Journal of Business Research 121 (2020) 364–374

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications

Christian Hildebrand [a,*], Fotis Efthymiou [b], Francesc Busquet [b], William H. Hampton [c], Donna L. Hoffman [d], Thomas P. Novak [e]

[a] Professor of Marketing Analytics & Director of the TechX Lab at the University of St. Gallen, Switzerland
[b] Doctoral candidates in marketing at the University of St., Gallen, Switzerland
[c] Post-Doctoral Research Fellow & Co-Director of the TechX Lab at the University of St. Gallen, Switzerland
[d] Louis Rosenfeld Distinguished Scholar and Professor of Marketing at The George Washington School of Business in Washington, D.C., USA
[e] Denit Trust Distinguished Scholar and Professor of Marketing at the George Washington School of Business in Washington, D.C., USA

ARTICLE INFO

Keywords:
Voice Analytics
Natural language processing
Voice-controlled interfaces
Emotion detection
Acoustic markers of emotion

ABSTRACT

Recent advances in artificial intelligence and natural language processing are gradually transforming how humans search, shop, and express their preferences. Leveraging the new affordances and modalities of human–machine interaction through voice-controlled interfaces will require a nuanced understanding of the physics and psychology of speech formation as well as the systematic extraction and analysis of vocal features from the human voice. In this paper, we first develop a conceptual framework linking vocal features in the human voice to experiential outcomes and emotional states. We then illustrate the effective processing, editing, analysis, and visualization of voice data based on an Amazon Alexa user interaction, utilizing state-of-the-art signal-processing packages in R. Finally, we offer novel insight into the ways in which business research might employ voice and sound analytics moving forward, including a discussion of the ethical implications of building multi-modal databases for business and society.

1. Introduction

The use of natural language in voice-controlled interfaces is gradually transforming how humans interact with technology (Dale, 2016; Hirschberg & Manning, 2015). Voice-controlled interfaces such as Amazon Alexa, Google Assistant, or Siri entail a new kind of interaction between humans and machines that could prove revolutionary, with some declaring voice-controlled interfaces to be the new "operating system in commerce" (Netzer, Feldman, Goldenberg, & Fresko, 2011; Suri, Elia, & van Hillegersberg, 2017). With over 100 million smart speakers already in homes worldwide and sales of voice assistant technologies predicted to increase to $31.82 billion by 2025, voice-controlled interfaces are poised to transform how humans search, shop, and express their preferences.

Despite their enormous potential, leveraging the modalities and affordances of voice-enabled technology will require both a nuanced understanding of the physics and psychology of speech formation as well as understanding how to extract and analyze vocal features from human voice data. Specifically, the human voice can be quantitatively parsed into a number of vocal features including pitch, loudness, and the presence and duration of speech pauses. Analysis of these features can reveal both state- and trait-level information about a speaker. For example, a user who is upset with Amazon's Alexa for repeatedly failing to understand a command might increasingly raise their voice as they become progressively more frustrated. A user engaged in stressful car navigation in a dense urban traffic situation might evince their state of anxiety and stress by speaking more quickly or in a higher-pitched voice. Conversely, an employee tired from a long, difficult day at work might ask their voice assistant in a monotonous, low variability vocal tone to play Mozart's Requiem. Each of these vignettes represents a facet of a broader psychophysiological phenomenon: the way in which we speak often accurately reveals our current emotional state (Jurafsky & Martin, 2014; Scherer, 1986, 2003).

Despite a corpus of research on the human voice and speech formation, research examining the impact of voice technology or voice-controlled interfaces on users is relatively scarce, and has

* Corresponding author.
E-mail addresses: christian.hildebrand@unisg.ch (C. Hildebrand), fotis.efthymiou@unisg.ch (F. Efthymiou), francesc.busquet@unisg.ch (F. Busquet), william.hampton@unisg.ch (W.H. Hampton), dlhoffman@gwu.edu (D.L. Hoffman), novak@gwu.edu (T.P. Novak).

https://doi.org/10.1016/j.jbusres.2020.09.020
Received 18 January 2020; Received in revised form 5 September 2020; Accepted 8 September 2020
Available online 29 September 2020

predominantly been concerned with optimizing the design features of voice-controlled interfaces (Santos, Ferreira, & Dias, 2015), examining the factors that relate to security issues of voice-controlled interfaces (Diao, Liu, Zhou, & Zhang, 2014), or general user acceptance (Portet, Vacher, Golanski, Roux, & Meillon, 2013). This paper takes a different route, providing a theory-driven model integrating vocal features and emotional states of a user and illustrating the practical use of and unique insight from voice analytics in business research.

The objectives of this paper are threefold. First, we provide a theory-driven conceptual framework building on prior work in phonetics and psychology, linking the vocal features in the human voice to the emotional states and traits of a speaker. Second, we practically illustrate how to operationalize the effective processing, analysis, and visualization of voice data using state-of-the-art signal processing packages in R. Third, we provide new directions on how to effectively employ voice and sound analytics in business research with an emphasis on voice-controlled interfaces, while highlighting the ethical implications of building multi-modal databases.

In what follows, we first provide a review of prior work on speech formation, integrating research from phonetics, acoustics, and psychology. We then develop a conceptual framework linking human vocal features to the trait and state level characteristics they reveal. Next, we illustrate the unique insight that voice analytics can provide based on an actual user interaction with Amazon Alexa, utilizing the dominant signal processing packages for practical voice and sound analytics in R. Finally, we discuss the theoretical, managerial, and policy implications of voice analytics for consumer welfare, business, and society at large.

2. Theoretical background

2.1. Physiological foundations, soundwaves, & psychoacoustics

Drawing inferences from the human voice entails the integration of concepts from phonetics (i.e., the physiological foundations of speech), acoustics (i.e., the physics of a soundwave), and cognitive psychology (i.e., the perception and recognition of speech signals). Fig. 1 illustrates the interplay of these domains during speech formation, transmission, and speech perception.

Imagine uttering the salutation “Hello” to a close friend. This vocalization begins when the speech production area of the brain signals the lungs to release a stream of air that passes through your trachea and larynx, causing your vocal folds to vibrate. Each time the vocal folds open, air is allowed to continue its ascent, ultimately passing through the oral and nasal cavity to produce an audible word such as “Hello” (Giegerich, 1992). The specific way that *you* say “Hello” depends upon the shape and length of your anatomical articulators such as your teeth, tongue, and the size of your oral and nasal cavity (Zhang, 2016). Together, these components determine to what extent air flow is compressed and oscillates in different frequencies and amplitudes that are captured in a soundwave and determine the unique sound of *your* “Hello”.

Although each person has a unique vocal signature, differences can also be identified at higher levels based on the average physical characteristics of a given cohort. For example, relative to men, women generally speak with a higher frequency, i.e. “pitch”, due to the shorter membranous length of their vocal folds (Latinus & Taylor, 2012; Titze, 1989). Similarly, people with larger body sizes (more in terms of body height than in terms of body weight) often have longer vocal tracts that result in lower and more closely spaced formant frequencies (Fitch, 1997). These vocal differences have evolutionary implications, as lower-frequency voices are generally perceived as more attractive (Collins, 2000; Hodges-Simeon, Gaulin, & Puts, 2011), relate to greater mating success in mammals (Puts, 2005), signal social dominance (Cheng, Tracy, Ho, & Henrich, 2016), and can alter status perceptions (Klofstad, Anderson, & Peters, 2012).

Vocal expression is not only affected by our relatively stable anatomy, but also fluctuates according to our current emotional state. For instance, contemplation is associated with slower speech with longer pauses (Dasgupta, 2017), anger is often associated with louder speech (Clark, 2005; Juslin & Laukka, 2003), and fear with greater pitch variability (Clark, 2005; Juslin & Laukka, 2003).

In short, human speech formation involves the interplay of individual traits (anatomical and psychological) and emotional states that together determine how we express ourselves verbally, and subsequently how we are perceived by those listening. In what follows, we focus on making inferences based on the vocal features that can be extracted from human voice data (i.e., the analysis of a soundwave). We first provide a conceptual framework that organizes the extractable features in the human voice along four distinct dimensions, and then provide a practical illustration using state-of-the-art signal processing packages in R.

2.2. A four-dimensional framework of speech: Linking vocal features and emotional states

Table 1 displays our conceptual framework linking the vocal features in the human voice and the emotional states and traits of a user. As articulated in the preceding sections, the focus of this research is on a theory-driven, yet practical, analysis of human voice data. We assume the presence of a voice recording (such as a WAV or MP3 file), which generates an abstract representation of sound captured by a receiving device (i.e., a microphone) that converts soundwaves in the air into an

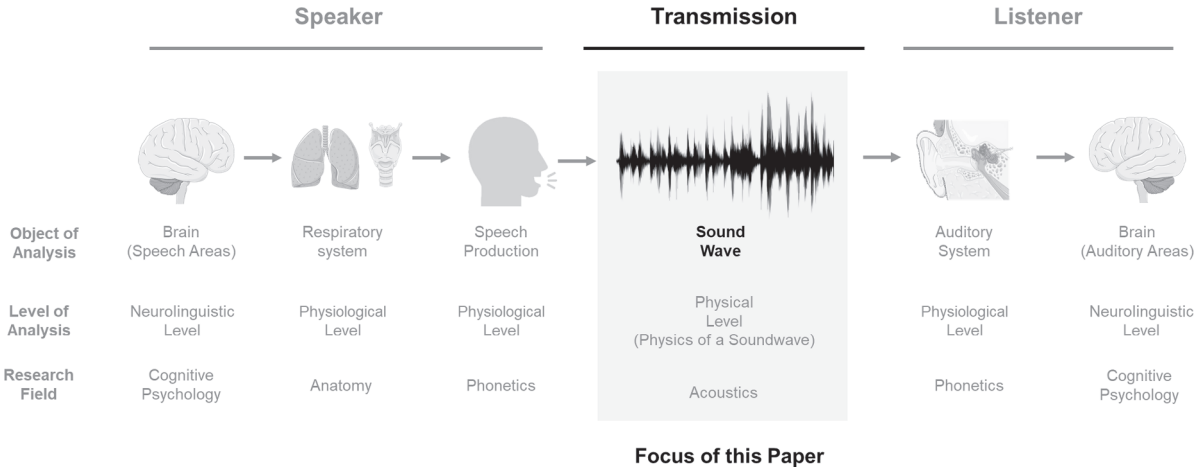

Fig. 1. From Speech Production and Soundwave Transmission to Perception.

Table 1
A Four-Dimensional Conceptual Framework of Speech: Linking Vocal Features to Speaker States and Traits.

SOUNDWAVE DOMAIN	PRIMARY VOCAL FEATURES (EXAMPLE METRIC)	LISTENER PERCEPTION	INFERRED STATES AND TRAITS BASED ON EXPRESSED SPEECH	SELECTEDRESEARCH
TIME				
	Duration (Milli-/Seconds)	Duration of an Utterance	Anger↑, Fear↑, Sorrow↑	Williams & Stevens, 1972
	Speech and Articulation Rate (Words per Second)	Velocity of Speech	Anger↑, Competence ↑, Contemplation↓, Dominance↓, Enthusiasm↑, Extraversion↑, Fear↑, Happiness↑, Persuasiveness↑, Sadness↓, Stress↑,Tenderness↓	Brenner, Doherty, & Shipp, 1994; Dasgupta, 2017; Juslin & Laukka, 2003; MacLachlan, 1982; Miller et al., 1976; Mohammadi & Vinciarelli, 2015; Scherer & Giles, 1979; Tusing & Dillard, 2000; Williams & Stevens, 1972
	Voice breaks (Percentage Unvoiced Frames)	Number and Duration of Pauses	Competence↓, Contemplation↑, Extraversion↓	Dasgupta, 2017; Mallory & Miller, 1958; Mohammadi & Vinciarelli, 2015; Scherer & Giles, 1979
AMPLITUDE				
	Intensity / Power (Sone)	Loudness of Speech	Aggression↑, Anger↑, Annoyance↑, Dominance↑, Extraversion↑, Fear↓, Happiness↑, Tenderness↓, Sadness↓, Shyness↓, Stress↑	Abelin & Allwood, 2000; Brenner et al., 1994; Johnstone & Scherer, 1999; Juslin & Laukka, 2003; Mallory & Miller, 1958; Scherer, 2003; Scherer & Giles, 1979; Tusing & Dillard, 2000
	Variability of Intensity / Power (Sone Variance)	Loudness Variability	Anger↑, Dominance↑, Fear↑, Happiness↑, Sadness↓, Tenderness↓	Juslin & Laukka, 2003; Tusing & Dillard, 2000
FREQUENCY				
	Fundamental Frequency (Hertz)	Pitch	Anger↑, Competence↓, Confidence↓, Empathy↓, Extraversion↑, Fear↑, Happiness↑, Nervousness↑, Persuasiveness↓, Sadness↓, Stress↑, Tenderness↓, Trustworthiness↓	Apple et al., 1979; Brenner et al., 1994; Guyer et al., 2019; Juslin & Laukka, 2003; Oleszkiewicz et al., 2017; Scherer & Giles, 1979; Williams & Stevens, 1972
	Variability of Fundamental Frequency (Hertz Variance)	Pitch Variability	Anger↑, Extraversion↑, Happiness↑, Sadness↓, Shyness↓, Sociability↑, Tenderness↓	Abelin & Allwood, 2000; Burgoon, Birk, & Pfau, 1990; Juslin & Laukka, 2003; Ray, 1986; Scherer & Giles, 1979
SPECTRAL				
	Vocal Shimmer (cycle to cycle deviation from mean amplitude)	Loudness Perturbations	Anger↑, Confidence↑, Joy↓, Stress↑,	Jacob, 2016; Jiang & Pell, 2017; Li et al., 2007
	Vocal Jitter (mean absolute difference between consecutive μs periods)	Pitch Perturbations	Anger↑, Annoyance↑, Happiness↑, Sadness↓, Stress↑	Johnstone & Scherer, 1999; Juslin & Laukka, 2003; Li et al., 2007
	HNR (additive noise in signal in dB)	Voice Quality	Confidence↑, Happiness↑, Interest↑, Lust↓, Pleasure↑	Jiang & Pell, 2017; Kamiloğlu et al., 2020
	Vocal Entropy (Shannon evenness of frequency spectrum)	Diversity of Vocal Transitions	Low mood↑	Yingthawornsuk, 2016

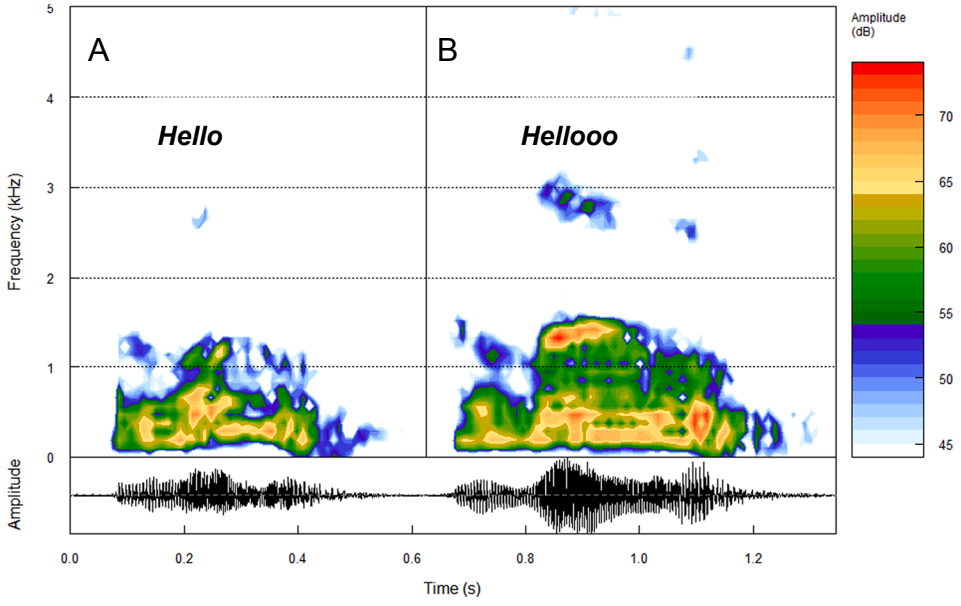

Fig. 2. Spectrogram of Two Exemplary Utterances of "Hello". *Note*: Panel A illustrates an utterance of saying "Hello" in a shorter and soft-spoken expression with a lower pitch (620 ms; 73.57 dB; 181 Hz) compared to the longer, louder, and higher-pitched expression of "Hello" in Panel B (720 ms; 76.64 dB; 204 Hz). Consistent with our conceptual framework in Table 1, the second "Hello" (Panel B) indicates arguably greater excitement compared to the first "Hello" (Panel A).

electric signal. Each and every soundwave can be completely described according to four distinct dimensions (Jurafsky & Martin, 2014; Sueur, 2018): a) time, b) amplitude, c) frequency, and d) spectrum, the joint distribution of frequency and amplitude, as indicated in the first column of Table 1. Consider the visual representations of the soundwaves for two different utterances of the word "Hello," as depicted in Fig. 2. Imagine again saying the word "Hello" in two different contexts. In the first, you are greeting a neighbor with a soft-spoken "Hello," and in the second, you are excitedly greeting a close friend whom you have not seen for a long time. Panel A illustrates the former and Panel B the latter.

The first dimension describing soundwaves is time, which captures the duration or length of a soundwave measured in seconds or milliseconds. The main vocal features of time used in prior research include duration, speech and articulation rate, and voice breaks, as shown in the first three rows of the second column of Table 1. As seen in Fig. 2, although the semantic content of speech is identical, the second soundwave (Panel B) is longer compared to the first (Panel A). Duration is important because it can reveal nuanced information about a speaker. There are three main reasons why two semantically identical waves might differ in duration: (1) duration of an utterance, (2) speech or articulation rate, and (3) length of voice breaks. As in the two distinct settings before, the speaker greeting a close friend expands the length of vowels *e* and *o* leading to generally longer duration, indicating greater excitement (Williams & Stevens, 1972). The second metric in the time domain is a speaker's speech rate often measured as the number of words spoken per minute. In conversational contexts, faster speech rate tends to increase the persuasiveness of the communicated message (Miller, Maruyama, Beaber, & Valone, 1976) and has been associated with positive speaker attributes such as competence (Ray, 1986), extraversion (Nass & Lee, 2001), enthusiasm, and overall "energy" (MacLachlan, 1982). Speech rate can also indicate a speaker's emotional state, with faster speech rate typically observed in high arousal states of both negative and positive valence (such as anger, fear and happiness), whereas slower speech rate is more often observed in low arousal states with predominantly negative valence (such as sadness) (Juslin & Laukka, 2003). In addition to the absolute length of utterances, further information can be gleaned from the third metric capturing the time domain, the number of voice breaks or unvoiced frames of a soundwave in a given time period. For example, speakers who exhibit more voice breaks during speech formation are generally perceived as being less extraverted (Mallory & Miller, 1958) and less competent (Mohammadi & Vinciarelli, 2015). The relationships between these vocal metrics, listeners' perceptions, and speakers' states and traits are listed in the second, third and fourth columns of Table 1.

The second dimension of our conceptual framework examines the amplitude of a soundwave, which is captured by the sound intensity level or sound pressure level, usually described as the amount of power per unit area (such as watt per square meter or centimeter) and commonly measured in decibels (dB). Put simply, greater sound intensity corresponds to a louder voice. This is illustrated by comparing the second soundwave (Panel B) in Fig. 2 to the first (Panel A), which is not only notably longer (in terms of duration in milliseconds) but also louder. Voices with greater sound intensity, relative to baseline, are typically perceived as being higher in dominance (Tusing & Dillard, 2000), aggression (Scherer, 2003), and extraversion (Abelin & Allwood, 2000; Scherer & Giles, 1979), as noted in Table 1. Conversely, lower sound intensity can indicate that a person is in a fearful, sad or tender state (Juslin & Laukka, 2003). Of course, human voice volume is not static, but rather tends to shift dynamically as we generate speech. Therefore, such variability in intensity serves as the second indicator of affective state, with higher variability associated with high arousal emotions such as anger and fear (Juslin & Laukka, 2003).

The third dimension captures the frequency of a soundwave measured in Hertz (Hz) and reflects the number of cycles of a soundwave per second of vibrating air particles. Human hearing typically ranges from 20 Hz to 20 kHz (kilohertz), corresponding to how "deep" (lower pitched) or "shrill" (higher pitched) a sound is perceived. As with other vocal characteristics, pitch also signals a variety of speaker characteristics, with lower pitch generally associated with higher perceived speaker competence, trustworthiness (Oleszkiewicz, Pisanski, Lachowicz-Tabaczek, & Sorokowska, 2017), empathy (Apple, Streeter, & Krauss, 1979), and persuasiveness (Guyer, Fabrigar, & Vaughan-Johnston, 2019). Pitch can also serve as a proxy for different affective states, with higher pitch indicating anger, fear, and happiness, and lower pitch indicative of sadness and tenderness (Juslin & Laukka, 2003). As previously noted, the pitch of the human voice is strongly related to biological sex, with woman typically speaking in higher pitches (approximately 150 to 350 Hz), relative to men (approximately 85 to 200 Hz). As with vocal intensity, frequency also varies during speech, and relates to the speaker's personality and emotional state. For instance, shier people tend to have lower pitch variability, which gives the impression of a steadier voice (Abelin & Allwood, 2000), while extroverts tend to exhibit higher pitch variability in their utterances (Scherer & Giles, 1979). Aside from this inter-individual variability, there are also intra-individual differences in pitch. As seen in Fig. 2, again taking a voice sample from a single speaker, we can see that frequency moderately increases in the second soundwave, likely due to a greater state of arousal. Indeed, increased pitch variability is associated with high arousal emotions such as anger or happiness, while the opposite is true for low arousal emotions such as sadness and tenderness (Juslin & Laukka, 2003). Table 1 summarizes the links between frequency and affective states and traits.

The fourth dimension captures perturbations of a soundwave by analyzing the spectral features of a soundwave. These spectral features of a soundwave assess the amount of perturbance or periodicity of a soundwave and summarize the extent of "vocal instability" during speech formation. Imagine a nervous speaker with a shaky voice (i.e., high perturbance) compared to a calm speaker with a steady and stable voice (i.e., low perturbance). This instability or perturbance can be quantified by several sub-measures including vocal jitter, shimmer, harmonics-to-noise ratio (HNR), or spectral entropy, each of which relates to a speaker's current emotional state. For example, vocal jitter is a measure of period-to-period variations of the fundamental frequency, indicating irregularity of the speaker's pitch (Farrús, Hernando, & Ejarque, 2007; Kamiloğlu, Fischer, & Sauter, 2020). Vocal jitter has been associated with both positive and negative emotional states. Specifically, higher jitter indicates emotional stress (Li et al., 2007; Scherer, 1986), annoyance (Johnstone & Scherer, 1999), anger (Jacob, 2016; Juslin & Laukka, 2003), but also happiness (Johnstone & Scherer, 1999; Juslin & Laukka, 2003) and verbal confidence (Jiang & Pell, 2017), while lower values of vocal jitter indicate more negative, low arousal emotions such as sadness (Juslin & Laukka, 2003). Vocal shimmer on the other hand, captures period-to-period variations of the soundwave amplitude (Farrús et al., 2007; Kamiloğlu et al., 2020). Shimmer is another important proxy for an individual's current emotional state, with higher shimmer signaling higher stress (Li et al., 2007) and anger (Jacob, 2016) and lower shimmer signaling manifestations of positive affect such as joy (Jacob, 2016). Both jitter and shimmer have been used in conjunction with machine learning algorithms as important components in several of the latest speech emotion recognition algorithms (Jacob, 2016).

The harmonics-to-noise ratio (HNR) is a vocal parameter that quantifies the amount of periodic (harmonic) signal over noise and is a principal determinant of "voice quality" (Ferrand, 2002). Higher HNR has been associated with feelings of pleasure, interest and happiness, while lower HNR has been associated with lust and lack of confidence (Jiang & Pell, 2017; Kamiloğlu et al., 2020). Finally, the concept of entropy measures the extent of disorganization or uncertainty of a quantity. The voiced regions of a soundwave typically present lower entropy, whereas silent or noisy regions present higher vocal entropy (Toh, Togneri, & Nordholm, 2005). Vocal entropy has been linked to states of negative mood and has been cautiously proposed as acoustical

indicator of depression in female speakers (Yingthawornsuk, 2016).

Our conceptual framework of speech in Table 1 summarizes how the vocal features in the human voice along the four dimensions of a soundwave (time, amplitude, frequency, and spectral domain) relate to speaker traits and experienced emotional states. Our framework is inspired by earlier work on the "speech chain" proposed by Denes and Pinson (Denes & Pinson, 1993), Scherer's seminal work on emotion perception in the human voice (Scherer, 1978, 2003), and classic information theory models such as Shannon and Weaver's model of communication (Shannon & Weaver, 1949). Our conceptualization integrates these previous models, providing a structured view along the four dimensions of a soundwave, the related objective vocal features of a soundwave, and the associated emotional states and speaker traits that they reflect.

In the next sections we illustrate the process and unique insight derived from analyzing a soundwave using state-of-the-art signal processing packages in R, followed by directions for future research.

3. Case: User interaction with voice-controlled interfaces

In what follows, we illustrate the practical analysis and unique insight that can be generated through the application of voice analytics in the all too common occurrence of a user struggling to communicate with a voice-controlled interface (see also Hildebrand, Hoffman, & Novak, 2020). Specifically, we analyze the audio from a viral video which has received millions of views and tens of thousands of likes and comments (Harwell, 2018; Newsflare, , 2018). This video features a female Scottish user attempting in vain to issue a command to Amazon Alexa to play a song on Spotify. Despite the humorous public press coverage, this inability to communicate effectively with a voice-controlled interface is often as frustrating as it is common. Although we focus here on a particular example for practical reasons, the insights generated from this case readily apply to a broad set of applications and contexts including those outside the realm of voice-controlled interfaces.

First, we extracted the publicly available audio file from YouTube as a waveform audio (WAV) file. We focus our analyses on two distinct key characteristics of the interaction: (1) the speech formation during the initiation of the so-called wake word of saying "Alexa" and (2) the syntactical and vocal changes when issuing the full command ("Alexa, play something is cooking in my kitchen on Spotify by Dana"). Specifically, we selected the range of audio containing the sequence of the three subsequent wake words and the two commands issued by the user (note that the third wake word of saying "Alexa" was not connected to issuing a command and was only followed by a string of pejorative language).

All data, code, and additional walk-through explanations are provided in a Data-In-Brief documentation accompanying the current manuscript.

3.1. Software environment & voice analytics production pipeline

The following use case focuses on the substantive insights generated by utilizing voice analytics. As highlighted earlier, the detailed code with further explanations is published in a separate, in-depth Data-In-Brief documentation. Before addressing the real-world case, we highlight the key software programs and review the main steps used to analyze and process audio files.

The Data-In-Brief code covers all analyses in the statistical computing environment R. This begs the question of why we focus on R? First, while standalone programs such as Audacity provide an easy to use graphical user interface (GUI) for editing soundwaves, they tend to offer very limited functions to extract vocal features, and few or no tools for automating the feature extraction process. Other programs such as Praat provide a powerful GUI for editing, feature extraction, and also visualizing soundwaves, but lack the option to automate parallel or serial processing of the hundreds or thousands of voice files typically necessary for research in this area. In contrast, R provides single platform for all steps of the voice and sound analytics process: (1) batch processing of multiple files in parallel, (2) extracting the vocal features using already established signal processing packages, (3) analyzing these features both descriptively (e.g., visualize key features of a soundwave) or using inference statistics (e.g., use vocal features as predictors or outcomes in statistical models), and (4) producing high-quality statistical graphs and data visualizations. We acknowledge the presence of excellent libraries in Matlab and Python, but focus on R due to its strong statistical computing environment and active developer community (Maronna, Martin, Yohai, & Salibián-Barrera, 2019; McElreath, 2016).

Fig. 3 provides a visual summary of the voice analytics pipeline from reading, editing, and visualizing sound files to extracting vocal features for further analysis. We leave all technical details for the interested reader to the Data-In-Brief documentation (see the Data-In-Brief for all technical package specifications and functions used in R). We wish to underscore that, as in any data science and analytics oriented project, the process of reading, editing, and further processing sound files often dwarfs the time required to conduct the focal analysis of the vocal characteristics of interest (García, Luengo, & Herrera, 2015; Luengo, García-Gil, Ramírez-Gallego, García, & Herrera, 2020). This is simply because prior to analysis, a researcher may desire to (1) remove irrelevant utterances from a soundwave, (2) remove periods of silence (often at the beginning or the end of a sound file, think of an interview for example), and (3) extract only specific sections of interest for further analysis.

3.2. Visualizing sound

After reading and editing a sound object as illustrated in Fig. 3, we can visualize key features of the soundwave. Visualization entails converting soundwave data into a statistical graph or image. Common elements of a soundwave visualization include: (1) amplitude, (2) frequency, and (3) a combination of amplitude and frequency over time. Oscillograms are often used to depict the amplitude, while spectrograms are typically used to depict the frequency and combinations of frequency and amplitude over time.

3.3. Oscillograms

Oscillograms display the amplitude of a soundwave over time. Oscillograms are particularly useful to identify potential differences in loudness. For example, Fig. 4 illustrates the oscillograms of the Alexa commands. We visualize the entire soundwave (Fig. 4, top panel) and then zoom-in on the articulation of the wake word only (i.e., "Alexa") (Fig. 4, bottom panel). The four depicted oscillograms provide initial evidence that the second command is longer in duration than the first command even though the two commands are semantically identical ("Alexa, play by Dana"). We can also see that the vocal breaks (i.e., the time between each spoken word) are longer in the second compared to the first command. Second, we observe several striking differences in the emphasis on individual words between the two commands. Per the lower panel of Fig. 4, we see larger overall amplitudes in the second command, with the wake word "Alexa" having a noticeably larger amplitude. Thus, without listening to the voice files or knowing what the person is saying, we can identify that the second command is substantially louder than the first. As summarized in our conceptual framework in Table 1, this pattern is suggestive of increased anger or a stressful interaction experience in the current usage context.

3.4. Visualizing fundamental frequency

Visualizing the user's fundamental frequency over time is often referred to as the f0 contour or pitch track. We care about f0 contour information because it reveals the range of the fundamental frequency of

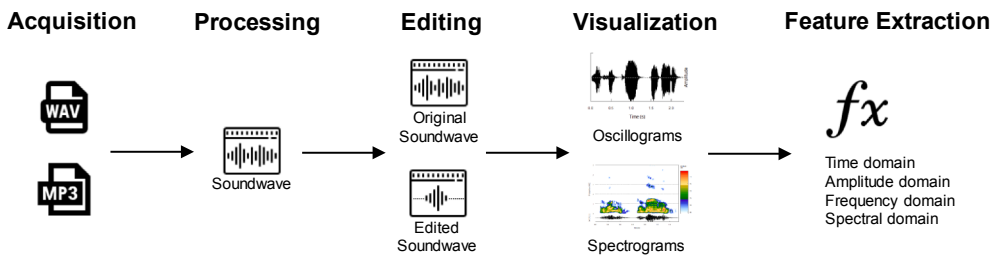

Fig. 3. Voice Analytics Production Pipeline. *Note*: A more technical in-depth description detailing the specific packages and functions in R to reproduce all analyses in the current paper is available in the accompanying Data-In-Brief documentation.

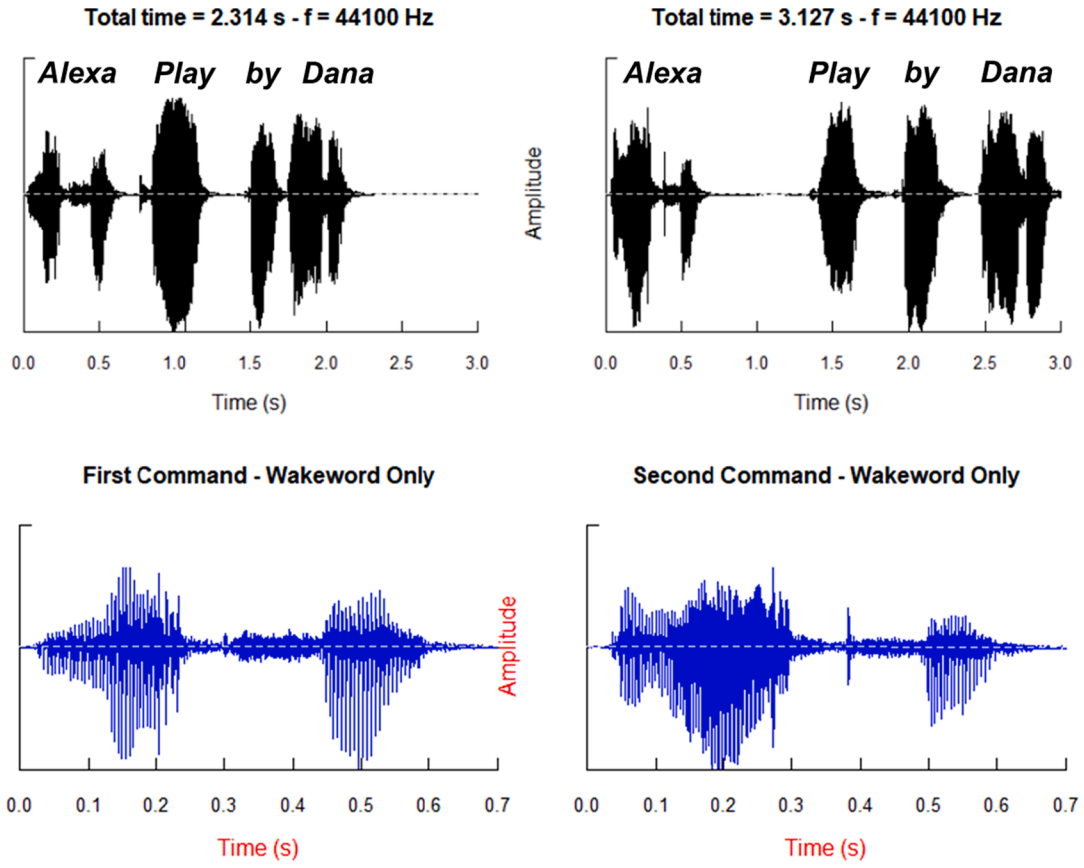

Fig. 4. Oscillograms of First and Second Commands (Upper Panel) and Isolated Wakewords (Lower Panel).

a speaker, the speaker's pitch variability over time, voiceless and voiced regions, as well as regular and irregular phonation. To illustrate its application and key insight, we again focus our analysis on the wake words of the speaker. Fig. 5 illustrates the first three frequency bands for each wake word respectively. These visualizations illustrate the general increase in frequency and also depict the increasing variability from the first to second to the third wake word, as well as a general increase in variability across time. Given the strong association between increased frequency levels and the experience of stress, anger, and annoyance, Fig. 5 likely represents a visual manifestation of the escalating frustration the user experienced during her failed interaction with Amazon Alexa.

3.5. Spectrograms

Spectrograms provide a multidimensional representation of a soundwave based upon amplitude, frequency, and time. Spectrograms generally display time on the x-axis, frequency on the y-axis and represent different amplitude levels (or loudness) by varying color gradients of the data points. Fig. 6 provides more nuanced insight, illustrating further key differences in the three consecutive wakewords spoken by the user. Across all spectrograms, we can observe (1) increasingly red-colored gradients that indicate an increase in power or loudness, (2) the tendency toward progressively longer voice breaks, and (3) the moderately lower and stable frequency in the first compared to the subsequent commands and wake words. Taken together, these findings suggest that the greater frequency and intensity are likely reflections of increased anger and experienced frustration during the interaction with Amazon Alexa.

3.6. Extracting vocal features

After importing, selecting, and visualizing sound data, researchers will likely want to extract several key vocal features for further analysis, so as to use them as predictors or outcome variables in statistical models (for example, to predict or classify levels of anger or stress or other constructs as outlined in Table 1). In what follows, we highlight some of the key vocal features to extract for further analysis and the unique

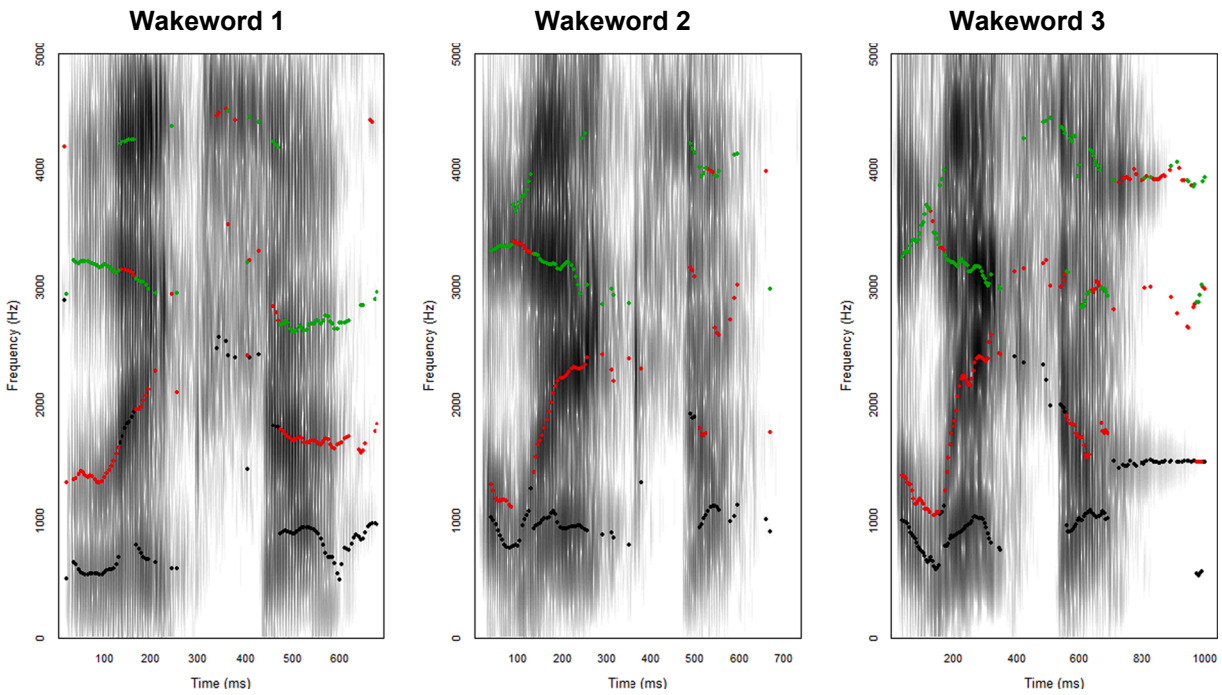

Fig. 5. Visualizing the First (Black), Second (Red), and Third (Green) Formant Frequencies Across Wakewords.

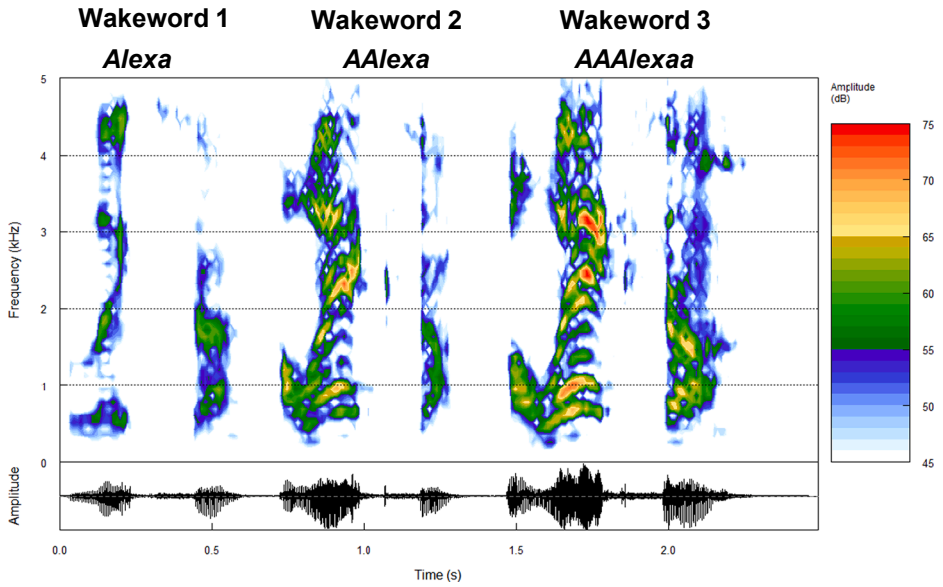

Fig. 6. Spectrogram of Consecutive Wakewords.

insights in the current user interaction example.

3.7. Time domain

The simplest measure in the time domain is the duration measured in seconds or milliseconds of a soundwave. Extracting the duration of the two commands reveals a duration of 2.31 s for the first and 3.13 s for the second command, corresponding to 1.296 words and 0.959 words per second respectively. This is consistent with our previous observation that the second command generally contained longer vocal breaks.

Analysis of the of the percentage of vocal breaks provides further evidence that vocal breaks were generally larger in the second (52.42%) compared to the first (45.05%) command, indicating lower verbal fluency during speech formation. Thus, these time domain measures, paired with the insight generated from the visualization of the soundwave, together provide converging evidence that the user spoke both longer and louder with each consecutive command and wake word.

3.8. Amplitude domain

The amplitude of a soundwave determines the power or loudness, with larger amplitudes indicating a louder sound. The amplitude of a soundwave refers to the extent to which the air particles are displaced from the position of equilibrium. One measure of loudness based on the subjective perception of sound pressure that is typically used in psychophysics and psychoacoustics research is the 'sone' metric (Stevens, 1936). Extracting the average loudness measured in sone corroborates our earlier observation that the second command was generally louder

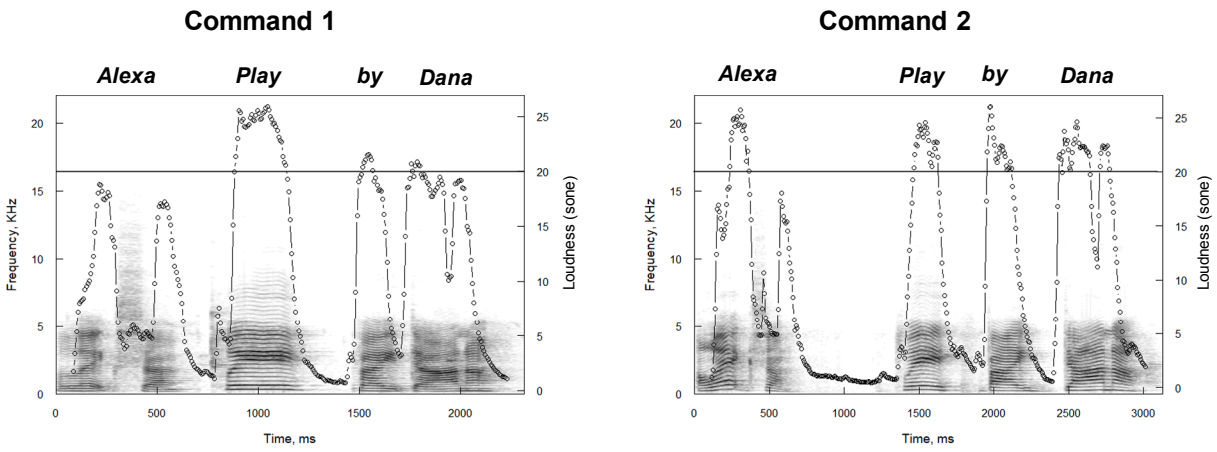

Fig. 7. Loudness of First and Second Command.

compared to the first command. Fig. 7 visualizes the loudness spectrum of the two soundwaves and reveals the dominant and overall higher sone distribution in the second compared to the first command. The grey scale visualization of the spectrum further highlights the darker colored areas indicating greater sound pressure levels (i.e., higher loudness). These analyses point to the fact that the user increases substantially in loudness from the first to the second command and provide supporting evidence for our initial observation of regions with greater sound pressure levels when analyzing the spectrogram in Fig. 6.

3.9. Frequency domain

Although there are many subcomponents of sound frequency, here we focus on the extraction of two key frequency features of our soundwaves: pitch and variability. As noted earlier, pitch is strongly related to biological sex, with men having generally lower (85 to 200 Hz), and women having generally higher vocal pitch (150 to 350 Hz). Given that the speaker in our samples is a woman, we would expect a pitch range above 150 Hz in our frequency analyses. Our visualizations of the soundwave also depict moderately higher frequencies in the second compared to the first command with overall greater variability (see also Fig. 5 depicting the different frequency bands). Consistent with our expectation, we find an increase from 222.45 Hz to 270.92 Hz in the average pitch. Thus, the extracted quantitative features confirm the "shrill" and aroused nature of the user's voice after repeatedly failing to make herself understood to Alexa.

These key differences are even more prominent when comparing only the wake words from the first (181.25 Hz), to the second (271.99 Hz), to the third (278.33 Hz) wake word "Alexa," with increasing variability in frequency levels from the first (21.96 Hz), second (49.56 Hz), and then the third wake word (93.40 Hz). Taken together, the extracted pitch and pitch variability demonstrate the consistent increase across wake words, and reveal a vocal pattern of both high frequency (i.e., shrill) and high variability, indicating the apparent anger and frustration of the Amazon Alexa user.

3.10. Spectral domain

Next, we examine the speaker's voice for potential differences in spectral qualities, which generally assess the amount of perturbance or periodicity of a soundwave. As highlighted in our theory section, greater vocal jitter, shimmer, and entropy, as well as lower harmonicity (i.e., lower HNR) have been associated with greater experiences of negative affect such as stress, fear, or anxiety. Consistent with this intuition and expanding the previous vocal features, we also find greater vocal jitter in the second (Jitter = 1.22) compared to the first command (Jitter = 1.04), greater vocal shimmer ($Shimmer_{Second}$ = 7.87; $Shimmer_{First}$ = 7.83), greater entropy ($Entropy_{Second}$ = 0.457; $Entropy_{First}$ = 0.448) and also lower harmonicity (HNR_{Second} = 5.51; HNR_{First} = 6.19). These findings corroborate our visual observations depicted in the spectrogram of Fig. 6 that the user's vocal speech pattern increases in variability from the first to the second command.

As summarized in Table 2, analyzing the vocal features of the Amazon Alexa user across the four key dimensions of speech (time, amplitude, frequency, and spectrum), demonstrates systematic increases in loudness, speech duration, systematically higher pitch, and greater spectral variability. This pattern offers consistent evidence of a user in a state of high arousal and negative affect due to an arguably frustrating interaction with Amazon Alexa.Table 3.

4. Directions for future research & applications

Although the above use case was specifically conceived to identify a user's experience with a voice-controlled virtual assistant from their vocal features, we can derive four important directions for the future of voice analytics applications in research and business practice. These include: (1) enhancing our understanding of the consumer-technology relationship, (2) defining the role of voice identification in consumer-firm interactions, (3) using voice analytics in service and sales automation, and (4) detecting fraudulent behaviors. All these applications will likely leverage extracted vocal features as inputs to supervised machine learning models, which will aim to discern discrete consumer emotional states, identify a person or customer, or predict shopping-relevant outcomes such as purchase likelihood during a sales call or the truthfulness of a communication message.

Consumer-Technology Relationships

The current work holds the potential to provide a novel lens to study and understand the dynamic nature of the modern, technology-augmented consumer (Melumad, Hadi, Hildebrand, & Ward, 2020). For example, building on Hoffman and Novak's (Hoffman & Novak, 2018; Novak & Hoffman, 2019) work on the emergent consumer-object experiences in the Internet of Things, future work might address whether, and to what extent, vocal features in the human voice reflect enabling versus constraining consumer-object experiences. Constrained consumer-object experiences might manifest through the use of

Table 2
Summary of Vocal Features from First and Second Command.

	First Command	Second Command
Duration (s)	2.31	3.13
Loudness (sone)	16.23	18.09
Pitch (Hz)	222.45	270.92
Harmonic-Noise-Ratio	6.19	5.51

Table 3
Voice as a Focal Object of Analysis versus Input or Output of Analysis.

Level of Analysis	Voice as Focal Object	Voice as Input	Voice as Output
Type of Analysis	- Voice	- Voice Transcription	- Voice
	- Analytics		- Synthesis
Methodological Focus	- Vocal Feature Extraction	- Speech-to-Text	- Text-to-Speech
			- Waveform Generation
	o Time	- Sentiment Analysis	
	o Domain		- Convolutional Neural Networks
		- Parts-of-Speech Tagging	
	o Amplitude Domain		- Deep Transfer
	o Frequency Domain	- Word Embeddings	- Learning
	o Spectral Domain		
Key Packages in R	- seewave	- quanteda	- googleLanguageR
	- tuneR	- tidytext	- Rtts
	- phonTools	- text2vec	- aws.polly
Key Packages in Python	- Parselmouth	- Natural Language Toolkit	- pyttsx3
	- pyAudioAnalysis		- gTTS
	- aubio	- Gensim	- boto3
	- librosa	- TextBlob	

syntactically different styles of language, as well as temporally shifting differences in vocal expression. Specifically, first-time users of Amazon Alexa might speak to Alexa in a fundamentally different manner, compared to later, more mature stages of usage. For example, initial use might be characterized by greater overall loudness and vocal breaks, reflecting consumers intentional, slow-paced issuing of a command to maximize understanding or intent-matching. With increasing experience, consumers might adjust to a more natural speech pattern, with faster-paced expression and fewer vocal breaks. Likewise, analyzing the vocal features of a user might provide novel insight into the process of product abandonement and terminating product usage. For example, a systematically increasing frequency paired with louder vocal expressions might indicate greater user frustration and ultimately help predict (and eventually prevent) product abandonement. As these subtle changes in vocal expression might be outside of consumers' conscious awareness, voice analytics might provide a powerful, objective tool to study the dynamics of continuously evolving consumer-technology experiences.

Consumer-firm interactions

The vocal features of an individual are already used to develop accurate individual-level speaker profiles for identification and customization purposes. While such "voiceprints" are currently predominantly used to optimize speaker understanding of voice-controlled interfaces (Tirumala, Shahamiri, Garhwal, & Wang, 2017) or applications in law enforcement settings (West, 2019), other industries such as the banking sector are beginning to use voice samples as a consumer identification tool. Aside from narrow consumer identification applications, future work might also address to what extent the use of such customer identification through voiceprints affects firm perception more broadly. It is conceivable, for example, advanced voice identification technology could lead to higher perceptions of firm innovativeness or conversely, and more likely given increasing consumer privacy concerns, to poorer perceptions of respect of consumer privacy. Relatedly, future work might also explore to which extent sharing consumer voice data is conceptually distinct from sharing other forms of data (e.g., social media data, location data, etc.) and the factors that shape consumer willingness to share voice identification data. Such voice-based identification might also provide greater levels of customer value by leveraging convenience, usability, and portability needs (a customer's voiceprint cannot be lost or forgotten as with a traditional password).

Service and Sales Automation

We see a wide range of applications in traditional customer service and sales automation settings, particularly in the realm of call center optimization. The use of voice analytics may help to gain novel insight into the passive assessment and tracking of customer call satisfaction. Specifically, voice analytics has the potential to signal the quality of service or sales agents, aid in predicting the risk of customer churn, or even serve as a coaching tool for frontline employees. The use of voice analytics as a coaching tool might also help to shift focus and perception toward the experiences of the customer, helping front office employees to be more attuned to the momentary emotions a customer might be exhibiting, and foster greater self-awareness during service or sales calls.

Fraud detection

If voice analytics could be further developed to serve as a "vocal truth algorithm", it could harbor massive potential as a tool to combat fraud or evaluate the veracity of communicated messages. While voice analytic truth detection has been primarily used in law enforcement and military contexts, we see a great potential for future use in insurance or financial fraud detection. For example, given the evidence that biomarkers of stress in the human voice have been associated with higher rates of untruthful responses (Sondhi, Vijay, Khan, & Salhan, 2016), further development of such vocal analytic techniques could be extended to the diversity of business scenarios such as assessing the veracity of insurance claims or venture capital investor pitches. However, consumer apprehension over technologies like facial recognition have the potential to extend to voice analytic truth detection, so firms are advised to proceed with caution.

5. Voice data as input, output, or focal object of analysis

The outlined directions for future research treat the vocal features, or voice more generally, as the focal object. However, and as highlighted in Table 2, we also see great potential in using voice data as (1) an input for other methods and (2) as an output that is generated based on predefined voice characteristics.

First, the extracted vocal features can serve as inputs for existing research methods that combine traditional content analysis with speech-to-text algorithms. These algorithms transform a vocal data object into a corpus of text data. Such corpuses can then be analyzed using a broad range of text analysis techniques such as sentiment analysis, parts-of-speech tagging, word embeddings, or more traditional qualitative content analysis (Yu & Deng, 2015).

Second, speech synthesis applications could be used to generate entirely new, artificial soundwaves from the original voice data. For example, convolutional neural networks and other deep learning methods such as transfer learning have opened the door to develop artificially-generated voices, which in turn will likely spark entirely new areas in both research and business. Such innovations will facilitate the development of personalized voice assistants that can be tailored to match individual user characteristics and preferences. In lieu of a prebuilt stock of male and female voices (such as those in today's GPS navigation systems), users will instead be able to create highly personalized "digital voice avatars" constructed from user-supplied vocal samples of their choosing (e.g. themselves, their spouse, or a celebrity).

Table 2 provides a summary of the diversity of research streams that could use voice data either as the focal or ancillary object of analysis. For readers who wish to delve deeper than what we can cover in this primer, we also highlight some of the key text processing and speech synthesis packages in R and Python applicable to each of these areas in Table 2.

6. General discussion

Analyzing the vocal features of the human voice has significant potential to better understand and ultimately predict human behavior. The key objectives of this paper were to: (1) provide a conceptual framework

demonstrating how speaker states and traits can be revealed by analyzing key vocal features (time, amplitude, frequency, and spectral domains), (2) offer a guided primer on how to operationalize processing, editing, visualization and analysis of sound files using the latest signal processing packages in R, and (3) suggest new directions for how to effectively leverage voice and sound analytics for future, cross-disciplinary research.

As in other areas of rapid technological development, voice analytics methods have quickly outpaced international regulatory policy. This disparity is already beginning to raise thorny ethical questions that will have to be addressed by researchers, business professionals, and ultimately public policy. One basic first step toward improving ethical business practices is to keep individuals informed as to what data is being collected and for what purposes (as has approximately been dictated by the General Data Protection Regulation in Europe) (Regulation, 2016).

With the proliferation of voice-controlled interfaces such as Amazon Alexa or Google Home as well as technologies that passively collect voice recordings on mobile devices or in public places (Porcheron, Fischer, Reeves, & Sharples, 2018), we believe that many individuals are completely unaware that their voice data is being collected both by companies and governments. Regarding the latter, recent reports from the Human Rights Watch indicate that China is already collecting citizens' voice samples with the goal of building a multi-modal biometric profiling system (Hincks, 2017).

Thus, despite the significant potential of voice analytics for future research, business practice, and public policy, we wish to emphasize the significant responsibilities associated with the analysis and extraction of vocal features from the human voice. Ideally, future voice analytic products and services should be developed in such a way that avoids any consumer privacy violations and algorithmic biases that could negatively affect and discriminate against vulnerable groups, and generally preclude its use as a surveillance tool.

7. Conclusion

The way in which we speak often reveals more about our personality and emotional state than most of us realize. To the best of our knowledge, the current paper provides the first non-technical introduction to the field of voice analytics in business research, including an integrative conceptual framework that links the vocal features of the human voice to distinct emotional states and speaker traits, proposes an outline for future research directions, the inherent business potential, and highlights the issues that will have to be addressed to ensure ethical practice of voice analytics.

Given the ubiquity of voice-controlled interfaces and the fascinating range of applications of voice analytics in both research and business practice, we hope that the current paper provides fruitful directions for future cross-disciplinary research, stimulates the development of novel research questions, and provides a new lens (and method) to study and understand human behavior.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbusres.2020.09.020.

References

Abelin, Å., & Allwood, J. (2000). Cross linguistic interpretation of emotional prosody. Retrieved from *International Tutorial and Research Workshop on Speech and Emotion*, 110–113 http://www.isca-speech.org/archive_open/speech_emotion/spem_110.html.

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology, 37*(5), 715–727. https://doi.org/10.1037/0022-3514.37.5.715

Brenner, M., Doherty, E. T., & Shipp, T. (1994). Speech measures indicating workload demand. Retrieved from *Aviation, Space, and Environmental Medicine, 65*(1), 21–26 http://www.ncbi.nlm.nih.gov/pubmed/8117221.

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research, 17*(1), 140–169. https://doi.org/10.1111/j.1468-2958.1990.tb00229.x

Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General, 145*(5), 536–547. https://doi.org/10.1037/xge0000166

Clark, A. V. (2005). *Psychology of moods*. New York: Nova Science Publishers Inc.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour, 60*(6), 773–780. https://doi.org/10.1006/anbe.2000.1523

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817. https://doi.org/10.1017/S1351324916000243

Dasgupta, P. B. (2017). Detection and analysis of human emotions through voice and speech pattern processing. International Journal of Computer Trends and Technology, 52(1), 1–3. 10.14445/22312803/IJCTT-V52P101.

Denes, P., & Pinson, E. (1993). *The speech chain*. Macmillan.

Diao, W., Liu, X., Zhou, Z., & Zhang, K. (2014). Your voice assistant is mine. Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices - SPSM '14, 63–74. 10.1145/2666620.2666623.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. *Eighth Annual Conference of the International Speech Communication Association*.

Ferrand, C. T. (2002). Harmonics-to-noise ratio. *Journal of Voice, 16*(4), 480–487. https://doi.org/10.1016/S0892-1997(02)00123-6

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America, 102*(2), 1213–1222. https://doi.org/10.1121/1.421048

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining.*. https://doi.org/10.1007/978-3-319-10247-4

Giegerich, H. J. (1992). *English phonology: An introduction*. Retrieved from: In Cambridge Textbooks in Linguistics. https://www.cambridge.org/core/books/english-phonology/6A07CFB3C4D7A6D0A41551FDC1DB8044.

Guyer, J. J., Fabrigar, L. R., & Vaughan-Johnston, T. I. (2019). Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion. *Personality and Social Psychology Bulletin, 45*(3), 389–405. https://doi.org/10.1177/0146167218787805

Harwell, D. (2018). *The accent gap*. The Washington Post: Retrieved from. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

Hildebrand, C., Hoffman, D. L., & Novak, T. P. (2020). Dehumanization in the IoT: Experiential consequences of syntactically constricted human-machine interaction with digital voice assistants. *Working Paper*.

Hincks, J. (2017). China Is creating a database of Its citizens' voices to boost its surveillance capability: report. Time. Retrieved from https://time.com/4992849/china-voice-database-surveillance/.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Hodges-Simeon, C. R., Gaulin, S. J. C., & Puts, D. A. (2011). Voice correlates of mating success in men: Examining "contests" versus "mate choice" modes of sexual selection. *Archives of Sexual Behavior, 40*(3), 551–557. https://doi.org/10.1007/s10508-010-9625-0

Hoffman, D. L., & Novak, T. P. (2018). Consumer and object experience in the internet of things: An assemblage theory approach. *Journal of Consumer Research, 44*(6), 1178–1204. https://doi.org/10.1093/jcr/ucx105

Jacob, A. (2016). Speech emotion recognition based on minimal voice quality features. *International Conference on Communication and Signal Processing (ICCSP), 2016*, 886–890. https://doi.org/10.1109/ICCSP.2016.7754275

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication, 88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. *Proceedings of the XIVth International Congress of Phonetic Sciences*.

Jurafsky, D., & Martin, J. (2014). Speech and Language Processing. In Speech and Language Processing.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770–814. https://doi.org/10.1037/0033-2909.129.5.770

Kamiloğlu, R. G., Fischer, A. H., & Sauter, D. A. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review, 27*(2), 237–265. https://doi.org/10.3758/s13423-019-01701-x

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences, 279*(1738), 2698–2704. https://doi.org/10.1098/rspb.2012.0311

Latinus, M., & Taylor, M. J. (2012). Discriminating male and female voices: Differentiating pitch and gender. *Brain Topography, 25*(2), 194–204. https://doi.org/10.1007/s10548-011-0207-9

Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K., & Newman, J. (2007). Stress and emotion classification using jitter and shimmer features. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, 4*. https://doi.org/10.1109/ICASSP.2007.367261

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). *Big data preprocessing.*. https://doi.org/10.1007/978-3-030-39105-8

MacLachlan, J. (1982). Listener perception of time-compressed spokespersons. *Journal of Advertising Research, 22*(2), 47–51.

Mallory, E. B., & Miller, V. R. (1958). A possible basis for the association of voice characteristics and personality traits. *Speech Monographs, 25*(4), 255–260. https://doi.org/10.1080/03637755809375240
Maronna, A. R., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). Robust Statistics: Theory and Methods (with R). *Retrieved from.* https://books.google.gr/books?id=K5RxDwAAQBAJ.
McElreath, R. (2016). Statistical rethinking : a Bayesian course with examples in R and Stan.
Melumad, S., Hadi, R., Hildebrand, C., & Ward, A. F. (2020). Technology-augmented choice: How digital innovations are transforming Consumer decision processes. *Customer Needs and Solutions, 1–12.* https://doi.org/10.1007/s40547-020-00107-4
Miller, N., Maruyama, G., Beaber, R. J., & Valone, K. (1976). Speed of speech and persuasion. *Journal of Personality and Social Psychology, 34*(4), 615–624. https://doi.org/10.1037/0022-3514.34.4.615
Mohammadi, G., & Vinciarelli, A. (2015). Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract. *International Conference on Affective Computing and Intelligent Interaction (ACII), 2015*, 484–490. https://doi.org/10.1109/ACII.2015.7344614
Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171
Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2011). Mine your own business: Market-structure surveillance through text mining. *SSRN Electronic Journal, 10*(3), 10–202. https://doi.org/10.2139/ssrn.1816494
Newsflare. (2018). Amazon alexa can't understand scottish accent. Retrieved from YouTube website: https://www.youtube.com/watch?v=CYvFxs32zvQ.
Novak, T. P., & Hoffman, D. L. (2019). Relationship journeys in the internet of things: A new framework for understanding interactions between consumers and smart objects. *Journal of the Academy of Marketing Science, 47*(2), 216–237. https://doi.org/10.1007/s11747-018-0608-3
Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review, 24*(3), 856–862. https://doi.org/10.3758/s13423-016-1146-y
Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, 1–12. 10.1145/3173574.3174214.
Portet, F., Vacher, M., Golanski, C., Roux, C., & Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects. *Personal and Ubiquitous Computing, 17*(1), 127–144. https://doi.org/10.1007/s00779-011-0470-5
Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice pitch. *Evolution and Human Behavior, 26*(5), 388–397. https://doi.org/10.1016/j.evolhumbehav.2005.03.001
Ray, G. B. (1986). Vocally cued personality prototypes: An implicit personality theory approach. *Communication Monographs, 53*(3), 266–276. https://doi.org/10.1080/03637758609376141
Regulation, E. U. (2016). 679 of the European parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. EC (General Data Protection Regulation).
Santos, B. S., Ferreira, B. Q., & Dias, P. (2015). Heuristic evaluation in information visualization using three sets of heuristics: an exploratory study. In Human-Computer Interaction - Design and Evaluation: Vol. HCII 2015, (pp. 259–270). 10.1007/978-3-319-20901-2_24.
Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*(4), 467–487. https://doi.org/10.1002/ejsp.2420080405
Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. Retrieved from *Psychological Bulletin, 99*(2), 143–165 http://www.ncbi.nlm.nih.gov/pubmed/3515381.
Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5
Scherer, K. R., & Giles, H. (1979). *Social markers in speech.* Cambridge; New York; Paris: Cambridge University Press. Éditions de la Maison des sciences de l'homme.
Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. *Retrieved from.* http://books.google.com/books?id=ZiYIAQAAIAAJ.
Sondhi, S., Vijay, R., Khan, M., & Salhan, A. K. (2016). Voice analysis for detection of deception. 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), 1–6. 10.1109/KICSS.2016.7951455.
Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review, 43*(5).
Sueur, J. (2018). Sound analysis and synthesis with R. *Springer International Publishing.* https://doi.org/10.1080/09524622.2019.1651507
Suri, V. K., Elia, M., & van Hillegersberg, J. (2017). Software bots - the next frontier for shared services and functional excellence. *Lecture Notes in Business Information Processing, 306*, 81–94. https://doi.org/10.1007/978-3-319-70305-3_5
Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications, 90*, 250–271. https://doi.org/10.1016/j.eswa.2017.08.015
Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America, 85*(4), 1699–1707. https://doi.org/10.1121/1.397959
Toh, A., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. *Proceedings of PEECS.*
Tusing, K. J., & Dillard, J. (2000). The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research, 26*, 148–171. https://doi.org/10.1093/hcr/26.1.148
West, E. (2019). Amazon: Surveillance as a Service. Surveillance & Society, 17(1/2), 27–33. 10.24908/ss.v17i1/2.13008.
Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America, 52*(4B), 1238–1250. https://doi.org/10.1121/1.1913238
Yingthawornsuk, T. (2016). Spectral entropy in speech for classification of depressed speakers. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)* (pp. 679–682). https://doi.org/10.1109/SITIS.2016.113
Yu, D., & Deng, L. (2015). *Automatic Speech Recognition.*. https://doi.org/10.1007/978-1-4471-5779-3
Zhang, Z. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America, 140*(4), 2614. https://doi.org/10.1121/1.4964509

Christian Hildebrand is director of the TechX Lab and Full Professor of Marketing Analytics at the University of St. Gallen. He had doctoral and post-doctoral visits at Stanford University, Duke University, and the University of Michigan. His work explores how new technologies change human decision making, cognition, and perception, with an emphasis on digital voice assistants, conversational interfaces such as chatbots, and mobile devices. His research has been published in leading marketing and information systems journals such as the *Journal of Marketing Research*, *Journal of the Academy of Marketing Science*, *Information Systems Research* or the *Journal of Management Information Systems*.

Fotis Efthymiou is a PhD candidate at the Institute of Marketing at the University of St. Gallen, and a member of the TechX Lab. His research mainly focuses on the interaction with voice assistants at the intersection of psychology and marketing. He received his B.Sc. from the University of Patras, Greece in 2015 and his M.Sc. from the University of Geneva, Switzerland in 2019.

Francesc Busquet is a PhD candidate at the Institute of Marketing at the University of St. Gallen and a member of the TechX Lab. His research focuses on the impact of new technologies on consumers with an emphasis on digital voice assistants. He received his B.Sc. from the University of Girona, Spain in 2016 and his M.Sc. from the Polytechnic University of Catalonia, Spain in 2018.

William H. Hampton is co-director of TechX Lab and International Postdoctoral Fellow of the Institute of Marketing at the University of St. Gallen. His research leverages a diverse toolset ranging from behavioral and self-report measures to neuroimaging to study real world decision-making behaviors, particularly those relating to reward in the context of human-computer interaction (HCI) and consumer behavior. He holds a BA in psychology from the University of Pennsylvania and an PhD in Decision Neuroscience from Temple University.

Donna L. Hoffman is the Louis Rosenfeld Distinguished Scholar, Professor of Marketing at The George Washington School of Business in Washington, D.C., and Co-Director of the Center for the Connected Consumer. Her research has appeared in top academic and managerial publications, such as *Marketing Science*, *Management Science*, the *Journal of Marketing Research*, the *Journal of Marketing*, the *Journal of Consumer Research*, and the *Journal of Consumer Psychology*. She currently serves on the editorial boards of leading academic publications in the marketing discipline, as well as serving as an associate editor at the Journal of Marketing.

Thomas P. Novak is the Denit Trust Distinguished Scholar and a Professor of Marketing at the George Washington School of Business in Washington, D.C., where he co-directs the Center for the Connected Consumer. His research is focused on consumer behavior in online environments and digital marketing. His current research interests deal with consumer experience of smart devices and the Internet of Things, using machine learning methods for natural language processing and visualization. His research has appeared in top marketing and managerial journals, such as *Marketing Science*, *Management Science*, the *Journal of Marketing Research*, the *Journal of Marketing*, the *Journal of Consumer Research*, and the *Journal of Consumer Psychology*.