JASA ARTICLE

An acoustic study of Cantonese alaryngeal speech in different speaking conditions

Steven R. Cox,[1,a] Ting Huang,[2] Wei-Rong Chen,[2] and Manwa L. Ng[3]
[1]*Department of Communication Sciences and Disorders, Adelphi University, Garden City, New York 11530, USA*
[2]*Haskins Laboratories, New Haven, Connecticut 06511, USA*
[3]*Speech Science Laboratory, Faculty of Education, University of Hong Kong, Hong Kong SAR, China*

ABSTRACT:
Esophageal (ES) speech, tracheoesophageal (TE) speech, and the electrolarynx (EL) are common methods of communication following the removal of the larynx. Our recent study demonstrated that intelligibility may increase for Cantonese alaryngeal speakers using clear speech (CS) compared to their everyday "habitual speech" (HS), but the reasoning is still unclear [Hui, Cox, Huang, Chen, and Ng (2022). Folia Phoniatr. Logop. **74**, 103–111]. The purpose of this study was to assess the acoustic characteristics of vowels and tones produced by Cantonese alaryngeal speakers using HS and CS. Thirty-one alaryngeal speakers (9 EL, 10 ES, and 12 TE speakers) read *The North Wind and the Sun* passage in HS and CS. Vowel formants, vowel space area (VSA), speaking rate, pitch, and intensity were examined, and their relationship to intelligibility were evaluated. Statistical models suggest that larger VSAs significantly improved intelligibility, but slower speaking rate did not. Vowel and tonal contrasts did not differ between HS and CS for all three groups, but the amount of information encoded in fundamental frequency and intensity differences between high and low tones positively correlated with intelligibility for TE and ES groups, respectively. Continued research is needed to understand the effects of different speaking conditions toward improving acoustic and perceptual characteristics of Cantonese alaryngeal speech. © *2023 Acoustical Society of America*.
https://doi.org/10.1121/10.0019471
(Received 20 October 2022; revised 30 April 2023; accepted 2 May 2023; published online 22 May 2023)
[Editor: Benjamin V Tucker] Pages: 2973–2984

I. INTRODUCTION

Alaryngeal speech is an alternative method of verbal communication following the removal of the larynx. Common alaryngeal communication methods include esophageal speech (ES), tracheoesophageal (TE) speech, and speech produced using an electrolarynx (EL). ES and TE speech are referred to as "intrinsic" methods as they both rely on the vibration of the pharyngoesophageal segment (PE) of the speaker (Graville *et al.*, 2019; Searl and Reeves, 2014). EL speech relies on the use of a handheld electronic device that is often placed on the neck. The vibrating head of an EL generates sound energy that is transmitted through the neck and into the vocal tract where articulation transforms it into speech (Cox, 2019; Nagle, 2019).

All alaryngeal communication methods are often described as 'atypical' across a number of parameters (e.g., fundamental frequency [F0], speech intelligibility, voice quality, speaking rate) and there is considerable variability across speakers (Cox *et al.*, 2020; Doyle and Eadie, 2005; Knollhoff *et al.*, 2021). The intelligibility of alaryngeal speech has been extensively investigated since the introduction of the EL in the 1950s [see Sleeth and Doyle (2019), and citations therein]. The most important consideration when investigating alaryngeal speakers' intelligibility is that it is multidimensional in nature (Doyle and Eadie, 2005). For example, Meltzner and Hillman (2005) found that the best voice "quality" for EL users was attained by improving low-frequency energy deficits, reducing radiating device noise, *and* varying F0. This was evident in depletion of low-frequency energy associated with Cantonese EL speech using long-term average spectrum analysis reported by Ng *et al.* (2009). It also is important to consider that the majority of research in this area has focused on English alaryngeal speakers, and as a result, it might not accurately reflect or translate to individuals who use an alaryngeal communication method and speak a tonal language(s). Their communication might be further impacted for a myriad of reasons, including aspects related to acoustic, phonological and/or phonetic aspects of tone-based languages.

Tonal contrasts are central to the intelligibility of tonal languages such as Mandarin, Thai, and Cantonese. These languages require speakers to produce a varying number of lexical tones that signal changes in meaning [e.g., four in Mandarin, five in Thai, and six[1] in Cantonese (Abramson, 1975; Whalen and Xu, 1992; Zee, 1991)]. There is a consensus in the literature that the contrasts of phonological tones are dominantly based on differences in the vocal pitch [i.e., F0 (Abramson, 1972)]. Other supplementary perceptual cues, such as intensity, duration and vowel quality, may also carry tonal information and enhance the recognition of tones (Fry, 1968; Tupper *et al.*,

[a]Electronic mail: scox@adelphi.edu

2021; Whalen and Xu, 1992). Whalen and Xu (1992) highlighted the importance of both F0 and amplitude (i.e., intensity) as important perceptual cues for identifying tonal contrasts in speakers of tonal languages. They created Mandarin tone stimuli that encoded only sound intensity information (F0 and duration were removed) and recruited 12 native Mandarin speakers to give tone judgements by listening to those stimuli; the mean (across four tones) accuracy of tone identifications was 65.5% (chance = 25%). Ching *et al.* (1994) studied tone identification in Cantonese alaryngeal speakers as judged by naive listeners. Twenty-two undergraduate students were asked to identify 48 stimulus items spoken by three ES speakers, two TE speakers, two EL users, and two pneumatic artificial laryngeal users. Stimulus items were embedded within a carrier phrase (/ˉni-ko ˌhai ___ [This word is __]) and they included all six Cantonese tones, including: high level, high rising, mid level, low falling, low rising, and low level (e.g., /ˉji/ [clothing], /ˈji/ [chair], /-ji/ [meaning], /ˊji/ [child], /ˌji/ [ears], and /ˍji/ [two], respectively) (Ching *et al.*, 1994). They found that Cantonese ES speakers produced the highest number of tones that were correctly identified (65.12%), followed by TE speakers (51.89%) and EL users (21.83%; chance = 16.7%). Further, the percent of correctly identified tones for alaryngeal speaker groups were considerably lower than scores for the normophonic Cantonese speakers (98.2%).

This prior work was supported by Yan *et al.* (2012) who demonstrated that 15 superior ES and 15 superior TE speakers can change F0 while reading passages in Cantonese or tasks involving pitch scaling. In fact, these speakers were found to produce a comparable or even higher mean F0 than normophonic speakers. Ng *et al.* (2001) examined F0, intensity and vowel duration associated with Cantonese alaryngeal speech (i.e., ES and EL in comparison with normophonic speakers). Findings suggested that, when producing the six Cantonese tones, ES speakers showed comparable F0 patterns to laryngeal speakers, while EL speech was associated with non-distinctive, flat F0 contours across all lexical tones. This was expected due to their inability to vary F0 using their EL devices. The lower intensity in ES (compared to laryngeal speakers) appeared to reflect reduced loudness produced by ES speakers, and lower intensity in EL speech might be related to, among other possibilities, the setting of the EL devices. The longer vowel duration found in both alaryngeal speaker groups suggested a reduced speaking rate. Vowel duration did not differ between the six tones in open syllables for all three groups (laryngeal, ES, and EL). The EL speakers showed no differences in mean intensity between tones, while laryngeal and ES speakers produced higher mean intensity in high level tone than low falling tone by 6.3 and 4.2 dB SPL, respectively, although this contrast in mean intensity between tones did not meet the significance level in their two-way (speaker group × tone) analysis of variance (ANOVA).

Several studies also have reported on vowels produced by Cantonese alaryngeal speakers (Law *et al.*, 2009; Ng and Chu, 2009; Ng *et al.*, 2001; Ng *et al.*, 1997; Ng *et al.*, 1998; Yan *et al.*, 2012). Ng and Chu (2009) found that proficient Cantonese ES and TE speakers consistently exhibited elevated vowel formants. In particular, the first two formant frequencies (F1 and F2) associated with all vowels were increased in ES and TE speakers, implying a shortened vocal tract when speaking with the neoglottis postlaryngectomy. The shorter effective vocal tract in ES and TE speech was attributed to the ascended location of the neoglottis. They reasoned that, according to earlier imaging studies, the neoglottis (or PE segment), was situated at the level of C3-C5 (third and fifth cervical vertebra), compared to C7 for vocal folds, which resulted in a markedly shortened vocal tract and a pronounced effect on resonances. In addition, they also reported similar changes in EL speech, except for F2 of /ɛ/. Despite the use of an external device, Cantonese EL users in general also produced vowels with higher formant values. These findings were consistent with early research focused on English alaryngeal speech. A reduction in vocal tract length following laryngectomy was reported by Diedrich and Youngstrom (1966) and increases in formant frequencies were demonstrated by Sisty and Weinberg (1972). However, formant analysis in alaryngeal speech is known to be challenging and error-prone, especially when measured by linear predictive coding (LPC), the most used algorithm in previous studies. Liao (2016) inspected LPC-measured (with default settings) formant frequencies in ES speech and found that 38% of the measurements were erroneous by the author's judgements of spurious values for a given vowel. Even when measuring manually on conventional spectrograms, vowel formants can be unmeasurable around 20% of time in ES [18% (Sisty and Weinberg, 1972)] and TE [19% (van As *et al.*, 1997)] speech, which may be due to the presence of noise in the source-function (Sisty and Weinberg, 1972). Recently, Whalen *et al.* (2022) established that reassigned spectrogram (RS) (Fulop and Fitz, 2006) is the most accurate formant measurement method and can provide the ground truth of vocal tract resonances. Their discussions about the advantages of RS over LPC focused on the fact that LPC is unable to measure formants in speech with high F0s (i.e., F0 > 200 Hz) while RS can. The only disadvantage of using the RS for formant measurement is that it is underdeveloped and measurements must be done manually. In this study, we will also demonstrate the usefulness of RS in analyzing alaryngeal speech, which is mostly produced with low F0s, and call for future efforts into developing automatic formant measurement by RS.

One important consideration for all therapeutic pursuits in speech-language pathology, including the rehabilitation of individuals undergoing removal of the larynx, is improving communication outcomes [see Doyle (2019) and the citations therein]. The intelligibility of Cantonese ES, TE, and EL speakers is often reported to be less than 80%, which is considered low relative to the intelligibility of normophonic speakers (Law *et al.*, 2009). Research that seeks to improve Cantonese alaryngeal speakers' intelligibility, then, is much needed and warranted.

https://doi.org/10.1121/10.0019471

Improving intelligibility in Cantonese alaryngeal speech might lie in the principles that underlie clear speech (CS). CS is a style of speaking that was originally proposed to improve speech intelligibility for hearing impaired listeners (Picheny *et al.*, 1985, 1986). The underlying principles of CS require individuals "…to speak as clearly as possible, as if he was trying to communicate in a noisy environment or with an impaired listener…" and "…to enunciate consonants more carefully and with greater (vocal) effort…" [Picheny *et al.* (1985), p. 97]. When compared to conversational speech (or, "habitual" speech [HS]), CS has been shown to improve intelligibility up to 34% for healthy and hard of hearing individuals in quiet and noisy conditions (Krause and Braida, 2002, 2004; Fergusson and Kewley-Port, 2002; Lam *et al.*, 2012; Lam and Tjaden, 2013; Payton *et al.*, 1994; Smiljanić and Bradlow, 2009; Uchanski, 2005). Alongside these improvements, clearly spoken words can positively affect word and sentence recall and recognition among native and non-native English listeners (Keerstock and Smiljanić, 2018, 2019).

Potential reasons for greater intelligibility of clearly spoken sentences have been hypothesized. For example, Picheny *et al.* (1986) found that sentences produced using CS were twice the duration of the same sentences spoken using HS. CS also has been shown to decrease the occurrence of vowel reduction, increase vowel durations and expand the vowel space area (VSA) (Ferguson and Kewley-Port, 2007; Lam *et al.*, 2012; Tjaden *et al.*, 2013; Picheny *et al.*, 1986). Larger VSAs have been observed with higher levels of speech intelligibility in healthy talkers (Bradlow *et al.*, 1996; Hazan and Markham, 2004) and English and non-English talkers with neurological conditions (Liu *et al.*, 2005; Tjaden *et al.*, 2014; Turner *et al.*, 1995). Our previous study demonstrated that Cantonese alaryngeal speakers spoke significantly slower in CS than in HS by 0.78 syllables per second (syll/s) and the intelligibility increased up to 9.1% while using CS (Hui *et al.*, 2022). However, the decreased speaking rate does not fully explain the increased intelligibility; the underlying mechanisms of CS in Cantonese alaryngeal speech and their contribution to intelligibility have yet to be studied.

At present, little is known about the effect of different speaking conditions (HS and CS) on the production of Cantonese alaryngeal speech and the acoustic properties that contribute to intelligibility. Therefore, this study assessed the acoustic correlates of CS produced by three groups of Cantonese alaryngeal speakers (ES, TE, and EL). Specifically, we hypothesized that acoustic features associated with CS (e.g., F0, intensity, duration, formant frequencies, and VSA) would significantly affect the intelligibility of Cantonese alaryngeal speakers.

II. METHOD

A. Participants

Thirty-one male laryngectomees ($M = 66.48$ years, $SD = 11.31$ years) participated in this study. All participants were alaryngeal speakers who attended the New Voice Club of Hong Kong and used their primary form of alaryngeal communication (9 EL, 10 ES, and 12 TE speakers) for an average of 7.15 years ($SD = 6.74$ years). All EL users used the Servox Digital neck-type EL (F0 $M = 79.6$ Hz; F0 $SD = 10.1$ Hz), and TE speakers digitally occluded their stoma during speech production.

All speakers were perceptually judged to be proficient by an experienced speech-language pathologist with more than 10 years of clinical experience in alaryngeal voice and speech rehabilitation. All alaryngeal speakers met the following inclusion criteria: (1) they were proficient in their primary alaryngeal speech mode as judged by the speech-language pathologist, (2) they were healthy at the time of the study with no other speech, language, and hearing problems, except those associated with laryngectomy, and (3) they were native speakers of Cantonese. This study was approved by the Faculty Research Ethics Committee of the Faculty of Education, University of Hong Kong.

B. Speech stimuli and recording

The speech stimuli analyzed in the current study included *The North Wind and the Sun* passage (see the Appendix). This passage contains a variety of sound structures such as different vowels and tones. Cantonese is known for having 11 vowel contrasts. The passage includes ten monophthong vowels /i, y, u, oe, ɔ, a, ɪ, ɵ, ʊ, ɐ/. For the present study, we focused on the three quantal vowels /i/, /a/, and /u/ and three long-short vowel pairs [i-ɪ], [a-ɐ], and [u-ʊ]. The vowels /ɪ, ɐ, ʊ/ have shorter durations than the other vowels and do not occur in an open syllable (Bauer and Benedict, 1997; Zee, 2003). The recorded passage comprised six *long* tones [tone 1 (high level), tone 2 (high rising), tone 3 (mid level), tone 4 (low falling), tone 5 (low rising), and tone 6 (low level)]. It is noted that three *short* tones (sometimes known as tone 7, tone 8, and tone 9) with an unreleased stop coda /p/, /t/, or /k/ were also present in the passage.

The North Wind and the Sun was recorded alongside sentences from the Cantonese Sentence Intelligibility Test (CSIT) (Lo, 2015) at the New Voice Club of Hong Kong [see Hui *et al.* (2022) for detailed information on the CSIT]. Stimuli were recorded using a Shure SM58 microphone (frequency response = 50 Hz to 15 kHz) that was placed 15 cm from each participant's mouth. The microphone was connected to an M-Audio USBPre preamplifier that was fed into a laptop. All recordings were digitized at a sampling rate of 44 kHz (bit depth = 16).

The recording began with each participant reading all stimuli in a manner they used in daily communication [or habitual speech (HS)]. This was followed by CS, which involved instructing the speakers "to overarticulate" (in Cantonese: 誇張咁讀) and "to slow down their speech" (in Cantonese: 減慢語速). Each speaker was allowed to practice a sample sentence not included in the research stimuli and a demonstration was provided if any errors were perceptually detected. Participants read all stimuli using HS, then

read all stimuli using CS to ensure conditions were maintained for all stimuli.

C. Measurements

This study used speech intelligibility data from our previous study (Hui *et al.*, 2022). Two healthy female listeners (aged 23 and 27 years) who were native speakers of Cantonese transcribed a total of 11 CSIT sentences containing 220 words. Words were typed into designated cells in an Excel file with the total number of words labeled. Each sentence was presented a total of 2 times and this procedure was repeated until all CSIT sentences were transcribed. Speech intelligibility was calculated by dividing the number of correctly identified words divided by the total number of words.

Three aspects of measures were obtained from *The North Wind and the Sun*: (1) variabilities directly induced by speaking condition (HS vs CS), including speaking rate and VSA; (2) tonal information, including F0 entropy and speech intensity; and (3) long-short vowel contrasts, including formant distance and durational differences between long and short vowel counterparts.

Speaking rate was defined as the number of syllables produced per second in a continuous utterance (i.e., manually removed interruptions, coughs, hesitations, and long pauses but kept natural silent portions in a fluent utterance) of *The North Wind and the Sun* to ensure consistency across speakers. The passage was transcribed by the authors; the syllable and phoneme boundaries were automatically aligned using the Montreal forced aligner (McAuliffe *et al.*, 2017) and then manually adjusted.

We analyzed the acoustic features, including F0, the first two formant frequencies (F1 and F2) and speech intensity, were extracted from the utterance. Formant frequencies were first estimated by using the linear predictive coding (LPC) algorithm (window size = 50 ms; pre-emp from 50 Hz; cut-off frequency = 5000 Hz) in PRAAT version 6.0.49 (Boersma and Weenink, 2001), and then were manually checked and corrected by using RS [codes available in Fulop (2011)]. Figure 1 demonstrates a comparison of the formants measured by LPC (with the default settings in PRAAT) and RS on a token of perceptually acceptable short vowel /ɐ/ produced by an esophageal speaker (ES8). The aperiodic waveform in Fig. 1(a) indicates that the vowel was largely voiceless. Formant structure is roughly apparent in the wideband spectrogram [Fig. 1(b)], but it is difficult to identify consistent formant frequencies. The LPC-measured F1 [solid line in Fig. 1(b)] was inaccurate and differed from the true first resonance at 763 Hz shown in the RS [Fig. 1(d)] by 220 Hz. The LPC-measured F3 (dotted line) was clearly erroneous. The energy concentrations in the RS [Figs. 1(c) and 1(d)], on the other hand, unambiguously reveal the first five formants under 5000 Hz.

VSA for each speaker was defined as the area encompassed by the three quantal vowels /i/, /a/, and /u/ (each vowel was represented by its median value) in a formant space where F1 is the *y* axis and F2 the *x* axis [in equivalent rectangular bandwidth (ERB)].

We further analyzed the formant distance and duration difference between long and short vowels for the three long-short vowel pairs [i-ɪ], [a-ɐ], and [u-ʊ]. For each vowel, the median values of F1, F2, and duration were calculated to represent the vowel. Formant distance was defined as the Euclidean distance between long and short vowels in F1 × F2 space where the F1 and F2 frequencies were converted to ERB, and the duration difference as subtracting the duration of the short vowel from that of the corresponding long vowel in a given pair.

Speech intensity, voicing detection, and F0 were calculated by the algorithms (window size = 3/pitch floor for F0 and 3.2/pitch floor for intensity; step size = 5 ms) implemented in PRAAT; the pitch floor and ceiling were manually set for each speaker by visual inspection of the speech signals. To evaluate the contrastivity of lexical tones in alaryngeal speech, we considered two measures: the amount of

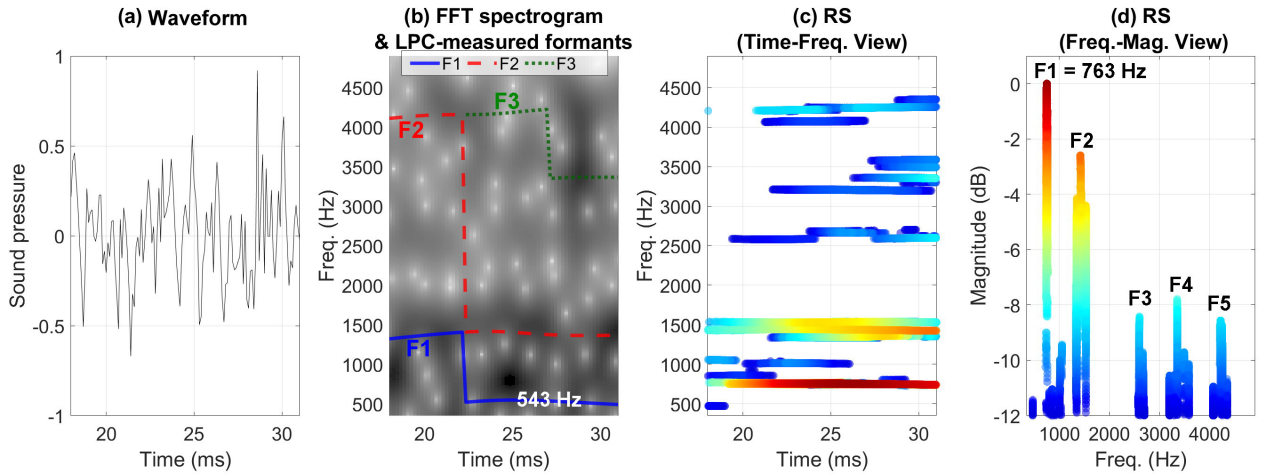

FIG. 1. (Color online) Comparison of (a) the waveform, (b) wideband spectrogram and LPC-measured formant frequencies (F1: solid line, F2: dashed line, F3: dotted line), (c) reassigned spectrogram (RS) in time-frequency view, and (d) RS in frequency-magnitude view, based on an example of the short vowel /ɐ/ produced by an esophageal speaker (ES8).

https://doi.org/10.1121/10.0019471

information in F0 signals and the intensity difference between high and low tones.

The method of quantifying the amount of information in a stream of signals is known as "entropy" as introduced in the information theory (Shannon, 1948). We calculated the entropy of F0 for each speaker as follows. First, the F0 signals of all voiced windows were attributed into 15 equal-ERB bins across a pitch range of 45–200 Hz. Unvoiced windows (i.e., no pitch value) were attributed to an additional bin. Then, the F0 entropy was calculated as

$$H = -\sum_{i=1}^{N} p_i * \log_2(p_i), \tag{1}$$

where p_i indicates the probability of the ith bin. The entropy of F0 for each utterance quantifies the amount of information encoded in the F0 signals or it can be interpreted as how "informative" the sequence of F0 signals is. In general, higher entropy indicates more contrastive tonal categories realized in F0 fluctuations. Figure 2 demonstrated three examples of F0 contours and the corresponding entropies. In the top panel, the speaker EL1 produced speech with constant F0 and therefore the entropy was zero (no information). In the middle panel, the speech produced by ES2 was mostly unvoiced, resulting in sporadic F0 signals carrying 0.89 bit of information. The lower panel in Fig. 2, on the other hand, showed an example of informative fluctuations of F0 signals (entropy = 2.31 bits) in the speech produced by TE8 and the tonal contrasts in the utterance were indeed perceptually salient.

To calculate the intensity difference between high and low tones, we first identified the peak intensity in each syllable and then the median of peak intensities of all syllables bearing the same lexical tone was calculated to represent the intensity of the lexical tone. For each speaker, the difference in median peak intensity between the tone category with the highest pitch, tone 1 (T1, high level), and the one with the lowest pitch, tone 4 (T4, low falling) (Gandour 1981, 1983; Vance, 1977), was our approximation of tonal contrast realized in speech intensity.

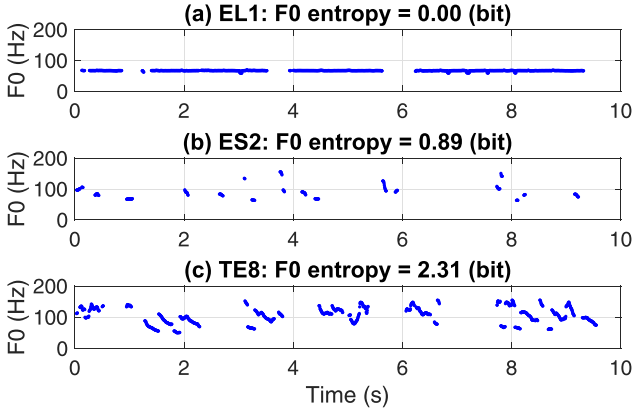

FIG. 2. (Color online) Signals of F0 contours and the corresponding entropies in three examples of (a) EL, (b) ES, and (c) TE speech.

D. Statistical analyses

One of the main purposes of this study is to compare the acoustic properties between HS and CS. To this end, we fitted three linear mixed effects models by using the "LME4" (Bates *et al.*, 2015) package in R (R Core Team, 2020). The dependent variable of the models was intelligibility (Intelli); the fixed effects included speaking rate (SpRate), vowel space area (VSA), alaryngeal groups (group with three levels: EL, ES, and TE), VSA difference between long and short vowels (VsaDiffLongShortV), durational difference between long and short vowels (DurDiffLongShortV), F0 entropy (F0Entropy), and peak intensity difference between high and low tones (IntensDiffHighLowTones). All fixed effects and their interactions went through stepwise model comparisons with likelihood ratio tests. The speaking condition (two levels: CS and HS) was not included due to strong collinearity with VSA and SpRate. Speaker was included as a random intercept. Models with random slopes failed to converge. For the interaction terms, only the interaction between Group and spRate significantly improved the model. The two factors that compare long and short vowel contrasts (VsaDiffLongShortV and DurDiffLongShortV) did not improve the models and were thus removed. The two acoustic features that measured tonal contrasts (F0Entropy and IntensDiffHighLowTones) encoded no information in EL group (causing rank deficiency), due to the fact that the EL devices in this study produced constant F0 and intensity in the output. Therefore, we did not include these two tonal factors in the optimal model across groups, but fitted two separate models with these factors for the ES and TE groups only.

Pair-wise t-tests were carried out to compare HS and CS in VSA, F0 entropy, and intensity. Corrections for p-values in multiple hypothesis testing were performed by using false discovery rate (FDR) (Benjamini and Yekutieli, 2005). We imported the speech intelligibility scores for the same speakers from those reported in Hui *et al.* (2022), as the dependent variable of the statistical models in this study. Significance level was set at 0.05 while marginally significant trends ($p < 0.1$) will also be marked up.

III. RESULTS

A. Variabilities directly induced by speaking condition

Hui *et al.* (2022) reported that the speaking rate was significantly slower for CS compared to HS for all three groups of the alaryngeal speakers. In the current study, we observed an increase in the vowel space area (VSA) to enhance the vowel contrasts in CS. Figure 3 shows that the VSAs were larger in CS than in HS for all three groups of alaryngeal speakers. Pairwise one-sample t-tests comparing the VSAs in CS and HS with FDR corrections suggested that the adjusted p-values are 0.002 ($t(8) = 5.24$), 0.013 ($t(9) = 3.09$), and 0.003 ($t(11) = 3.95$) for EL, ES, and TE speakers, respectively.

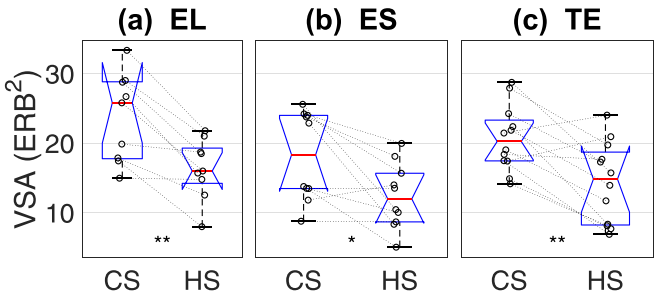

FIG. 3. (Color online) Vowel space area (VSA) in clear speech (CS) and habitual speech (HS) conditions for Cantonese alaryngeal speakers (EL = electrolaryngeal, ES = esophageal, and TE = tracheoesophageal speakers). (**: $p < 0.01$; *: $p < 0.05$.)

B. Tonal contrasts

Figure 4 represents the F0 entropies for the three alaryngeal groups in two speaking conditions. The F0 entropies for EL speakers are zeros because they produced speech with constant F0s. There are no observable differences in F0 entropy between CS and HS for both ES and TE speakers.

The peak intensity differences between high and low tones for the three alaryngeal groups are summarized in Fig. 5. All positive values in Figs. 5(b)–5(c) indicate that the tonal contrasts (between high and low tones) were realized in peak intensity of the syllables for ES and TE speakers. EL speakers, on the other hand, did not show contrastive intensity differences between high and low tones. For all three alaryngeal groups, speaking in CS condition did not increase the tonal contrast in F0 and intensity, as compared to HS condition.

C. Vowel contrasts

The long-short vowel contrasts in formants for the three alaryngeal groups are shown in Fig. 6. Most of the measured formant distances are above 1 ERB, which indicates that the long-short vowel contrasts in formants were maintained for most vowel pairs produced by alaryngeal speakers. Pairwise t-tests with FDR corrections indicate that there are non-significant differences between CS and HS in terms of long-short vowel contrasts in formants.

On the other hand, Fig. 7 presents the duration differences (i.e., the duration of long vowel minus that of short vowel) between long and short vowels; positive values indicate the expected contrasts in vowel duration. For all three groups, the duration contrasts between long and short vowels were maintained for [i-ɪ] and [a-ɐ] (duration differences > 0) but not for [u-ʊ]. Such durational contrasts between long and short vowels were increased in CS for [i-ɪ] [$t(9) = 3.62$; FDR-adjusted $p = 0.009$] and [a-ɐ] pairs [$t(9) = 3.53$; FDR-adjusted $p = 0.009$], produced by ES speakers.

D. Statistical models

The final optimal model without tonal factors was fitted with the formula: Intelli ~ SpRate * Group + VSA + (1 | Speaker). A summary of the coefficients of the fixed effects in the optimal model is given in Table I. The main effect for VSA was significant, and SpRate was marginally significant for the TE group only but not for the EL and ES groups. Marginal means and pairwise comparisons of the group effect based on the predictions of the optimal model were carried out by using the "emmeans" (Lenth, 2022) package in R, as shown in Fig. 8. The optimal model predicted that the intelligibility for the EL group was lower than that for the ES group (marginally significant, $p = 0.06$), when all other effects were set at the means.

Scatter plots of intelligibility as a function of SpRate and VSA are shown in Figs. 9 and 10, respectively. Regression lines were calculated across CS and HS for each group by using least squares method and the associated correlation coefficients (Pearson's r) and p-values are shown at the southwest corners of each panel. In Fig. 9, there are no correlations between intelligibility and speaking rate in both the EL and ES groups and there is a positive correlation in the TE group, reflecting the interaction terms between SpRate and Group in the optimal model shown in Table I. In Fig. 10, there are positive correlations between intelligibility and VSA in all three groups; this pattern is consistent with the fixed effect VSA in the optimal model.

Table II summarizes an additional LME model with the inclusion of the two tonal factors, fitted with the data of the ES group only. The estimated coefficients for the fixed effects SpRate and VSA in this model fitted with subset data are comparable with those in the optimal model fitted with the full dataset. The two tonal factors F0Entropy and IntensDiffHighLowTones did not significantly improve the

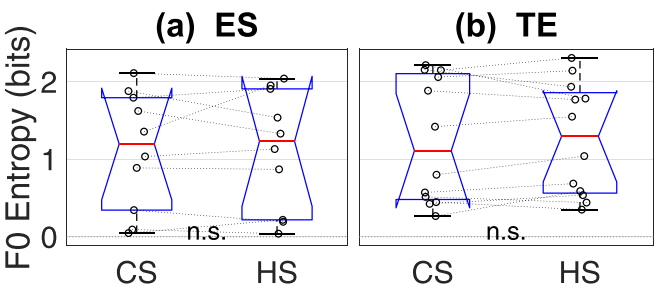

FIG. 4. (Color online) The entropies of F0 for (a) esophageal (ES) and (b) tracheoesophageal (TE) speakers, separately in clear (CS) and habitual speech (HS). (n.s.: non-significant.)

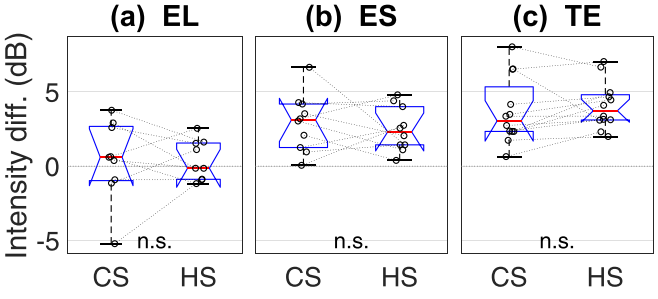

FIG. 5. (Color online) The peak intensity differences between high and low tones for (a) electrolaryngeal (EL), (b) esophageal (ES), and (c) tracheoesophageal (TE) speakers. (n.s.: non-significant.)

https://doi.org/10.1121/10.0019471

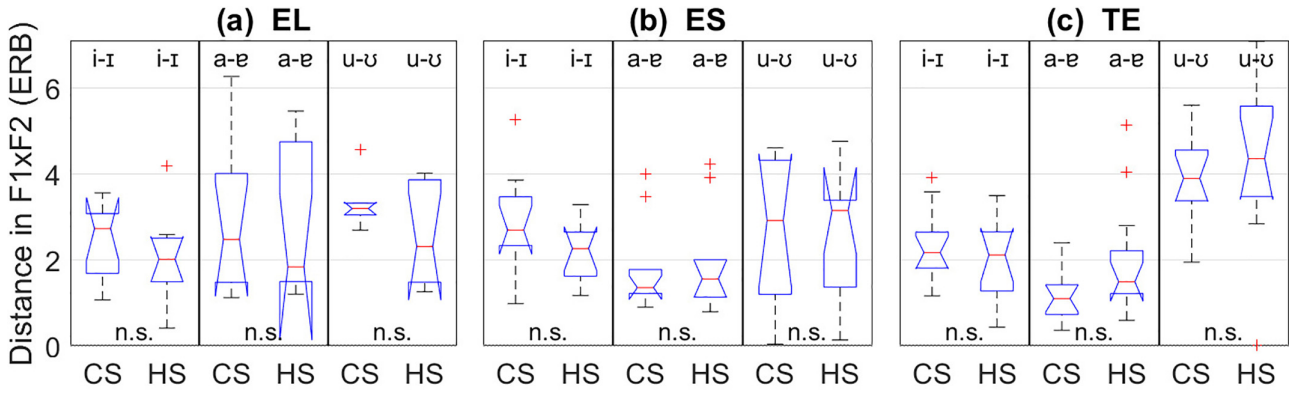

FIG. 6. (Color online) Formant distance in F1 × F2 space (in ERB) between long and short vowels for the three alaryngeal groups. (n.s.: non-significant.)

model. The other LME model with the same formula but fitted with the subset data of the TE group is summarized in Table III. The coefficients for SpRate and VSA are also consistent with the optimal model with the full dataset. The tonal factor F0Entropy significantly improved the model, whereas IntensDiffHighLowTones did not.

Examinations of the scatter plots of intelligibility as a function of F0 entropy and peak intensity difference between high and low tones (IntensDiffHighLowTones) for each group, as shown in Figs. 11 and 12, respectively, revealed positive correlations of intelligibility with F0Entropy and IntensDiffHighLowTones for both the ES and TE groups, but only the correlations between intelligibility and F0Entropy for the TE group and the correlation between intelligibility and IntensDiffHighLowTones for the ES group were significant. The patterns of correlation scatter plots in Figs. 11 and 12 are mostly consistent with our LME models presented in Tables II and III, except that the fixed effect of IntensDiffHighLowTones (on Intelli) for the ES group did not meet the significance level in the LME model. However, there is a strong and significant correlation of intelligibility and IntensDiffHighLowTones for the ES group [Fig. 12(a)].

IV. DISCUSSION

This study investigated the interaction between speaking condition, VSA, tonal contrasts, long/short vowel contrasts, and their potential effects on the intelligibility of 31 Cantonese alaryngeal speakers (9 EL, 10 ES, and 12 TE speakers). Several important findings were revealed.

The general pattern of intelligibility for Cantonese alaryngeal speakers involved ES speakers having the highest intelligibility, followed by TE speakers and EL users, although only the comparison between EL and ES was marginally significant ($p = 0.06$; Fig. 8). This pattern might be explained by findings from Ching *et al.* (1994), who studied the identification of tones produced by Cantonese alaryngeal speakers. Three ES speakers had 65.12% of their tones correctly identified, followed by 51.89% of tones for two TE speakers and 21.83% (nearly chance) for two EL users. In the present study, the amount of information encoded in the F0 and the peak intensity difference were positively correlated with intelligibility for TE and ES groups, respectively. Overall, the pattern of intelligibility in the present study was consistent with the pattern of tone identification, and Cantonese alaryngeal speakers' intelligibility appears to be positively affected when they are able to convey tonal information.

The present study also saw a significant expansion of the VSAs in all three alaryngeal groups during CS. An expanded VSA has been shown to be a defining characteristic of CS (Ferguson and Kewley-Port, 2007). Ferguson and Kewley-Port (2007) studied vowels produced by 12 talkers using CS and found that CS had a significant effect on VSA expansion alongside vowel durations. Other studies also have reported relationships between larger VSAs and longer vowel durations with higher levels of intelligibility in healthy speakers (Bradlow *et al.*, 1996; Hazan and Markham, 2004) and English and non-English speakers with neurological conditions (Liu *et al.*, 2005; Tjaden *et al.*, 2014; Turner *et al.*, 1995). According to Ng and Woo (2021), EL, ES, and TE speakers of Cantonese consistently exhibited comparable VSA values (either based on three or five vowels) when compared with laryngeal speakers in conversational (or "habitual") speech. The present findings demonstrated that Cantonese alaryngeal speakers had larger VSAs during CS, which significantly correlated with increased intelligibility. Overall, our findings concerning VSAs showed that Cantonese alaryngeal speakers consistently exhibited an expanded VSA while using CS, allowing listeners to more easily distinguish among the vowels, and thus yielding better intelligibility.

CS is known to result in a slower speaking rate, so increases in intelligibility might have been achieved at the expense of time. Speaking more slowly did not improve the intelligibility of all three alaryngeal groups. The maximum speaking rate in all three groups was approximately four syllables per second, which is a reasonably slow speaking rate when compared to typical everyday "conversational" speech [e.g., 5.1 syllables/s (Wong, 2004)]. More surprisingly, in our TE group, those who spoke faster were rated with better intelligibility scores [Fig. 9; see also Hui *et al.* (2022)]. Prior research has demonstrated that a reduced rate of speech is not the only important factor for improving

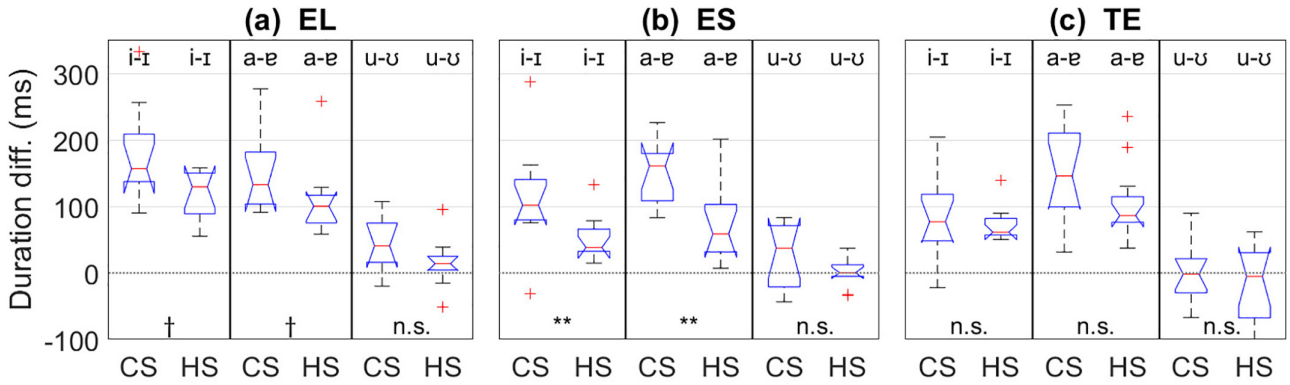

FIG. 7. (Color online) Duration difference between long and short vowels for the three alaryngeal groups. (**: $p < 0.01$; *: $p < 0.05$; †: $p < 0.1$; n.s.: non-significant.)

intelligibility while using CS (Lam *et al.*, 2012; Lam and Tjaden, 2013; Krause and Braida, 2002, 2004). In fact, Krause and Braida (2002) demonstrated that a CS benefit can be extended to faster speaking rates with training. They proposed that there must be inherent acoustic properties in CS that increase intelligibility without modifying speaking rate. Several studies have shown that a myriad of factors other than speaking rate contribute to a CS benefit including over-articulating (Lam *et al.*, 2012; Lam and Tjaden, 2013) and increasing mouth opening (Picheny *et al.*, 1985). In the current study, it is possible that some alaryngeal speakers might not have focused on one (or more) of the components involved in the initial instructions (e.g., …"to over-articulate" [in Cantonese: 誇張咁讀] and "to slow down their speech" [in Cantonese: 減慢語速 (Hui *et al.*, 2022)]. Further, Krause and Braida (2002) provided their participants with one hour of practice with CS after a thorough discussion of the technique, whereas our participants produced CS after a brief "practice session" involving the accurate production of a practice sentence.

It should be noted that vowel duration associated with ES and EL has been found to be markedly longer than laryngeal speakers while using their everyday conversational speech [e.g., Christensen and Weinberg (1976), Gandour *et al.* (1980), and Ng *et al.* (2001)]. The lengthened vocalic duration might indicate that more time is needed by alaryngeal speakers for articulating different phonemes. This seems to be more apparent when they were using the CS speaking style, which often involves hyperarticulation. Overall, more information is needed regarding the relative duration between consonants and vowels associated with different types of alaryngeal speech when using HS and CS, especially when making comparisons with laryngeal speakers.

A closer inspection of formant and durational contrasts between long and short vowels indicated that they were well-maintained in both HS and CS conditions among Cantonese alaryngeal speakers. This is a reasonable finding given that laryngectomy should not affect both the control of the articulators in the upper vocal tract and realization of durational contrasts for vowels. The long-short vowel contrasts, in either formant or duration, were not significantly increased in the CS condition to improve intelligibility (Figs. 6 and 7). Previous research investigating vowel production in HS and CS demonstrated no significant differences in vowel intelligibility between HS (e.g., 85.4%) and CS (82.7%) in English alaryngeal speakers (Cox *et al.*, 2020). This is likely due to the finding that English and Cantonese alaryngeal speakers have been consistently shown to have relatively high levels of vowel intelligibility in conversational speech [e.g., >70% (Cox *et al.*, 2020; Ng and Chu, 2009)].

Despite that the utterances produced by ES and TE speakers were largely unvoiced (cf. fully voiced in EL speech), Cantonese tones were perceivable in ES and TE speech but not

TABLE I. Summary of the fixed effects in the optimal linear mixed effect model. The formula of LME model is (Intelli ~ SpRate * Group + VSA + (1 | Speaker)). SpRate: speaking rate; VSA: vowel space area; Semicolon (:) indicates interaction.

Term	Estimate	SE	df	t-value	p-value
(Intercept)	39.2	17.0	54.6	2.297	0.026
SpRate (Group = EL)	2.6	4.6	51.9	0.572	0.57
GroupES	17.7	18.0	53.3	0.984	0.330
GroupTE	−17.5	16.3	53.2	−1.069	0.290
VSA	1.5	0.3	55.0	4.472	0.000
SpRate:GroupES	−0.8	6.8	51.1	−0.113	0.911
SpRate:GroupTE	10.8	5.6	49.1	1.913	0.062

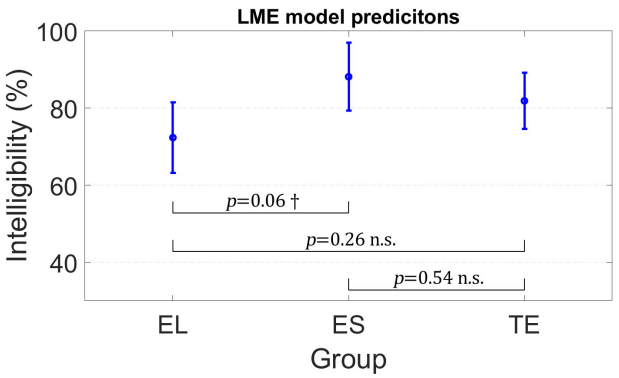

FIG. 8. (Color online) LME model predicted intelligibility for EL, ES, and TE groups with pairwise comparisons. Error bars indicate 95% confidence intervals of the mean. The comparison between EL and ES groups nearly met the significance level ($p = 0.06$) but other comparisons did not. (†: $p < 0.1$; n.s.: non-significant.)

https://doi.org/10.1121/10.0019471

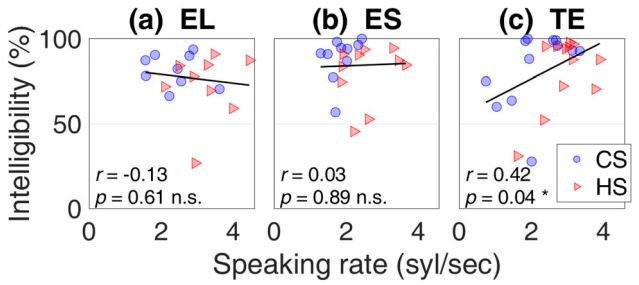

FIG. 9. (Color online) Scatter plots of intelligibility as a function of speaking rate (circle: clear speech; triangle: habitual speech). Regression lines indicate that slower speaking rate does not necessarily co-occur with higher intelligibility. (*: $p < 0.05$; n.s.: non-significant.)

in EL speech. Such devoicing may be due to insufficient closure of neoglottis while forming the sound source. Whalen and Xu (1992) pointed out that while the F0 contour is the main perceptual cue for tonal contrasts, F0 and intensity are strongly correlated; in the absence of F0, intensity plays a significant role in tone perception. Based on our measurements of the amount of the information (i.e., entropy) encoded in F0 and the intensity differences between high and low tones, tonal contrast did not change across speaking conditions (see Figs. 4 and 5). Linear mixed effect (LME) models separately trained for ES and TE groups revealed that the entropy of F0 significantly explained the variance in intelligibility for the TE group, but not in the ES group. Further, the intensity difference between high and low tones was not a significant predictor in both the ES and TE groups, despite a significant correlation between intelligibility and the intensity difference in the ES group. These "intrinsic" methods of alaryngeal speech production rely on the pulmonary system, whereas EL users do not rely on the pulmonary system to produce voice and speech. As a result, EL speakers do not show contrastive intensity differences between high and low tones. Our ES and TE speakers, on the other hand, consistently showed higher levels of intensity in high tones than in low tones [Figs. 5(b) and 5(c)] without exception. These findings suggest that Cantonese ES and TE speakers maintained reasonable tonal contrasts.

There are several limitations in the present study that must be acknowledged. First, there was a lack of normophonic Cantonese speakers to compare with our alaryngeal speakers. There are numerous studies detailing the effects of

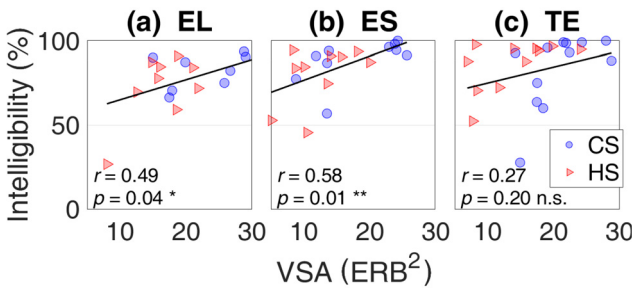

FIG. 10. (Color online) Scatter plots of intelligibility as a function of VSA (circle: clear speech; triangle: habitual speech). Regression lines indicate that there are positive correlations between intelligibility and VSA for each group. (**: $p < 0.01$; *: $p < 0.05$; n.s.: non-significant.)

TABLE II. Summary of the fixed effects in the LME model fitted with the ES group only. The formula of LME model is Intelli ~ F0Entropy + SpRate + VSA + IntensDiffHighLowTones + (1|Speaker). SpRate: speaking rate; VSA: vowel space area; IntensDiffHighLowTones: peak intensity difference between high and low tones.

Term	Estimate	SE	df	t-value	p-value
(Intercept)	42.2782	16.4538	15	2.5695	0.0214
F0Entropy	−3.978	5.4825	15	−0.7256	0.4793
SpRate	7.4329	6.1407	15	1.2104	0.2448
VSA	1.4389	0.5962	15	2.4133	0.0291
IntensDiffHighLowTones	2.9142	1.9386	15	1.5033	0.1535

CS on healthy English speakers, but it should be noted that there is a dearth of literature that attempts to understand many of the variables we investigated in Cantonese speakers (e.g., tonal contrasts). Second, the EL speakers recruited in this study used EL devices with a constant F0, and therefore, our study lacked EL speakers using devices with a variable pitch function. A variable F0 is known to positively influence intelligibility (Cox *et al.*, 2020; Watson and Schlauch, 2009), so proficient EL users who can vary their F0 could provide more information as it relates to tonal contrasts in Cantonese alaryngeal speech. They also could provide insights into the potential relationships between tonal contrasts and intelligibility in HS and CS. Such relationships were found for TE speakers, who had positive correlation between F0 entropy and intelligibility but no information was available for EL speakers (see Fig. 11). Watson and Schlauch (2009) demonstrated improved intelligibility for EL users when they can vary F0, in addition to Cox *et al.* (2020) who found positive correlations between F0 standard deviation and intelligibility in the presence of background noise. However, future research should expand upon this body of work to understand the effect of tonal contrasts on intelligibility of EL speech with varying F0s. It also should be noted that, although the intelligibility ratings and acoustic measures were assessed for the same participants, they were not analyzed using the same stimuli. This was partly because a substantial portion of our acoustic analyses focused on the prosodic characteristics of sentences, which required analyzing *The North Wind and the Sun* passage. Last, but not least, we have analyzed the acoustic properties for tones and vowels in Cantonese alaryngeal speech but not for consonants. Certainly, the goodness of

TABLE III. Summary of the fixed effects in the LME model fitted with the TE group only. The formula of LME model is: Intelli ~ F0Entropy + SpRate + VSA + IntensDiffHighLowTones + (1|Speaker). SpRate: speaking rate; VSA: vowel space area; IntensDiffHighLowTones: peak intensity difference between high and low tones.

Term	Estimate	SE	df	t-value	p-value
(Intercept)	18.8896	18.7196	16.8854	1.0091	0.3272
F0Entropy	12.7211	5.4882	10.9329	2.3179	0.0409
SpRate	11.2319	4.5793	17.9648	2.4527	0.0246
VSA	1.1196	0.6156	18.9596	1.8186	0.0848
IntensDiffHighLowTones	−0.1655	1.9933	17.8886	−0.083	0.9348

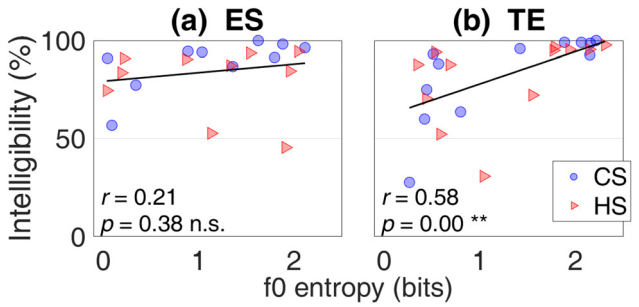

FIG. 11. (Color online) Scatter plots of intelligibility as a function of F0 entropy (circle: clear speech; triangle: habitual speech). (**: $p < 0.01$; n.s.: non-significant.)

consonant production will also have impacts on intelligibility; this may be explored by analyzing spectral shape [envelope, peak and tilt (Stevens and Blumstein, 1978)], center of gravity (Forrest *et al.*, 1988), formant transitions (Delattre *et al.*, 1955), locus equations (Lindblom, 1963), etc., in future research.

V. CONCLUSIONS

The findings of this study demonstrate that differences between Cantonese alaryngeal speakers become apparent in HS and CS when factoring in intelligibility, speech rate, and vowel characteristics. The communication functions of Cantonese, a language with a very complex tonal system, rely heavily on pitch contrasts, and thus, alaryngeal speech is more intelligible when the utterance encodes more F0 information, at least for the TE speakers. EL speakers were more intelligible in CS, which may be mediated by larger VSAs. They also produced larger VSAs with slower speaking rates compared to ES and TE speakers during CS. When all studied variables were considered, the estimated marginal means based on our linear mixed effect model indicated that the intelligibility of Cantonese alaryngeal speakers was in the order: ES > TE > EL, although only the (*post hoc*) comparison between ES and EL was marginally significant.

This study also highlights the importance of how clinicians and researchers choose to analyze alaryngeal speech. Accurate acoustic measurements of alaryngeal speech are particularly challenging with traditional voice/speech analysis techniques. For example, LPC (the most common formant estimation algorithm, as used in PRAAT) failed to identify formants in a large portion of alaryngeal speech, and as a result, required us to measure them manually by using RS. Previous studies reported that the advantages of RS over LPC were most prominent in speech with high F0; they were mostly based on synthesized speech (Whalen *et al.*, 2022). The LPC algorithm, as an all-pole model, has many assumptions (e.g., the signal is quasi-stationary, absence of anti-resonances). These assumptions are mostly met in synthesized speech, moderately met in natural speech produced with modal voice, but may not be quite so in speech with an irregular sound source, such as alaryngeal speech. Thus, LPC likely produces inaccurate formant measurements in pathological speech where those assumptions are not met, even with low F0. The representation of speech in RS, on the other hand, does not have those assumptions and provides faithful information about the resonances. More efforts should be put into developing automatic methods of RS in order to accurately measure formants in atypical speech. Direct comparison of F0 contours across tokens is infeasible with largely unvoiced utterances, as in alaryngeal speech. Alternatively, quantifying the amount of information in F0 provides a robust and useful tool. Future research should continue to examine the effects of different speaking conditions toward improving acoustic and perceptual characteristics of Cantonese alaryngeal speech, in addition to the potential nonlinear relationships between speaking rate and intelligibility. However, researchers should proceed cautiously when attempting to analyze alaryngeal voice and speech using traditional acoustic analysis methods (i.e., LPC). We encourage clinicians and researchers to consider alternatives to avoid inaccuracies in measuring alaryngeal voice and speech.

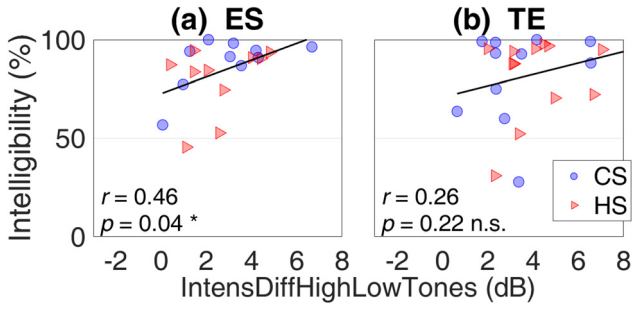

FIG. 12. (Color online) Scatter plots of intelligibility as a function of peak intensity difference between high and low tones (IntensDiffHighLowTones) (circle: clear speech; triangle: habitual speech). (*: $p < 0.05$; n.s.: non-significant.)

ACKNOWLEDGMENTS

This work was partially supported by NIH Grant No. DC-002717 to Haskins Laboratories. The authors would like to acknowledge Tak Fai Hui, formerly of the Speech Science Laboratory, University of Hong Kong, Hong Kong SAR, China, for his assistance with data collection.

APPENDIX

(1) The North Wind and the Sun.

Orthographic Version	Transcription of recorded passage
有一天	jau^{5} jɐt^{7} tʰin^{7}
北風和太陽爭論說	pɐk^{7} fʊŋ^{1} wɔ^{4} tʰai^{3} joeŋ^{4} t͡sɐŋ^{1} lɵn^{6} syt^{8}
到底誰的本領高	tou^{3} tɐi^{2} sɵy^{4} tɪk^{7} pun^{2} lɪŋ^{5} kou^{1}
當他們爭論的時候	tɔŋ^{1} tʰa^{1} mun^{4} tsɐŋ^{1} lɵn^{6} tɪk^{7} si^{4} hɐu^{6}
有一個人經過	jɐu^{5} jɐt^{7} kɔ^{3} jɐn^{4} kɪŋ^{1} kwɔ^{3}
他正穿著一件	tʰa^{1} t͡sɪŋ^{3} t͡sʰyn^{1} t͡soek^{8} jɐt^{7} kin^{6}
厚厚的黑色外衣	hɐu^{5} tɪk^{7} hɐk^{7} sɪk^{7} ŋɔi^{6} ji^{1}

The Cantonese tones are represented with superscript numbers, corresponding to Chao's (1968) five-letter tone system (5 = highest pitch; 1 = lowest pitch). The pitch height is 55 for T(one) 1, 25 for T2, 33 for T3, 21 for T4, 23 for T5, 22 for T6, 5 for T7, 3 for T8, and 2 for T9.

[1]Although earlier linguists suggested nine tones in Cantonese, more recent researchers adopted a six-tone system, as they believed the extra tones (tones 7, 8, and 9) are simply entering counterparts (syllables ending with unreleased /p, t, k/) of tones 1, 3, and 6 [e.g., Bauer and Benedict (1997) and Matthews and Yip (2011)]. This is later confirmed by a psychological study (Chu and Taft, 2011) (see also the Sec. II).

Abramson, A. S. (**1972**). "Tonal experiments with whispered Thai," in *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, edited by V. Albert (Mouton, The Hague), pp. 31–44.
Abramson, A. S. (**1975**). "The coarticulation of tones: An acoustic study of Thai," Status. Rep. Speech. Res. **SR-44**, 119–125.
Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.
Bauer, R. S., and Benedict, P. K. (**1997**). *Modern Cantonese Phonology* (Mouton De Gruyter, Berlin).
Benjamini, Y., and Yekutieli, D. (**2005**). "False discovery rate–Adjusted multiple confidence intervals for selected parameters," J. Am. Stat. Assoc. **100**, 71–81.
Boersma, P., and Weenink, D. (**2001**). "PRAAT, a system for doing phonetics by computer," Glot Int. **5**(9/10), 341–345.
Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech. Commun. **20**(3), 255–272.
Ching, T. Y., Williams, R., and Hasselt, A. V. (**1994**). "Communication of lexical tones in Cantonese alaryngeal speech," J. Speech. Lang. Hear. Res. **37**(3), 557–563.
Christensen, J. M., and Weinberg, B. (**1976**). "Vowel duration characteristics of esophageal speech," J. Speech. Hear. Res. **19**, 678–689.
Chu, P. C. K., and Taft, N. (**2011**). "Are there six or nine tones in Cantonese?," in *Proceedings of the Psycholinguistic Representation of Tone Conference*, Hong Kong, China.
Cox, S. R. (**2019**). "A review of the electrolarynx: The past and present," Perspect. ASHA SIGs. **4**, 118–129.
Cox, S. R., McNicholl, K., Shadle, C. H., and Chen, W.-R. (**2020**). "Variability of electrolaryngeal speech intelligibility in multi-talker babble," Am. J. Speech. Lang. Pathol. **29**(4), 2012–2022.
Cox, S. R., Raphael, L. J., and Doyle, P. C. (**2020**). "Production of vowels by electrolaryngeal speakers using clear speech," Folia Phoniatr. Logop. **72**(4), 250–256.
Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**(1), 769–773.
Diedrich, W. M., and Youngstrom, K. A. (**1966**). *Alaryngeal Speech* (Charles C. Thomas, Springfield, IL).
Doyle, P. C. (**2019**). "Documenting voice and speech outcomes in alaryngeal speakers," in *Clinical Care and Rehabilitation in Head and Neck Cancer*, edited by P. C. Doyle (Springer, New York), pp. 281–297.
Doyle, P. C., and Eadie, T. L. (**2005**). "The perceptual nature of alaryngeal voice and speech," in *Contemporary Considerations in the Treatment and Rehabilitation of Head and Neck Cancer*, edited by P. C. Doyle and R. L. Keith (Pro-Ed, Austin), pp. 113–140.
Ferguson, S. H., and Kewley-Port, D. (**2002**). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **112**(1), 259–271.
Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: Acoustic characteristics of vowels," J. Speech. Lang. Hear. Res. **50**(5), 1241–1255.
Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N. (**1988**). "Statistical analysis of word-initial voiceless obstruents: Preliminary data," J. Acoust. Soc. Am. **84**(1), 115–123.
Fry, D. B. (**1968**). "Prosodic phenomena," in *Manual of Phonetics*, edited by B. Malmberg (North Holland Publishing Company, Amsterdam), pp. 365–410.
Fulop, S. A., and Fitz, K. (**2006**). "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram with applications," J. Acoust. Soc. Am. **119**(1), 360–371.
Fulop, S. A. (**2011**). "Speech Spectrum Analysis," http://zimmer.csufresno.edu/~sfulop/SpeechSpecmfiles.zip (Last viewed: February 13, 2023).
Gandour, J. (**1981**). "Perceptual dimensions of tone: Evidence from Cantonese," J. Chin. Ling. **9**, 20–36, available at https://www.jstor.org/stable/23753516?casa_token=MjlphPcwU6sAAAAA%3AVJXatx2g7T6Nhil1ApO7GSuCyttHa5srJieTxWw4ZzOChYgAlR2p0SLwkmmHQXpCCDHPYOSncFPyR3pDLjBnjtQb48YktNBxGtMGIoWp2ZTIYzW62rk-.
Gandour, J. (**1983**). "Tone perception in Far Eastern languages," J. Phon. **11**(2), 149–175.
Gandour, J., Weinberg, B., and Rutkowski, D. (**1980**). "Influence of postvocalic consonants on vowel duration in esophageal speech," Lang. Speech. **23**, 149–158.
Graville, D. J., Palmer, A. D., and Bolognone, R. K. (**2019**). "Voice restoration with the tracheoesophageal voice prosthesis: The current state of the art," in *Clinical Care and Rehabilitation in Head and Neck Cancer*, edited by P. C. Doyle (Springer, New York), pp. 163–187.
Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.
Hui, T. F., Cox, S. R., Huang, T., Chen, W. R., and Ng, M. L. (**2022**). "The effect of clear speech on Cantonese alaryngeal speakers' intelligibility," Folia Phoniatr. Logop. **74**, 103–111.
Keerstock, S., and Smiljanić, R. (**2018**). "Effects of intelligibility on within- and cross-modal sentence recognition memory for native and non-native listeners," J. Acoust. Soc. Am. **144**(5), 2871–2881.
Keerstock, S., and Smiljanić, R. (**2019**). "Clear speech improves listeners' recall," J. Acoust. Soc. Am. **146**(6), 4604–4610.
Knollhoff, S. M., Borrie, S. A., Barrett, T. S., and Searl, J. P. (**2021**). "Listener impressions of alaryngeal communication modalities," Int. J. Speech Lang. Pathol. **23**(5), 540–547.
Krause, J. C., and Braida, L. D. (**2002**). "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**(5), 2165–2172.
Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**(1), 362–378.
Lam, J., and Tjaden, K. (**2013**). "Intelligibility of clear speech: Effect of instruction," J. Speech. Lang. Hear. Res. **56**(5), 1429–1440.
Lam, J., Tjaden, K., and Wilding, G. (**2012**). "Acoustics of clear speech: Effect of instruction," J. Speech. Lang. Hear. Res. **55**(6), 1807–1821.
Law, I. K., Ma, E. P., and Yiu, E. M. (**2009**). "Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers," Arch. Otolaryngol. Head Neck Surg. **135**(7), 704–711.
Lenth, R. V. (**2022**). "emmeans: Estimated marginal means, aka least-squares means," R package version 1.8.1-1.
Liao, J.-S. (**2016**). "An acoustic study of vowels produced by alaryngeal speakers in Taiwan," Am. J. Speech. Lang. Pathol. **25**(4), 481–492.
Lindblom, B. (**1963**). "On vowel reduction," Report No. 29 (Royal Institute of Technology, Speech Transmission Laboratory, Stockholm).
Liu, H., Tsao, F., and Kuhl, P. K. (**2005**). "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," J. Acoust. Soc. Am. **117**, 3879–3889.
Lo, A. (**2015**). "Intelligibility and acceptability measures of Cantonese dysarthric speech," Dissertation, University of Hong Kong, Hong Kong.
Matthews, S., and Yip, V. (**2011**). *Cantonese: A Comprehensive Grammar*, 2nd ed. (Routledge, London).
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of Interspeech 2017*, pp. 498–502.
Meltzner, G. S., and Hillman, R. E. (**2005**). "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," J. Speech. Lang. Hear. Res. **48**(4), 766–769.
Nagle, K. F. (**2019**). "Elements of clinical training with the electrolarynx," in *Clinical Care and Rehabilitation in Head and Neck Cancer*, edited by P. C. Doyle (Springer, New York), pp. 129–143.
Ng, M. L., and Chu, R. (**2009**). "An acoustical and perceptual study of vowels produced by alaryngeal speakers of Cantonese," Folia Phoniatr. Logop. **61**(2), 97–104.

Ng, M. L., Gilbert, H. R., and Lerman, J. W. (**2001**). "Fundamental frequency, intensity, and vowel duration characteristics related to perception of Cantonese alaryngeal speech," Folia Phoniatr. Logop. **53**(1), 36–47.
Ng, M. L., Kwok, C. L. I., and Chow, S. F. W. (**1997**). "Speech performance of adult Cantonese-speaking laryngectomees using different types of alaryngeal phonation," J. Voice **11**(3), 338–344.
Ng, M. L., Lerman, J. W., and Gilbert, H. R. (**1998**). "Perceptions of tonal changes in normal laryngeal, esophageal, and artificial laryngeal male Cantonese speakers," Folia Phoniatr. Logop. **50**(2), 64–70.
Ng, M. L., Liu, H., Zhao, Q., and Lam, P. K. Y. (**2009**). "Long-term average spectral characteristics of Cantonese alaryngeal speech," Auris Nasus Larynx **36**(5), 571–577.
Ng, M. L., and Woo, H. K. (**2021**). "Effect of total laryngectomy on vowel production: An acoustic study of vowels produced by alaryngeal speakers of Cantonese," Int. J. Speech. Lang. Pathol. **23**, 652–661.
Payton, K. L., Uchanski, R. M., and Braida, L. D. (**1994**). "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am. **95**, 1581–1592.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). "Speaking clearly for the hard of hearing. I. Intelligibility differences between clear and conversational speech," J. Speech. Lang. Hear. Res. **28**, 96–103.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech," J. Speech. Lang. Hear. Res. **29**, 434–446.
R Core Team (**2020**). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria.
Searl, J. P., and Reeves, S. (**2014**). "Nonsurgical voice restoration following total laryngectomy," in *Head and Neck Cancer—Treatment, Rehabilitation, and Outcomes*, edited by E. C. Ward and C. J. van As-Brooks (Plural Publishing, Chicago), pp. 263–300.
Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423.
Sisty, N. L., and Weinberg, B. (**1972**). "Formant frequency characteristics of esophageal 'speech,'" J. Speech. Lang. Hear. Res **15**(2), 439–448.
Sleeth, L. E., and Doyle, P. C. (**2019**). "Intelligibility in postlaryngectomy speech," in *Clinical Care and Rehabilitation in Head and Neck Cancer*, edited by P. C. Doyle (Springer, New York), pp. 231–246.
Smiljanić, R., and Bradlow, A. R. (**2009**). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," Lang. Linguist. Compass **3**(1), 236–264.
Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.
Tjaden, K., Lam, J., and Wilding, G. (**2013**). "Vowel acoustics in Parkinson's disease and Multiple Sclerosis: Comparison of clear, loud, and slow speaking conditions," J. Speech. Lang. Hear. Res. **56**, 1485–1502.
Tjaden, K., Sussman, J. E., and Wilding, G. E. (**2014**). "Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and Multiple Sclerosis," J. Speech. Lang. Hear. Res. **57**, 779–792.
Tupper, P., Leung, K. W., Wang, Y., Jongman, A., and Sereno, J. A. (**2021**). "The contrast between clear and plain speaking style for Mandarin tones," J. Acoust. Soc. Am. **150**(6), 4464–4473.
Turner, G. S., Tjaden, K., and Weismer, G. (**1995**). "The influence of speaking rate on vowel space and speech intelligibility for individuals with Amyotrophic Lateral Sclerosis," J. Speech. Lang. Hear. Res. **38**(5), 1001–1013.
Uchanski, R. M. (**2005**). "Clear speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 207–235.
van As, C. J., van Ravesteijn, A. M. A., Beinum, F. J. K-v., Hilgers, F. J. M., and Pols, L. C. W. (**1997**). "Formant frequencies of Dutch vowels in tracheoesophageal speech," in *Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21*, pp. 143–153.
Vance, T. J. (**1977**). "Tonal distinctions in Cantonese," Phonetica **34**(2), 93–107.
Watson, P. J., and Schlauch, R. S. (**2009**). "Fundamental frequency variation with an electrolarynx improves speech understanding: A case study," Am. J. Speech. Lang. Pathol. **18**(2), 162–167.
Whalen, D. H., Chen, W.-R., Shadle, C. H., and Fulop, S. A. (**2022**). "Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986)," J. Acoust. Soc. Am. **152**, 933–994.
Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica **49**(1), 25–47.
Wong. (**2004**). "Syllable fusion and speech rate in Hong Kong Cantonese," *Paper Presented at Speech Prosody 2004*.
Yan, N., Lam, P. K., and Ng, M. L. (**2012**). "Pitch control in esophageal and tracheoesophageal speech of Cantonese," Folia Phoniatr. Logop. **64**(5), 241–247.
Zee, E. (**1991**). "Chinese (Hong Kong Cantonese)," J. Int. Phon. Assoc. **21**(1), 46–48.
Zee, E. (**2003**). "Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers," Paper presented at the *Proceedings of the 15th ICPhS*.