Confidential

Audio Engineering Society

Convention Paper 10181

Presented at the AES 157th Convention

2024 October 8-10, New York, NY, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Decoding Emotions: Lexical and Acoustical Cues in Vocal Affects

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 Eunmi Oh (eunmioh@yonsei.ac.kr)

ABSTRACT

This study investigated listeners' ability to detect emotions from a diverse set of speech samples, including spontaneous conversations and actor-posed speech. The study explored the influence of lexical content and acoustic properties when native listeners rated seven pairs of affective attributes. Two experimental conditions were employed: a text condition where participants evaluated emotional attributes from written transcripts without vocal information, and a voice condition where participants listened to audio recordings to assess emotions. The results showed that the importance of lexical and vocal cues varied across 14 affective states of posed and spontaneous speech. Vocal cues enhanced expressions of sadness and anger, but had less impact on conveying happiness in posed speech. Vocal cues tended to mitigate the negative emotions conveyed by lexical content in spontaneous speech. Further, correlational analyses revealed that lexical meanings suggesting anger or hostility could be interpreted as positive affective states such as intimacy or confidence. Linear regression analyses indicated that emotional ratings by listeners were predicted by up to 59% of lexical content and up to 26% of vocal cues. Listeners relied more on vocal cues to perceive emotions when the lexical content was ambiguous regarding feeling and attitude. Finally, the analysis identified statistically significant basic acoustic parameters and non/para-linguistic information after controlling for the effects of lexical content.

1 Introduction

This study is concerned with listeners' natural ability to identify an anonymous speaker's emotions in spoken language. The communication of emotions and attitudes relies, in addition to verbal messages, on non-verbal factors such as tone of voice and body language. In everyday life, we use our natural skills to detect and recognize a speaker's emotions through vocal tone, which encompasses paralinguistic and acoustic information [1,2]. This study focuses on how listeners perceive and infer speakers' feelings and attitudes from vocal tones, which is known as vocal affect. Vocal affect plays a crucial role in interpersonal interaction and communication [3]. Numerous studies of vocal affect have demonstrated a relationship between vocal cues and emotions. These studies often use synthetic speech samples [4-7], as well as recordings by actors and professionals portraying posed emotions [8-14].

One of the major challenges in evaluating vocal affect is controlling for the influence of semantics on emotional ratings, while preserving prosodic features and vocal information. One approach to minimize the effect of lexical meaning is to use speech in a foreign language [15]. Other methods include employing pseudo-random words [16] or applying low-pass filtering [17], which inherently distort acoustic parameters in the analysis.

Our previous studies have demonstrated that non-native listeners can identify the emotional states of communicators through vocal expressions [18,19]. However, non-native listeners might not fully utilize these vocal cues to recognize emotions in a foreign language [20-23]. For a more accurate assessment of emotions, it is preferable to use speech stimuli in the

Confidential

listeners' native languages. However, in the native language, it is challenging to remove the effect of literal meaning while maintaining the acoustic properties intact. Thus, this study examines how lexical content interacts with vocal cues to create the perception of emotion in spoken language by employing two experimental conditions: text and voice conditions.

Previous studies have addressed the intertwined nature of semantic and vocal cues in perceiving emotions and attitudes from vocal cues. When the tone of voice and the meaning of a word are inconsistently combined (e.g., a negative tone with a positive word), the inference of emotion is mainly determined by vocal cues relative to lexical content [2,24]. In a Stroop task, affective words are spoken with congruent or incongruent emotions (e.g., a word with a sad semantic meaning spoken with happy prosody) [25]. When participants are instructed to identify emotions conveyed by semantic content, incongruent prosody severely deteriorates their ability to accurately identify emotions.

The goal of this study was to explore the interaction between lexical content and vocal cues and how these factors contribute to the perception of emotion. The interaction between lexical and vocal cues was examined across a diverse set of dialogue excerpts with various affective attributes, reflecting vocal affects in more naturalistic communications. Furthermore, this study investigated the relationship between emotional evaluation and non/para-linguistic information in addition to basic acoustical parameters. This approach allowed us to understand how listeners utilize vocal cues to perceive emotions in native speech.

2 Methods

2.1 Stimuli

This study utilized the same dataset used in our previous study [19], comprising 317 Korean speech samples, which included 180 spontaneous and 137 posed samples. Spontaneous speech samples were sourced from a Korean Spontaneous Speech corpus [26]. The posed samples were obtained from the AI Hub Emotion Classification dataset [27].

From this dataset, we randomly selected 180 spontaneous clips, extracting 2-second segments from random points within each clip. For the posed samples, we chose 137 clips representing four emotions (anger, happiness, sadness, and neutral) and extracted 2-second segments from the beginning of each clip. The transcripts of the selected clips showed no overlap. Owing to the 2-second truncation, some clips contained incomplete utterances with abrupt starts or ends. All samples were saved in waveform audio format with a 16 kHz sampling rate, and their loudness levels were normalized to match one sample's root mean square (RMS). A 150-millisecond sigmoidal fade was applied to both the beginning and end of each sample.

The non/para-linguistic information of the selected speech samples was further analyzed and is listed in Table 1. The voice-gender ratio for all 317 samples was approximately 1:2. Most spontaneous samples featured incomplete sentences, whereas posed samples exhibited the opposite pattern. Spontaneous samples also included more filler words and interjections. The higher occurrence of incomplete sentences, along with a more frequent use of filler words and interjections, reflects the degree of naturalness and spontaneity experienced in everyday conversations. The posed samples composed with predetermined scripts had fewer incomplete sentences and fewer instances of filler words and interjections. Both sample types had similar laughter rates and a dominance of declarative sentences, with interrogative sentences occurring at similar frequencies. Exclamatory sentences appeared only in the posed samples, which contained breathier and tearful sounds. This reflected the intentional conveyance of stronger emotions, particularly sadness, in the posed samples.

		Spontaneous Speech		Posed Speech	
		Sample (#)	Sample (%)	Sample (#)	Sample (%)
Gender	Male	68	38	45	33
	Female	112	62	92	67
Sentence Completion	Complete	34	19	103	75
	Incomplete	146	81	34	25
Sentence Type	Declarative	154	86	104	76
	Interrogative	26	14	19	14
	Exclamatory	0	0	14	10
Filler Word		29	16	14	10
Repetitive word		11	6	1	1
Breathy sound	Inhale	14	8	3	2
	Exhale	5	3	24	18
	Inhale&Exhale	1	1	1	1
Laughter		7	4	2	1
Extension		42	23	32	23
Tearful Sound		0	0	14	10

Table 1. Non/para-linguistic information of speech samples.

2.2 Participants

This study was reviewed and approved by the ethics committee review board at Yonsei University. All participants provided informed consent before participating in the experiments. Native Korean participants were recruited through the Department of Psychology Research Management System of Yonsei University or Surveasy [28], a survey response collection website. After a 20- to 30-minute session, participants recruited through the Research Management System of Yonsei University received

Confidential

one credit for psychology courses at Yonsei University, while those recruited through Surveasy were compensated with approximately 5,000 won.

To ensure participant engagement, post-screening trials to recognize two images were included in each experimental condition. The post-screening rule was to exclude the data of those participants who did not pass both post-screening trials. In the voice condition, 211 participants were recruited and 5 participants were excluded based on the post-screening rule. This resulted in 206 participants (147 females, 57 males, and 2 unspecified). In the text condition, 174 participants were recruited, and 10 participants were excluded, resulting in 164 participants (90 females, 71 males, and 3 unspecified).

2.3 Procedure

To compare native Korean participants' evaluations of emotional expressions in both voice and text format, participants were assigned to either the text or voice condition. In the text condition, participants read the transcriptions of 2-second speech samples and rated the emotions they perceived. Each participant assessed 106 stimuli from a randomly selected batch of 317 speech samples divided into three batches. In the voice condition, participants listened to speech samples and rated the corresponding emotions, with each participant evaluating 79 stimuli from one of four randomly assigned batches of 317 samples. Each sample was rated by a minimum of 50 participants.

The experimental procedure was the same as that described in our previous study [19]. The participants in the voice condition used earphones or headphones and adjusted the volume to a reference speech sample with the same RMS level as the test stimuli. At the beginning of each trial, a speech sample was played on a blank screen for the voice condition, whereas a speech transcript without sound was displayed for the text condition. The same sample was repeated with a questionnaire page showing seven opposite pairs of affective attributes on a 7-point scale, with 0 as the midpoint. These pairs included *Relaxed-Stressed, Content-Angry, Friendly-Hostile, Sad-Happy, Bored-Interested, Intimate-Formal,* and *Timid-Confident* [4]. This method enables participants to assess emotions that are characterized by opposite attributes in valence and arousal on the two dimensional emotion wheel [29,30]. To prevent bias, the order of presentation of the attribute pairs was counterbalanced.

Participants could take a short break of at least one minute, if needed. The online experiment was conducted in Pavlovia [31] and programmed using the jsPsych library [32]. All statistical analyses were conducted using R software [33].

2.4 Acoustic Parameters

Seven acoustic parameters were extracted from each sample: pitch (mean, maximum, and minimum) in Hz, standard deviation (SD) of pitch, jitter as a percentage measuring frequency disturbance, shimmer in dB measuring amplitude disturbance, and harmonic-to-noise ratio (HNR) in dB.

The parameters for the analysis were extracted using Praat [34], using the software's default setting. In the default settings for all parameters, the silence threshold was set to 0.03, designating frames lacking amplitudes above this threshold as silent, whereas the voicing threshold was set to 0.45. The maximum period factor used to calculate the jitter was set to 1.3, and the maximum amplitude factor was set to 1.6. Pitch was measured within a 40ms analysis window, with 75 Hz as the pitch floor and 600 Hz as the pitch ceiling.

3 Results

3.1 Evaluation of emotional attributes

We computed the rating scores for the seven affective pairs by averaging the participants' response scores across 317 voice samples. Figure 1 presents the distribution of the participants' rating scores for the seven affective pairs with spontaneous and posed samples. Overall, the ratings exhibited a distribution centred around zero, without skewing towards the extremes. The responses to posed samples displayed higher mean ratings than those to spontaneous samples, indicating that the posed samples exhibited more intense emotions. The mean ratings were also higher for pairs that included the intended emotions such as anger, happiness, and sadness. In contrast, affective pairs with fewer associations with acted emotions, such as *Interested-Bored*, received lower mean ratings.

We then classified the samples into 14 affective attributes based on participants' response scores. For example, a sample with a *Relaxed-Stressed* score higher than 0 was categorized as a *Stressed* stimulus, whereas a score less than 0 was classified as a *Relaxed* stimulus. Samples scoring exactly 0 were not assigned any emotion. To compare participants' perceptions of emotions between spontaneous and posed samples in both text and voice conditions, the mean ratings and standard errors for the 14 affective attributes were computed using the absolute values of the participants' ratings. Figure 2 illustrates

Confidential

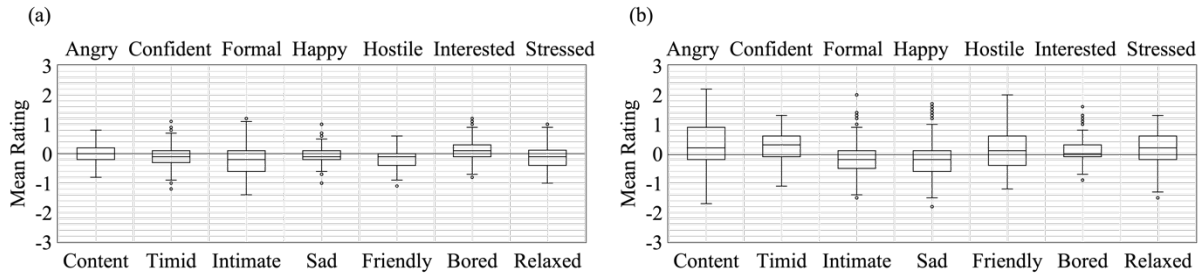


Figure 1. Evaluation results in the voice condition. Mean response scores for affective pairs with (a) spontaneous and (b) posed speech samples. Boxplots show interquartile range (box), median (horizontal line in box) with minimum and maximum values (vertical line), and extreme values (dots).

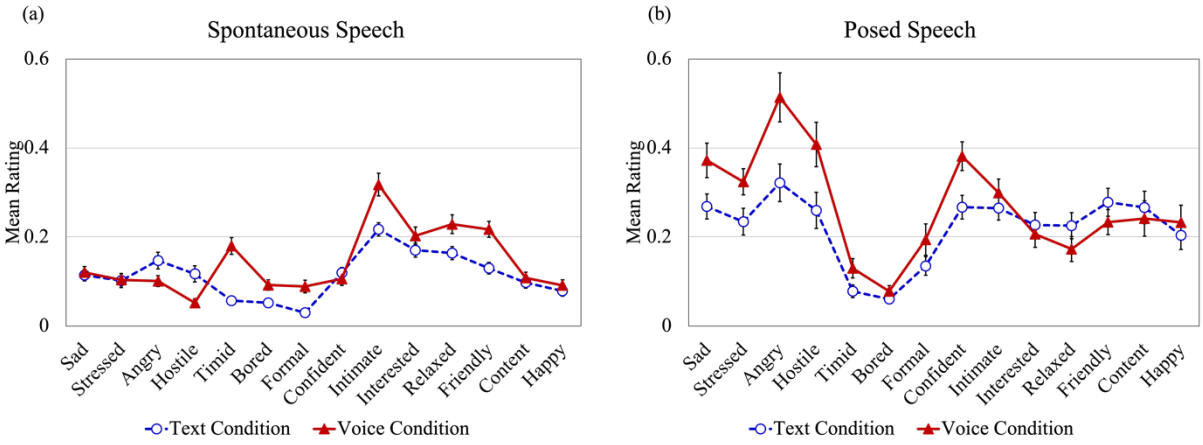


Figure 2. Mean ratings with standard errors for spontaneous and posed samples in text and voice conditions.

the mean ratings with standard errors for both the spontaneous and posed samples within each affective category, arranged by valence, from sadness (on the far left) to happiness (on the far right). For spontaneous speech, most emotion rating scores were higher in the voice condition than in the text condition, which was expected because speech generally conveys emotions more intensely than lexical cues alone. Interestingly, angry and hostile texts had higher emotion rating scores than the voice condition, suggesting that negative content may be expressed in a less negative tone. In the posed samples, scores in the voice condition tended to be higher than those in the text condition for angry, hostile, and confident emotions. However, there was no significant increase in the voice condition rating scores for happiness. Emotions that were less associated with the intended emotions, such as relaxation and boredom, showed little difference in scores between the two experimental conditions.

3.2 Relationship between perceived emotions in the text and voice conditions

Pearson's correlation coefficients were computed to examine the relationship between participants' perceived emotions in the text and voice conditions. This analysis aimed to evaluate how lexical information in the text condition was correlated with vocal cues in the voice condition, which varied across emotions.

Tables 2 and 3 present the correlations between the response scores in the text and voice conditions, for the spontaneous and posed samples, respectively. Overall, emotions perceived in the text condition were positively correlated with emotions of similar valence in the voice condition. However, this trend is less pronounced in spontaneous samples, where fewer significant correlations are observed. The results for spontaneous speech samples in Table 2 reveal two notable findings. First, no positive correlations were found between the text and voice conditions for specific emotional attributes such as stress, hostility, and timidity. Second, there were instances where negative valences in the text condition exhibited positive correlations with positive valences in the voice condition, as highlighted in Table 2. For

Confidential

example, angry and hostile texts were positively correlated with angry and hostile voices, but also showed a positive correlation with emotions of opposite valence in the voice condition, such as intimacy and confidence. In the case of posed samples highlighted in Table 3, some negative emotions in the text were perceived as positive in the voice condition. Specifically, angry and hostile texts were correlated with angry and hostile voices, while also showing positive correlations with confident voices, an emotion of opposite valence.

Text Condition	**Voice Condition for Spontaneous Speech Samples**													
	Sad	Stre.	Ang.	Host.	Tim.	Bor.	Form.	Conf.	Inti.	Inter.	Relax.	Frie.	Cont.	Hap.
Sad	**0.39****		**0.36****	0.15*										
Stressed	**0.29****		**0.29****						0.17*					
Angry	**0.38****		**0.35****	0.17*					**0.2****					
Hostile	**0.26****		**0.27****					0.16*	**0.27****					
Timid								**-0.20****						
Bored	**0.25****		0.17*	0.17*		**0.26****				**-0.21****			-0.16*	
Formal							**0.31****		-0.16*					
Confident					-0.16*			**0.29****		0.16*				**0.24****
Intimate							**-0.25****		**0.26****		0.18*		**0.21****	**0.32****
Interested	**-0.23****		-0.16*			**-0.24****		0.17*		**0.35****			**0.26****	**0.32****
Relaxed	**-0.27****	**-0.21****	**-0.2****		-0.17*						**0.21****	0.19*	**0.33****	**0.37****
Friendly	**-0.27****	-0.19*	**-0.26****	-0.18*		-0.16*						**0.22****	**0.3****	**0.3****
Content	-0.19*	-0.16*	**-0.19****							0.17*		**0.2****	**0.33****	**0.32****
Happy	**-0.21****		**-0.2****							**0.28****	**0.2****	**0.26****	**0.39****	**0.43****

*$p < .05$, **$p < .01$

Table 2. For spontaneous samples, correlations between participants' perceived emotions in the text (vertical) and voice condition (horizontal). Numbers shown in blue and red indicate positive and negative correlations respectively. Dotted lines indicate borders between negative and positive valence. Only statistically significant correlations are displayed.

Text Condition	**Voice Condition for Posed Speech Samples**													
	Sad	Stre.	**Ang.**	Host.	Tim.	Bor.	Form.	Conf.	Inti.	Inter.	Relax.	Frie.	Cont.	**Hap.**
Sad	**0.67****	**0.31****	**0.37****	**0.27****	**0.48****	0.2*	**-0.22****	**-0.31****		**-0.39****	**-0.4****	**-0.42****	**-0.43****	**-0.42****
Stressed	**0.32****	**0.63****	**0.64****	**0.6****						**-0.29****	**-0.35****	**-0.42****	**-0.34****	**-0.34****
Angry	**0.23****	**0.62****	**0.81****	**0.76****			-0.17*	0.18*		**-0.27****	**-0.35****	**-0.45****	**-0.34****	**-0.34****
Hostile		**0.59****	**0.76****	**0.74****				**0.24****		-0.21*	**-0.29****	**-0.39****	**-0.28****	**-0.28****
Timid	**0.41****				**0.66****	**0.34****		**-0.37****		-0.21*	-0.21*	-0.18*	-0.21*	-0.2*
Bored	0.21*	0.19*	**0.24****	0.2*		**0.56****				**-0.22****	**-0.22****	**-0.26****	**-0.23****	**-0.22****
Formal	-0.18*	**0.22****					**0.79****		**-0.38****	**-0.24****	**-0.27****		**-0.23****	**-0.25****
Confident	**-0.41****				**-0.37****	**-0.22****		**0.53****	**0.29****	**0.45****	**0.34****	**0.32****	**0.46****	**0.47****
Intimate		**-0.37****	**-0.25****	**-0.25****		-0.2*	**-0.37****		**0.64****	**0.54****	**0.48****	**0.38****	**0.49****	**0.53****
Interested	**-0.33****	**-0.36****	**-0.31****	**-0.27****	-0.19*	**-0.26****		**0.3****	**0.5****	**0.69****	**0.57****	**0.47****	**0.68****	**0.7****
Relaxed	**-0.37****	**-0.52****	**-0.49****	**-0.44****	-0.21*	**-0.22****		0.19*	**0.53****	**0.66****	**0.76****	**0.65****	**0.73****	**0.71****
Friendly	**-0.37****	**-0.49****	**-0.56****	**-0.51****	-0.17*		**0.25****		**0.4****	**0.49****	**0.53****	**0.73****	**0.56****	**0.54****
Content	**-0.43****	**-0.5****	**-0.49****	**-0.43****	**-0.25****	**-0.23****		**0.32****	**0.59****	**0.76****	**0.78****	**0.67****	**0.83****	**0.82****
Happy	**-0.4****	**-0.44****	**-0.41****	**-0.36****	**-0.23****	-0.21*		**0.38****	**0.56****	**0.76****	**0.71****	**0.6****	**0.81****	**0.82****

*$p < .05$, **$p < .01$

Table 3. For posed samples, correlations between participants' perceived affect in the text (vertical) and the voice condition (horizontal).

Confidential

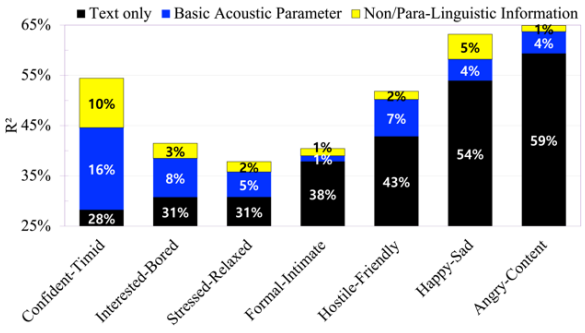


Figure 3. Linear regression results: R^2 of text, basic acoustic parameters, and non/para-linguistic information to predict emotional rating scores.

Regression \ Affective Pairs		Stressed -Relaxed	Hostile -Friendly	Angry -Content	Formal -Intimate	Confident -Timid	Interested -Bored	Happy -Sad
R^2 (Text Only)		31%	43%	59%	38%	28%	31%	54%
R^2		36%	50%	64%	39%	45%	39%	58%
Regression Coefficients	Text	0.5912**	0.6993**	0.8703**	0.9134**	0.6527**	0.6058**	0.8843**
	Mean Pitch	**0.0026****	**0.003****	**0.0025****				**-0.0018****
	Min. Pitch	**-0.0033****	**-0.0026****	**-0.0031****				**0.0031****
	SD Pitch	**-0.006****						**0.0067****
	Jitter					**-0.1507****	**-0.0744***	**-0.128****
	HNR		**-0.0435****	**-0.0338***		**-0.0917****	**-0.0252***	**-0.0288***

*$p < .05$, **$p < .01$

Table 4. Simple linear regression (R^2 text only) predicted by lexical meaning and multiple linear regression (R^2) predicted by basic acoustic parameters after controlling lexical meaning. Only statistically significant coefficients are displayed. The sign of the regression coefficient is based on the first attribute of the affective pair in bold.

Regression \ Affective Pairs		Stressed -Relaxed	Hostile -Friendly	Angry -Content	Formal -Intimate	Confident -Timid	Interested -Bored	Happy -Sad
R^2		38%	52%	65%	40%	54%	42%	63%
Regression Coefficients	Text	0.5792**	0.6807**	0.8509**	0.9155**	0.5757**	0.606**	0.7894**
	Mean Pitch	**0.0031****	**0.0033****	**0.0026****				**-0.0019****
	Min. Pitch	**-0.0032****	**-0.0023****	**-0.0028****				**0.0032****
	SD Pitch	**-0.006****						**0.0052****
	Jitter					**-0.1558****	**-0.0768***	**-0.1176****
	HNR		**-0.0315***			**-0.0806****	**-0.0226***	
	Male	**0.1544***	**0.152***					
	Complete Sentence					**0.1142***		
	Declarative						**-0.1002***	
	Exclamatory		**0.2659***	**0.2595***	**-0.2822***	**0.2575***		
	Repetitive Word					**-0.3815****		
	Breathy Exhale					**-0.1428***	**-0.1382***	
	Breathy Inhale					**-0.2122***		
	Laughter	**-0.2724***					**0.2303***	**0.271***
	Tearful					**-0.3393****		**-0.5****

*$p < .05$, **$p < .01$

Table 5. Multiple linear regression analysis of emotion perception on vocal cues. Forward stepwise selection was used to identify non/para-linguistic information.

In summary, speech samples containing negative lexical content were perceived as expressing positive emotions. Notably, some spontaneous angry speech was interpreted as intimacy owing to the vocal expressions. This phenomenon represents an incongruent cue that may occur in natural, everyday dialogues to express intimacy or familiarity between speakers.

3.3 Effects of lexical and vocal cues on emotion perception

To investigate how lexical information, basic acoustic parameters, and non/para-linguistic information influence perceived emotion, we conducted linear regression analyses. Figure 3 provides an overview of the linear regression results comparing the R^2 values of the text, basic acoustic parameters, and non/para-linguistic information. The lowest explained variance attributable to text was 28% for the *Confident-Timid* pair and the highest was 59% for the *Angry-Content* pair. This indicates that the script of the acting voice, including happiness, sadness, and anger, conveys strong emotions in its lexical content. The basic acoustic parameters explained variance from 1% for the *Formal-Intimate* pair to 16% for the *Confident-Timid* pair. Non/para-linguistic information had an explanatory power ranging from 1% for *Formal-Intimate* to 10% for *Confident-Timid.* Interestingly, the pair with the lowest explanatory power for lexical information (*Confident-Timid*) had the highest explanatory power for both acoustic parameters and non/para-linguistic information. These findings suggest that, when emotions and attitudes in lexical cues are ambiguous, native listeners rely more on vocal cues to infer the emotional tone of the voice that they are listening to.

Table 4 presents detailed results from the regression analyses investigating the relationship between perceived emotion and both lexical information and basic acoustic parameters. Simple linear regression analyses were conducted using mean rating scores in the voice condition as the dependent variable and mean rating scores in the text condition as the independent variable. Additionally, multiple linear regression analyses shown were performed using mean rating scores in the voice condition as the dependent variable and basic acoustic parameters including mean pitch, minimum pitch, maximum pitch, pitch variation, jitter, shimmer, and HNR as the independent variable, while controlling for the text's influence. Non-significant coefficients, such as those for maximum pitch and shimmer, were excluded from the regression table. The model using only the text condition score could explain 28% (for the *Confident-*

Confidential

Timid pair) to 59% (for the *Angry-Content* pair) of the variance in the listeners' ratings of the speech samples. When basic acoustic parameters were included while controlling for the influence of text, the explanatory power of the model increased, explaining between 36% (for the *Stressed-Relaxed* pair) and 64% (for the *Angry-Content* pair) of the variance in the listeners' ratings of the speech samples. Notice that the sign of the regression coefficient is based on the first attribute of the affective pair. The value of the regression coefficient quantifies the extent to which the dependent variable (emotional rating) varies with each one-unit change in the independent variables (acoustic parameters). Given that the acoustic parameters are measured in different units—such as pitch in Hz, jitter in %, and HNR in dB—the magnitude of the regression coefficient is expressed in these respective units. Stressed, hostile, and angry voices scored higher with a higher mean pitch and lower minimum pitch. In contrast, happy voices scored higher with a lower mean pitch and higher minimum pitch. For the *Stressed-Relaxed* pair, relaxed voices were perceived with lower SD pitch. For the *Hostile-Friendly* and *Angry-Content* pairs, hostile and angry voices were perceived with lower HNR. For the *Happy-Sad* pair, happy voices were perceived as both jitter and HNR decreased. For the *Confident-Timid* and *Interested-Bored* pairs, the same sign of significant coefficients as the *Happy-Sad* pair for jitter and HNR were observed, but without significant coefficients in pitch parameters. These findings suggest that listeners perceived happiness when the mean pitch values decreased and the minimum pitch values increased, while they perceived anger and hostility when these parameters varied in the opposite direction. Listeners appeared to utilize little pitch information to perceive confidence and interest.

Table 5 displays detailed results from multiple linear regression analysis investigating the relationship between perceived emotion and vocal cues while controlling for the effect of text. Mean rating scores in the voice condition served as the dependent variable, with basic acoustic parameters and non/para-linguistic information as independent variables. A forward selection methodology was employed, initially including only the seven basic acoustic parameters and then adding the non/para-linguistic information variables one by one based on their ability to improve the model's performance, as evaluated using the Akaike information criterion (AIC). Variables that improved the model were retained, with new additions ceasing once no further improvement was observed. The model, incorporating both basic acoustic parameters and non/para-linguistic information while controlling for the text condition, explained between 38% (*Stressed-Relaxed*) and 65% (*Angry-Content*) of the variance in listeners' ratings in the voice condition. As expected, listeners perceived happy voices with laughter and sad voices with tearful sounds. For the *Confident-Timid* pair, confident voices were perceived with complete or exclamatory sentences, and timid voices were perceived by vocal cues such as repetitive words, breathy sounds, or tearful sounds. For the *Interested-Bored* pair, interested voices were perceived with laughter, and bored voices were perceived with declarative sentences and exhaled breathy sounds. Overall, non/para-linguistic information had a greater influence on perception for the *Confident-Timid* and *Interested-Bored* pairs.

4 Conclusions

The main purpose of this study was to measure Korean native listeners' ability to perceive subtle emotions from spontaneous speech and basic emotions, such as happiness, sadness, and anger, posed by professionals. This study describes the interplay between lexical meanings and vocal cues across seven pairs of opposing affective attributes.

First, the mean rating results showed that the relative importance of the lexical and vocal cues varied across the 14 affective states of posed and spontaneous speech. Vocal cues enhanced sad and angry expressions conveyed by lexical content in posed speech, but had little effect on happy expressions. This finding aligns with previous research on the interaction between facial and vocal affect [35]. Correlation analysis revealed that lexical meanings indicating anger or hostility could be vocally expressed in positive affective states. In some cases, vocal cues may override angry or hostile lexical meaning, leading listeners to perceive intimacy in spontaneous speech and confidence in posed speech. This incongruity between lexical content and vocal expressions can occur in everyday conversation. For example, someone might use angry content but may still be perceived as intimate or caring because of the context or familiarity between the speakers.

Second, a linear regression was performed to estimate the contributions of lexical meaning and vocal cues. Vocal cues are divided into two categories: basic acoustic parameters and non/para-linguistic information. The regression results showed that emotional ratings were predicted by lexical content, ranging from 28% (*Confident-Timid* pairs) to 59% (*Angry-Content* pairs). Additionally, vocal cues, including both basic acoustic parameters and

Confidential

non/para-linguistic information, can estimate 2%−26% of native listeners' emotional ratings. Listeners tended to rely more on vocal cues to perceive emotional voices when lexical content appeared ambiguous, as in the *Confident-Timid* pair.

Finally, regression analyses revealed statistically significant basic acoustical parameters and other non/para-linguistic information after controlling for the effects of lexical content. Native listeners perceived hostile or angry speech with a higher mean pitch, whereas they perceived happy speech with a lower mean pitch and frequency perturbation (lower jitter). Confident emotions were recognized with several paralinguistic features such as lower breathing and complete exclamatory sentences.

In summary, this study quantified the influences of lexical content and vocal cues on emotional perception. Lexical content predicted native listeners' emotional ratings by up to 59%, whereas vocal cues predicted ratings by up to 26%. Moreover, this study demonstrated that the influence of vocal cues can vary across different emotional attributes, and identified specific emotions where vocal cues can override lexical contents. Future research will explore how non-native listeners utilize vocal cues. The current research underscores the complex interactions between lexical and vocal cues in emotional communication and offers insights into how emotions are perceived from speech, even when there is a mismatch between verbal and nonverbal information.

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 and 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] 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

[4] 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). https://doi.org/10.1016/S0167-6393(02)00082-1

[5] 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

[6] 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

[7] 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). https://doi.org/10.21437/Interspeech.2020-1938

[8] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–636 (1996). https://doi.org/10.1037//0022-3514.70.3.614

[9] K. R. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research,"

Confidential

Psychol. Bull., vol. 99, no. 2, pp. 143–165 (1986). https://doi.org/10.1037//0033-2909.99.2.143

[10] 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). https://doi.org/10.1007/s10936-017-9526-7

[11] 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/10.1016/j.biopsycho.2011.02.010

[12] 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). https://doi.org/10.1109/T-AFFC.2011.14

[13] 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 (2014). https://doi.org/10.1007/s10919-013-0165-x

[14] 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

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

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

[17] 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). https://doi.org/10.1016/j.specom.2008.08.001

[18] 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), paper 10016.

[19] E. Oh and J. Suhr, "Vocal Affects Perceived from Spontaneous and Posed Speech," presented at the 155*th Convention of the Audio Engineering Society Convention* (2023), paper 10671.

[20] 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). https://doi.org/10.1038/s41562-019-0533-6

[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). https://doi.org/10.1080/02699931.2013.812033

[22] 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). https://doi.org/10.1007/s10919-008-0065-7

[23] T. Waaramaa and T. Leisiö, "Perception of Emotionally Loaded Vocal Expressions and its Connection to Responses to Music. A Cross-Cultural Investigation: Estonia,

Confidential

Finland, Sweden, Russia, and the USA," *Front. Psychol.*, vol. 4, article 344 (2013). https://doi.org/10.3389/fpsyg.2013.00344

[24] A. Mehrabian, "Decoding of Inconsistent Communication," *J. Pers. Soc. Psychol.*, vol. 6, no. 1, pp. 109–114 (1967). https://doi.org/10.1037/h0024532

[25] 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

[26] 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). https://doi.org/10.3390/app10196936

[27] 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).

[28] Surveasy. https://gosurveasy.com (accessed Oct. 25, 2023)

[29] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia & S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, no. 6, pp. 907-929, (2019). https://doi.org/10.1007/s11263-019-01158-4

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

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

[32] 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, (2016).

[33] R Core Team. "R: A language and environment for statistical computing. R Foundation for Statistical Computing", Version. 4.2.1, (2022). https://www.R-project.org/

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

[35] 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). https://doi.org/10.3389/fnint.2015.00001