Voice Quality in Telephone Interviews: A preliminary Acoustic Investigation

Timothy Pommée, and Dominique Morsomme, *Belgium*

Summary: Objectives. To investigate the impact of standardized mobile phone recordings passed through a telecom channel on acoustic markers of voice quality and on its perception by voice experts in normophonic speakers.

Methods. Continuous speech and a sustained vowel were recorded for fourteen female and ten male normophonic speakers. The recordings were done simultaneously with a head-mounted high-quality microphone and through the telephone network on a receiving smartphone. Twenty-two acoustic voice quality, breathiness and pitch-related measures were extracted from the recordings. Nine vocologists perceptually rated the G, R and B parameters of the GRBAS scale on each voice sample. The reproducibility, the recording type, the stimulus type and the gender effects, as well as the correlation between acoustic and perceptual measures were investigated.

Results. The sustained vowel samples are damped after one second. Only the frequencies between 100 and 3700Hz are passed through the telecom channel and the frequency response is characterized by peaks and troughs. The acoustic measures show a good reproducibility over the three repetitions. All measures significantly differ between the recording types, except for the local jitter, the harmonics-to-noise ratio by Dejonckere and Lebacq, the period standard deviation and all six pitch measures. The AVQI score is higher in telephone recordings, while the ABI score is lower. Significant differences between genders are also found for most of the measures; while the AVQI is similar in men and women, the ABI is higher in women in both recording types. For the perceptual assessment, the interrater agreement is rather low, while the reproducibility over the three repetitions is good. Few significant differences between recording types are observed, except for lower breathiness ratings on telephone recordings. G ratings are significantly more severe on the sustained vowel on both recording types, R ratings only on telephone recordings. While roughness is rated higher in men on telephone recordings by most experts, no gender effect is observed for breathiness on either recording types. Finally, neither the AVQI nor the ABI yield strong correlations with any of the perceptual parameters.

Conclusions. Our results show that passing a voice signal through a telecom channel induces filter and noise effects that limit the use of common acoustic voice quality measures and indexes. The AVQI and ABI are both significantly impacted by the recording type. The most reliable acoustic measures seem to be pitch perturbation (local jitter and period standard deviation) as well as the harmonics-to-noise ratio from Dejonckere and Lebacq. Our results also underline that raters are not equally sensitive to the various factors, including the recording type, the stimulus type and the gender effects. Neither of the three perceptual parameters G, R and B seem to be reliably measurable on telephone recordings using the two investigated acoustic indexes. Future studies investigating the impact of voice quality in telephone conversations should thus focus on acoustic measures on continuous speech samples that are limited to the frequency response of the telecom channel and that are not too sensitive to environmental and additive noise.

Key Words: Voice—Telephone—Acoustics—Auditory perception—Telecommunication channel.

INTRODUCTION

The labor market is undergoing significant changes, particularly following an increase in job losses exacerbated by the COVID-19 health crisis.[1,2] Among these changes and thanks to technological advances, a rise in telework is observed, along with an increasing popularity of digital recruitment procedures (eg video CVs, online and telephone interviews).[3−8]

Among the emerging digital selection procedures, online interviews can consist of live interviews involving both the applicant and the recruiter, but also of asynchronous interviews where the jobseeker records themselves answering a prespecified set of questions.[6] Similarly, video resumes are short recordings of the applicant presenting their experiences and skills. Unlike their paper counterpart, videos provide visual and auditory clues about the applicant's personality, social skills and mental capacity.[9] These can lead to non-job-related biases in the selection process, eg based on the applicant's perceived attractiveness. Next to videos, telephone interviews, including only the audio channel, are also used to screen/preselect candidates.[10−12] Studies comparing interview modalities have shown that telephone interviews filter visual cues that can have a negative impact on the applicant evaluation and thereby yield more favorable evaluations as compared to face-to-face or video

Accepted for publication August 25, 2022.

This project has received funding from the Commission permanente facultaire à la recherche (CPFR) under the grant number R.CFRA.3906-J-F.

From the Research Unit for a life-Course perspective on Health and Education, Voice Unit, University of Liège, Belgium.

Address correspondence and reprint requests to Timothy Pommée University of Liège – Voice Unit (B38), Rue de l'Aunaie, 30, 4000 - Sart Tilman, Belgium. E-mail: timothy_pommee@hotmail.com

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2022.08.027

meetings, especially for “less attractive applicants”.[12] Indeed, attractiveness is assimilated to other socially and professionally desirable features, which is commonly referred to as the “what is beautiful is good” stereotype or “halo effect”;[13] its opposite is known as the “horn effect”, where a negative aspect radiates over other ones.[14] In addition to visual parameters, another factor influencing the recruiter’s perception is the applicant’s voice.[11,15,16] Voice characteristics induce personality judgments (eg kindness, trustworthiness), both in voice-disordered[17−20] and in healthy talkers.[21−23] This cognitive voice-induced bias thus occurs in our daily verbal interactions and is of great societal importance, including in recruitment.[11] Some organizations have used the influence of voice in the personnel selection as a commercial argument (eg voicesense and precire; https://www.voicesense.com/, https://precire.com/human-resources/?lang=en). The tuning of certain voice parameters can be coached to induce positive impressions, eg by avoiding a slow speech rate and high pitch (indicating stress[24]) and using low pitch (leadership[25]) and a loud voice with large volume variations (confidence[26]).

Various studies have investigated the effects of vocal attractiveness on the listener, using perceptual ratings (eg[17]) or acoustic pitch, volume and speech rate measures.[15,27] Voice signal manipulation studies have also been carried out, showing that reducing a voice’s acoustic distance to the average timbre and pitch increases its attractiveness.[28] In the social signal processing domain, prosody is a favored factor[29] shown to have a higher influence than linguistic and visual aspects on the applicant’s appreciation.[27] However, these studies mostly rely on multimodal deep learning models trained on databases to provide automatic predictions of hireability.[30−32] While they are valuable prediction tools, they are less suitable to identify the fine-grained voice parameters that influence the hireability ratings (“black box” issue). Moreover, few studies[32] have investigated the contribution of voice quality acoustics to vocal attractiveness, let alone their importance in the hiring process.[11] Most studies indeed focus on prosody-related voice parameters such as pitch and loudness.[33,34] While extensive literature is available on attractiveness in the personnel selection domain, the present study is preliminary to a project that specifically focuses on the aspect of voice in the recruitment process.

Passetti et al.[35] have analyzed the effect of telephone transmission on the perception of voice quality, demonstrating higher ratings on supralaryngeal (eg nasality, overall muscular tension), laryngeal (eg creaky, harsh voice) and suprasegmental settings (eg prosody, temporal organization). The audio signal received by the employer might thus influence the perception of the candidate’s vocal attractiveness − either in a negative or in a positive way (eg if the high-frequency noise in breathy voices is not transmitted accurately,[36] similarly to what has been described in voice signals filtered by a facemask[37]).

Numerous authors have analyzed the validity of speech and voice quality measures on mobile phone recordings with low- and high-end phones being used as a digital recorder.[38−47] Some also investigated the effect of the transmission through telecommunication channels on the prediction accuracy of automatic detection systems in various pathologies (eg vocal fold paralysis,[48,49] Parkinson’s disease,[50,51] vocal fold carcinoma[49]). Fewer have concentrated on measure-by-measure comparisons between telephone-quality and traditional high-quality recordings[52,53] using standardized telephone calls, while Mundt et al.[53] demonstrated that the telephone standardization is critical to obtain reliable voice data.

To our knowledge, no study addressed the impact of standardized mobile phone calls on normophonic (ie without vocal complaints) voices, using measure-by-measure comparison to characterize the distortions in normophonic voice samples before investigating pathological samples. Yet, telephone conversations involve a double source of direct signal alteration: the characteristics of the internal microphone (usually microelectricalmechanical system [MEMS] or electret microphones[54]) and the frequency response of the telecom channel (which is generally said to range from 250-300 to 3000-3800 Hz[55,56]). It thus seems necessary to first characterize this signal alteration without the additional inherent particularities of pathological voice signals.

Hence, the present preliminary study aims to prepare the acoustic part of the methodology for the larger project by investigating the impact of standardized mobile phone recordings passed through a telecom channel on acoustic markers of voice quality and on its perception by voice experts in normophonic speakers. Our main hypothesis is that voice signals passed through this modality will be significantly altered and added with noise components which will be picked up by acoustic voice quality and breathiness indexes. Furthermore, we hypothesize that this acoustic alteration will significantly impact the perception of voice parameters by experienced vocologists, leading either to worse ratings due to the added noise components, or, on the contrary, to less severe ratings as the telecom channel filters out important frequencies for voice perception.

This investigation will then be useful when considering the acoustic analysis of dysphonic voice signals in telephone conversations,eg for screening purposes[51,52,56] or, in our case, when studying the impact of voice quality in telephone interviews.

METHODS

Dataset collection

Speakers

Recruitment was done via word-to-mouth. Recruited subjects had to be adult, non-smoking, native French speakers, with no past or present self-reported nor perceived voice disorder and no voice-related profession.

The final dataset is composed of 14 female and 10 male speakers, with a median age of 29 years (inter-quartile range [IQR]: 26.75-34, min=18, max=57).

Voice samples

Each participant read aloud two sentences of Harmegnies and Landercy's phonetically balanced text[57] – used in the validation study for the Acoustic Voice Quality Index 03.01 in French[58] – followed by a sustained vowel [a] for at least three seconds. This sequence was repeated three times in a row to allow investigating the reproducibility of acoustic and perceptual evaluations.

The recordings were done in a quiet room, simultaneously with two microphones:

- Directly, with a head-mounted AKG C544 cardioid condenser microphone (frequency response 20 Hz-20 kHz) at about 3-6 cm from the speaker's lips, using a Focusrite Scarlett 2i2 3rd generation USB audio interface and an Apple MacBook Pro.
- Through the telephone network, with an iPhone 13. At the receiving end, an Asus Zenfone 4 Max ZC520KL smartphone was used with its microphone disabled, in a different room, to record the incoming signal.

All recordings were digitized at a sample rate of 48 kHz, and 16 bits of resolution (mono-recording), as suggested by Chial.[59] The sound files were recorded in the lossless WAV format.

White noise was also recorded with both microphones to provide a baseline of their frequency response for the subsequent acoustic analysis. The white noise was played for 60 seconds on a single Audioengine A5+ speaker (frequency response 50 Hz-22 kHz +- 1.5 dB) in a quiet room, with both microphones placed at 6 cm from the output.

It was observed that the sustained vowel /a/, when passed through the telecom channel, is affected by the noise-reducing algorithms. Indeed, after one to two seconds, the sound is damped (eg Figure 1). Hence, the acoustic and perceptual analyses could not be carried out on the whole three seconds; in order to allow for comparisons between the values obtained on each recording type, both the AKG and the telephone recordings were trimmed to keep only the initial undamped second of the sustained vowel.

Acoustic analysis

White noise

A Fast Fourier Transform spectrum was obtained via Praat for the white noise recordings by both microphones, with a subsequent cepstral smoothing (bandwidth of 500 Hz).

Voice quality

The Acoustic Voice Quality Index (AVQI)[60,61] and the Acoustic Breathiness Index (ABI)[62,63] are both recent indexes based on literature reviews of valid acoustic voice quality measures. They correspond to the G and B scores of the GRBAS scale,[64] respectively,[65] and have been validated in many languages (eg[66–71]), among which in French[58] for the AVQI. Our own literature review confirmed that their constituent measures cover the most common and recent measures, which were also used in a recent validity investigation[72] and most of which have been found to be the most robust measures of vocal roughness and breathiness.[73] In the numerous studies validating both tools for dysphonic voices, normophonic voice samples were used concomitantly to statistically determine the pathological threshold. Hence, scores below the pathological threshold predict a G0 rating. The most investigated among the two indexes, the AVQI, has also been used in studies on normophonic voices[65,74,75] and as a screening tool on slightly dysphonic voices.[76]

In addition, pitch measures were also extracted in our study,[52,77] resulting in a total of 22 measures listed in Table 1.

Auditory-perceptual assessment

Nine vocologists with at least ten years of experience in the perceptual assessment of voice[60] were recruited. Most of their patients consult for a voice disorder (between 50 and 100%). Five of them have more than 20 years of experience in the field of voice and its perceptual assessment, the remaining four have more than 10 years of experience. By virtue of their profile, all nine vocologists are frequently associated with voice studies as scientific collaborators. Eight of the nine vocologists are both active in voice pathology and in coaching for the singing voice (ie experience with normophonic voices/voice optimization).

These experts perceptually assessed all 288 voice samples, ie the three repetitions by each of the 24 participants of both a sustained vowel and continuous speech, recorded by the AKG microphone and by the telephone.

Each sample was perceptually assessed using the three most reliable parameters of the GRBAS scale: G, R and B.[78–81] While high scores were not expected for our normophonic voice samples, it has been shown that the average overall grade (G) of normophonic speakers is close to one on the 0-3 scale,[82] highlighting the inherently variable nature of what can be considered as a normophonic voice. Therefore, normophonic speech and voice samples should not be presumed to be 100% accurate or "perfect".

The listeners were provided with detailed instructions to carry out the experiment remotely, using headphones in a quiet room. The experiment was implemented through Praat using a script adapted from Mayer[83] to present all the sound files in a random order and allow the expert to rate each parameter on a scale from 0 (normal voice) to 3 (severely dysphonic/rough/breathy voice); replay of the sound files was also available.

The listening experiment lasted about one to one and a half hour.

Statistical analysis

All statistical analyses were carried out using Stata/MP software (version 14, StataCorp, College Station, Texas). The significance level was set to 5% (P<0.05).

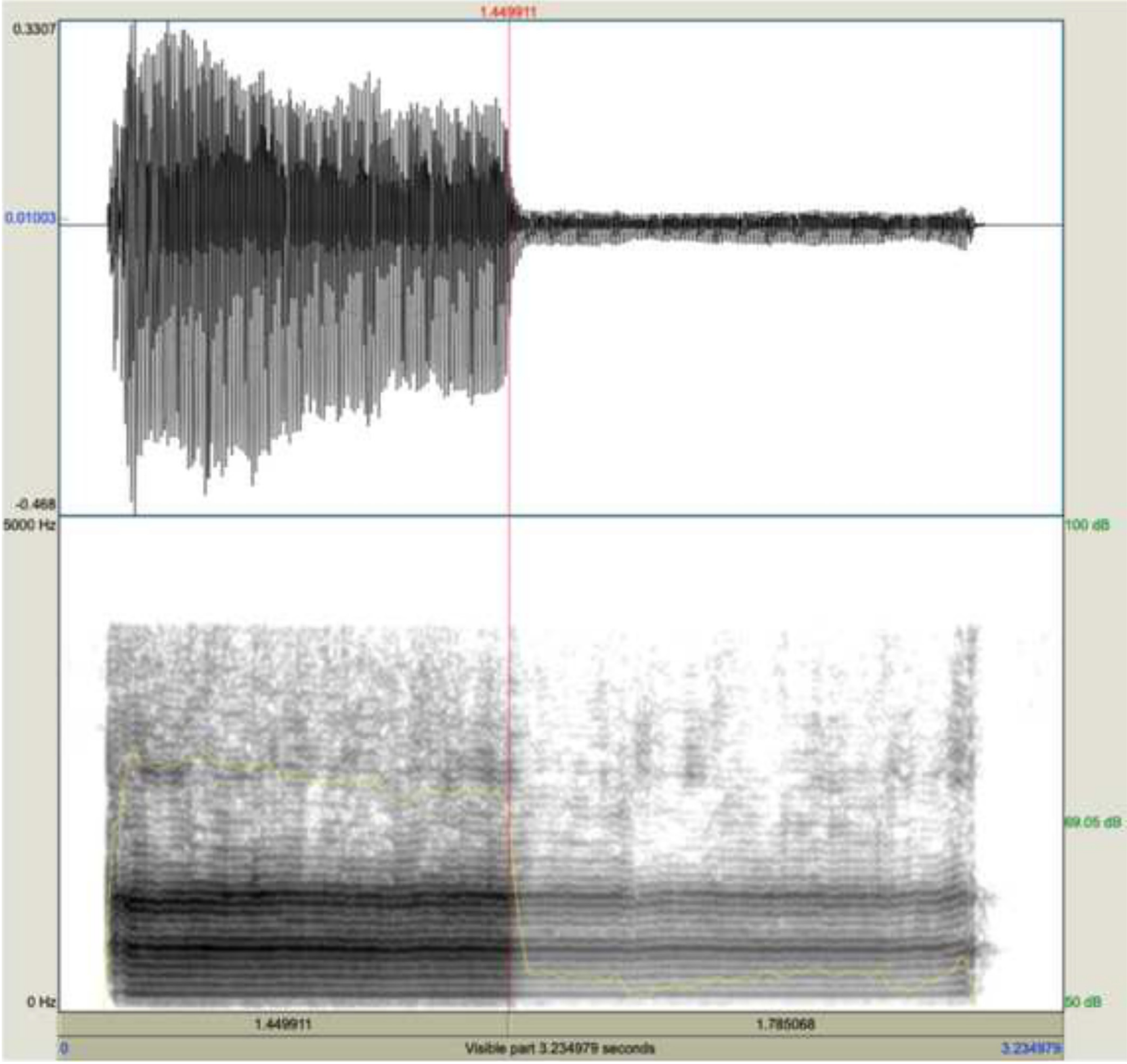

FIGURE 1. Example of a sustained vowel /a/ from the telephone recording of an 18-year-old man, with an abrupt damping around 1.45 seconds.

Results of the Shapiro-Wilk normality test led to the rejection of the Gaussian distribution null hypothesis for most of the acoustic measures and for all perceptual ratings. Hence, nonparametric tests were used for all statistical analyses.

The following statistical comparisons were all carried out separately on the AKG and on the telephone recordings, and on the continuous speech and sustained vowel samples for the perceptual ratings.

The reproducibility of both acoustic measures and perceptual ratings over the three repetitions was assessed using Friedman's test. Bonferroni-corrected Durbin-Conover's pairwise comparison was used to identify the significant differences between the repetitions ($P<0.017$ for three pairwise comparisons).

The interrater agreement for the perceptual ratings was assessed using Kendall's coefficient of concordance for the three parameters (G, R, B). The coefficients of concordance were interpreted according to[84]: 0-0.20=slight, 0.21-0.40=fair, 0.41-0.60=moderate, 0.61-0.80=substantial and 0.81-1=almost perfect/perfect agreement.

To investigate the influence of the recording type (AKG *vs.* telephone recordings) on the acoustic measures and on the perceptual ratings, Wilcoxon's signed-rank test for paired samples was used.

TABLE 1.
Acoustic Measures of Voice Quality Used in the Present Study

	Type	Measure	Abbreviation	Definition
AVQI	Index	Acoustic Voice Quality Index 03.01 (AVQI)	AVQI	Six-variable acoustic index for quantitative assessment of the overall severity of dysphonia (0: normophonic voice – 10: severely dysphonic); based on the recordings of a sustained vowel and read sentences
	Short-term amplitude perturbation	Shimmer local	SHIM	Amplitude perturbation measure: mean absolute deviation of the amplitudes of consecutive periods divided by the average amplitude
		Shimmer local dB	SHIM-DB	Absolute mean of the base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20
	LTAS	LTAS tilt	TILT	Energy difference of the regression line between 0-1000 Hz and 1000-10000 Hz across the long-term average spectrum (LTAS)
		LTAS slope	SLOPE	Energy of the low frequencies (0-1000 Hz) divided by that of the high frequencies (1000-10000 Hz) across the LTAS
	Additive noise and harmonicity	HNR	HNR	Base-10 logarithm of the ratio between the periodic energy and the noise energy, multiplied by 10
ABI	Index	Acoustic Breathiness Index (ABI)	ABI	Nine-variable acoustic index for quantitative assessment of vocal breathiness (0: normophonic voice – 10: very breathy voice); based on the recordings of a sustained vowel and read sentences
	Short-term amplitude perturbation measures	Shimmer local	SHIMLOC	Amplitude perturbation measure: mean absolute difference of the amplitudes of consecutive periods divided by the average amplitude
		Shimmer local dB	SHIMLOC-DB	Absolute mean of the base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20
	Short-term frequency perturbation measures	Jitter local	JITTERLOC	Frequency perturbation measure: mean absolute difference between consecutive periods divided by the average period
		Period standard deviation	PSD	Variation in the standard deviation of periods
	Additive noise and harmonicity	GNEmax-4500 HZ	GNE	Glottal-to-noise-excitation ratio with a maximum frequency of 4500 Hz
		High-frequency noise 0-6 kHz/6-10 kHz	HFNO	Relative level of high-frequency noise between energy from 0 to 6 kHz and energy from 6 to 10 kHz
		HNR-D	HNR-D	Harmonics-to-noise ratio from Dejonckere and Lebacq, analyzes the harmonic emergence between 500 Hz and 1500 Hz (formant zone of [a]), using a cepstral analysis to determine F0 and to localize the harmonic structure in the LTAS
		Amplitude difference H1-H2	H1-H2	Amplitude difference between the first and second harmonics; a higher H1 – H2 value (dominance of the first harmonic) has been linked to a breathier voice

(*Continued*)

TABLE 1. ***(Continued)***

	Type	Measure	Abbreviation	Definition
AVQI & ABI	Additive noise and harmonicity	Cepstral peak prominence smoothed	CPPS	Distance between the amplitude of the cepstral peak and the amplitude of the point having the same quefrency on the regression line through the smoothed cepstrum; evaluates the vocal signal without relying on identifying the fundamental period
Pitch measures	Fundamental frequency	Mean pitch continuous speech	MEANPITCH_CS	Average F0 over the continuous speech recording
		Standard deviation pitch continuous speech	SDPITCH_CS	Standard deviation of F0 over the continuous speech recording
		Pitch variability continuous speech	PITCHVAR_CS	Standard deviation of F0 divided by the average F0 over the continuous speech recording
		Mean pitch sustained vowel	MEANPITCH_SV	Average F0 over the sustained vowel recording
		Standard deviation pitch sustained vowel	SDPITCH_SV	Standard deviation of F0 over the sustained vowel recording
		Pitch variability sustained vowel	PITCHVAR_SV	Standard deviation of F0 divided by the average F0 over the sustained vowel recording

Notes: The shimmer measures in the AVQI and ABI are identical, except for one Praat setting: pitch floor and ceiling (respectively 50 Hz and 400 Hz in the AVQI, 70 Hz and 600 Hz in the ABI).

The Wilcoxon signed-rank test was also used to compare the perceptual ratings on the continuous speech and on the sustained vowel samples, for each rater. This analysis was carried out on the second repetition.

Gender differences in the acoustic measures and in the perceptual ratings were assessed using the Wilcoxon Mann-Whitney test.

Eventually, spearman correlations were carried out between the AVQI and ABI and the G, R and B ratings, separately for each judge. As the perceptual ratings were made separately on the continuous speech and sustained vowel samples while the acoustic indexes are computed on their concatenation, the perceptual ratings were averaged over each pair. The spearman correlations were interpreted according to Prion and Haerling[85]: r_s 0-0.20=negligible, 0.21-0.40=weak, 0.41-0.60=moderate, 0.61-0.80=strong and 0.81-1.00 = very strong correlation.

Ethical Considerations

This study was approved by the Ethics Committee of the Faculty of Psychology, Speech Therapy, and Education at the University of Liège. All the participants whose data were used for this study had given their informed consent (signed form) for the use of their data for research purposes. This study was carried out in accordance with the Code of ethics for speech therapists.[86]

RESULTS

White noise

Figure 2 shows the long-term average spectra for the 60 seconds of white noise recorded by the AKG head-set microphone, by the iPhone 13 as a direct voice recorder, and by the Asus Zenfone through the telecommunication channel. The resulting frequency ranges for each recording type are shown in Table 2.

On the telephone recordings, as noted for the sustained vowel, the white noise is damped after one second.

Acoustic voice quality

The median and interquartile range (IQR) for each acoustic measure are reported by repetition, by recording type and by gender (Tab. A.1 (Appendix)) as well as for each repetition x recording type x gender combination (Tab. A.2 (Appendix)), in Appendix A.

Reproducibility

Table 3 shows the median and interquartile range for all acoustic measures over the three repetitions.

For the **AKG recordings**, only the SHIMLOC measure significantly differs between the three repetitions (Q [2] = 6.13, P=0.047), more specifically between the first and second repetitions (T_2=2.509, P=0.016).

For the **telephone recordings**, two measures significantly differ:

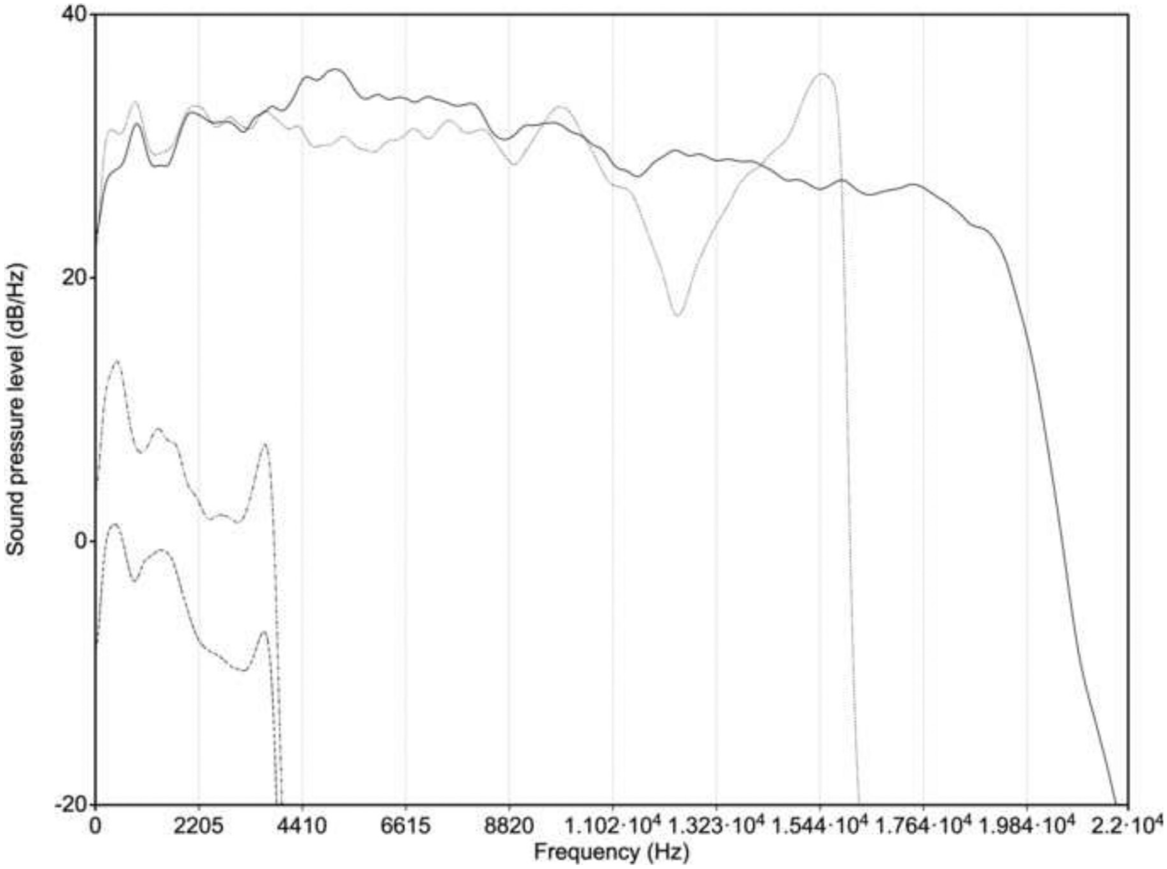

FIGURE 2. Long-term average spectra for the 60 seconds of white noise recorded by the AKG headset microphone (solid line), the iPhone 13 (dotted line), the Asus Zenfone (dashed line), as well as for the first, undamped second of the Asus Zenfone (dashed-dotted line).

TABLE 2.
Frequency Range and Response for the Three Recording Types

Source	Frequency range	Frequency response	Observations
AKG headset microphone	50-19000 Hz	23.5 – 35.8 dB	Slightly sloping frequency response
iPhone 13 as direct voice recorder	150-15700 Hz	29.8 – 35.5 dB	Dip from 11300 to 13700 Hz, peak at 15482 Hz
Asus Zenfone (through telecom channel)	100-3700 Hz	-8.4 – 1.3 dB	Three peaks (417 Hz, 1392 Hz, 3596 Hz)

- cpps (Q [2] = 6.08, P=0.048), more specifically the first and third repetitions (T_2=2.57, P=0.013);
- TILT (Q [2] = 8.33, P=0.016): more specifically the first and third repetitions (T_2=3.11, P=0.003).

Recording type

In light of the reproducibility of the values, the comparisons between the recording types were made on the second repetition only, to avoid unnecessary averaging. The median and IQR for all acoustic measures on the AKG and on the telephone recordings are shown in Table 3.

The statistical analysis showed that all measures significantly differ between the recording types, except for JITTERLOC, HNR-D, PSD and all six pitch measures (mean, standard deviation and variability on continuous speech and on the sustained vowel).

The following measures are **higher in AKG recordings**: SLOPE, HNR, ABI, GNE, H1-H2 and CPPS.

The following measures are **higher in telephone recordings**: AVQI, SHIM, SHIM-DB, TILT, SHIMLOC, SHIMLOC-DB and HFNO.

Gender effect

The comparisons between men and women were also made on the second repetition only. The median and IQR values for each acoustic measure in men and women for the AKG and for the telephone recordings are shown in Table 4.

The statistical analysis showed that all measures differ between men and women, except for the following: CPPS,

TABLE 3.
Median (Med) and IQR for Each Acoustic Measure Over the Three Repetitions, in the AKG and in the Telephone Recordings. Asterisks Indicate the Values that Significantly Differ (*=*P*<0.05, **=<0.01) Over the Repetitions (Underlined) or Between Recording Types (Bold); in Gray, the Values for the Second Repetition Used to Compare the Recording Types

	AKG						Telephone					
Measure	Repetition 1 (N=24)		Repetition 2 (N=24)		Repetition 3 (N=24)		Repetition 1 (N=24)		Repetition 2 (N=24)		Repetition 3 (N=24)	
	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR	med	IQR
AVQI												
AVQI	2.36	0.83	**2.41****	0.99	2.23	0.79	3.00	0.68	**2.89****	0.61	2.81	0.88
SHIM	6.55	2.05	**6.23****	2.07	6.20	1.16	9.28	1.78	**9.59****	2.14	9.20	2.47
SHIM-DB	0.67	0.13	**0.63****	0.15	0.63	0.12	0.93	0.13	**0.94****	0.15	0.94	0.16
TILT	-11.65	1.00	**-11.55****	0.94	-11.36	0.92	-13.75*	0.76	**-13.75****	0.88	-13.84*	0.56
SLOPE	-24.02	4.92	**-24.52****	4.48	-24.52	4.28	-19.89	4.27	**-20.27****	5.01	-19.29	4.83
HNR	17.52	2.81	**17.99****	2.13	17.93	3.07	14.73	2.66	**14.86****	2.66	14.94	2.76
ABI												
ABI	3.11	1.29	**3.02****	1.02	2.77	1.05	1.67	0.85	**1.65****	1.08	1.69	1.50
SHIMLOC	6.56*	1.72	**5.80*/****	2.25	6.07	1.33	9.21	2.19	**9.54****	1.46	9.09	2.18
SHIMLOC-DB	0.63	0.09	**0.61****	0.13	0.61	0.14	0.90	0.13	**0.94****	0.13	0.91	0.14
JITTERLOC	1.98	0.67	1.90	0.58	1.91	0.65	1.88	0.37	1.81	0.40	1.86	0.50
PSD	0.001	0.0003	0.0009	0.00025	0.00085	0.0003	0.001	0.0003	0.0008	0.0002	0.00095	0.0003
GNE	0.89	0.06	**0.89****	0.05	0.89	0.06	0.88	0.04	**0.86****	0.04	0.86	0.09
HFNO	2.20	0.17	**2.16****	0.22	2.18	0.23	4.07	0.55	**4.13****	0.44	4.05	0.55
HNR-D	24.56	7.22	24.89	7.74	24.10	6.24	25.77	8.20	25.18	7.63	25.80	6.94
H1-H2	2.87	4.14	**3.34****	4.84	3.49	4.46	-1.64	11.75	**-2.03****	11.00	-1.64	9.56
AVQI & ABI												
CPPS	12.76	1.26	**12.81****	1.63	13.34	1.51	11.99*	0.80	**11.90****	0.91	12.14*	1.13
Pitch												
MEANPITCH_CS	204	107	199	101	194	103	206	102	200	102	194	104
SDPITCH_CS	30	26	30	23	27.24	22	31	21	29	20	31	23
PITCHVAR_CS	0.16	0.07	0.16	0.06	0.16	0.07	0.15	0.08	0.17	0.06	0.16	0.08
MEANPITCH_SV	168	88	166	87	163	90	167	88	156	85	163	90
SDPITCH_SV	1.38	0.85	1.19	0.97	1.21	0.79	1.41	0.92	1.22	1.06	1.20	0.83
PITCHVAR_SV	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00

TABLE 4.
Median (Med) and IQR for Each Acoustic Measure in Men and Women on the Second Repetition, in the AKG and in the Telephone Recordings. Asterisks Indicate the Values that Significantly Differ Between Genders (*=*P*<0.05, **=<0.01)

Measure	AKG				Telephone			
	Women (rep 2, N=14)		Men (rep 2, N=10)		Women (rep 2, N=14)		Men (rep 2, N=10)	
	median	IQR	median	IQR	median	IQR	median	IQR
AVQI								
AVQI	2.19	0.97	2.52	0.75	2.77	0.50	3.18	0.50
SHIM	5.35**	2.09	7.23**	1.76	9.21	1.69	10.18	1.28
SHIM-DB	0.56**	0.13	0.70**	0.20	0.91*	0.13	1.00*	0.12
TILT	-11.38	0.79	-11.74	1.03	-13.63	0.83	-13.80	0.70
SLOPE	-24.52	4.49	-24.61	5.57	-20.52	4.84	-19.42	5.55
HNR	18.69**	2.26	16.58**	2.67	15.72**	1.71	13.20**	2.18
ABI								
ABI	3.44**	0.57	2.47**	0.56	1.98**	0.67	1.01**	0.66
SHIMLOC	5.37**	1.05	7.51**	2.10	9.07*	1.24	10.01*	1.36
SHIMLOC-DB	0.58**	0.11	0.71**	0.16	0.91	0.14	0.98	0.14
JITTERLOC	1.70**	0.39	2.34**	0.44	1.74*	0.45	2.06*	0.45
PSD	0.00085*	0.0003	0.001*	0.0001	0.0008*	0.0002	0.001*	0.0002
GNE	0.88	0.04	0.91	0.08	0.85	0.07	0.88	0.03
HFNO	2.11	0.19	2.28	0.23	4.12	0.46	4.17	0.47
HNR-D	26.83**	2.22	18.93**	1.36	26.61**	1.86	19.17**	1.77
H1-H2	4.89*	4.23	1.87*	3.06	1.33**	5.10	-9.47**	3.04
AVQI & ABI								
CPPS	12.88	1.95	12.59	0.84	12.22	0.88	11.75	0.62
Pitch								
MEANPITCH_CS	215**	10	113**	7	215**	14	112**	11
SDPITCH_CS	38**	6	14**	10	38**	19	15**	11
PITCHVAR_CS	0.18*	0.03	0.12*	0.07	0.18	0.08	0.14	0.07
MEANPITCH_SV	191**	36	103**	8	190**	40	103**	8
SDPITCH_SV	1.50	1.24	1.00	0.39	1.73*	1.16	0.98*	0.45
PITCHVAR_SV	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.00

GNE, HFNO, SHIMLOC-DB(telephone), SHIM(telephone), SLOPE, TILT, AVQI, PITCHVAR_CS(telephone), SDPITCH_SV(akg) and PITCHVAR_SV.

The following acoustic measures show **higher values in men**: SHIM(AKG), SHIM-DB, SHIMLOC, SHIMLOC-DB(AKG), JITTERLOC and PSD.

The following acoustic measures show **higher values in women**: HNR, ABI, HNR-D, H1-H2, MEANPITCH_CS, SDPITCH_CS, PITCHVAR_CS(AKG), MEANPITCH_SV and SDPITCH_SV(telephone).

Auditory-perceptual assessment

Interrater agreement

Kendall's coefficient of concordance ranged from 0.15 to 0.35, indicating only a slight to fair interrater agreement, which will be discussed hereafter. Hence, no averaging was carried out across the raters and the following analyses were carried out considering each rater independently.

Reproducibility

The analysis of the reproducibility of the perceptual ratings (grade, roughness, breathiness) over the three repetitions for each rater showed no significant difference, neither for the telephone recordings nor for the AKG recordings (see Table 5 for rating frequencies over all raters), neither for the continuous speech recordings nor for the sustained vowels.

Recording type

Again, in light of the reproducibility of the ratings, the comparisons between the recording types were made on the second repetition only.

The results, synthesized in Table 6, show few significant differences between the ratings on the AKG and the telephone recordings. One exception is observed for breathiness, which is rated significantly less severely on telephone recordings by six of the nine raters.

The overall percentages for all raters on repetition two on the AKG and on the telephone recordings are shown in Table 5.

Stimulus type

The statistical analysis, synthesized in Table 7, showed that the G ratings are significantly more severe on the sustained

TABLE 5.
Voice Quality Ratings Over the Three Repetitions in AKG and in Telephone Recordings

		AKG			Telephone		
		Rep 1 (N=432)	Rep 2 (N=432)	Rep 3 (N=432)	Rep 1 (N=432)	Rep 2 (N=432)	Rep 3 (N=432)
Grade	0	44%	41%	48%	50%	53%	52%
	1	43%	46%	39%	36%	34%	35%
	2	11%	11%	11%	12%	11%	11%
	3	2%	2%	2%	2%	2%	2%
Roughness	0	60%	60%	62%	55%	53%	51%
	1	28%	28%	27%	30%	33%	34%
	2	8%	9%	7%	10%	8%	10%
	3	4%	3%	4%	5%	6%	5%
Breathiness	0	46%	44%	47%	67%	69%	70%
	1	40%	41%	40%	22%	21%	21%
	2	9%	10%	10%	10%	9%	8%
	3	5%	5%	3%	1%	1%	1%

TABLE 6.
Frequencies of Higher G, R and B Ratings (*P*<0.05) on AKG Recordings, on Telephone Recordings and Non-Significant Differences, Both in Continuous Speech and in Sustained Vowels

	N=9	AKG score > telephone	Telephone score > AKG	No difference
Continuous speech	G	1	0	8
	R	0	0	9
	B	2	0	7
Sustained vowel	G	2	0	7
	R	0	2	7
	B	6	0	3

vowel samples for almost half of the raters, on both recording types. The roughness and breathiness parameters are slightly less equivocal on the AKG recordings.

On the telephone recordings, roughness is more severely rated on sustained vowels by a majority of the raters, while breathiness ratings seem to be less influenced by the stimulus type.

The overall percentages for all raters on repetition two on the continuous speech and on the sustained vowel samples are shown in Table 8.

Gender effect

The statistical analysis showed that the G parameter was rated significantly higher in men by three raters on the telephone recordings, while no significant difference was measured for AKG recordings. Roughness was rated significantly higher in men by four raters on the AKG recordings and by five raters on the telephone recordings. No difference was found between genders for breathiness, neither on the AKG nor on the telephone recordings.

TABLE 7.
Frequencies of Higher G, R and B Ratings (*P*<0.05) on Continuous Speech, on Sustained Vowel and Non-Significant Differences, Both in AKG and in Telephone Recordings

	N=9	CS score > SV	SV score > CS	No difference
AKG	G	0	4	5
	R	1	3	5
	B	1	3	5
Telephone	G	0	4	5
	R	0	6	3
	B	1	1	7

TABLE 8.
Voice Quality Ratings on the Continuous Speech and on the Sustained Vowel for the Second Repetition in AKG and in Telephone Recordings

		AKG (Rep2)		Telephone (Rep2)	
		Continuous speech (N=216)	Sustained vowel (N=216)	Continuous speech (N=216)	Sustained vowel (N=216)
Grade	0	53%	29%	63%	43%
	1	40%	53%	30%	37%
	2	5%	17%	6%	16%
	3	2%	1%	1%	4%
Roughness	0	65%	53%	64%	42%
	1	25%	31%	29%	37%
	2	7%	12%	4%	11%
	3	3%	4%	3%	10%
Breathiness	0	47%	41%	68%	71%
	1	39%	42%	20%	20%
	2	10%	11%	12%	7%
	3	4%	6%	0%	2%

Correlations between acoustic indexes and perceptual ratings

The number of raters for which a significant correlation was found between the acoustic indexes and the G, R and B parameters are shown in Table 9, by recording type.

The results show that neither the AVQI nor the ABI yield strong correlations with any of the perceptual parameters.

The AVQI and the ABI both show weak to moderate positive correlations with breathiness ratings in the AKG recordings for most raters.

In telephone recordings, this trend disappears for both indexes, while the ABI shows a weak to moderate negative correlation with roughness in most raters.

DISCUSSION

The aim of the present study is to investigate the impact of standardized mobile phone recordings passed through a telecom channel on acoustic markers of voice quality and on its perception by voice experts in normophonic speakers.

Frequency ranges

The white noise recordings already reveal a first important observation: only the frequencies between 100 and 3700 Hz are passed through the telecom channel and the frequency response is characterized by peaks and troughs. This is in accordance with previous studies, although the filtering effect is slightly less important in our results; indeed, previous studies have reported frequencies between 300 and 3000 Hz only.[53,55,56]

Reproducibility

Both the acoustic measures and the perceptual ratings show a good reproducibility across the repeated recordings, for the AKG as well as for the telephone recordings. These results indicate that, as long as the same recording instrument and parameters are used, the acoustic measures and perceptual ratings are stable over time and thus reliable.

Interrater reliability

The interrater reliability analysis showed that, while they were consistent over time, the nine experts who participated in the present study did not provide similar ratings – neither for the AKG recordings, nor for the telephone recordings. This has been observed in several previous studies.[60,87] As has been demonstrated by Delvaux et al.,[82] the rating reliability when using the GRBAS scale is affected by speaker-, listener- and task-related factors. In the present study, so as not to bias the ratings, the experts were not told beforehand that they would listen to alleged normophonic voices, nor that they would hear both high-quality and telephone recordings. The lack of prior specification of the types of samples to be evaluated may have led to differential processing among raters, thus increasing the already widely described interrater variability.

The resulting ratings involve each level of severity, and the scale is activated in its entirety. We do thus not observe a floor effect, which would have induced a higher interrater reliability. While high severity ratings were not expected for our normophonic voice samples, a little less than half of the samples yield G and B scores of 1, while still a considerable amount (about 10%) yields a rating of 2. These results confirm previous findings[82] and highlight the importance of moving away from the assumption that normal voices are inherently “perfect”. The complex nature of voice production and perception as well as their inherent variability warrants the need to assess the voices of control groups when studying acoustic and perceptual voice measures of dysphonia. In addition to the “imperfect” nature of normal voices, another hypothesis to explain the rating distributions is that when assessing normophonic voices, expert raters tend to

TABLE 9.
Significant Correlations (*P*<0.05) Between the Acoustic Indexes AVQI and ABI and the Perceptual Ratings, by Recording Type (in Bold, the Correlations that Were Found in at Least Seven Out of the Nine Raters)

Recording type	Perceptual parameter	Acoustic measure	Raters (/9)	Spearman coefficient (r_S) range	Interpretation (n)	*P*-value range
AKG	G	AVQI	3	[0.25;0.40]	weak (3)	[0.0006;0.04]
		ABI	3	[-0.27;0.38]	weak (3)	[0.001;0.02]
	R	AVQI	1	0.24	weak (1)	0.04
		ABI	6	[-0.25;-0.56]	weak (3) to moderate (3)	[<0.0001;0.03]
	B	AVQI	**7**	**[0.27;0.44]**	weak (6) to moderate (1)	**[0.0001;0.04]**
		ABI	**9**	**[0.24;0.56]**	weak (6) to moderate (3)	**[<0.0001;0.04]**
Telephone	G	AVQI	3	[0.25;0.36]	weak (3)	[0.001;0.03]
		ABI	4	[-0.25;-0.42]	weak (2) to moderate (2)	[0.0002;0.04]
	R	AVQI	2	[0.24;0.32]	weak (2)	[0.006;0.04]
		ABI	**7**	**[-0.24;-0.60]**	weak (4) to moderate (3)	**[<0.0001;0.04]**
	B	AVQI	3	[0.25;0.40]	weak (3)	[0.0006;0.04]
		ABI	1	0.33	weak (1)	0.004

seek even slight deviations to distribute their ratings over the entire scale. This induces more severe ratings than if the samples had been presented along with dysphonic voices.

Hence, while the interrater reliability is rather low in both recording types, the perceptual ratings were nevertheless analyzed, bearing in mind that our experts cannot be considered as a homogeneous group. Indeed, the detailed observation of the perceptual ratings highlighted that the raters are not equally sensitive to the various factors, including the recording type and the stimulus type effects. This raises caution for future studies employing telephone recordings.

Moreover, our results further support the importance of perceptually assessing normophonic speech and voice samples, eg in control groups, instead of assuming a “perfect” speech or voice quality.[88]

Gender effect

All amplitude and pitch perturbation measures are higher in men, whereas the HNR measures are higher in women. Previous studies on the matter have shown contradictory results and it has been concluded by several authors that the jitter and shimmer measures are highly dependent on each individual’s sound pressure level and f_o.[89,90] Our results indicate that our cohort of men includes voices of lesser acoustic quality than our cohort of women. This is confirmed by the perceptual ratings, several raters perceiving a higher roughness both in the AKG and in the telephone recordings of men, as well as a higher G in men on telephone recordings. The perceived breathiness, however, does not differ between genders, on either recording type. This is surprising in light of both the acoustic measures in our study − the ABI and HNR measures being significantly higher in women − as well as previous knowledge of gender differences in perceived breathiness.[91,92] This may be due to our expert panel, which was exclusively composed of women. Indeed, as explained in the theoretical framework proposed in,[87] perceptual ratings are influenced among others by listener-related factors, which shape their individual internal standards and affect interrater reliability more than intrarater reliability. Several studies demonstrate that listeners show a voice perception bias based on their own auditory experience,[93] eg showed that dysphonic women rated other dysphonic women less severely, while[94] observed an own-age bias in age estimations from voices in older people. Hence, we hypothesize that women might rate other women’s breathiness less severely based on their own auditory experience, thereby smoothing out the gender difference.

Considering the acoustic indexes, the AVQI score does not differ by gender, independently of the recording type, which confirms previous conclusions in the literature.[95−97] The ABI, however, is higher in women − both in the AKG and in the telephone recordings −, which is consistent with the knowledge of an incomplete posterior glottal closure in women,[98−101] hence inducing a breathier voice quality. Surprisingly, in the original validation study of the ABI,[62] the authors include both men and women in their cohort, but do not investigate possible gender differences. The absence of a gender difference has only been reported in one study, on Japanese-speaking patients.[68] It is to be noted that while H1-H2 is significantly higher for women in our results, indeed indicating a higher degree of breathiness, Simpson[102] has shown that the H1-H2 measure is not a reliable measure to identify gender-related breathiness differences. Moreover, Pépiot[103] showed that these differences are language dependent. The gender effect on the ABI thus needs to be further investigated.

Recording type

When comparing the acoustic measures on the AKG and on the telephone recordings, we observe that the **pitch measures** on the sustained vowel and on the continuous speech samples do not differ, nor do the period perturbation measures relying on pitch tracking (JITTERLOC and PSD). This is in accordance with several studies demonstrating that pitch-related measures are more robust and reliable on phone recordings than eg amplitude perturbation measures.[38,39,41,46,47] It is to be noted that the present study employed the periodicity-to-pitch autocorrelation function in Praat for the pitch measures, which is not dependent on the actual presence of a fundamental frequency and is thus less impacted by potential effects of the low-frequency filtering in telephone recordings.[52,53]

The **composite voice quality indexes AVQI and ABI** are both significantly impacted by the recording type.

The AVQI score is higher in telephone recordings, associated with higher amplitude perturbation measures, a steeper tilt of the regression line through the LTAS, a diminished HNR and a lower CPPS, indicating a less periodic and noisier spectrum. Several considerations should be taken into account when analyzing these acoustic measures in telephone recordings.

First, the amplitude perturbation and HNR measures have previously been shown to be less reliable on telephone recordings.[38,46,104] They rely on the identification of the fundamental period,[105] which is highly impacted by the telecom channel filtering out the low frequencies. The CPPS is said to be a more reliable alternative, as it does not rely on the pitch tracking[105]; however, as explained in,[46] the CPPS was developed to be particularly sensitive to noise (glottal noise or perceived breathiness) and it is thus not surprising that it is also influenced by the recording conditions in the present study – the directional AKG microphone reducing the environmental noise, while the telephone recording most likely adds instrumental noise both due to its microphone and due to the transmission channel. The reduced ratio of harmonics to noise components can also be related to this instrumental noise induced by the telecom channel.

Second, the filtering out of frequencies above 3700 Hz also has to be kept in mind when analyzing the acoustic measures on telephone recordings. Indeed, it is in these high frequencies that the noise components are mainly concentrated,[50] hence the filter directly impacts both the HNR and the TILT measures. While the steeper tilt of the regression line through the LTAS is easily explained by the filter effect, the less important SLOPE measure is a somewhat surprising result; we suspect the SLOPE measure to be less reliable than the trendline measure, in light of the peaks and troughs that are observed in the telephone white noise recordings.

The ABI score, on the other hand, is lower in telephone recordings. Yet, as in the AVQI, higher amplitude perturbations and a lower CPPS are measured, and the lower GNE measure indicates a higher breathiness. This is inconsistent with the higher HFNO and lower H1-H2 values, indicating a lower breathiness[32] in telephone recordings. Again, these observations can be explained by the particularities of the acoustic signal passed through the telecom channel. The higher HFNO can be related to the filtering of the frequencies above 3700 Hz, as this measure compares high frequency noise components (>6000 Hz) to low frequencies (<6000 Hz). The maximum frequency considered by the GNE measure, on the other hand, is 4500 Hz. This measure might thus be less affected by the telephone-related 3700 Hz low-pass frequency filter and thereby better represent the actual noise as compared to HFNO. Furthermore, the low-bandwidth spectral tilt measure H1-H2 might be influenced by the troughs observed in the low frequencies of the white noise telephone recordings, where the first harmonics are situated. A negative H1-H2 measure could eg result from a first harmonic situated in a trough and a second harmonic situated in a peak.

Finally, unlike the AVQI's HNR measure, the ABI's HNR-D is limited to the frequency region between 500 and 1500 Hz, which is within the 100-3700 Hz frequency response of the telephone recordings. This explains why the HNR-D is not influenced by the recording type; this measure should thus be preferred when analyzing voice quality on telephone recordings.

Overall, most of these results are consistent with those from studies using direct smartphone recordings. The observations that contradict previous findings can be explained by the fact that the telephone recordings in the present study were passed through a telecom channel, thereby adding frequency filtering and additive noise effects. Indeed, van der Woerd[104] eg also found lower HNR and higher shimmer values in telephone recordings, with no significant impact of the telephone recording on F0 and jitter measures. However, they also measured higher cpp values in telephone recordings, while our results show a reduced CPPS. This supports the assumption that passing the telephone recording through a telecom channel adds noise components to the signal, which significantly impact the CPPS measure. Furthermore,[41] also measured no difference for F0 and jitter, as well as a higher shimmer and a lower H1-H2 ratio. However, a significantly lower TILT was measured for telephone recordings, while our results show a higher TILT. This again can be related to the use of the telecom channel in our study, which filters out all frequencies above 3700 Hz and thereby impacts the overall shape of the LTAS.

Hence, to conclude on the use of acoustic voice quality measures on telephone recordings, the acoustic voice quality indexes cannot be used as such in our subsequent study aimed at investigating the influence of voice quality in telephone interviews. Indeed, the breathiness as measured by the ABI is underrated in telephone recordings mainly because of the low-pass filtering as well as instrumental and ambient noise. To measure additive noise in the voice signal through telephone recordings, the HNR-D and GNE measures should be preferred, as they concentrate on the frequencies that are not filtered out. The overall voice quality as assessed by the AVQI is underrated for similar reasons. The CPPS, used in both indexes, is overly sensitive to noise, which is artificially increased in telephone recordings. Furthermore, both indexes use a sustained vowel, which when passed through the telephone channel is identified as "noise" and damped after one second. This duration is not ideal either for acoustic or for perceptual assessments. Hence, future studies on voice samples passed through the telecom channel should focus on voice quality measures on continuous speech.

Considering the perceptual ratings of the voice samples on both recording types, we observe very few significant differences. One exception is noted for breathiness, which is rated less severely on the telephone recordings by most raters in both stimulus types. We hypothesize that this is because the source-related high-frequency noise is filtered out in telephone recordings, while instrumental noise is added to the signal. This might lead to two phenomena: on the one hand, the source-related noise in the high frequencies is canceled out, thus possibly reducing the perceived breathiness; on the other hand, the instrumental noise which is added in telephone recordings might be perceived and identified as such by the listener, thereby inducing a cognitive assimilation. The rater hence assimilates the potential source-related noise or perceptual deviance with the low quality of the telephone recordings and thus rates the samples less severely.

Stimulus type

Overall, in sustained vowels, G and R are globally more severely rated, both on AKG and on telephone recordings − which is consistent with previous studies.[106,107] One hypothesis, based on observations made during the recording sessions, is that sustaining a vowel is an uncommon task for most speakers which induces various phenomena, eg overly forced phonation, aggravated pitch/vocal fry, or breathiness due to timidity, contrasting with a normophonic voice quality in continuous speech. This has been observed in other studies as well, eg in,[106] the authors concluding that "normal speakers frequently produce sustained vowels that are more dysphonic than continuous speech".

Breathiness ratings seem to be less influenced by the stimulus type, which can be linked to the above-described hypothesis regarding the filtering of the source-related high-frequency noise and the addition of instrumental noise, clouding the perception of breathiness.

Correlations between acoustic indexes and perceptual ratings

The AVQI and the ABI were created and validated as indexes of overall voice quality and breathiness, respectively. Our results, however, show that neither the AVQI nor the ABI yield strong correlations with any of the perceptual parameters.

The AVQI, for which a high correlation with G was expected, rather seems to correlate with breathiness in AKG recordings. In telephone recordings, the AVQI does not reach satisfactory correlations with either of the perceptual parameters.

For the ABI, although the correlations with perceived breathiness in the AKG recordings are weak to moderate, they are observed for all the nine judges as expected. In telephone recordings, however, this correlation disappears, while a negative link with roughness is observed. Voices that are perceived as rougher thus seem to correspond to a lower ABI score in telephone recordings.

These results must be interpreted with caution. Indeed, as discussed above, most of the acoustic measures composing the AVQI and the ABI are highly impacted by the particularities of the acoustic signal passes through the telecom channel. Furthermore, both indexes have been validated for the use on concatenated samples composed of continuous speech and three seconds of a sustained vowel. In our study, only one second of the sustained vowel was kept, because of the damping observed in telephone recordings. This sample additionally includes the vocal attack, while the original indexes are intended to use three seconds without the voice onset and offset. The resulting median scores in our study for the AVQI are pathological both in the AKG and in the telephone recordings if we consider the cut-off score of 2.33 which was computed on the same French sample used in the present study.[58] For the ABI, no validation study has yet been carried out in French.

Hence, while the AVQI and ABI indexes have been shown to be valid and reliable on high-quality recordings in numerous studies, neither of the three perceptual parameters seem to be reliably measurable on telephone recordings using these indexes.

Perspectives for future studies on the impact of voice quality in telephone recordings

The results of this preliminary study allow us to conclude that upcoming studies should focus on acoustic measures on continuous speech samples that are limited to the frequency response of the telecom channel and that are not too sensitive to environmental and additive noise. We recommend the use of the here investigated pitch-related measures as well as the harmonics-to-noise ratio from Dekonckere and Leback, in addition to modified versions of the spectral tilt and glottal-to-noise excitation ratio, both to be limited to the frequency response of the telecom channel. An extensive literature review could also allow identifying other potential measures that could apply to this context.

The present study used the standardized speech and voice materials on which the AVQI 03.01 has been validated in French, in order to be able to interpret the composite score. Using semi-spontaneous speech would have resulted in highly variable samples, especially concerning vowels and voiced segments that are extracted for the analyses. Comparing the resulting scores extracted from non-standardized samples would hence have induced an important supplementary source of variability. Therefore, in the present study, this factor was controlled by using the standardized samples – at the expense of ecological representativeness. Indeed, these voice samples are not the most reflective of how individuals typically communicate over the telephone. Hence, as this study allowed to conclude a low reliability of the two indexes on telephone recordings, further studies should focus on the isolated acoustic features that were found to be most reliable. These can be used on more representative speech samples, as no standardized procedure exists for their extraction as opposed to the indexes.

Intra-rater agreement by repetition of some of voice samples to each rater was not calculated in this study. This is a limitation, as rating normophonic voice samples is not a common task for vocologists. Further exploration of rater performance under these conditions is necessary to allow future studies investigating measurements in normophonic voices before they can be applied to pathological voices. Indeed, in order to be able to make good use of measurements, be they acoustic or perceptual, on samples that are highly variable due to voice pathology, it is necessary to describe the already inherent variability of the complex and multidimensional phenomenon ie the normophonic voice.

Furthermore, the inter-rater agreement results from the perceptual ratings in the present study highlight the importance of considering each rater's individual sensitivity to various factors such as vocal parameters, recording type and gender. Therefore, in future studies on telephone recordings, interrater agreement between the raters assessing the voice samples should be analyzed in depth. Some of the factors that need to be studied are the influence of the scale used for the perceptual ratings (eg Likert *vs* visual analog scales, direct magnitude estimation or pairwise comparison), the types of voice samples (eg continuous speech *vs* sustained vowel), the raters' expertise in rating normoponic and pathological voices (eg laypersons *vs* vocologists). We hypothesize that groups of raters with similar cognitive and perceptual strategies could be identified. To test this hypothesis, large samples of raters would be necessary to allow reaching a high statistical power, to identify potential groupings and to be able to interpret correlations between acoustic and perceptual voice parameters despite a potentially low overall interrater agreement.

CONCLUSION

This study aimed to investigate the impact of standardized mobile phone recordings passed through a telecom channel on acoustic markers of voice quality and on its perception by voice experts in normophonic speakers. Our results show that passing a voice signal through a telecom channel induces filter and noise effects that limit the use of common acoustic voice quality measures and indexes. The AVQI and the ABI do not prove to be reliable on telephone recordings. The most reliable acoustic measures seem to be pitch perturbation (JITTERLOC and PSD) as well as the harmonics-to-noise ratio from Dejonckere and Lebacq.

Our results also underline that raters react individually to the different factors when perceptually assessing voice samples, including the recording and the stimulus types as well as the gender effect.

Neither of the three perceptual parameters G, R and B seem to be reliably measurable on telephone recordings using the investigated acoustic indexes.

This preliminary study allows us to conclude that future studies investigating the impact of voice quality in telephone conversations should thus focus on acoustic measures on continuous speech samples that are limited to the frequency response of the telecom channel and that are not too sensitive to environmental and additive noise.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

The authors thank all the speakers and vocologists who gave their time to participate in this study.

APPENDIX A – MEDIANS AND INTERQUARTILE RANGES FOR EACH ACOUSTIC MEASURE BY REPETITION, RECORDING TYPE AND GENDER.

.

TABLE A.1.
Medians and Interquartile Ranges (IQR) for Each Acoustic Measure by Repetition, by Recording Type and by Gender

Measure	Repetition 1		Repetition 2		Repetition 3		AKG (rep 2)		Telephone (rep2)		Women (rep 2)		Men (rep 2)	
	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR	median	IQR
AVQI														
avqi	2,66	0,84	2,63	0,82	2,50	0,92	2,41	0,99	2,89	0,61	2,57	0,83	2,78	0,77
shim	7,91	2,87	8,01	3,36	7,84	3,12	6,23	2,07	9,59	2,14	7,29	3,87	8,47	2,96
shdb	0,78	0,26	0,82	0,30	0,81	0,31	0,63	0,15	0,94	0,15	0,70	0,34	0,85	0,30
tilt	-12,64	2,10	-12,95	2,20	-12,54	2,48	-11,55	0,94	-13,75	0,88	-12,57	2,25	-12,98	2,06
slope	-22,46	4,72	-22,52	6,21	-22,43	6,17	-24,52	4,48	-20,27	5,01	-22,30	5,85	-22,76	6,79
hnr	15,97	3,45	16,26	3,56	16,43	3,23	17,99	2,13	14,86	2,66	17,39	2,97	14,51	3,38
ABI														
abi	2,22	1,50	2,33	1,48	2,40	1,16	3,02	1,02	1,65	1,08	2,63	1,46	1,98	1,46
shloc	7,96	2,78	8,09	3,74	7,87	3,07	5,80	2,25	9,54	1,46	7,13	3,71	8,34	2,50
shlocdb	0,78	0,28	0,79	0,32	0,77	0,31	0,61	0,13	0,94	0,13	0,72	0,33	0,84	0,27
jitterlocal	1,93	0,51	1,89	0,45	1,86	0,54	1,90	0,58	1,81	0,40	1,71	0,42	2,16	0,53
psd	0,001	0,0003	0,0009	0,0002	0,0009	0,0003	0,0009	0,00025	0,0008	0,0002	0,0008	0,00025	0,001	0,0001
gnemax	0,88	0,06	0,88	0,05	0,89	0,06	0,89	0,05	0,86	0,04	0,87	0,05	0,89	0,06
hfno6000	2,99	1,87	2,93	1,97	3,04	1,87	2,16	0,22	4,13	0,44	2,92	2,01	2,94	1,89
hnrd	25,07	7,45	25,10	7,70	24,77	6,86	24,89	7,74	25,18	7,63	26,77	1,96	19,13	1,60
h1h2	1,24	5,80	1,36	6,95	1,14	6,10	3,34	4,84	-2,03	11,00	3,36	5,22	-3,57	11,34
AVQI & ABI														
cpps	12,33	1,14	12,35	1,19	12,55	1,70	12,81	1,63	11,90	0,91	12,38	1,22	12,19	1,22
Pitch														
meanpitch_cs	205	107	200	101	194	104	199	101	200	102	215	13	113	9
sdpitch_cs	31	22	29	22	29	23	30	23	29	20	38	11	14	11
pitchvariab_cs	0,15	0,08	0,17	0,06	0,16	0,07	0,16	0,06	0,17	0,06	0,18	0,05	0,12	0,07
meanpitch_sv	167	88	164	85	163	90	166	87	156	85	191	38	103	8
sdpitch_sv	1,39	0,91	1,19	1,01	1,20	0,83	1,19	0,97	1,22	1,06	1,60	1,15	1,00	0,42
pitchvariab_sv	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00

TABLE A.2.
Medians and Interquartile Ranges (IQR) for Each Acoustic Measure, for Each Repetition x Recording Type x Gender Combination

Measure	Repetition 1								Repetition 2								Repetition 3							
	AKG				Telephone				AKG				Telephone				AKG				Telephone			
	Women (N=14)		Men (N=10)		Women (N=14)		Men (N=10)		Women (N=14)		Men (N=10)		Women (N=14)		Men (N=10)		Women (N=14)		Men (N=10)		Women (N=14)		Men (N=10)	
	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR	M	IQR
AVQI																								
avqi	2,41	1,39	2,25	0,41	2,90	0,63	3,07	0,57	2,19	0,97	2,52	0,75	2,77	0,50	3,18	0,50	2,14	1,25	2,23	0,52	2,66	1,07	2,98	0,54
shim	6,14	1,72	7,34	1,50	8,88	1,51	10,31	1,36	5,35	2,09	7,23	1,76	9,21	1,69	10,18	1,28	6,01	0,83	7,01	2,56	9,04	2,46	9,20	2,64
shdb	0,64	0,16	0,72	0,13	0,91	0,14	1,01	0,16	0,56	0,13	0,70	0,20	0,91	0,13	1,00	0,12	0,59	0,09	0,71	0,17	0,93	0,13	0,95	0,18
tilt	-11,16	1,30	-11,80	0,59	-13,75	0,54	-13,70	0,82	-11,38	0,79	-11,74	1,03	-13,63	0,83	-13,80	0,70	-11,30	0,90	-11,97	1,14	-13,93	0,51	-13,74	0,66
slope	-23,66	4,45	-24,56	4,94	-20,01	6,42	-19,76	4,47	-24,52	4,49	-24,61	5,57	-20,52	4,84	-19,42	5,55	-24,52	3,91	-24,38	5,26	-20,10	5,54	-18,72	4,49
hnr	18,48	3,08	15,97	2,63	15,70	1,63	13,07	1,58	18,69	2,26	16,58	2,67	15,72	1,71	13,20	2,18	19,24	2,98	16,45	2,59	15,80	2,08	13,78	2,39
ABI																								
abi	3,41	0,82	2,36	0,41	1,79	0,54	1,11	0,99	3,44	0,57	2,47	0,56	1,98	0,67	1,01	0,66	3,31	0,83	2,33	0,58	1,91	0,91	0,85	0,80
shloc	5,93	1,86	7,31	1,62	8,95	1,38	10,35	1,49	5,37	1,05	7,51	2,10	9,07	1,24	10,01	1,36	5,95	1,15	7,13	2,31	8,75	1,85	9,24	2,34
shlocdb	0,62	0,13	0,67	0,15	0,88	0,06	0,97	0,13	0,58	0,11	0,71	0,16	0,91	0,14	0,98	0,14	0,59	0,05	0,73	0,13	0,91	0,13	0,92	0,14
jitterlocal	1,79	0,49	2,42	0,55	1,72	0,28	2,01	0,40	1,70	0,39	2,34	0,44	1,74	0,45	2,06	0,45	1,63	0,43	2,27	0,30	1,80	0,62	1,94	0,37
psd	0,0008	0,0004	0,001	0	0,0007	0,0004	0,001	0,001	0,00085	0,0003	0,001	0,0001	0,0008	0,0002	0,001	0,0002	0,0008	0,0003	0,001	0,0002	0,00075	0,0003	0,0015	0,001
gnemax	0,88	0,05	0,91	0,05	0,88	0,04	0,88	0,05	0,88	0,04	0,91	0,08	0,85	0,07	0,88	0,03	0,89	0,05	0,91	0,05	0,84	0,09	0,88	0,05
hfno6000	2,13	0,16	2,29	0,17	4,17	0,56	3,86	0,60	2,11	0,19	2,28	0,23	4,12	0,46	4,17	0,47	2,14	0,16	2,31	0,16	4,05	0,53	4,10	0,87
hnrd	26,36	2,16	19,51	1,21	26,96	2,21	18,65	3,03	26,83	2,22	18,93	1,36	26,61	1,86	19,17	1,77	25,63	3,01	19,47	2,87	26,47	1,87	19,03	2,17
h1h2	4,34	4,64	1,57	2,93	1,28	4,38	-9,84	5,55	4,89	4,23	1,87	3,06	1,33	5,10	-9,47	3,04	4,79	4,44	1,70	3,53	1,14	4,48	-8,59	3,42
AVQI & ABI																								
cpps	12,71	1,50	12,76	1,22	11,99	0,77	11,95	0,97	12,88	1,95	12,59	0,84	12,22	0,88	11,75	0,62	13,37	1,94	13,34	1,27	12,21	1,26	12,14	0,84
Pitch																								
meanpitch_cs	220	12	112	14	221	20	113	18	215	10	113	7	215	14	112	11	213	19	113	8	215	23	113	13
sdpitch_cs	37	15	13	7	37	14	18	9	38	6	14	10	38	18	15	11	37	13	13	12	38	10	13	10
pitchvariab_cs	0,18	0,07	0,11	0,03	0,18	0,07	0,15	0,09	0,18	0,03	0,12	0,07	0,18	0,08	0,14	0,07	0,17	0,05	0,11	0,07	0,18	0,05	0,11	0,06
meanpitch_sv	194	32	102	10	194	32	102	10	191	36	103	8	190	40	103	8	193	34	103	12	193	34	103	12
sdpitch_sv	1,77	0,78	0,99	0,46	1,82	0,75	1,00	0,49	1,50	1,24	1,00	0,39	1,73	1,16	0,98	0,45	1,41	0,97	0,86	0,73	1,41	1,09	0,86	0,70
pitchvariab_sv	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00	0,01	0,00

REFERENCES

1. International Labour Organization. ILO Monitor: COVID-19 and the world of work. Seventh edition. Updated estimates and analysis. International Labour Organization. Available at: https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/documents/briefingnote/wcms_740877.pdf. Accessed March 25, 2022.
2. Institut national de la statistique et des études économiques. Emploi salarié - deuxième trimestre 2020. Available at: https://www.insee.fr/fr/statistiques/4653055. Accessed December 15, 2021.
3. Lo Giudice C. Une IA peut-elle enrichir le processus de sélection ? *HRSquare*. 2020;36:49.
4. Hiemstra AMF, Derous E. Video résumés portrayed : findings and challenges. In: Nikolaou I, Oostrom J, eds. *Employee Recruitment, Selection, and Assessment: Contemporary Issues for Theory and Practice*. Sussex, UK: Routledge/Taylor & Francis Group; 2015:45–60.
5. Oostrom JK, Van Der Linden D, Born MP, et al. New technology in personnel selection: how recruiter characteristics affect the adoption of new selection technology. *Comput Human Behav*. 2013;29:2404–2415. https://doi.org/10.1016/j.chb.2013.05.025.
6. Woods SA, Ahmed S, Nikolaou I, et al. Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges. *Eur J Work Organ Psychol*. 2020;29:64–77. https://doi.org/10.1080/1359432X.2019.1681401.
7. Baker M. Gartner HR survey shows 86% of organizations are conducting virtual interviews to hire candidates during Coronavirus pandemic. Available at: https://www.gartner.com/en/newsroom/press-releases/2020-04-30-gartner-hr-survey-shows-86−of-organizations-arecond. Accessed January 18, 2022.
8. Walters People. Video interviews spike by 67% − according to recruitment firm. Available at: https://www.walterspeople.co.uk/news/video-interviews-spike.html. Accessed January 11, 2022.
9. Waung M, Hymes RW, Beatty JE. The effects of video and paper resumes on assessments of personality, applied social skills, mental capability, and resume outcomes. *Basic Appl Soc Psych*. 2014;36:238–251. https://doi.org/10.1080/01973533.2014.894477.
10. Basch JM, Melchers KG. The use of technology-mediated interviews and their perception from the organization's point of view. *Int J Sel Assess*. 2021;29:495–502. https://doi.org/10.1111/ijsa.12339.
11. Tylečková L, Prokopová Z, Skarnitzl R. The effect of voice quality on hiring decisions. *AUC Philol*. 2017;2017:109–120. https://doi.org/10.14712/24646830.2017.37.
12. Straus SG, Miles JA, Levesque LL. The effects of videoconference, telephone, and face-to-face media on interviewer and applicant judgments in employment interviews. *J Manage*. 2001;27:363–381. https://doi.org/10.1177/014920630102700308.
13. Katopol P. The halo effect and bounded rationality: limits on decision-making. *Libr Leadersh Manag*. 2018;32:1–5. https://doi.org/10.5860/llm.v32i3.7312.
14. Lievens F. *Handboek Human Resource Management: Back to Basic*. Den Haag: Lannoo Campus; 2011.
15. DeGroot T, Kluemper D. Evidence of predictive and incremental validity of personality factors, vocal attractiveness and the situational interview. *Int J Sel Assess*. 2007;15:30–39. https://doi.org/10.1111/j.1468-2389.2007.00365.x.
16. Isetti DD, Baylor CR, Burns MI, et al. Employer reactions to adductor spasmodic dysphonia: exploring the influence of symptom severity and disclosure of diagnosis during a simulated telephone interview. *Am J Speech-Language Pathol*. 2017;26:469–482. https://doi.org/10.1044/2016_AJSLP-16-0040.
17. Verduyckt I, Morsomme D. Vocal beauty: a mediating variable in the negative stereotyping of dysphonic speakers. *Logop Phoniatr Vocology*. 2020;45:164−171. https://doi.org/10.1080/14015439.2019.1694697.
18. Blood GW, Mahan BW, Hyman M. Judging personality and appearance from voice disorders. *J Commun Disord*. 1979;12:63–67. https://doi.org/10.1016/0021-9924(79)90022-4.
19. Isetti D, Xuereb L, Eadie TL. Inferring speaker attributes in adductor spasmodic dysphonia: ratings from unfamiliar listeners. *Am J Speech-Language Pathol*. 2014;23:134–145. https://doi.org/10.1044/2013_AJSLP-13-0010.
20. Nagle KF, Eadie TL, Yorkston KM. Everyday listeners' impressions of speech produced by individuals with adductor spasmodic dysphonia. *J Commun Disord*. 2015;58:1–13. https://doi.org/10.1016/j.jcomdis.2015.07.001.
21. Mahrholz G, Belin P, McAleer P. Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS One*. 2018;13: e0204991. https://doi.org/10.1371/journal.pone.0204991.
22. McAleer P, Todorov A, Belin P. How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One*. 2014;9:e90779. https://doi.org/10.1371/journal.pone.0090779.
23. Anderson RC, Klofstad CA, Mayew WJ, et al. Vocal fry may undermine the success of young women in the labor market. *PLoS One*. 2014;9:1–8. https://doi.org/10.1371/journal.pone.0097506.
24. Pisanski K, Sorokowski P. Human stress detection: cortisol levels in stressed speakers predict voice-based judgments of stress. *Perception*. 2021;50:80–87. https://doi.org/10.1177/0301006620978378.
25. Oleszkiewicz A, Pisanski K, Lachowicz-Tabaczek K, et al. Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychon Bull Rev*. 2017;24:856–862. https://doi.org/10.3758/s13423-016-1146-y.
26. Van Zant AB, Berger J. How the voice persuades. *J Pers Soc Psychol*. 2020;118:661–682. https://doi.org/10.1037/pspi0000193.
27. Hodges-Simeon CR, Gaulin SJC, Puts DA. Different vocal parameters predict perceptions of dominance and attractiveness. *Hum Nat*. 2010;21:406–427. https://doi.org/10.1007/s12110-010-9101-5.
28. Bruckert L, Bestelmeyer P, Latinus M, et al. Vocal attractiveness increases by averaging. *Curr Biol*. 2010;20:116–120. https://doi.org/10.1016/j.cub.2009.11.034.
29. Naim I, Tanveer MI, Gildea D, et al. Automated prediction and analysis of job interview performance: the role of what you say and how you say it. *2015 11th IEEE Int Conf Work Autom Face Gesture Recognition, FG 2015*. 2015. https://doi.org/10.1109/FG.2015.7163127.
30. Hemamou L, Felhi G, Vandenbussche V, et al. HireNet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews. *arXiv*. 2019. https://doi.org/10.1609/aaai.v33i01.3301573.
31. Lee T, Ziegler M. Forewarned is forearmed: using AI to detect SDR tendencies as personnel selection tools. 2021. https://doi.org/10.31234/osf.io/7htfs.
32. Babel M, McGuire G, King J. Towards a more nuanced view of vocal attractiveness. *PLoS One*. 2014;9:1–10. https://doi.org/10.1371/journal.pone.0088616.
33. Imhof M. Listening to voices and judging people. *Int J List*. 2010;24:19–33. https://doi.org/10.1080/10904010903466295.
34. Claeys A-S, Cauberghe V. Keeping control: the importance of nonverbal expressions of power by organizational spokespersons in times of crisis. *J Commun*. 2014;64:1160–1180. https://doi.org/10.1111/jcom.12122.
35. Passetti RR, Constantini AC. The effect of telephone transmission on voice quality perception. *J Voice*. 2019;33:649–658. https://doi.org/10.1016/j.jvoice.2018.04.018.
36. Chhetri DK, Merati AL, Blumin JH, et al. Reliability of the perceptual evaluation of adductor spasmodic dysphonia. *Ann Otol Rhinol Laryngol*. 2008;117:159–165. https://doi.org/10.1177/000348940811700301.
37. Nguyen DD, McCabe P, Thomas D, et al. Acoustic voice characteristics with and without wearing a facemask. *Sci Rep*. 2021;11:1–11. https://doi.org/10.1038/s41598-021-85130-8.
38. Vogel AP, Rosen KM, Morgan AT, et al. Comparability of modern recording devices for speech analysis: smartphone, landline, laptop, and hard disc recorder. *Folia Phoniatr Logop*. 2014;66:244–250. https://doi.org/10.1159/000368227.
39. Uloza V, Padervinskis E, Vegiene A, et al. Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *Eur Arch Oto-Rhino-Laryngology*. 2015;272:3391–3399. https://doi.org/10.1007/s00405-015-3708-4.

40. Manfredi C, Lebacq J, Cantarella G, et al. Smartphones offer new opportunities in clinical voice research. *J Voice*. 2017;31:111.e1–111.e7. https://doi.org/10.1016/j.jvoice.2015.12.020.
41. Lin E, Hornibrook J, Ormond T. Evaluating iPhone recordings for acoustic voice assessment. *Folia Phoniatr Logop*. 2012;64:122–130. https://doi.org/10.1159/000335874.
42. Lebacq J, Schoentgen J, Cantarella G, et al. Maximal ambient noise levels and type of voice material required for valid use of smartphones in clinical voice research. *J Voice*. 2017;31:550–556. https://doi.org/10.1016/j.jvoice.2017.02.017.
43. Kojima T, Fujimura S, Hori R, et al. An innovative voice analyzer "VA" smart phone program for quantitative analysis of voice quality. *J Voice*. 2019;33:642–648. https://doi.org/10.1016/j.jvoice.2018.01.026.
44. Kojima T, Hasebe K, Fujimura S, et al. A new iPhone application for voice quality assessment based on the GRBAS scale. *Laryngoscope*. 2021;131:580–582. https://doi.org/10.1002/lary.28796.
45. Munnings AJ. The current state and future possibilities of mobile phone "voice analyser" applications, in relation to otorhinolaryngology. *J Voice*. 2020;34:527–532. https://doi.org/10.1016/j.jvoice.2018.12.018.
46. Maryn Y, Ysenbaert F, Zarowski A, et al. Mobile communication devices, ambient noise, and acoustic voice measures. *J Voice*. 2017;31:248.e11–248.e23. https://doi.org/10.1016/j.jvoice.2016.07.023.
47. Marsano-Cornejo M-J, Roco-Videla Á. Comparison of the acoustic parameters obtained with different smartphones and a professional microphone. *Acta Otorrinolaringol (English Ed*. 2022;73:51–55. https://doi.org/10.1016/j.otoeng.2020.08.009.
48. Wormald RN, Moran RJ, Reilly RB, et al. Performance of an automated, remote system to detect vocal fold paralysis. *Ann Otol Rhinol Laryngol*. 2008;117:834–838. https://doi.org/10.1177/000348940811701107.
49. Moran RJ, Reilly RB, De Chazal P, et al. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Trans Biomed Eng*. 2006;53:468–477. https://doi.org/10.1109/TBME.2005.869776.
50. Tsanas A, Little MA, Ramig LO. Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network. *IEEE Access*. 2021;9:11024–11036. https://doi.org/10.1109/ACCESS.2021.3050524.
51. Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *J Acoust Soc Am*. 2019;145:2871–2884. https://doi.org/10.1121/1.5100272.
52. Cannizzaro MS, Reilly N, Mundt JC, et al. Remote capture of human voice acoustical data by telephone: a methods study. *Clin Linguist Phonetics*. 19:649−658.
53. Mundt JC, Snyder PJ, Cannizzaro MS, et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics*. 2007;20:50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001.
54. Zawawi SA, Hamzah AA, Majlis BY, et al. A review of MEMS capacitive microphones. *Micromachines*. 2020;11:1–26. https://doi.org/10.3390/MI11050484.
55. Kent R, Read C. *Acoustic Analysis of Speech*. 2nd ed. Canada: Singular; 2002.
56. Johnson DM, Hapner ER, Klein AM, et al. Validation of a telephone screening tool for spasmodic dysphonia and vocal fold tremor. *J Voice*. 2014;28:711–715. https://doi.org/10.1016/j.jvoice.2014.03.009.
57. Harmegnies B, Landercy A. Intra-speaker variability of the long term speech spectrum. *Speech Commun*. 1988;7:81–86. https://doi.org/10.1016/0167-6393(88)90023-4.
58. Pommée T, Maryn Y, Finck C, et al. Validation of the acoustic voice quality index, version 03.01, in French. *J Voice*. 2018. https://doi.org/10.1016/j.jvoice.2018.12.008.
59. Chial MR. Suggestions for computer-based audio recording of speech samples for perceptual and acoustic analyses. *(Tech. Rep. No. 13). Phonology Project, Waisman Center, University of Wisconsin-Madison*. 2003.
60. Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *J Voice*. 2010;24:540–555. https://doi.org/10.1016/j.jvoice.2008.12.014.
61. Barsties B, Maryn Y. External validation of the acoustic voice quality index Version 03.01 with extended representativity. *Ann Otol Rhinol Laryngol*. 2016;125:571–583. https://doi.org/10.1177/0003489416636131.
62. Barsties v. Latoszek B, Maryn Y, Gerrits E, et al. The acoustic breathiness index (ABI): a multivariate acoustic model for breathiness. *J Voice*. 2017;31:511.e11–511.e27. https://doi.org/10.1016/j.jvoice.2016.11.017.
63. Barsties v, Latoszek B, Kim GH, et al. The validity of the acoustic breathiness index in the evaluation of breathy voice quality: a meta-analysis. *Clin Otolaryngol*. 2021;46:31–40. https://doi.org/10.1111/coa.13629.
64. Hirano M. *Clinical Examination of Voice*. New York, NY: Springer Verlag; 1981.
65. Lehnert B, Herold J, Blaurock M, et al. Reliability of the acoustic voice quality index AVQI and the acoustic breathiness index (ABI) when wearing CoViD-19 protective masks. *Eur Arch Oto-Rhino-Laryngology*. 2022;279:4617–4621. https://doi.org/10.1007/s00405-022-07417-4.
66. Uloza V, Petrauskas T, Padervinskis E, et al. Validation of the acoustic voice quality index in the lithuanian language. *J Voice*. 2017;31:257.e1–257.e11. https://doi.org/10.1016/j.jvoice.2016.06.002.
67. Kankare E, Barsties V, Latoszek B, et al. The acoustic voice quality index version 02.02 in the Finnish-speaking population. *Logop Phoniatr Vocology*. 2020;45:49–56. https://doi.org/10.1080/14015439.2018.1556332.
68. Hosokawa K, von Latoszek BB, Ferrer-Riesgo CA, et al. Acoustic breathiness index for the Japanese-speaking population: validation study and exploration of affecting factors. *J Speech, Lang Hear Res*. 2019;62:2617–2631. https://doi.org/10.1044/2019_JSLHR-S-19-0077.
69. Barsties v. Latoszek B, Lehnert B, Janotte B. Validation of the acoustic voice quality index version 03.01 and acoustic breathiness index in German. *J Voice*. 2020;34:157.e17–157.e25. https://doi.org/10.1016/j.jvoice.2018.07.026.
70. Delgado Hernández J, León Gómez NM, Jiménez A, et al. Validation of the acoustic voice quality index version 03.01 and the acoustic breathiness index in the Spanish language. *Ann Otol Rhinol Laryngol*. 2018;127:317–326. https://doi.org/10.1177/0003489418761096.
71. Kim G-H, Lee Y-W, Bae I-H, et al. Validation of the acoustic voice quality index in the Korean language. *J Voice*. 2019;33:948.e1–948.e9. https://doi.org/10.1016/j.jvoice.2018.06.007.
72. Englert M, Latoszek BB v, Behlau M. Exploring the validity of acoustic measurements and other voice assessments. *J Voice*. 2022. https://doi.org/10.1016/j.jvoice.2021.12.014.
73. Ben BB, Maryn Y, Gerrits E, et al. A meta-analysis: acoustic measurement of roughness and breathiness. *J Speech, Lang Hear Res*. 2018;61:298–323. https://doi.org/10.1044/2017_JSLHR-S-16-0188.
74. Laukkanen A-M, Rantala L. Does the acoustic voice quality index (AVQI) correlate with perceived creak and strain in normophonic young adult finnish females? *Folia Phoniatr Logop*. 2022;74:62–69. https://doi.org/10.1159/000514796.
75. Batthyany C, Maryn Y, Trauwaen I, et al. A case of specificity: how does the acoustic voice quality index perform in normophonic subjects? *Appl Sci*. 2019;9:2527. https://doi.org/10.3390/app9122527.
76. Faham M, Laukkanen A-M, Ikävalko T, et al. Acoustic voice quality index as a potential tool for voice screening. *J Voice*. 2021;35:226–232. https://doi.org/10.1016/j.jvoice.2019.08.017.
77. Li G, Hou Q, Zhang C, et al. Acoustic parameters for the evaluation of voice quality in patients with voice disorders. *Ann Palliat Med*. 2021;10:130–136. https://doi.org/10.21037/apm-20-2102.
78. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *Eur Arch Oto-Rhino-Laryngology*. 2001;258:77–82. https://doi.org/10.1007/s004050000299.

79. Dejonckere PH, Obbens C, de Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr Logop*. 1993;45:76–83. https://doi.org/10.1159/000266220.
80. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80. https://doi.org/10.1016/S0892-1997(97)80026-4.
81. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Oto-Rhino-Laryngology*. 2004;261:429–434. https://doi.org/10.1007/s00405-003-0707-7.
82. Delvaux V, Pillot-Loiseau C. Perceptual judgment of voice quality in nondysphonic French speakers: effect of task-, speaker- and listener-related variables. *J Voice*. 2020;34:682–693. https://doi.org/10.1016/j.jvoice.2019.02.013.
83. Mayer J. Praat Skripte : auditive Stimmanalyse (GRBAS) mit dem Demo- Window. Available at: http://praatpfanne.lingphon.net/downloads/demo_GRBAS.txt. Accessed January 14, 2022.
84. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
85. Prion S, Haerling KA. Making sense of methods and measurement: spearman-rho ranked-order correlation coefficient. *Clin Simul Nurs*. 2014;10:535–536. https://doi.org/10.1016/j.ecns.2014.07.005.
86. Belgian National Institute for Health and Disability Insurance. Code éthique et déontologique des logopèdes. Available at: https://www.inami.fgov.be/fr/professionnels/sante/logopedes/Pages/logopedes-code-ethique-deontologique.aspx. Accessed December 11, 2021.
87. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40. https://doi.org/10.1044/jshr.3601.21.
88. Pommée T, Balaguer M, Pinquier J, et al. Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech: a systematic review. *Speech, Lang Hear*. 2021;24:105–132. https://doi.org/10.1080/2050571X.2021.1913300.
89. Brockmann M, Drinnan MJ, Storck C, et al. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice*. 2011;25:44–53. https://doi.org/10.1016/j.jvoice.2009.07.002.
90. Brockmann M, Storck C, Carding PN, et al. Voice loudness and gender effects on jitter and shimmer in healthy adults. *J Speech, Lang Hear Res*. 2008;51:1152–1160. https://doi.org/10.1044/1092-4388(2008/06-0208.
91. Van Borsel J, Janssens J, De Bodt M. Breathiness as a feminine voice characteristic: a perceptual approach. *J Voice*. 2009;23:291–294. https://doi.org/10.1016/j.jvoice.2007.08.002.
92. Hejná M, Šturm P, Tylečková L, et al. Normophonic breathiness in Czech and Danish: are females breathier than males? *J Voice*. 2021;35:498.e1–498.e22. https://doi.org/10.1016/j.jvoice.2019.10.019.
93. Paz KE da S, de Almeida AAF, Almeida LNA, et al. Auditory perception of roughness and breathiness by dysphonic women. *J Voice*. 2022. https://doi.org/10.1016/j.jvoice.2022.01.005.
94. Moyse E, Beaufort A, Brédart S. Evidence for an own-age bias in age estimation from voices in older persons. *Eur J Ageing*. 2014;11:241–247. https://doi.org/10.1007/s10433-014-0305-0.
95. Barsties B, Maryn Y. Test-Retest-Variabilität und interne Konsistenz des acoustic voice quality index. *HNO*. 2013;61:399–403. https://doi.org/10.1007/s00106-012-2649-0.
96. Barsties v. Latoszek B, Ulozaitė-Stanienė N, Maryn Y, et al. The influence of gender and age on the acoustic voice quality index and Dysphonia severity index: a normative study. *J Voice*. 2019;33:340–345. https://doi.org/10.1016/j.jvoice.2017.11.011.
97. Jayakumar T, Benoy JJ, Yasin HM. Effect of age and gender on acoustic voice quality index across lifespan: a cross-sectional study in Indian population. *J Voice*. 2022;36:436.e1–436.e8. https://doi.org/10.1016/j.jvoice.2020.05.025.
98. Henton CG, Bladon RAW. Breathiness in normal female speech: inefficiency versus desirability. *Lang Commun*. 1985;5:221–227. https://doi.org/10.1016/0271-5309(85)90012-6.
99. Södersten M, Lindestad P-Å. Glottal closure and perceived breathiness during phonation in normally speaking subjects. *J Speech, Lang Hear Res*. 1990;33:601–611. https://doi.org/10.1044/jshr.3303.601.
100. Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am*. 1990;87:820–857. https://doi.org/10.1121/1.398894.
101. Hanson HM, Stevens KN, Kuo H-KJ, et al. Towards models of phonation. *J Phon*. 2001;29:451–480. https://doi.org/10.1006/jpho.2001.0146.
102. Simpson A. Breathiness differences in male and female speech. Is H1-H2 an appropriate measure? In: Proceedings Fonetik. *Stockholm University*. 2009. doi: 10.1.1.623.3294.
103. Pépiot E. Voice, speech and gender. *Corela*. 2015:(HS-16):0–13. https://doi.org/10.4000/corela.3783.
104. van der Woerd B. VOice analysis with Iphones: a low Cost Experimental Solution. 2019. Electronic Thesis and Dissertation Repository. 6719. https://ir.lib.uwo.ca/etd/6719.
105. Heman-Ackah YD, Michael DD, Baroody MM, et al. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol*. 2003;112:324–333. https://doi.org/10.1177/000348940311200406.
106. Wolfe V, Cornell R, Fitch J. Sentence/vowel correlation in the evaluation of dysphonia. *J Voice*. 1995;9:297–303. https://doi.org/10.1016/S0892-1997(05)80237-1.
107. Maryn Y, Roy N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. *J Soc Bras Fonoaudiol*. 2012;24:107–112. https://doi.org/10.1590/S2179-64912012000200003.