Acoustic evaluation of voice signal distortion by videoconferencing platforms and devices used in telepractice for cleft palate

Tajiri, Shiho
Section of Oral and Maxillofacial Surgery, Division of Maxillofacial Diagnostic and Surgical Sciences, Faculty of Dental Science, Kyushu University

Hidaka, Shunsuke
Graduate School of Design, Kyushu University

Takehisa, Shuhei
Graduate School of Design, Kyushu University

Hasegawa, Sachiyo
Department of Oral and Maxillofacial Surgery, Kyushu University Hospital

他

出版情報：Congenital Anomalies. 64 (6), pp.242-253, 2024-10-08. 日本先天異常学会
バージョン：

Received: 18 October 2023 | Revised: 1 August 2024 | Accepted: 20 September 2024

DOI: 10.1111/cga.12584

ORIGINAL ARTICLE

Congenital Anomalies | WILEY

Acoustic evaluation of voice signal distortion by videoconferencing platforms and devices used in telepractice for cleft palate

Shiho Tajiri[1,2] | Shunsuke Hidaka[3] | Shuhei Takehisa[3] | Sachiyo Hasegawa[4] | Yukiko Ohyama[1] | Tomohiro Yamada[1]

[1]Section of Oral and Maxillofacial Surgery, Division of Maxillofacial Diagnostic and Surgical Sciences, Faculty of Dental Science, Kyushu University, Fukuoka, Japan

[2]International Medical Department, Kyushu University Hospital, Fukuoka, Japan

[3]Graduate School of Design, Kyushu University, Fukuoka, Japan

[4]Department of Oral and Maxillofacial Surgery, Kyushu University Hospital, Fukuoka, Japan

Correspondence
Shiho Tajiri, 3-1-1, Maidashi, Higashi-ku, Fukuoka, Japan.
Email: shiho.t@dent.kyushu-u.ac.jp

Abstract

The usefulness and effectiveness of telepractice have been reported in recent years. Treatment of cleft palate patients with compensatory articulation is based on perceptual identification. Telepractice using videoconferencing platforms causes voice signal distortion and impacts auditory-perceptual perception. This study aimed acoustically examine voice signal distortion and determine the optimal videoconferencing platforms, in addition to the phonemes that can be discriminated with the same quality as in face-to-face interactions. ATR503 with 50 phoneme-balanced Japanese speech sentences was used as a reference corpus. Four videoconferencing platforms, —Zoom, Cisco Webex, Skype, and Google Meet, —and five devices, —iPhone, Android, iPad Air, Windows, and MacBook Pro were used as transmission conditions to examine voice signal distortions with the objective measure log-spectral distortion (LSD). Tukey's test was conducted to evaluate the degree of consonant distortion related to voicings (voiceless and voiced), places of articulation (bilabial, alveolar, alveolo-palatal, palatal, velar, labial-velar, and glottal), and manners of articulation (plosive, fricative, affricate, tap or flap, nasal, and approximant). With statistically significant differences, voiced, bilabial, labial-velar, nasal, and plosive consonants exhibited smaller distortions. In contrast, voiceless, alveolo-palatal, fricative, and affricate consonants exhibited larger distortions. Google Meet exhibited the lowest distortion among videoconferencing platforms and MacBook exhibited the lowest distortion among devices. This study provides significant insights into the telepractice strategies with the appropriate videoconferencing platform and device, and useful settings for cleft palate patients with compensatory articulations with respect to acoustics.

KEYWORDS
articulation disorder, cleft palate, speech therapy, telehealth, telemedicine

1 | INTRODUCTION

Cleft lip and palate are among the most common birth defects. The international prevalence of cleft lip with or without cleft palate is 7.94 per 10 000, especially high in Japan (19.05 per 10 000).[1] Cleft palate patients receive palatal surgery to obtain velopharyngeal function,[2] and generally require early speech intervention. The most critical objective of early intervention is to aid oral speech development by increasing consonant inventories.[3] After palate repair, articulation disorders are observed in approximately 50% of cleft palate patients.[4] Approaches to treatment for articulation disorders associated with cleft palate are dependent on whether it is related to anatomical or structural defects or is learned. Despite the presence of adequate velopharyngeal mechanisms, learned behaviors can lead to the emergence of compensatory articulations.[2] There is a high probability that articulation disorders caused by velopharyngeal function or fistula require surgical intervention, whereas learned compensatory articulations are responsive to speech therapy with a speech-language pathologist. At present, assessment and diagnosis are based on perceptual identification, as performed by speech-language pathologists. Speech-language pathologist assessments differentially diagnose speech errors and determine targets that are appropriate for intervention.[5] Cleft palate patients with compensatory articulations require speech therapy to correct the wrong place or manner of articulation, and to establish appropriate speech function.[6,7]

The COVID-19 pandemic recently increased the demand for telemedicine.[8] In the field of speech-language pathology, the American Speech-Language-Hearing Association has advocated for telepractice, which is a speech-language pathology service that uses telecommunication and internet technology to remotely connect clinicians to clients for screening, assessment, intervention, and consultation.[9] COVID-19 can be transmitted via droplets or physical contact.[10,11] Speech therapy exhibits a high risk of COVID-19 infection because speech-language pathologists are generally required to communicate face-to-face with their patients. Additionally, the close distance between medical staff and patients, the requirement to check the patient mouth and eyes, and the inability to use personal protective equipment such as medical masks and face shields increase the risk.[12] Therefore, multiple researchers reported that telepractice was beneficial during the COVID-19 pandemic.[12,13] Telepractice have the following benefits; alleviate the travel burden on families in more rural settings and getting to physicians, education for other medical professionals internationally, shortage of specialist.[14] Numerous reports of the usefulness and effectiveness of telepractice for cleft palate patients were reported.[14–16]

In telepractice using videoconferencing platforms, the voice signals can be distorted by the following factors: the microphone on the sending device, method of encoding voice data using the videoconferencing platform, noise, loudspeaker of the receiving device, various features such as speech enhancement algorithms, noise-canceling, and an automatic volume adjustment function equipped, videoconferencing platforms,[17–19] and internet bandwith.[20,21] Distortion of the clinician's speech may influence the patient's perception and capacity to accurately reproduce speech. In contrast, distortion of the patient's speech may influence the clinician's assessment of the patient capacity to reproduce speech. The aim of this study was to determine the degree of voice signal distortion caused by videoconferencing platforms with respect to acoustics, whether there is an optimal videoconferencing platform and device, and the phonemes that can be perceived with the same quality as in face-to-face contact.

2 | MATERIALS AND METHODS

2.1 | Recording experiments: Collection of data for analysis

2.1.1 | Voice Data

The 503-sentence set digital audio database (ATR503, ATR Promotions, Kyoto, Japan) was used as reference speech content. ATR503 is a phonetically balanced text corpus that includes 402 two-phoneme sequences, all of which may be present in Japanese speech, and 223 three-phoneme sequences that can be influenced by preceding and subsequent phonemes, thus totalling 625 items.[22,23] ATR503 is divided into 10 sets (from A to J), where Set A consists of 50 sentences read by the same male Japanese speaker with normal articulation and no history of speech or hearing impairments. Normal articulation was selected to avoid the confusion caused by cleft palate speech, including compensatory articulation, and to examine the effects of voice signal distortion. Since Nakaichi reported that there are no significant differences in the formant frequencies of all consonants in Japanese between males and females,[24] a single male voice was selected. The utterances were recorded in a soundproof room with an omnidirectional microphone (Type 4191-L-001, Brüel & Kjær, Virum, Denmark) and microphone amplifier (Type 2690-0S2, Brüel & Kjær, Virum, Denmark). The voice data were digitized at a sampling rate of 48 kHz and 16-bit floating-point resolution.

2.1.2 | Video conferencing platforms and devices

Four videoconferencing platforms—Zoom version 5.10.4 (Zoom Video Communications Japan, Tokyo, Japan), Cisco Webex version 42.5.0/42.5.1 (Cisco Systems, Tokyo, Japan), Skype version 8.83 (Microsoft Japan, Tokyo, Japan), and Google Meet version 89.0.0 (Google Japan, Tokyo, Japan)—were verified. Telemedicine guidelines in Japan clearly state that the encryption (Transport Layer Security 1.2 or higher) of communications with server certificates issued by a reputable organization should be implemented as a security measure[25]; thus, we selected it based on its compliance with the security level. The videoconferencing platforms were downloaded applications, except for Google Meet for personal computers (PCs), which had no application and was run on the Google Chrome web browser version 101.0.4951.67 (Google Japan, Tokyo, Japan). Most applications allow for volume and noise adjustments. To verify all the devices

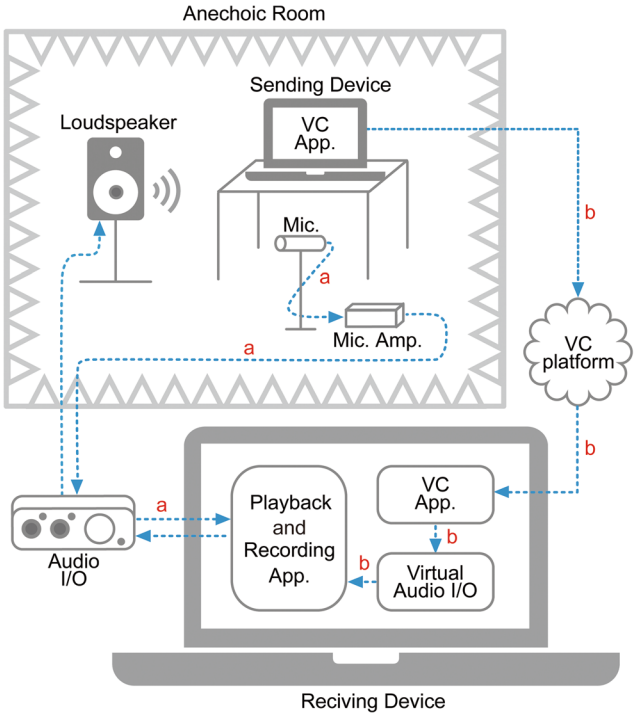

FIGURE 1 Equipment setup, signal chain, and recording of each audio data (VC, videoconferencing. App. Appreciation Mic.; microphone. Mic Amp.; Microphone amplifier. Audio I/O; audio interface): (A) the signal chain of the "input waveform," which is the audio data received from the microphone, and (B) the signal chain of the "output waveform," which is the audio data received from videoconferencing platform.

were used under the same conditions, all the videoconferencing platforms were set to their default settings (i.e., automatic volume adjustment and noise reduction). These features could not be selected using the free version of Google Meet®. Five sending devices—two laptops (Apple MacBook Pro 13-inch M1 2020; macOS Monterey 12.4, Apple Japan, Tokyo, Japan, and Microsoft Surface Go; Windows 10, Microsoft Japan, Tokyo, Japan), a tablet Apple iPad Air 4th generation iOS 15.5; Apple Japan, Tokyo, Japan, and two smartphones (Samsung Galaxy A21; Android 12.0, Samsung Japan, Tokyo, Japan and Apple iPhone 8; iOS 15.5, Apple Japan, Tokyo, Japan)— were verified. The selection was based on the popularity of these operating systems, in addition to the household ownership rates of information and communication devices in Japan, which were 86.8% for smartphones, 70.1% for PCs, and 38.7% for tablets.[26]

2.1.3 | Equipment setup and signal chain

The sending device was one of the abovementioned laptops, tablets, or smartphones, and the receiving device was an Apple MacBook Pro; macOS Monterey 12.4. Additionally, the receiving device was utilized for playback and recording using the programming language (Python; https://www.python.org/). Equipment related to the air transmission path of the sound was placed in an anechoic room. The anechoic room was noise-free, and the temperature and humidity were maintained at 20.3–22.6°C and 27%–28%, respectively. The gains of the audio interface (UAC-2, Zoom, Tokyo, Japan), loudspeaker (8050A, GENELEC Japan, Tokyo, Japan), and microphone amplifier (Type 2690-0S2) were fixed throughout the experiments. The sending and receiving devices were physically connected to the internet to avoid the impact of packet loss due to Wi-Fi connections.[27] The internet speeds of the devices were measured using an internet speed test application (Speedtest; https://www.speedtest.net/ja) when the devices were changed. The internet speed was stable, with a download speed of >150 Mbps and an upload speed of >100 Mbps throughout the experiments. To reduce unnecessary network load, the camera and loudspeaker on the sending device were turned off, in addition to the camera and microphone on the receiving device.

Figure 1 presents the sound flow. First, audio data were transmitted from the receiving device to the loudspeaker through the audio interface. Subsequently, the sound wave emitted from the loudspeaker was transferred to the sending device and omnidirectional microphone (Type 4191-L-001). The sending device and microphone were placed on a concentric circle with a radius of 0.5 m centered on the loudspeaker and at the same height as the loudspeaker position. When the sending device was a laptop, it was placed on a desk in an open position. However, when the sending device was a tablet or smartphone, it was placed on a tablet stand on a desk, with the bottom microphone on the left. The sound input to the microphone via a microphone amplifier and audio interface was recorded by the receiving device, and the sound input to the sending device was transferred to the receiving laptop via the videoconferencing platform. The audio signal on the receiving device was recorded directly through a virtual audio interface (BlackHole, Existential Audio; https://github.com/ExistentialAudio/BlackHole) on the same device. The virtual audio interface had the advantage of zero transmission latency.

2.1.4 | Recording and storage

In the pre-processing step, the signal amplitude of the source audio data for each sentence of ATR503 was normalized to obtain an equal average signal power by digital signal processing. Due to the normalization of the signal power and not sound power, the sound power differed for each audio data. When white noise with the same average signal power was emitted from the loudspeaker, the A-weighted sound pressure level measured close to the microphone was 68.2 dB SPL.

The audio data received from the microphone was referred to as the "input waveform," and the audio data received from the videoconferencing platform was referred to as the "output waveform" (see Figure 1). The normalized audio data of 50 sentences were played twice consecutively for each sentence. The first was used to adjust the volume to prevent unexpected volume changes due to the automatic volume adjustment of the videoconferencing platforms, and the second was used to record the input and output waveforms. During recording, the audio data of the output waveform was checked for significant auditory and visual anomalies such as delays and noise by listening to the sound through the loudspeakers of the receiving device and viewing the log-amplitude spectrograms on the display of

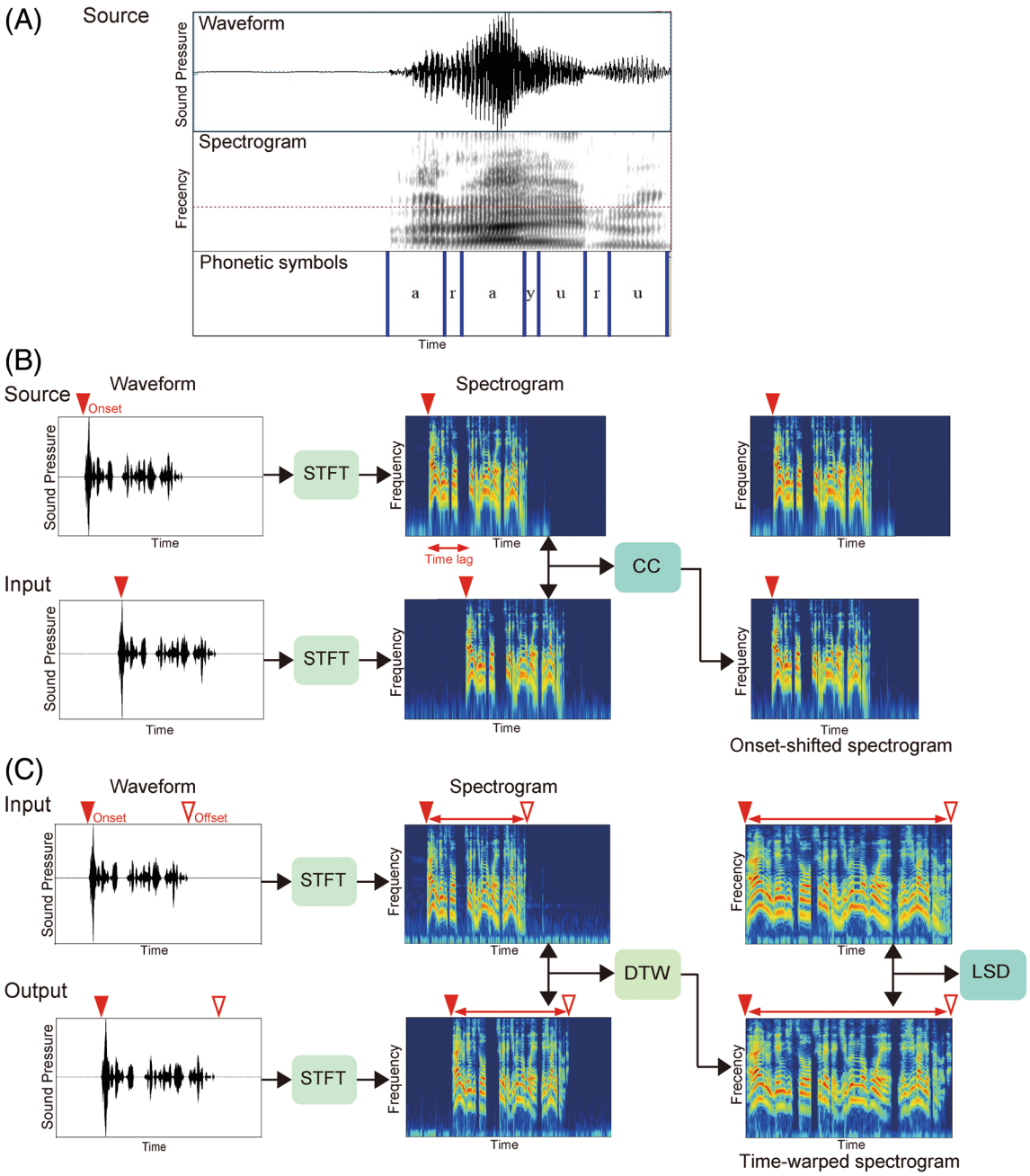

FIGURE 2 Overview of data analysis (STFT, short-time Fourier transform; CC, correlation coefficient; DTW; dynamic time warping; LSD, log-spectral distortion; Onset, start point of voice data; Offset, end point of voice data); example of the annotation for source audio data in Japanese "arayuru" (which translates to "each" in English) using Praat. (A) Onset of input waveform shifted to match souce for mapping of phenetic symbols. (B) Alignment of input and output waveforms for the log-spectral distortion calculation (C).

the receiving device. The input and output waveforms were saved as WAV files with a sampling rate of 48 kHz and 32-bit floating-point resolution.

2.2 | Data analysis

First, the source audio data was manually annotated with mono-phonemes using the speech analysis and phonetics software (Praat; https://www.fon.hum.uva.nl/praat/), while listening to the sound and viewing the sound waveform shape and log amplitude spectrogram. Figure 2A presents an audio sample in Japanese "arayuru" (which translates to "each" in English) of phonetic annotation using Praat. The assigned phonetic symbols were based on the International Phonetic Alphabet.[28] Each consonant phoneme was further classified into two voicings (voiceless and voiced), seven places of articulation in a consonant (bilabial, alveolar, alveolo-palatal, palatal, velar, labial-velar, and glottal), and six manners of articulation in a consonant (plosive,

TABLE 1 Classification and number of International Phonetic Alphabet (IPA) in the ATR503 phoneme-balanced sentences of this study.

	Place of articulation						
	Bilabial	**Alveolar**	**Alveolo-palatal**	**Palatal**	**Velar**	**Labial-velar**	**Glottal**
Manner of articulation							
Plosive							
Voiceless	p (115)	t (505)			k (700)		
Voiced	b (190)	d (280)			g (375)		
Fricative							
Voiceless	ɸ (80)	s (350)	ɕ (240)	ç (140)			h (125)
Voiced			z (80)				
Affricate							
Voiceless		ts (130)	tɕ (145)				
Voiced		dz (130)	dz (140)				
Tap or flap							
Voiced		ɾ (585)					
Nasal							
Voiced	m (385)	n (355)		ɲ (220)			
Approximant							
Voiced				j (235)		j (235)	

fricative, affricate, tap or flap, nasal, and approximant). Table 1 presents the classification of the phonetic symbols and numbers assigned to the consonants.

Second, to map the phoneme annotations to the input waveform, the time difference between the source and input waveforms was determined by the time of the maximum cross-correlation (CC), and the onset of the input waveform was matched to the onset of the source (see Figure 2B). The CC was defined as follows:

$$\mathrm{CC}(k) = \sum_{m} S(m,n) I(m,n+k) \tag{1}$$

where S and I are the amplitude spectrograms of the source and input waveforms, respectively; m and n are the frequency and time indices, respectively, and k is the time lag. To eliminate the effects of phase distortion, CC was defined based on amplitude spectrograms rather than on waveforms. When calculating the short-time Fourier transform (STFT) for the amplitude spectrogram, the window length and discrete Fourier transform (DFT) points were set to 2400 samples (50 ms), and the shift length was set to one sample (1/48 ms).

Third, the onset and offset of the input and output waveforms were matched, and alignment was performed between the input and output waveforms (see Figure 2C). The output waveform passed through the videoconferencing platform, which may induce time distortions beyond a simple time lag, for example, expansion and contraction in the time direction. Time distortions lead to the inability to make quantitative evaluations. Therefore, dynamic time warping (DTW) was used for alignment in the time direction. In particular, DTW is generally used for temporal alignment in speech recognition.[29] The Euclidean distance (ED) between two log-amplitude spectra was used as the distance function for DTW. In this study, it was defined as follows:

$$\mathrm{ED}(i,j) = \sqrt{\sum_{m=1}^{M} \left(20\log_{10} O(m,i) - 20\log_{10} I(m,j)\right)^2} \tag{2}$$

where O is the amplitude spectrogram of the output waveform, and M is the number of frequency bins from the direct current component to the Nyquist frequency. The input and output waveforms were downsampled to 16 kHz for the ED calculation because the output waveform was band-limited to a minimum of 8 kHz by the videoconferencing platforms. Each log-amplitude spectrogram (20 $\log_{10} O$ or 20 $\log_{10} I$) was normalized to a maximum value of 0 dB. Subsequently, the values less than −80 dB were set as −80 dB. During the STFT calculation, the window length was set to 800 samples (50 ms), the DFT points were set to 4000 samples (250 ms), and the shift length was set to 200 samples (12.5 ms).

The log-amplitude spectrogram of the output waveform was stretched or contracted in the time direction based on alignment by DTW. Subsequently, the log-spectral distortion (LSD) between the time-warped log-amplitude spectrogram of the output waveform and the log-amplitude spectrogram of the input waveform was calculated per frame to quantify the distortion caused by transmission. The LSD is a measure of the distortion between two log-amplitude spectra.[30,31] Smaller LSD values indicate less distortion. In this study, the LSD was defined as follows:

TABLE 2 Statistically significant ANOVA followed by Tukey-post hoc test results among videoconferencing platforms and devices.

	ANOVA p-value	η^2	Tukey's test		
			Multiple comparisons		Cohen's d
Videoconferencing platform	<0.001	0.093	Zoom	Cisco Webex	0.41
				Skype	0.72
				Google Meet	0.87
			Cisco Webex	Skype	0.27
				Google Meet	0.38
			Skype	Google Meet	0.09
Device	<0.001	0.083	MacBook	Windows	0.49
				iPad	0.58
				Android	0.69
				iPhone	0.84
			Windows	iPad	0.14
				Android	0.24
				iPhone	0.43
			iPad	Android	0.10
				iPhone	0.29
			Android	iPhone	0.20

Note: η^2 effect size: 0.01 = small, 0.09 = medium, 0.25 = large. Cohen's d effect size: 0.2 = small, 0.5 = medium, 0.8 = large.

$$\mathrm{LSD}(j) = \sqrt{\frac{1}{M-\tilde{M}+1}\sum_{m=\tilde{M}}^{M}\left(20\log_{10}\tilde{O}(m,j) - 20\log_{10}I(m,j)\right)^2} \quad (3)$$

where $\tilde{O}$ is the time-warped amplitude spectrogram of the output waveform, and $\tilde{M}$ was set to exclude frequency components <64 Hz from the LSD calculation. The removal of components <64 Hz was conducted to eliminate electrical noise in the input waveform and the effects of different low-frequency cut-offs for different videoconferencing platforms. Electrical noise was included down to approximately 64 Hz, and the highest cut-off frequency was approximately 32 Hz for Zoom. In particular, the frequency bands were visually confirmed from the spectrograms. The downsampling, normalization, threshold processing, and STFT were performed as in the ED calculation. The degrees of distortion of all the phonemes, including vowels, consonants, and special mora, were verified for each videoconferencing platform, device, and combination. For each consonant phoneme, the degree of distortion within the groups was compared with respect to the classified voicing, places of articulation, and manner of articulation.

2.3 | Statistical analysis

Statistical analyzes were performed using JMP Pro version 16 (SAS Institute Japan, Tokyo, Japan). All statistically significant differences were set at $p < 0.05$. For all phonemes, an analysis of variance (ANOVA) was conducted to measure the effects of videoconferencing platforms and devices on distortion. If the transmission condition was a statistically significant factor, the partial eta squared (η^2) was calculated to determine its effect size, and post hoc Tukey's tests were conducted to evaluate the differences. Consonants of each phoneme were compared within groups by conducting a T-test for the voicings and Tukey's test for the places and manners of articulation. Moreover, the effect size was calculated using Cohen's d (d).

3 | RESULTS

3.1 | Effect of transmission conditions on distortion for all phonemes

The result of ANOVA was statistically significant differences for both videoconferencing platform and device and medium effect size for videoconferencing platform and small effect size for device (see Table 2). Figure 3A reveals that Zoom exhibited the largest mean LSD, followed by Cisco Webex, Skype, and Google Meet. Additionally, it presents the Tukey's test results and significant differences among all videoconferencing platforms. A large effect was observed only between Zoom and Google Meet (see Table 2). Figure 3B revealed that iPhone exhibited the largest mean LSD, followed by Android, iPad, Windows, and MacBook; in addition to the Tukey's test results and significant differences among all devices. A large effect was observed only between iPhone and MacBook (see Table 2).

Figure 4 presents the audio waveform and log-amplitude spectrogram for the same audio sample in Japanese *futsuu*/ɸ/, /ɯ/, /ts/, /ɯ/, /ɯ/(which translates to "*general*" in English). In the audio waveforms at the upper part of each combination, when (A) and (B) were compared with the other samples via the videoconferencing platform, the

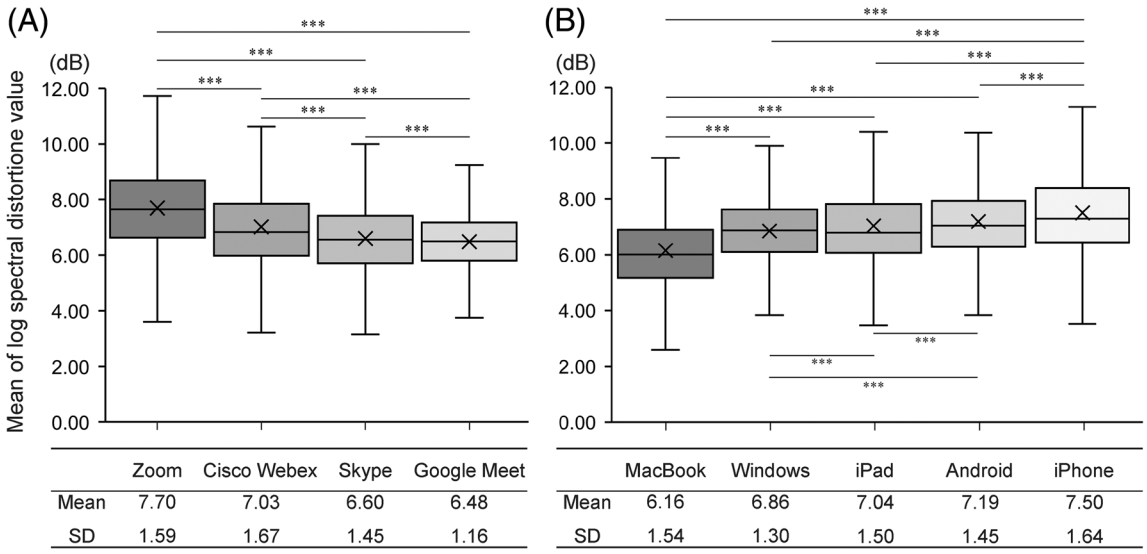

FIGURE 3 Comparison of mean LSDs of all phonemes with respect to four videoconferencing platforms ($n = 12\,925$) (A) and five devices ($n = 10\,340$) (B). The whiskers extend to points that lie within 1.5 interquartile ranges (IQRs) of the lower and upper quartile. Statistical significance was observed for all videoconferencing platforms and devices based on Tukey's test ($p < 0.001$).

low-amplitude area decreased in size, whereas the high-amplitude area was maintained. The beginning of the/ϕ/for the combinations of Zoom and Android (F), Cisco Webex and iPad (J), and Skype and Android (P) exhibited an audio waveform loss and a mean LSD of >18.00 dB. In the log-amplitude spectrograms at the lower part for each combination, the low-frequency range from 0 to 4000 Hz exhibited slight changes, whereas the high-frequency range from 4000 to 8000 Hz was attenuated. In the case of Zoom combined with MacBook (C) and Windows (D), and iPhone (G), the signals at frequencies of 3800 and 7000 Hz disappeared.

3.2 | Effect of transmission conditions on distortion for each phoneme in consonants

Table 3 presents the mean and standard deviation of the LSD with respect to the consonants measures for each videoconferencing platform. The mean LSD of voiceless consonants was larger than that of voiced consonants for all videoconferencing platforms. Figure 5A presents the Tukey's test results and significant differences between them; however, the effect sizes were not significant for all videoconferencing platforms: Zoom ($p < 0.0001$, $d = 0.376$), Cisco Webex ($p < 0.0001$, $d = 0.280$), Skype ($p < 0.0001$, $d = 0.374$), and Google Meet ($p < 0.0001$, $d = 0.366$). With respect to the place of articulation, Table 3 reveals that alveolo-palatal consonants exhibited the largest mean LSD for all videoconferencing platforms. Bilabial consonants in Zoom, Skype, and Google Meet, and labial-velar consonants in Cisco Webex exhibited the smallest mean LSDs. Figure 5B presents the Tukey's test results and revealed that the alveolo-palatal consonants were significantly different from the other places of articulation, in addition to large effects between the alveolo-palatal and bilabial consonants, and between the alveolo-palatal and labial-velar consonants for all videoconferencing platforms (see Data S1). With respect to the manner of articulation, Table 3 reveals that fricative consonants exhibited the largest mean LSD among all videoconferencing platforms. In Zoom, Cisco Webex, and Skype, followed by affricate, approximant, tap or flap, plosive and nasal, and in Google Meet, followed by affricate, tap or flap, approximant, plosive, and nasal. Figure 5C presents the Tukey's test results, which revealed significant differences between the fricative and affricate consonants from other places of articulation and large effects between the plosive and fricative, plosive and affricate, fricative and nasal, affricate and tap or flap, and affricate and nasal consonants for all videoconferencing platforms (see Data S2).

4 | DISCUSSION

The degree of voice distortion for various videoconferencing platforms and device combinations was investigated in this study. The results revealed that all combinations exhibited varied signal distortions, and the degree of variation was dependent on the combination of the videoconferencing platform and device. Furthermore, a comparison of the distortion for each phoneme revealed that the degree of distortion due to the videoconferencing platforms varied depending on the type of phoneme. In this study, valuable insights were gained regarding the method of voice distortion caused by videoconferencing platforms, devices, and combinations, in addition to the phoneme most influenced by sound distortion due to videoconferencing platforms.

4.1 | Possible factors affecting variations between videoconferencing platforms

Zoom exhibited the largest mean LSD and most significant distortion between the input and output waveforms. Although the frequency

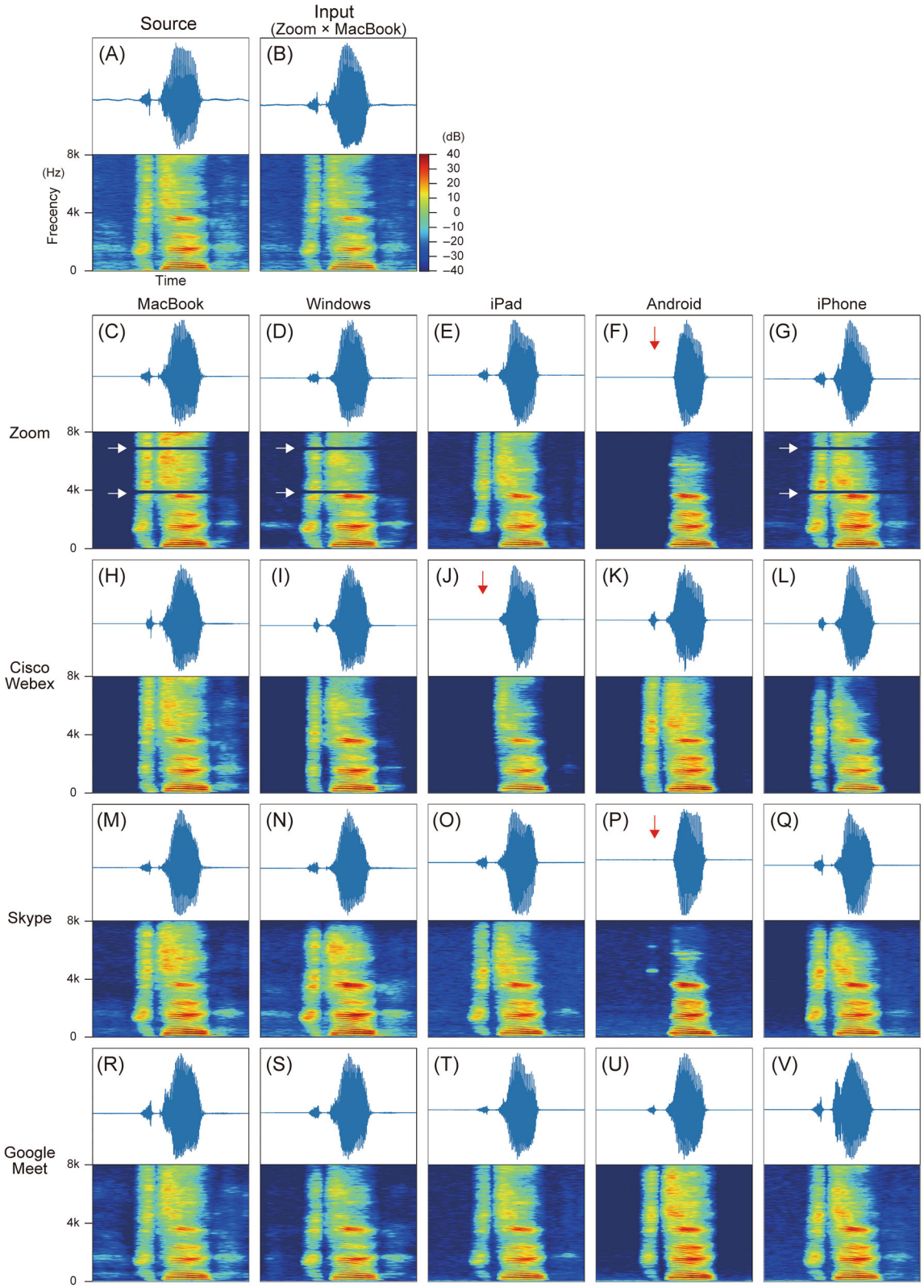

FIGURE 4 Example of the audio waveform and amplitude spectrogram of the same voice sample in Japanese /ɸ/, /ɯ/, /ts/, /ɯ/, /ɯ/ (which translates to "*general*" in English) for all combinations of videoconferencing platform and devices. For each combination, the upper part is the audio waveform and the lower part is the amplitude spectrogram. Given that the "input waveforms" were almost the same regardless of the transmission condition, only the results for one transmission condition are illustrated in this figure. Sub-figures F, J, and P present the audio waveform loss in the low-amplitude area (↓: Loss of audio waveform); and C, D, and G present the signal attenuation (→: Signal attenuation). [Correction added on 18 October 2024, after first online publication: Figure 4 caption has been corrected.]

TABLE 3 The mean LSD of consonants for each videoconferencing platform.

		Zoom		Cisco Webex		Skype		Google Meet	
	n	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Voicing									
Voiceless	2530	7.95	1.72	7.25	1.93	6.91	1.69	6.70	1.31
Voiced	3160	7.35	1.53	6.76	1.59	6.35	1.34	6.25	1.13
Place of Articulation									
Bilabial	770	6.82	1.45	6.58	1.69	6.08	1.46	6.02	1.17
Alveolar	2335	7.71	1.59	7.05	1.76	6.64	1.52	6.57	1.21
Alveolo-palatal	605	8.70	1.67	7.70	1.73	7.62	1.42	6.99	1.12
Palatal	595	8.08	1.51	7.39	1.60	6.64	1.36	6.73	1.05
Velar	1075	7.28	1.46	6.80	1.76	6.34	1.56	6.29	1.29
Labial-velar	185	7.19	1.39	6.49	1.30	6.21	1.12	6.06	0.94
Glottal	125	7.88	1.31	7.40	1.61	6.62	1.30	6.79	1.09
Manner of Articulation									
Plosive	2165	7.16	1.46	6.67	1.70	6.26	1.51	6.25	1.26
Fricative	1015	8.95	1.50	8.09	1.85	7.55	1.52	7.28	1.07
Affricate	545	8.89	1.25	8.02	1.64	7.51	1.41	7.23	0.98
Tap or Flap	585	7.38	1.41	6.75	1.35	6.30	1.19	6.37	1.01
Nasal	960	6.95	1.36	6.48	1.50	6.04	1.26	5.99	1.07
Approximant	420	7.60	1.50	6.90	1.44	6.56	1.24	6.36	0.96

Abbreviation: LSD, log-spectral distortion.

response of each videoconferencing platform was not published, differences in the degrees of attenuation within the investigated frequency range were observed. As shown in Figure 4, a signal loss close to 3800 and 7000 Hz was observed for Zoom, which influenced the increase in the LSD value. Google Meet exhibited the smallest mean LSD and lowest distortion between the input and output waveforms. Speech enhancement algorithms and noise-canceling features of videoconferencing platforms may have affected increasing LSD.[17,18] Given that the current experiment was conducted in an anechoic room with audio samples and recorded in a soundproof room, the excessive suppression caused by noise cancellation distorted the required voice data. The lowest mean LSD in Google Meet can be attributed to the lack of noise cancellation in the default settings.

Weerathunge et al. and Tran et al. reported that internet bandwidth is a critical factor in maintaining signal quality, including the sampling frequency.[20,21] This study was conducted at an internet speed of >100 Mbps for both uploads and downloads. However, signal distortion was observed, which indicates that the degree of distortion variation is dependent on the algorithm in each videoconferencing platform, regardless of the differences in internet speed.

4.2 | Possible factors affecting variations between devices

Smartphones such as iPhone and Android exhibited larger mean LSDs than PCs such as MacBook and Windows. As shown in Figure 4, different devices exhibited different degrees of frequency attenuation in the amplitude spectrograms for the same videoconferencing platform; thus, different combinations of videoconferencing systems and devices exhibited different audio compression algorithms for audio transmission. The differences in the LSDs may be also due to the differences in specifications of the central processing units and built-in microphones. In addition, all devices were set 0.5 m from the loudspeaker in this experiment. An automatic volume adjustment function of videoconferencing platforms may have affected increasing LSD.[30] Given that the smartphones were at a greater distance than expected, an automatic volume adjustment function of the videoconferencing platforms may have required amplifying the volume.

4.3 | Effect of transmission on the degree of distortion of each phoneme

The significant differences between the manners of articulation can be attributed to the differences in amplitude and frequency. Voiceless consonants have a lower amplitude than voiced consonants, and they do not exhibit vocal fold vibrations, which are at lower frequencies. Therefore, as shown in Figure 4, the signal was attenuated and influenced the increase in LSD. The fricative and affricate consonants contain noise components at higher frequencies. However, the plosive consonants exhibit stronger bursts of pressure and are lower in frequency, and the nasal consonants cavity even lower frequencies.[32] Sounds such as /ɸ/ with high frequencies were considered as noise, and the sound was completely lost, as shown in Figure 4F,J,P.

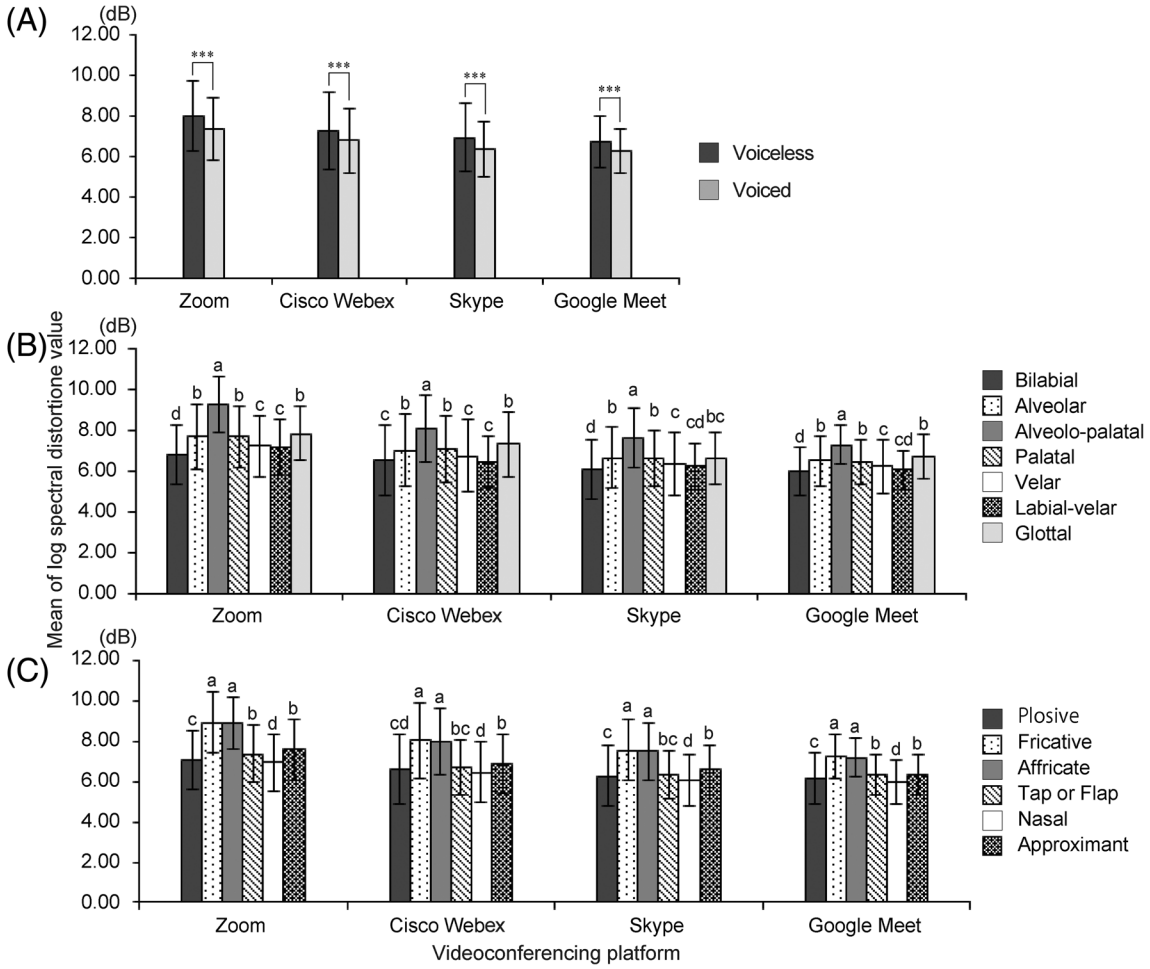

FIGURE 5 The mean log-spectral distortion (LSD) for each videoconferencing platform with respect to the voicing (A), places of articulation (B), and manners of articulation in consonants (C). The error bars indicate the standard deviation. (A) Voiceless consonants exhibited a larger mean LSD than voiced consonants and significant differences among all videoconferencing platforms, that is, (***$p < 0.01$; Tukey's test). (B) Alveolo-palatal consonants exhibited the largest mean LSD, whereas bilabial or labial-velar consonants exhibited the smallest mean LSD. (C) Fricative and affricate consonants exhibited the largest mean LSD, whereas nasal and plosive consonants exhibited the smallest mean LSD. Different lowercase letters represent statistically significant differences ($p < 0.05$; Tukey's test).

4.4 | Clinical implications

Dahl et al. reported that telepractice transmission does not substantially reduce the reliability or accuracy of auditory-perceptual voice assessments based on vowels as speech samples.[33] This study was focused on consonants for validation, which is critical in telepractice for cleft palate patients with compensatory articulations who require target selection and modification. The voice data used in this study was that of a single male with normal articulation. Numerous reports have been published that the voices of females and children have higher pitch and vowel formant frequencies than males.[34,35] However, there are no significant differences in the formant frequencies of all consonants in Japanese between males and females,[24] therefore we suggest that analysis using the female voice will show similar results to this study. In previous studies, the usefulness and effectiveness of telepractice for cleft palate patients were empirically reported based on the therapeutic effects and surveying satisfaction of clinicians, patients, and their parents.[14–16] Cleft palate speech have been reported to the following acoustical characteristics; hypernasality voice have the addition of the nasal cavity to the oral cavity in the vocal tract causes antiresonance characteristics and fluctuation in formant intensity in vowels with hypernasality.[36] Glottal stop in cleft palate patients has the loss of consonant components and a shortening of the voice onset time.[37] Palatalized articulation in cleft palate patients have different frequency of the burst from normal articulate, and the variability of the frequency distribution of the voice onset time depending on the place of articulation.[38] Cleft patients with velopharyngeal incompetence have an increase in the fundamental frequency, which is characteristic of pitch, and an increase in shimmer and jitter, which is measurement of the fundamental frequency and amplitude cycle-to-cycle variations.[39,40] Considering these acoustic features that differ from normal articulation, the analysis of cleft palate speech as voice data may be perceived as noise on a videoconferencing platform, resulting in larger LSD values, which may have a

significant impact on the auditory-perceptual perception. This study investigated the voice signal distortion introduced by telepractice and validated the accuracy of voice signals used in telepractice The LSD used in this study may serve as a standard screening and objective measure to determine the appropriate devices, VCPs, settings, and so forth, for telepractice of cleft palate patients with compensatory articulations and improve the quality of telepractice. As shown in Figure 5, nasal and plosive consonants exhibited smaller distortions, which suggests that telepractice can achieve the same quality of audio discrimination as face-to-face interactions. Moreover, this paper first presents the usefulness of telepractice considering the targeting of specific sounds with respect to acoustics. Fricative and affricate consonants exhibited larger distortions, and as shown in Figure 4, the loss of speech waveforms and attenuation in the high-frequency range may influence perceptual- auditory discrimination. This may be resolved by considering the combination of videoconferencing platform and device, in addition to the adjustment of settings such as noise reduction to reduce voice signal distortion. In addition, this study can be useful for telepractice for not only cleft palate patients with compensatory articulations but with also any kind of speech disorder including congenital velopharyngeal incompetence, ankyloglossia, congenital hearing deafness, Down syndrome and Kabuki syndrome,[41,42] and language development delay since this study used nomal male speech. Furthermore, telepractice can develop into speech therapy with additional visual information, such as the real-time display of speech waveforms, in addition to auditory information.

4.5 | Limitations

The determination of the combination of videoconferencing platform and device with the least distortion could not be achieved in this study, and the further examination of the effects of other transmission conditions is required. For example, the effects of the settings of Zoom, Cisco Webex, and Skype without volume adjustments and noise reduction. In addition, the effect of ambient noise and internet speed should be considered in actual telepractice. The voice data used in this study was that of a single male with normal articulation. The topic of whether speech with cleft palate is equally or more influenced by signal distortion was not assessed. Finally, a degree of temporal signal distortion (e.g., unexpected stretching of the output signal) due to digital time modification may have been present, thus influencing the LSD values.

5 | CONCLUSIONS

This study provides significant insights into the telepractice strategies with the appropriate videoconferencing platform and device, and useful settings for cleft palate patients with compensatory articulations with respect to acoustics. Given that auditory-perceptual is most attributable to subjective evaluation, in future research, the voice signal distortion assessed acoustically using LSD should be corroborated with subjective evaluations by clinicians and patients.

ACKNOWLEDGMENTS

This study was conducted at the Department of Acoustic Design, Kyushu University, Fukuoka, Japan. The authors would like to thank Shunta Tomimatsu for assistance in this study.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ETHICS STATEMENT

The manuscript does not contain human studies.

ORCID

Shiho Tajiri https://orcid.org/0000-0003-3168-7629
Tomohiro Yamada https://orcid.org/0000-0002-6485-406X

REFERENCES

1. Tanaka SA, Mahabir RC, Jupiter DC, Menezes JM. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg*. 2012;129(3):511e-518e. doi:10.1097/PRS.0b013e3182402dd1
2. Peterson-Falzone SJ, Trost-Cardamone JE, Karnell MP, Hardin-Jones MA. *The Clinician's Guide to Treating Cleft Palate Speech*. 2nd ed. Elsevier; 2016:71-82.
3. Hardin-Jones M, Chapman K, Scherer NJ. Early intervention in children with cleft palate. *ASHA Lead*. 2006;11(8):8-32. doi:10.1044/leader.FTR3.11082006.8
4. Broen PA, Moller KT, Carlstrom J, Doyle SS, Devers M, Keenan KM. Comparison of the hearing histories of children with and without cleft palate. *Cleft Palate Craniofac J*. 1996;33(2):127-133. doi:10.1597/1545-1569_1996_033_0127_cothho_2.3.co_2
5. Kotlarek KJ, Krueger BI. Treatment of speech sound errors in cleft palate: a tutorial for speech-language pathology assistants. *Lang Speech Hear Serv Sch*. 2023;54(1):171-188. doi:10.1044/2022_LSHSS-22-00071
6. Lohmander-Agerskov A, Söderpalm E, Friede H, Persson EC, Lilja J. Pre-speech in children with cleft lip and palate or cleft palate only: phonetic analysis related to morphologic and functional factors. *Cleft Palate Craniofac J*. 1994;31(4):271-279. doi:10.1597/1545-1569_1994_031_0271_psicwc_2.3.co_2
7. American Cleft Palate-Craniofacial Association. Parameters for evaluation and treatment of patients with cleft lip/palate or other craniofacial differences. *Cleft Palate Craniofac J*. 2018;55(1):137-156. doi:10.1177/1055665617739564
8. Smith AC, Thomas E, Snoswell CL, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). *J Telemed Telecare*. 2020;26(5):309-313. doi:10.1177/1357633X20916567
9. American Speech-Language-Hearing Association. *Telepractice*. https://www.asha.org/Practice-Portal/Professional-Issues/Telepractice/.
10. Peng X, Xu X, Li Y, Cheng L, Zhou X, Ren B. Transmission routes of 2019-nCoV and controls in dental practice. *Int J Oral Sci*. 2020;12(1):1-6. doi:10.1038/s41368-020-0075-9
11. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. *Int J Antimicrob Agents*. 2020;55(3):105924. doi:10.1016/j.ijantimicag.2020.105924

12. Tohidast SA, Mansuri B, Bagheri R, Azimi H. Provision of speech-language pathology services for the treatment of speech and language disorders in children during the COVID-19 pandemic: problems, concerns, and solutions. *Int J Pediatr Otorhinolaryngol*. 2020; 138:110262. doi:10.1016/j.ijporl.2020.110262
13. Castillo-Allendes A, Contreras-Ruston F, Cantor-Cutiva LC, et al. Voice therapy in the context of the COVID-19 pandemic: guidelines for clinical practice. *J Voice*. 2021;35(5):717-727. doi:10.1016/j.jvoice.2020.08.001
14. Bedi G, Vyas KS, Chung MT, Morrison SD, Asaad M, Mardini S. Telemedicine in international cleft care: a systematic review. *Cleft Palate Craniofac J*. 2021;58(12):1547-1555. doi:10.1177/1055665621989140
15. Palomares-Aguilera M, Inostroza-Allende F, Solar LR. Speech pathology telepractice intervention during the COVID-19 pandemic for Spanish-speaking children with cleft palate: a systematic review. *Int J Pediatr Otorhinolaryngol*. 2021;144:110700. doi:10.1016/j.ijporl.2021.110700
16. Hayakawa T, Imura H, Inoue C, et al. Efficacy of telepractice, an alternative therapy tool during the coronavirus disease 2019 pandemic, for speech disorders related to congenital anomalies. *Congenit Anom*. 2023;63(6):206-210. doi:10.1111/cga.12543
17. Zoom Video Communications. *Configuring Professional Audio Settings for Zoom Meetings*. https://support.zoom.us/hc/en-us/articles/360046244692-Background-noise-suppression
18. Cisco Webex. *Remove Background Noise during Webex Meetings or Webinars*. https://help.webex.com/en-us/article/n70a8os/Remove-background-noise-during-Webex-meetings-or-webinars. 2023.
19. Cisco Webex. *Adjust your computer mic and speaker volume in a Webex Meeting*. https://help.webex.com/en-us/article/zf86fe/Adjust-your-computer-mic-and-speaker-volume-in-a-Webex-Meeting. 2024.
20. Weerathunge HR, Segina RK, Tracy L, Stepp CE. Accuracy of acoustic measures of voice via telepractice videoconferencing platforms. *J Speech Lang Hear Res*. 2021;64(7):2586-2599. doi:10.1044/2021_JSLHR-20-00625
21. Tran K, Xu L, Stegmann G, Liss J, Berisha V, Utianski RL. Investigating the Impact of Speech Compression on the Acoustics of Dysarthric Speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2022. doi:10.21437/Interspeech.2022-10817
22. Iso K, Watanabe T, Kuwahara H. Design of a Japanese Sentence list for a speech database, *Report of the Spring Meeting the Acoustical Society of Japan*. 1988 89–90.
23. Itahashi S. On recent speech corpora activities in Japan. *J Acoust Soc Jpn*. 1999;20(3):163-169. doi:10.1250/ast.20.163
24. Nakaichi K, Wasano K. [study of consonant formants using monosyllabic sound sources used in clinical settings] Rinshogenba de mochiirareru tan-onsetsuongen wo mochiita shiin horumanto no kento. *Audiology Japan*. 2021;64(5):386. doi:10.4295/audiology.64.386
25. Ministry of Health, Labor and Welfare. Guideline for the Appropriate Implementation of Telemedicine. (in Japanese). https://www.mhlw.go.jp/content/10800000/001233212.pdf.
26. Ministry of Internal Affairs and Communications. Household ownership rates for ICT devices. *White Paper Information and Communications in Japan*. 2022. https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2021/key-points.pdf
27. Mondal A, Cutler R, Huang C. SureCall: Towards Glitch-Free Real-Time Audio/Video Conferencing. Section V. *IEEE 18th International Workshop on Quality of Service (IWQoS)*. 2010. doi:10.1109/IWQoS.2010.5542727
28. International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press; 1999.
29. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Signal Process*. 1978;26(1):43-49. doi:10.1109/TASSP.1978.1163055
30. Wu YJ, Tokuda K. Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH. 2008. doi:10.21437/interspeech.2008-170
31. Bulut AE, Koishida K. Low-latency single channel speech dereverberation using U-net convolutional neural networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 2020 10.21437/Interspeech.2020-2421
32. Ladefoged P, Johnson KA. *Course in Phonetics*. 7th ed. Cengage Learning; 2014:197-226.
33. Dahl KL, Weerathunge HR, Buckley DP, et al. Reliability and accuracy of expert auditory-perceptual evaluation of voice via telepractice platforms. *Am J Speech Lang Pathol*. 2021;30(6):2446-2455. doi:10.1044/2021_AJSLP-21-00091
34. Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). *J Voice*. 1996;10(1):59-66. doi:10.1016/S0892-1997(96)80019-1
35. Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J Acoust Soc Am*. 1999;105(3):1455-1468. doi:10.1121/1.426686
36. Hirai S, Okazaki K, Arai T. A quantitative evaluation of hypernasality in children. *Jpn J Logop*. 1994;35(2):199-206. doi:10.5112/jjlp.35.199
37. Wang G, Takahashi K, Wakumoto M, Michia K. Acoustical characteristics evaluation of Japanese glottal stops. *J Cleft palate*. 1991;16(1):37-55. doi:10.11224/cleftpalate1976.16.1_37
38. Ozawa Y, Okazaki K. Articulatory tongue movements and acoustic characteristics of palatalized articulation in cleft palate patients. *Jpn J Logop Phoniatr*. 1994;35(4):322-330. doi:10.5112/jjlp.35.322
39. Villafuerte-Gonzalez R, Valadez-Jimenez VM, Hernandez-Lopez X, Ysunza PA. Acoustic analysis of voice in children with cleft palate and velopharyngeal insufficiency. *Int J Pediatr Otorhinolaryngol*. 2015;79(7):1073-1076. doi:10.1016/j.ijporl.2015.04.030
40. Segura-Hernández M, Valadez-Jiménez VM, Ysunza PA, et al. Acoustic analysis of voice in children with cleft lip and palate following vocal rehabilitation. Preliminary report. *Int J Pediatr Otorhinolaryngol*. 2019;126(May):109618. doi:10.1016/j.ijporl.2019.109618
41. Chapman RS, Hesketh LJ. Behavioral phenotype of individuals with down syndrome. *Ment Retard Dev Disabil Res Rev*. 2000;6(2):84-95. doi:10.1002/1098-2779(2000)6:23.0.CO;2-P
42. Morgan AT, Mei C, da Costa A, et al. Speech and language in a genotyped cohort of individuals with kabuki syndrome. *Am J Med Genet A*. 2015;167(7):1483-1492. doi:10.1002/ajmg.a.37026

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tajiri S, Hidaka S, Takehisa S, Hasegawa S, Ohyama Y, Yamada T. Acoustic evaluation of voice signal distortion by videoconferencing platforms and devices used in telepractice for cleft palate. *Congenit Anom*. 2024;64(6):242-253. doi:10.1111/cga.12584