Sound-to-Touch Crossmodal Pitch Matching for Short Sounds

Dong-Geun Kim, Jungeun Lee, *Graduate Student Member, IEEE*,
Gyeore Yun, *Graduate Student Member, IEEE*, Hong Z. Tan, *Fellow, IEEE*,
and Seungmoon Choi, *Senior Member, IEEE*

***Abstract*—This article explores how to relate sound and touch in terms of their spectral characteristics based on crossmodal congruence. The context is the audio-to-tactile conversion of short sounds frequently used for user experience improvement across various applications. For each short sound, a single-frequency amplitude-modulated vibration is synthesized so that their intensive and temporal characteristics are very similar. It leaves the vibration frequency, which determines the tactile pitch, as the only variable. Each sound is paired with many vibrations of different frequencies. The congruence between sound and vibration is evaluated for 175 pairs (25 sounds × 7 vibration frequencies). This dataset is employed to estimate a functional relationship from the sound loudness spectrum of sound to the most harmonious vibration frequency. Finally, this sound-to-touch crossmodal pitch mapping function is evaluated using cross-validation. To our knowledge, this is the first attempt to find general rules for spectral matching between sound and touch.**

***Index Terms*—Audio-to-tactile conversion, congruence, crossmodal, spectral matching, vibrotactile pitch.**

I. Introduction

Haptic effects can improve the realism, immersiveness, and user experiences of various applications in virtual reality, multimedia, and gaming [1], [2], [3], [4]. Among the many approaches to generating haptic effects, the study we report in this paper concerns the *multisensory* effects that combine sound and touch. When manually designing or automatically generating tactile stimuli for sounds, we need criteria for the adequacy of the tactile stimuli. One unanimous metric is *crossmodal congruence:* how harmonious tactile stimuli feel with sounds [5]. A critical barrier against achieving high crossmodal auditory-tactile congruence stems from the auditory and tactile channels' disparate frequency ranges to which humans are sensitive. The perceptible frequency band for sound is 20–20,000 Hz, whereas that for touch is 0 to 1,000 kHz [6]. Moreover, the perceptual sensitivity to frequency differences is markedly better for sound than touch [6]. These fundamental perceptual differences make it arduous to find general rules or guidelines for matching the spectral content between sound and touch to a high degree of congruence.

This article explores methods to determine the *pitch* of a vibrotactile stimulus most congruent with a short sound. Such knowledge is essential for the auditory-to-tactile conversion of short sounds—a core technology for haptic user experience enhancement in interactive applications, such as video games, multimedia, and broadcasting; see Section II for related work. Given a sound signal, we prepared many vibratory stimuli in the form of a single-frequency amplitude-modulated sinusoidal function (Section III). They had similar intensive and temporal properties, leaving the frequency as the only variable affecting the tactile pitch. We collected many short sound samples in five categories from games and paired each sound sample with seven vibrations of different frequencies. We conducted a perceptual experiment in which participants assessed the crossmodal congruence between the sound and vibration stimuli by ranking. Then, we used this crossmodal congruence data to estimate a regression function that predicts the most congruent vibration frequency to a sound sample (Section IV). This was done by stepwise linear regression after representing the sound samples by their loudness spectra. Finally, the prediction performance of the *crossmodal pitch matching function* was evaluated by standard cross-validation (Section V). It is followed by conclusions along with a discussion on limitations and future work (Section VI).

II. Related Work

A. Multisensory Sound and Tactile Effects

In sound-to-touch conversion, tactile stimuli are usually generated from sounds for simultaneous presentation. The conversion considers the sound's intensive, temporal, and spectral properties and maps them to tactile features. To this end, early methods relied on signal processing, beginning with a low-pass filter that transforms only the low-frequency bass information in sound to tactile sensations [7]. Using a filter bank enables the expression of both the bass and treble components in sound by two vibrations of respective pitches [8], [9]. We can also combine the spectral components in selected audio bands into one vibration stimulus [10]. Alternatively, the entire sound spectrum can be compressed into the frequency band for tactile perception by frequency shifting [11]. Using a spatiotemporal coding scheme, the spectral energies of many frequency bands in a sound can excite different body sites, representing the auditory spectral content by full-body tactile patterns [12]. These methods are generally appropriate for expressing music with tactile stimuli and were demonstrated to enhance music listening experiences [8], [9], [11], [12].

Another approach extracts physical or perceptual features from sounds and uses them to synthesize tactile stimuli. For instance, pitch, loudness, brightness, and envelope computed from music can determine the properties of vibration effects to be played back together [13].

Manuscript received 25 September 2023; revised 20 November 2023; accepted 27 November 2023. Date of publication 1 December 2023; date of current version 21 March 2024. This work was supported in part by the Research Gift from Google for "Semantic Sound-to-Haptic Conversion: Algorithms and Performance Metrics", in part by IITP, Republic of Korea under Grant 2022-0-01005, and in part by NST, Republic of Korea under Grant CRC23021-400. This paper was recommended for publication by Associate Editor Mounia (GE HC 24) Ziat and Editor-in-Chief D. Prattichizzo upon evaluation of the reviewers' comments. *(Corresponding author: Seungmoon Choi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at POSTECH under Application No. PIRB-2022-119 E036.

Dong-Geun Kim, Jungeun Lee, Gyeore Yun, and Seungmoon Choi are with the Interaction Laboratory, Department of Computer Science and Engineering, POSTECH, Pohang 37673, Republic of Korea (e-mail: dgkim94@postech.ac.kr; jungeunlee@postech.ac.kr; ykre0827@postech.ac.kr; choism@postech.ac.kr).

Hong Z. Tan is with Google LLC, Mountain View, CA 94043 USA (e-mail: hongtan@google.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TOH.2023.3338224, provided by the authors.

Digital Object Identifier 10.1109/TOH.2023.3338224

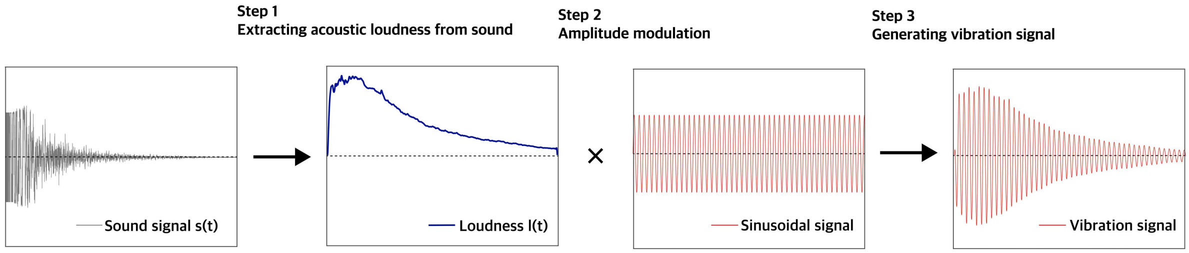


Fig. 1. Sound-to-touch conversion process preserving the intensive and temporal properties.

Perceptual features are also useful for selective conversion, where haptic effects are provided only for target sound classes. For example, haptic effects are desired for gunshot sounds but not for background music in gaming applications. Such selective conversion can be achieved by a perceptual mapping associating sound and tactile signals using their roughness and loudness [14]. This method can be extended to using more psychoacoustical features, loudness, booming, sharpness, and low-frequency energy, to determine when to present tactile effects from sounds [15]. Machine learning can also effectively control the classes of sounds for tactile conversion, using a neural network [16] and a random forest [17]. It was demonstrated that some of these methods can improve user experiences related to gaming [14], [16], [17] and movie viewing [14], [15].

Despite these efforts, we still need ways to explicitly address the spectral (pitch) harmony between sound and touch.

B. Multisensory Congruence

Congruence has been a popular and effective notion for multisensory design. For example, audio and tactile crossmodal icons were designed for mobile devices considering their congruence [18]. This idea was extended to include visual aspects [5]. Also, design guidelines were studied based on the emotional congruence between visual and tactile icons [19]. The effects of audio-tactile congruence were investigated for its extent of music experience enhancement [20]. Technologies and methods to provide harmonious musical audio and tactile experiences are surveyed in [21].

In comparison, the present work concentrates on only congruence in the spectral content between sound and touch. It seeks a functional relationship between them for short sounds useful for automatic auditory-to-tactile conversion.

III. Perceptual Experiment

We conducted a perceptual experiment to investigate the spectral relationship between congruent sound and tactile stimuli. As described earlier, the focus was on short sounds emphasizing event occurrences in multimedia content, such as games and movies.

A. Methods

1) Participants: Twenty volunteers (11 males and 9 females; 19–30 years old, $M = 24.5$, and $SD = 3.4$) participated in this experiment. None of the participants reported known sensorimotor disorders. The experiment took around 60 min, and each participant was paid KRW 20,000 ($\simeq$ USD 16) after the experiment.

2) Stimuli: We collected 25 sound samples from video games of five different types (gunshot, glass break, hitting, explosion, and sword clashing). All samples were shorter than 1 s and copyright-free. They are available in the supplemental video. Each sound sample $s(t)$ was converted to a vibration signal, as depicted in Fig. 1. First, an auditory loudness function $l(t)$ was computed from $s(t)$ using the ISO 532-1 standard [22]. Second, a vibration signal $v(t)$ was determined by

$$v(t) = l(t)\sin(2\pi f t), \tag{1}$$

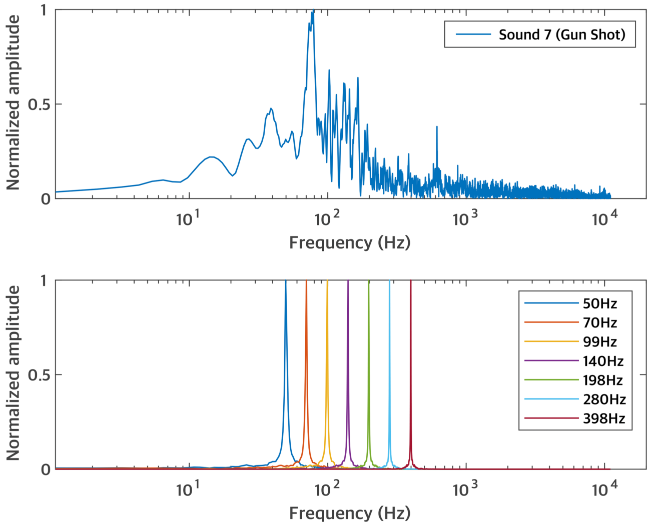


Fig. 2. Power spectra of a sound sample (top) and vibration stimuli (bottom). The vibration's peak frequency varied from 50 to 398 Hz to find the best match for crossmodal pitch perception.

where $v(t)$ is an amplitude-modulated sinusoid with the envelope $l(t)$ and the frequency f. For f, we chose seven values, 50, 70, 99, 140, 198, 280, and 398 Hz, at 0.5-octave spacing. Consequently, each sound sample was paired with seven vibrations; see Fig. 2.

This single-frequency amplitude-modulated waveform was selected for a few merits. First, it is widely used to model real collision events making sounds [23], [24]. Second, it allows us to predict the perceived tactile pitch clearly, even for short vibrations. Third, it can be generated by most vibration actuators if the vibration frequency falls within the actuator bandwidth. Therefore, the waveform is adequate for this study about crossmodal pitch matching.

In contrast, using more complex vibration waveforms in this study does not offer apparent advantages. For example, consider tactile vibrations consisting of two frequency components, which show systematic perceptual behaviors about consonance [25], dissimilarity [26], and perceived intensity [27]. However, nothing is known about the pitch perception of tactile vibrations that include two or more frequency components. It is more so for those of continuous spectra, which can result from applying frequency shifting to sounds [11]. Using stimuli with unknown pitch perception characteristics is unsuitable for crossmodal perception studies. Furthermore, producing such complex vibrations requires wideband actuators and frequency-turned rendering schemes, which may be overkill for the target applications using short sound-to-touch conversion.

We also scaled the sound and vibration signals as

$$s^*(t) = c_s s(t) \text{ and } v^*(t) = c_v v(t). \tag{2}$$

Each participant tuned the scaling constants c_s and c_v; see Section III-A3. Sound and vibration stimuli were produced using $s^*(t)$ and $v^*(t)$ as the input signals.

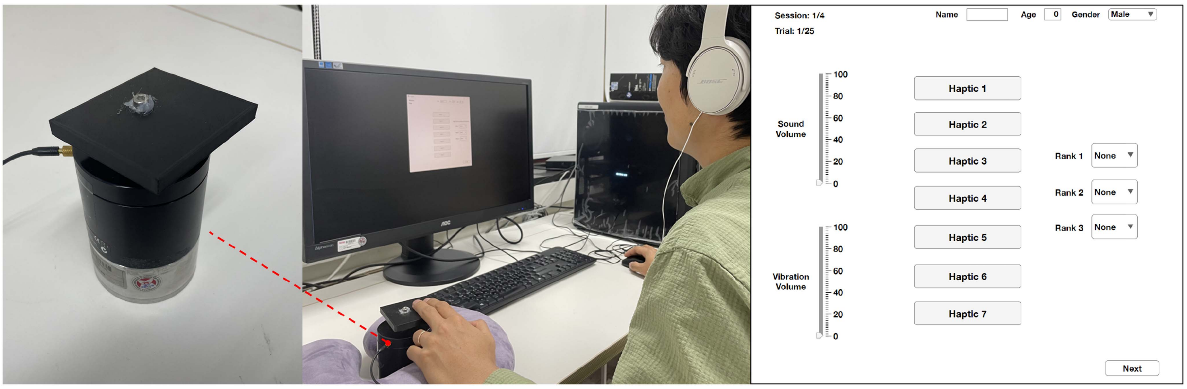

Session: 1/4
Trial: 1/25
Name
Age 0
Gender Male
Rank 1 None
Rank 2 None
Rank 3 None

Fig. 3. Experimental setup. The mini-shaker for stimulus production (left). A participant placed the index and middle fingers on a 3D-printed plate fastened to the shaker while wearing noise-canceling headphones (middle). Program interface (right).

$s^*(t)$ and $v^*(t)$ were designed to have identical, or at least very similar, properties in intensity and time. However, the frequency ranges of sound and tactile stimuli are disparate, and we paired each $s^*(t)$ with $v^*(t)$ of different frequencies varying widely from 50 to 400 Hz. Consequently, the essential task was matching the spectral characteristics between auditory and tactile stimuli.

3) Procedure: Participants sat in a chair in front of a computer monitor placed on a table (Fig. 3). Vibrations were generated by a mini-shaker (4810, Brüel Kjær) that had 18 g of effective moving mass. A 3D-printed plate (80 $\times$ 70 $\times$ 7 mm and 24 g) was fastened atop the shaker. The total moving mass for vibration stimulation was 42 g.[1] Participants put their left index and middle fingers on the plate for vibration perception. They wore earphones that played the sound samples and also noise-canceling headphones that generated while noise over the earphones. This setup allowed us to block all external sounds produced by the mini-shaker. Participants controlled the program's graphical user interface (GUI) using their right hands.

The experiment consisted of four sessions of the same procedure. The first session was to help participants become familiar with the stimuli and procedure. Each participant experienced all sound and vibration stimuli and then adjusted the sound and tactile volumes to comfortable levels using the sliders in the GUI (Fig. 3, right). This volume selection determined the conversion gains c_s and c_v in (2). These gains were used for the participant in the next three main sessions. Participants rested for 2 min between sessions.

Each session had 25 trials. Each trial was for one sound sample out of the 25. The order of the sound samples was randomized per session and participant. The GUI provided seven buttons in the center (Fig. 3, right). When participants clicked one button, a vibration with the frequency assigned to the button was played back with the sound sample. The assignment between buttons and vibration frequencies was randomly made for each trial. Participants could perceive the (sound + vibration) stimuli repeatedly. Then, they ranked the top three vibrations perceived as the most congruent with the sound sample using the drop-down lists on the GUI. Participants clicked the 'Next' button to proceed to the next trial.

We also tested a magnitude estimation task, i.e., representing the degree of congruence between sound and vibration using a number instead of ranking. However, many participants were inexperienced with the vibrotactile stimuli varying significantly in frequency and showed substantial response inconsistency. Hence, we decided to use the ranking task to improve response consistency while giving up the advantages of interval-scale data.

4) Data Analysis: Only the data collected in the three main sessions were used for analysis. The ranking data of all participants were pooled and then represented as follows. We denote the sound sample $i \in \{1, 2, \ldots, 25\}$ by s_i and the vibration signal $j \in \{1, 2, \ldots, 7\}$ for the sound sample i by $v_{i,j}$. $v_{i,j}$ has a frequency of f_j in the frequency vector $\boldsymbol{f} = (50, 70, 99, 140, 198, 280, 398\,\text{Hz})^T$. For each pair of s_i and $v_{i,j}$, we obtained three counts, $r^1_{i,j}$, $r^2_{i,j}$, and $r^3_{i,j}$, that summarized how many times the participants ranked the pair as the first, second, or third, respectively, for congruence. These counts are represented by three matrices, such that

$$\boldsymbol{R}^k = \left[r^k_{i,j} \right]. \tag{3}$$

Here, $\boldsymbol{R}^k$ has a dimension of 25$\times$7, and $k \in \{1, 2, 3\}$. Also, the row sums satisfy

$$\sum_{j=1}^{7} r^k_{i,j} = 60, \tag{4}$$

as we had 20 participants and 3 repetitions per condition.

We also used weights $\boldsymbol{w} = (w_1, w_2, w_3)^T$ to combine $\boldsymbol{R}^1$, $\boldsymbol{R}^2$, and $\boldsymbol{R}^3$, such that

$$\boldsymbol{P} = \sum_{k=1}^{3} w_k \boldsymbol{R}^k. \tag{5}$$

Each $p_{i,j}$ in $\boldsymbol{P}$ represents the extent of congruence of vibration j that has frequency f_j to sound sample i. For $\boldsymbol{w}$, we tested three choices: $\boldsymbol{w}_1 = (3, 2, 1)^T$, $\boldsymbol{w}_2 = (5, 3, 1)^T$, and $\boldsymbol{w}_3 = (7, 4, 1)^T$, where the latter weights emphasize the higher ranked vibrations more. It was, although empirical, to choose the weighting scheme that yielded the most accurate predictions for the congruent vibration frequencies.

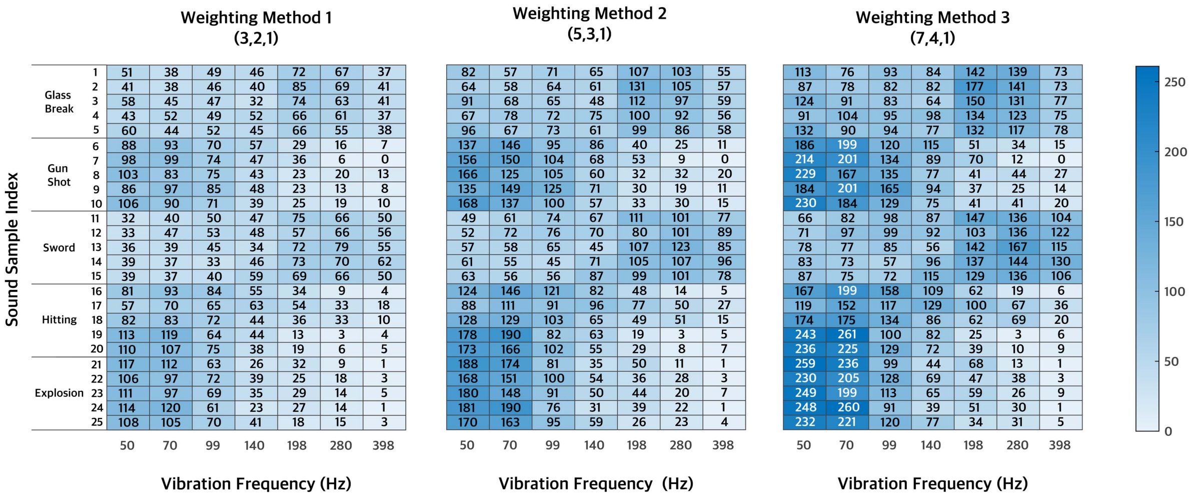
B. Results

The congruence score matrices $\boldsymbol{P}$ obtained by the experiment are shown in Fig. 4 for the three weighting methods. The sounds for glass break (index 1–5) and sword (11–15) showed higher congruence scores with the high-frequency ($\geq$ 198 Hz) vibrations. Contrarily, the sounds for gunshot (6–10), hitting (16–20), and explosion (21–25) were more congruent with the low-frequency ($\leq$ 100 Hz) vibrations. These observations were common for all weighting methods.

C. Discussion

We examined the specific loudness plots of the sound samples for their association with the experimental results. Examples are shown in Fig. 5, where the frequency axis is in the widely-used Bark scale [29].

[1] In vibrotactile perception, the weight of a contactor affects the perceived intensity; the heavier, the stronger [28].



Weighting Method 1 (3,2,1)

Sound Sample Index		50	70	99	140	198	280	398
Glass Break	1	51	38	49	46	72	67	37
	2	41	38	46	40	85	69	41
	3	58	45	47	32	74	63	41
	4	43	52	49	52	66	61	37
	5	60	44	52	45	66	55	38
Gun Shot	6	88	93	70	57	29	16	7
	7	98	99	74	47	36	6	0
	8	103	83	75	43	23	20	13
	9	86	97	85	48	23	13	8
	10	106	90	71	39	25	19	10
Sword	11	32	40	50	47	75	66	50
	12	33	47	53	48	57	66	56
	13	36	39	45	34	72	79	55
	14	39	37	33	46	73	70	62
	15	39	37	40	59	69	66	50
Hitting	16	81	93	84	55	34	9	4
	17	57	70	65	63	54	33	18
	18	82	83	72	44	36	33	10
	19	113	119	64	44	13	3	4
	20	110	107	75	38	19	6	5
Explosion	21	117	112	63	26	32	9	1
	22	106	97	72	39	25	18	3
	23	111	97	69	35	29	14	5
	24	114	120	61	23	27	14	1
	25	108	105	70	41	18	15	3

Vibration Frequency (Hz)

Weighting Method 2 (5,3,1)

Sound Sample Index		50	70	99	140	198	280	398
Glass Break	1	82	57	71	65	107	103	55
	2	64	58	64	61	131	105	57
	3	91	68	65	48	112	97	59
	4	67	78	72	75	100	92	56
	5	96	67	73	61	99	86	58
Gun Shot	6	137	146	95	86	40	25	11
	7	156	150	104	68	53	9	0
	8	166	125	105	60	32	32	20
	9	135	149	125	71	30	19	11
	10	168	137	100	57	33	30	15
Sword	11	49	61	74	67	111	101	77
	12	52	72	76	70	80	101	89
	13	57	58	65	45	107	123	85
	14	61	55	45	71	105	107	96
	15	63	56	56	87	99	101	78
Hitting	16	124	146	121	82	48	14	5
	17	88	111	91	96	77	50	27
	18	128	129	103	65	49	51	15
	19	178	190	82	63	19	3	5
	20	173	166	102	55	29	8	7
Explosion	21	188	174	81	35	50	11	1
	22	168	151	100	54	36	28	3
	23	180	148	91	50	44	20	7
	24	181	190	76	31	39	22	1
	25	170	163	95	59	26	23	4

Vibration Frequency (Hz)

Weighting Method 3 (7,4,1)

Sound Sample Index		50	70	99	140	198	280	398
Glass Break	1	113	76	93	84	142	139	73
	2	87	78	82	82	177	141	73
	3	124	91	83	64	150	131	77
	4	91	104	95	98	134	123	75
	5	132	90	94	77	132	117	78
Gun Shot	6	186	199	120	115	51	34	15
	7	214	201	134	89	70	12	0
	8	229	167	135	77	41	44	27
	9	184	201	165	94	37	25	14
	10	230	184	129	75	41	41	20
Sword	11	66	82	98	87	147	136	104
	12	71	97	99	92	103	136	122
	13	78	77	85	56	142	167	115
	14	83	73	57	96	137	144	130
	15	87	75	72	115	129	136	106
Hitting	16	167	199	158	109	62	19	6
	17	119	152	117	129	100	67	36
	18	174	175	134	86	62	69	20
	19	243	261	100	82	25	3	6
	20	236	225	129	72	39	10	9
Explosion	21	259	236	99	44	68	13	1
	22	230	205	128	69	47	38	3
	23	249	199	113	65	59	26	9
	24	248	260	91	39	51	30	1
	25	232	221	120	77	34	31	5

Vibration Frequency (Hz)

Fig. 4. Congruence scores between sound and vibration. The scores are color-coded.

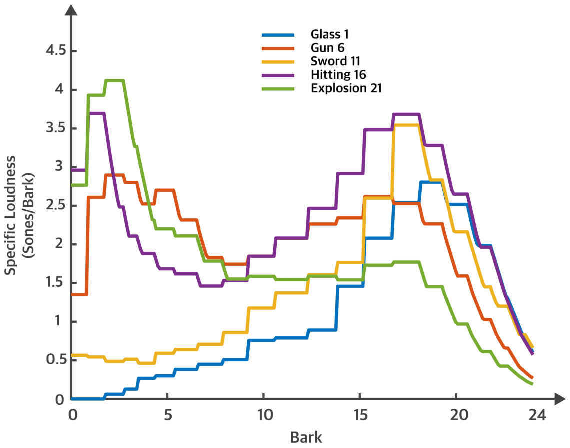


Fig. 5. Specific loudness plots of five sound samples.

Specific loudness values computed from the 24 Bark intervals are shown as plots for five sound samples selected for the five sound categories. Three sound samples, Gun 6, Hitting 16, and Explosion 21, had strong specific loudness values in the low-frequency regions below Bark 10 (1080–1270 Hz). In Fig. 4, the three sound samples showed high congruence with the low-frequency vibrations below 70 Hz. The other sound samples, Glass 1 and Sword 11, had prominent energies only in the high-frequency regions above Bark 10. Their congruence scores in Fig. 4 were large with the high-frequency vibrations above 198 Hz.

We can also obtain correlations between the specific loudness values of the sound samples (Fig. 5) and their congruence scores with different frequency vibrations (Fig. 4). For a sound sample, the specific loudness values of the 24 Bark bands are arranged in a sequence l_n in increasing order of sound frequency. The congruence scores of the sound sample are also put into a sequence in increasing order of vibration frequency. This sequence has a length of 7. Its frequency scale is transformed to a log scale, and the result is interpolated to another sequence q_n with a length of 24 for consistency with the Bark scale. Then, we compute a correlation between l_n and q_n. Pearson's correlation coefficient averaged over the 24 sound samples was $r = 0.631$, indicating a strong correlation [30].

The above results suggest that the loudness of a sound and its congruence to vibration are correlated in the frequency domain to an extent sufficient for further study.

IV. Crossmodal Pitch Matching

In this section, we explore how to relate the spectral features of a sound to the frequency of an amplitude-modulated vibration based on the congruence data obtained from the perceptual experiment.

A. Crossmodal Pitch Matching Function

We aimed to find a crossmodal pitch matching function $\psi(\cdot)$:

$$f = \psi(\boldsymbol{\lambda}), \tag{6}$$

where $\boldsymbol{\lambda}$ is a feature vector representing the spectral characteristics of a sound sample, and f is the frequency of the vibration in the form of (1) that feels the most congruent to the sound sample. Based on the results in Section III-C, we use the specific loudness values in the Bark scale to constitute the sound feature vector $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = (l_1, l_2, \ldots, l_{24})^T, \tag{7}$$

where l_i is the sound's specific loudness in the Bark band i.

For function fitting, we represent the congruence score distribution of each sound sample in Fig. 4 to one vibration frequency ζ by

$$\zeta_i = \frac{1}{\sum_{j=1}^{7} p_{i,j}} \sum_{j=1}^{7} p_{i,j} f_j \tag{8}$$

where $p_{i,j}$ is the (i, j) element of the congruence score matrix $\boldsymbol{P}$, and f_j is the j-th element of the frequency vector $\boldsymbol{f}$. ζ_i is the frequency averaged using the congruence scores as the weights. We regard ζ_i as the most congruent vibration frequency to the sound sample i. The values of ζ are shown in Fig. 6.

The above procedure provides a dataset $\boldsymbol{\Gamma}$, such that

$$\boldsymbol{\Gamma} = \{(\boldsymbol{\lambda}_i, \zeta_i) \mid i = 1, 2, \ldots, 25\} \tag{9}$$

for the 25 sound samples. We apply stepwise linear regression to learn the crossmodal spectral matching function ψ on $\boldsymbol{\Gamma}$. This method can reduce the input variables to only significant ones, resulting in a relatively simple linear relationship.

B. Results and Discussion

An optimal regression model found by the stepwise linear regression had the following form:

$$\zeta = \psi(\boldsymbol{\lambda}) = \sum_{i \in I} a_i \lambda_i, \tag{10}$$

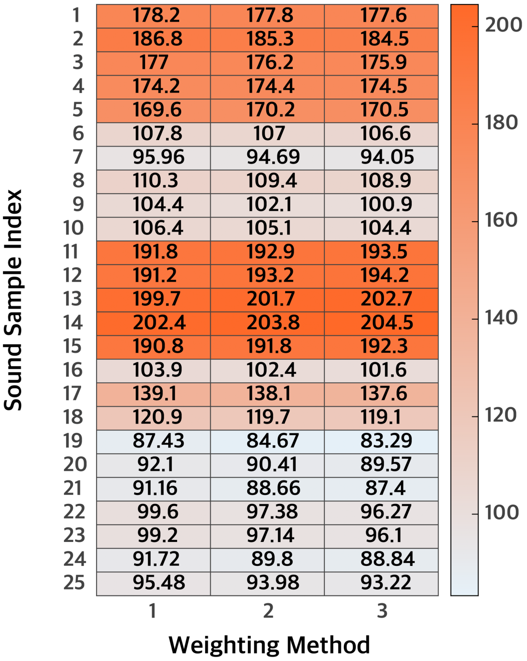


Fig. 6. Most congruent vibration frequencies to the sound samples.

TABLE I
RESULTS OF STEPWISE REGRESSION FOR CROSSMODAL MATCHING

	Weighting Method		
	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$	$\boldsymbol{w}_3$
R^2	0.920	0.935	0.934
p-value	1.12e-10	1.34e-10	1.38e-10
Significant Bark bands	2, 9, 12, 24	2, 3, 9, 12, 24	2, 3, 9, 12, 24

TABLE II
REGRESSION COEFFICIENTS a_i

Bark Band (i)	Frequency (Hz)	Weighting Method		
		$\boldsymbol{w}_1$	$\boldsymbol{w}_2$	$\boldsymbol{w}_3$
2	100–200	-0.003	-0.005	-0.005
3	200–300	—	0.003	0.003
9	920–1080	-0.013	-0.015	-0.015
12	1480–1720	0.009	0.008	0.008
24	12000–15500	0.006	0.008	0.008

where $I \subset \{1, 2, \ldots, 24\}$ is an index set for significant terms. The regression results and the coefficients for the three weighting methods $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, and $\boldsymbol{w}_3$ are shown in Tables I and II, respectively. The coefficient $a_i = 0$ for insignificant terms. Also, no interaction terms were significant, so they are not included in the regression model (10).

According to Table I, only $\lambda_2, \lambda_9, \lambda_{12}$, and λ_{24} were significant for all three weighting methods. λ_3 was significant only when the weighting method was $\boldsymbol{w}_2$ and $\boldsymbol{w}_3$. R^2 ranged from 0.920 to 0.934, with very small p-values. These results demonstrate that, for short sounds, we can obtain a well-defined function from a sound spectrum to a vibration pitch accounting for their perceptual congruence.

Table II shows the coefficients of the significant terms. A negative coefficient means that the spectral energy in the corresponding Bark band decreases the most congruent vibration frequency ζ. This role is shared by the Bark band 2 (100–200 Hz) and 9 (920–1080 Hz). The band 2 is critical for bass sound, and the band 9 is within the midrange of sound frequency in which upper harmonics frequently begin [31]. The band 9 has 4–5 times more influence in decreasing ζ than the band 2 according to their coefficients. Contrarily, the Bark bands 12 (1480–1720 Hz) and 24 (12000–15500 Hz) have positive coefficients. Thus, their spectral energies increase ζ. The band 12 is included in the upper midrange region, which adds volume to a midrange sound or gives articulation to a vocal [32]. The band 24 is the highest Bark band within the region often called the "air" or "openness" of sound [32]. Interestingly, the linear regression chose the mid- and extremely high-frequency bands instead of other high-frequency Bark bands (15–20) that transmit treble sound. The contributions of bands 12 and 24 are similar, as assessed from their similar coefficients. Lastly, the Bark band 3, significant only for $\boldsymbol{w}_2$ and $\boldsymbol{w}_3$ emphasizing higher contrast of congruence, had a positive coefficient. Its coefficient was 0.003, decreasing the effect of the adjacent band 2 with a coefficient of -0.005.

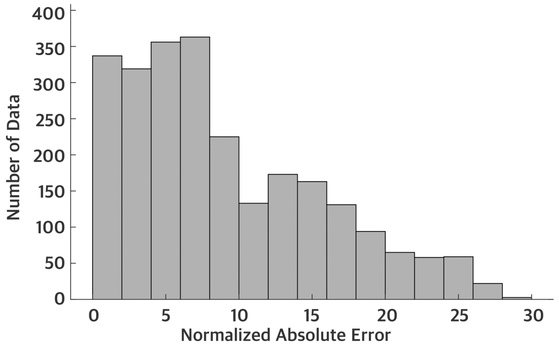


Fig. 7. Histogram for the normalized absolute error e.

The above results indicate that only partial spectral information of sound is required for conversion to a tactile stimulus that preserves congruence in pitch. This computational efficiency can be an important merit for real-time applications.

V. VALIDATION

In this section, we evaluate the performance of the crossmodal pitch matching function estimated in Section IV using cross-validation. The predicted frequencies of pitch matching functions using only the five Bark bands in Table II were compared with the most congruent frequencies obtained in the perceptual experiment (Section III).

A. Methods

We used 5-fold cross-validation on the dataset $\boldsymbol{\Gamma}$ in (9). $\boldsymbol{\Gamma}$ consisted of the five sound samples for each of the five categories. From $\boldsymbol{\Gamma}$, we randomly made five subsets, $\boldsymbol{\Gamma}_m (m = 1, 2, \ldots, 5)$, so that each $\boldsymbol{\Gamma}_m$ includes only one sample in each of the five categories. We selected and combined four subsets into a reduced training dataset, denoted by $\boldsymbol{\Gamma}_T$. The other subset, denoted by $\boldsymbol{\Gamma}_V$, was used for validation. This procedure made five cases for each of the five subsets.

We estimated a crossmodal pitch matching function ψ using linear regression on the reduced dataset $\boldsymbol{\Gamma}_T$. We used the weighting method $\boldsymbol{w}_2$ and its five significant terms, λ_2, λ_3, λ_9, λ_{12}, and λ_{24}, shown in Table I, for the regression. This configuration was chosen owing to its highest R^2. Then, for the sound samples in the validation set $\boldsymbol{\Gamma}_V = \{(\boldsymbol{\lambda},\ \zeta)\}$, we computed

$$\tilde{\zeta} = \psi(\boldsymbol{\lambda}). \tag{11}$$

This estimated vibration frequency $\tilde{\zeta}$ for the best congruence was compared with the measured frequency ζ using a normalized absolute error in percentage:

$$e = 100 \left| \frac{\tilde{\zeta} - \zeta}{\zeta} \right| (\%). \tag{12}$$

e was computed for each of the five sound samples in $\boldsymbol{\Gamma}_V$.

B. Results

The 5-fold cross-validation procedure for one random composition of the five sub-datasets ($\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \cdots, \mathbf{\Gamma}_5$) generated 25 values of the normalized absolute error e (5 regression functions × 5 errors per regression function). This procedure was repeated 100 times, resulting in 2500 values of e. This error distribution had $M = 9.03\%$ and $SD = 6.66\%$. It is also represented by a histogram in Fig. 7.

These errors can be compared with the JND (just noticeable difference) of vibrotactile frequency, which is between 17 and 21% in the frequency range of 50–300 Hz [33]. According to Fig. 7, 84.2% of the errors are below the JND (= 20%). 15.8% of the errors are over the JND, but they are all lower than 30% and not easy to perceive in actual applications. This analysis allows us to conclude that our method of crossmodal pitch matching from sound to tactile vibration has competent performance.

VI. Conclusion

This study allows us to provide the following conclusions. For a short sound matched with a single-frequency mechanical vibration that has very similar intensive and temporal properties, 1) the vibration frequency determines the congruence between the sound and vibration; 2) the loudness spectrum of the sound is correlated with the most congruent vibration frequency; 3) the most congruent vibration frequency can be predicted by a linear function of the sound loudness values in several frequency bands; 4) the high-frequency energy in sound generally increases the most congruent vibration frequency, whereas the low-frequency energy decreases it; and 5) the prediction error of the most congruent vibration frequency is acceptable for human tactile frequency discrimination.

These findings should be accompanied by a few remarks. First, our experiment used a ranking task. Therefore, given a short sound, our results allow prediction of the most congruent frequency, but not the extent of congruence. Second, our results need to be carefully interpreted for applications using ERM (Eccentric Rotation Mass) motors. These actuators have correlated amplitude and frequency in vibration output, and amplitude control often has a higher priority. Third, the contact location was the fingertips, and the vibration device was grounded in our experiment. How our results generalize to other contact sites or ungrounded cases, e.g., holding a mobile phone in the hand, require further examination.

References

[1] C. Krogmeier, C. Mousas, and D. Whittinghill, “Human–virtual character interaction: Toward understanding the influence of haptic feedback,” *Comput. Animation Virtual Worlds*, vol. 30, no. 3/4, 2019, Art. no. e1883.

[2] F. Danieau, A. Lécuyer, P. Guillotel, J. Fleureau, N. Mollet, and M. Christie, “Enhancing audiovisual experience with haptic feedback: A survey on HAV,” *IEEE Trans. Haptics*, vol. 6, no. 2, pp. 193–205, Apr.–Jun. 2013.

[3] S.-Y. Kim and K.-Y. Kim, “Interactive racing game with graphic and haptic feedback,” in *Proc. Int. Workshop Haptic Audio Interac. Des.*, Springer, 2007, pp. 69–77.

[4] M. Khamis, N. Schuster, C. George, and M. Pfeiffer, “Electrocutscenes: Realistic haptic feedback in cutscenes of virtual reality games using electric muscle stimulation,” in *Proc. ACM Symp. Virtual Reality Softw. Technol.*, 2019, pp. 1–10.

[5] E. Hoggan, T. Kaaresoja, P. Laitinen, and S. Brewster, “Crossmodal congruence: The look, feel and sound of touchscreen widgets,” in *Proc. Int. Conf. Multimodal Interfaces*, 2008, pp. 157–164.

[6] E. B. Goldstein, *Sensation and Perception*, 6th ed. Pacific Grove, CA, USA: Wadsworth-Thomson Learn., 2002.

[7] K. A. Li, T. Y. Sohn, S. Huang, and W. G. Griswold, “Peopletones: A system for the detection and notification of buddy proximity on mobile phones,” in *Proc. Int. Conf. Mobile Syst.*, 2008, pp. 160–173.

[8] I. Hwang, H. Lee, and S. Choi, “Real-time dual-band haptic music player for mobile devices,” *IEEE Trans. Haptics*, vol. 3, no. 3, pp. 340–351, Jul.–Sep. 2013.

[9] I. Hwang and S. Choi, “Improved haptic music player with auditory saliency estimation,” in *Proc. Int. Conf. Hum. Haptic Sens. Touch Enabled Comput. Appl.*, 2014, pp. 232–240.

[10] J.-M. Lim, J.-U. Lee, K.-U. Kyung, and J.-C. Ryou, “An audio-haptic feedbacks for enhancing user experience in mobile devices,” in *Proc. IEEE Consum. Electron.*, 2013, pp. 49–50.

[11] R. Okazaki, H. Kuribayashi, and H. Kajimoto, “The effect of frequency shifting on audio–tactile conversion for enriching musical experience,” in *Haptic Interaction: Perception, Devices and Applications*. Berlin, Germany: Springer, 2015, pp. 45–51.

[12] M. Karam, F. A. Russo, and D. I. Fels, “Designing the model human cochlea: An ambient crossmodal audio-tactile display,” *IEEE Trans. Haptics*, vol. 2, no. 3, pp. 160–169, Jul.-Sep. 2009.

[13] D. M. Birnbaum and M. M. Wanderley, “A systematic approach to musical vibrotactile feedback,” in *Proc. Int. Comput. Music Conf.*, 2007, pp. 397–404.

[14] J. Lee and S. Choi, “Real-time perception-level translation from audio signals to vibrotactile effects,” in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, 2013, pp. 2567–2576.

[15] Y. Li, Y. Yoo, A. Weill-Duflos, and J. Cooperstock, “Towards context-aware automatic haptic effect generation for home theatre environments,” in *Proc. ACM Symp. Virtual Reality Softw. Technol.*, 2021, pp. 1–11.

[16] G. Yun, H. Lee, S. Han, and S. Choi, “Improving viewing experiences of first-person shooter gameplays with automatically-generated motion effects,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–14.

[17] G. Yun, M. Mun, J. Lee, D.-G. Kim, H. Z. Tan, and S. Choi, “Generating real-time, selective, and multimodal haptic effects from sound for gaming experience enhancement,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–17.

[18] E. Hoggan and S. Brewster, “Designing audio and tactile crossmodal icons for mobile devices,” in *Proc. Int. Conf. Multimodal Interfaces*, 2007, pp. 162–169.

[19] T. Yoo, Y. Yoo, and S. Choi, “An explorative study on crossmodal congruence between visual and tactile icons based on emotional responses,” in *Proc. Int. Conf. Multimodal Interfaces*, 2014, pp. 96–103.

[20] S. C. Aker, H. Innes-Brown, K. F. Faulkner, M. Vatti, and J. Marozeau, “Effect of audio-tactile congruence on vibrotactile music enhancement,” *J. Acoust. Soc. Amer.*, vol. 152, no. 6, pp. 3396–3409, 2022.

[21] B. Remache-Vinueza, A. Trujillo-León, M. Zapata, F. Sarmiento-Ortiz, and F. Vidal-Verdú, “Audio-tactile rendering: A review on technology and methods to convey musical information through the sense of touch,” *Sensors*, vol. 21, no. 19, 2021, Art. no. 6575.

[22] *Methods for Calculating Loudness*, ISO Standard 532-1:2017, 2017. [Online]. Available: https://www.iso.org/standard/63077.html

[23] A. M. Okamura, M. R. Cutkosky, and J. T. Dennerlein, “Reality-based models for vibration feedback in virtual environments,” *IEEE/ASME Trans. Mechatron.*, vol. 6, no. 3, pp. 245–252, Sep. 2001.

[24] G. Park and S. Choi, “A physics-based vibrotactile feedback library for collision events,” *IEEE Trans. Haptics*, vol. 10, no. 3, pp. 325–337, Jul.–Sep. 2017.

[25] Y. Yoo, I. Hwang, and S. Choi, “Consonance of vibrotactile chords,” *IEEE Trans. Haptics*, vol. 7, no. 1, pp. 3–13, Jan.–Mar. 2014.

[26] I. Hwang, J. Seo, and S. Choi, “Perceptual space of superimposed dual-frequency vibrations in the hands,” *PLoS One*, vol. 12, no. 1, 2017, Art. no. e01695702016.

[27] Y. Yoo, I. Hwang, and S. Choi, “Perceived intensity model of dual-frequency superimposed vibration,” *IEEE Trans. Haptics*, vol. 15, no. 2, pp. 405–415, Apr.-Jun. 2022.

[28] H.-Y. Yao, D. Grant, and M. Cruz, “Perceived vibration strength in mobile devices: The effect of weight and frequency,” *IEEE Trans. Haptics*, vol. 3, no. 1, pp. 56–62, Jan.–Mar. 2010.

[29] E. Zwicker, “Subdivision of the audible frequency range into critical bands (frequenzgruppen),” *J. Acoust. Soc. Amer.*, vol. 33, no. 2, pp. 248–248, 1961.

[30] Complete Dissertation, “Pearson’s correlation coefficient.” Accessed: Dec. 7, 2023. [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/

[31] F. Koulakos, “Frequency band characteristics (part 1).” Accessed: Dec. 7, 2023. [Online]. Available: https://www.musical-u.com/learn/frequency-band-characteristics-part-1/

[32] F. Koulakos, “Frequency band characteristics (part 2).” Accessed: Dec. 7, 2023. [Online]. Available: https://www.musical-u.com/learn/frequency-band-characteristics-part-2/

[33] H. Pongrac, “Vibrotactile perception: Examining the coding of vibrations and the just noticeable difference under various conditions,” *Multimedia Syst.*, vol. 13, no. 4, pp. 297–307, 2008.