28, no. 2, pp. 230–244 (2014 Feb.). https://doi.org/10.1177/1754073919897295

[22] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The Primacy of Categories in the Recognition of 12 Emotions in Speech Prosody Across Two Cultures," *Nat. Hum. Behav.*, vol. 3, no. 4, pp. 369–382 (2019 Apr.). https://doi.org/10.1038/s41562-019-0533-6

[23] T. Waaramaa and T. Leisiö, "Perception of Emotionally Loaded Vocal Expressions and its Connection to Responses to Music. A Cross-Cultural Investigation: Estonia, Finland, Sweden, Russia, and the USA," *Front. Psychol.*, vol. 4, article 344 (2013 Jun.). https://doi.org/10.3389/fpsyg.2013.00344

[24] E. Oh, J. Lee, and D. Lee, "Mapping Voice Gender and Emotion to Acoustic Properties of Natural Speech," presented at the *150th Convention of the Audio Engineering Society Convention* (2021 May), paper 10016.

[25] J.-U. Bang, S. Yun, S.-H. Kim, M.-Y. Choi, M.-K. Lee, Y.-J. Kim, D.-H. Kim, J. Park, Y.-J. Lee, and S.-H. Kim, "KSPONSPEECH: Korean Spontaneous Speech Corpus for Automatic Speech Recognition," *Appl. Sci.*, vol. 10, no. 19, p. 6936 (2020 Oct.). https://doi.org/10.3390/app10196936

[26] National Information Society Agency (NIA), "AI Hub Speech Database," (2022), https://aihub.or.kr/aihubdata/data/view.do?currMenu=120&topMenu=100&aihubDataSe=extrldata&dataSetSn=259 (accessed Aug. 27, 2022).

[27] E. Uldall, "Dimensions of Meaning in Intonation," in D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. D. Scott, and J. L. M.Trim (Eds.), *In Honour of Daniel Jones: Papers Contributed on the Occasion of his Eightieth Birthday* (Longman, London, 1964, pp. 271–279)

[28] Open Science Tools Ltd., "Pavlovia," (2022), https://pavlovia.org/ (accessed Apr. 21, 2023)

[29] J. R. de Leeuw, "JsPsych: A Javascript Library for Creating Behavioral Experiments in a Web Browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12 (2015 Mar.). https://doi.org/10.3758/s13428-015-0567-2

[30] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," Version. 6.1.42, (2021). https://www.fon.hum.uva.nl/praat/

[31] R. Plutchik, "Emotions: A general psychoevolutionary theory," in K. R. Scherer and P. Ekman (Eds.), *Approaches to Emotion* (Erlbaum, Hillsdale, 1984, pp. 197–219).

[32] D. Kollias et al. "Deep Affect Prediction In-The-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond," *Int. J. Comput. Vis.*, vol. 127, pp. 907–929 (2019 Jun.). https://doi.org/10.1007/s11263-019-01158-4

[33] P. Laukka and H. A. Elfenbein, "Cross-Cultural Emotion Recognition and In-Group Advantage in Vocal Expression: A Meta-Analysis," *Emot. Rev.*, vol. 13, no. 1, 3–11 (2020 Feb.). https://doi.org/10.1177/1754073919897295

[34] I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Cross-Language Differences in How Voice Quality and *f*0 Contours Map to Affect," *J. Acoustic. Soc. Am.*, vol. 144, no. 5, pp. 2730–2750 (2018 Nov.). https://doi.org/10.1121/1.5066448

[35] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The Expression and Recognition of Emotions in the Voice Across Five Nations: A Lens Model Analysis Based on Acoustic Features," *J. Pers. Soc. Psychol.*, vol. 111, no. 5, pp. 686–705 (2016). https://doi.org/10.1037/pspi0000066

[36] D. M. Schuller and B. W. Schuller, "A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice," *Emot. Rev.*, vol. 13, no. 1, pp. 44–50 (2021 Apr.). https://doi.org/10.1177/1754073919898526

[37] Y. Xue, Y. Hamada, and M. Akagi, "Voice Conversion for Emotional Speech: Rule-Based Synthesis with Degree of Emotion Controllable in Dimensional Space," *Speech Commun.*, vol. 102, pp. 54–67 (2018 Jul.). https://doi.org/10.1016/j.specom.2018.06.006

[38] S. Takagi et al., "Multisensory Perception of the Six Basic Emotions is Modulated by Attentional Instruction and Unattended Modality," *Front. Integr. Neurosci.,* vol. 9, article 1 (2015 Feb.). https://doi.org/10.3389/fnint.2015.00001