Audio and Visual Analytics in Marketing

Abstract

With the ever-cheaper digital equipment and the prevalent digital platforms such as Facebook and YouTube, more and more human behaviors and activities are digitalized in the form of images, videos, and audio. However, due to the information's unstructured nature, there has been a lack of useful framework and tools that can help businesses to effectively leverage this information to improve their practices. As a result, businesses are missing out on the opportunities to use this information to gain better customer insights, understand customer preference, improve customer experience, discover unmet needs and optimize marketing effectiveness. In this monograph, the authors present an overview of audio and visual analytics and discuss how they can be used by marketers to improve business practices. This monograph first introduced a framework named Artificial Empathy (AE) to illustrate different contexts where the audio and/or visual information emitted by or presented to an individual are used to improve business decision making. Next, it presented a review of the cutting-edge techniques and methods used to mine valuable information and make useful inferences from the audio and visual data. Finally, it reviewed the use of A/V analytics in business practices and concluded with a discussion on the trends in applying audio and visual data analytics in business. This monograph aims to help readers understand how the new forms of rich data will affect the way we do business and gain insights into harnessing the power of audio, image, and video data to make useful inferences and improve business practices.

Keywords: Artificial Empathy, Audio Analytics, Image Analytics, Video Analytics

1. *INTRODUCTION*

With the ever-cheaper digital recording equipment and the prevalent use of online channels for interpersonal interactions such as Facebook, YouTube, TikTok and Zoom, more and more human behaviors and activities are digitalized in the form of images, videos and audio. In fact, one major trend in the digital economy is the exponentially growing number of images, videos and audios generated and consumed every day. For example, in one minute there are 700,000 hours of videos watched and 500 hours of videos uploaded on YouTube, 243,000 photos uploaded on Facebook and 1,000,000 swipes on Tinder, and 400,000 hours of music listened on Spotify[1]. In addition, the creation and consumption of the audio and visual content have been amplified by the global pandemic during the last two years. The ubiquity of audio and visual information has given birth to the development of technologies and applications to transform both online and traditional businesses and marketing practices. For example, face recognition technology has now been widely used in online and mobile payment systems.[2] Over 5 million face videos have been used by Affectiva, an emotion-analysis start-up, to understand how consumers react to video ads.[3] The audio, image, and video information contains much valuable information on many topics of interest to businesses, including advertising, branding, content marketing, retailing, e-commerce, customer relationship management, product recommendation, and data privacy etc. However, it's almost impossible for human beings to process this amount of information on a daily basis. In order to derive useful insights from this ever-growing ocean of information, organizations need more efficient and effective ways to process and analyze the rich audio/visual data and new models to assist business decision making related to these data.

Over the past decade, there has been growing interest among re- searchers across many disciplines including business management, computer science, neuroscience, and social science to use automated audio and visual data analytics for more efficient and effective decision making in various contexts. However, the applications of audio and visual data analytics are still very limited in practice and the majority of organizations have yet to unlock the potential of

[1] https://www.go-globe.com/blog/things-that-happen-every-60-seconds/ .
[2] https://www.bbc.co.uk/news/business-55748964.
[3] https://blog.affectiva.com/the-worlds-largest-emotion-database-5.3-million-faces-and-counting.

audio/visual data analytics for business decision making in the post-digital economy. To this end, many important questions have yet to be answered in this domain, such as, in what contexts should marketers consider using audio/visual data analytics? What types of business problems can be and are best addressed by audio/visual data analytics? What models are available for different types of problems and what skills should researchers and/or practitioners develop to better manage and leverage the potential of audio/visual data in various business contexts? This monograph aims to address these questions by (1) developing a framework for under- standing the internal states of individuals based on audio/visual signals and incorporating works in the domain of audio/visual (A/V) data analytics into marketing research in order to identify future research opportunities; (2) providing an overview of methodologies that are commonly used in conducting research with A/V data; (3) providing a review of A/V analytics-based research in various business contexts; and (4) reviewing the business practices using A/V analytics and identifying the future trends in both research and business applications.

The rest of this monograph is organized as follows: we first propose a framework for A/V data-based research in the business domain and discuss how A/V data analytics can be used to support business decision making in various contexts. We then provide an overview of the key techniques and tools used in A/V data analytics and discuss the procedures and key methodological questions. Finally, we discuss how the A/V analytics has been used in business practices and its trend and future development.

2. A/V DATA AND ARTIFICIAL EMPATHY

We begin by discussing the nature of audio and visual data, and its role in interpersonal interactions in the business contexts. We then discuss the value added by A/V data along the dimensions that are important for business practice and research. Next, we propose the notion of Artificial Empathy that is grounded in neuroscience, psychology and computer science, and a research framework for developing Artificial Empathy models using A/V data. Finally, we review the Artificial Empathy models in the marketing domain and discuss how A/V data-based Artificial Empathy models can be beneficial in various business contexts.

2.1. A/V data as Empathy Cues in Business Contexts

Audio/visual information is an integral part of interpersonal interactions. The audio and visual signals emitted by others in interpersonal interactions are important cues used by humans to make inferences on others' emotions, feelings and thoughts based on observation, memory, knowledge and reasoning (Ickes, 1997). Humans often communicate and make decisions based on the inferences of other's internal states (e.g., emotional, cognitive and physical states) from the audio/visual signals emitted by the person, such as facial expressions, body gestures, voice and words. This ability and capacity of human beings to infer and predict others' thoughts, feelings and experiences without having the thoughts, feelings and experience communicated explicitly is often referred to as "(human) Empathy" (Buie, 1981; Decety and Lamm, 2006; Goldman, 1993; Ickes, 1997). It's been found that human empathy plays an important role in interpersonal interactions to explain and predict others' feelings and behaviors (Decety and Lamm, 2006; for a de- tailed review of the human empathy research, see Duan and Hill, 1996). Considering the prevalence of interpersonal interactions in business contexts, especially in marketing communications, this ability to make psychological inference and understand others' inner experience from the emitted signals is critical for marketers. For example, it's been found that salespeople or service providers' affective states as perceived by the customers have significant impact on customers' purchase intentions and satisfaction levels (e.g., DeShields et al., 1996; Hennig-Thurau et al., 2006; Kahle and Homer, 1985; Wang et al., 2017).

With the development in digital technology and the prevalence of digital devices, more and more interpersonal communications and content consumptions are conducted online, where the digitalized audio/visual signals are sent and interpreted during the information exchange process. Table 2.1 summarizes the types of A/V data commonly used as human empathy cues in business contexts. Here we focus on four main parties in the business contexts: consumers, firms, investors, and third parties (Berger et al., 2020). Based on the main senders and interpreters of the signals, the majority of A/V data falls into one of the following five categories.

[Insert Table 1 about here]

Consumer-Consumer. Consumers often share their profiles, experience, thoughts, and opinions in the forms of images and/or videos in addition to texts on social media platforms such as Instagram (Klostermann et al., 2018; Nanne et al., 2020), YouTube, LinkedIn (Malik et al., 2019), Tumblr (Shin et al., 2020), Flickr (Liu et al., 2020), Twitter (Hartmann et al., 2021; Li and Xie, 2020), and Facebook (Smith et al., 2012), online dating platforms such as Tinder, and online review platforms such as Amazon, TripAdvisor and Yelp (Zhang and Luo, 2022). The information shared by one consumer usually provides rich information about their experiences, perceptions, and judgements about the focal product/brand, thus can be used by others during their decision-making processes.

Consumer-Firm. Consumers can communicate their thoughts and feelings with firms directly in the form of conversations through phone calls, face-to-face interactions at service encounters or in customer interviews, or through the use of online meeting apps such as Zoom and Teams. In addition, customers' offline shopping process (e.g., product trials) can also be captured by smart devices (e.g., smart mirrors) or in-store surveillance videos and used by firms to understand their behaviors and responses during the shopping process (Hui et al., 2013; Lu et al., 2016; Zhang et al., 2014). For example, Hui et al. (2013) and Zhang et al. (2014) used in-store video cameras to track and analyze customers' shopping path and activities during the store visit. Lu et al. (2016) used recordings of customer's product evaluation processes to analyze customer preferences on garment features.

Firm-Consumer. The main ways firms communicate with consumers are through marketing mix. The audio and/or visual information is an indispensable component of the design of marketing mix. This includes (but not limited to) the design of the product (Burnap et al., 2021; Liu et al., 2018), package and brand visuals (e.g., brand logo, font style) (Luffarelli et al., 2019a,b), the design of owned media (e.g., firm websites) and social media content (Kumar et al., 2016), the design of promotional content (e.g., product videos, advertisements) (Jia et al., 2020; Liaukonyte et al., 2015; Teixeira et al., 2012, 2014; Tellis et al., 2019), the design of e-commerce channels and/or offline stores and shelf spaces, and the design of interpersonal selling/shopping processes (Bharadwaj et al., 2022).

Third party-Consumer. Consumers are susceptible to third-party content such as influencers, social media platforms, and mass media platforms (e.g., news, broadcasts, and magazines). The audio and visual products designed and provided by these content providers not only provide experiential and entertainment values to consumers, but also provide ad spaces and play an important role in forming consumers' attitudes toward brands, new products, or new ideas (Hughes et al., 2019; Zhang et al., 2021). For example, Hughes et al. (2019) found that the influencers' characteristics and content characteristics affect the customer engagement and responses differently across different types of social platforms.

Firm-Investor. How managers communicate their ideas in board meetings or funding pitches has a great impact on the funding result (Clarke et al., 2019; Jiang et al., 2019; Li et al., 2019b; Wang et al., 2021). The audio and visual cues from the funding pitches or board meetings can be used by the investors in their funding decisions. For example, using the visual data from 1,460 pitch videos, Jiang et al. (2019) studied the impact of displaying positive emotions on an entrepreneur's chances of gaining more financial support during a funding pitch. Li et al. (2019b) analyzed 6822 crowdfunding videos from Kickstarter and identified key video features that lead to higher project success rates.

2.2.The Value of A/V data in Consumer and Business Decision Making

The growth in the volume and variety of data has led to the increasing use of data and analytics-based decision making in both business and consumer contexts. Digital technologies such as e-commerce and social platforms, video streaming, mobile, and the internet-of-things have generated unprecedentedly large amount of A/V data and have made it easily accessible by both individuals and organizations. This has not only changed the way consumers make purchase and consumption decisions, but also changed how organizations make business decisions (Wedel and Kannan, 2016).

The A/V data can add value to consumer and business decision

making from several aspects. First, each unit of the A/V data (e.g., a photo, a video clip, or a recording of a phone call) contains rich information about the individuals who emitted the signal, which can be used to gain valuable insights about their internal states (emotional, cognitive, and physical), make predictions about their behaviors or responses to new stimulations (e.g., contents or behaviors) or their providers (e.g., individuals, organizations, third party content providers). For example, the images or videos posted by a customer about their experience in a hotel provide more valuable insights into the customer's preference than ratings and can be used by the firm to understand the reasons why a customer gives a certain rating and how to improve its business practice (e.g., Zhang and Luo, 2022). The audio, images or videos contain valuable information that can't be fully captured by the structured/semi-structured data or the text data, thus enabling more valuable insights and diagnostic and prescriptive analyses that would otherwise be impossible.

Second, the A/V data allow organizations to make inferences or predictions about the individuals who emitted the signals without asking them explicitly. There are many reasons why asking explicitly may not be the best way to understand an individual's internal states, especially in business contexts. It may not be in the person's best interest to reveal their true thoughts, feelings, or emotions or it may be costly to do so, or it may be difficult for them to articulate. In addition, structured numeric data such as rankings or ratings are known to be susceptible to biases introduced in the data collections process, for example, users may interpret the rating scales differently, and the design of rating system has been proven to have a significant impact on the rating results (Eslami et al., 2017). In contrast, the A/V data are generated naturally with no pre-defined data structure, thus minimizes the cognitive effort as well as bias in the data collection process caused by imposed data structure.

Last but not the least, compared to numeric or semi-structured data, it's more intuitive for the human brain to process the A/V data and extract useful information from them even with no prior knowledge of how the data are generated or collected. The human brain is constructed to understand the world around us through audio/visual signals. What we see and what we hear carry important information we need for social interactions and survival. We learn the meanings of texts or numbers via their audio/visual representations, so the processing of A/V information underlies the processing of other types of data commonly used in decision making, including structured or semi-structured data and text data. In addition, people don't need additional information to interpret the A/V data. However, one caveat is that human interpretation of the A/V data can be subjective and costly, and it’s not intuitive for traditional analytical models to process A/V data automatically. This has been the main challenge in applying the A/V data in business practices.

2.3 A Framework of Artificial Empathy Using A/V analytics

One way to leverage the benefits of A/V data for business decision making is to scale up the analytical process by using research models and analytical methods that are designed for A/V data. The development in artificial intelligence and machine learning techniques in recent years gives rise to multidisciplinary research on developing machine-based models to fulfil human empathy tasks (Balducci and Marinova, 2018). To this end, Xiao et al. (2013) proposed a notion called Artificial Empathy (AE), which is “the ability of nonhuman models to predict a person’s internal states (e.g., cognitive, affective, physical) given the signals he or she emits (e.g., facial expression, voice, gesture) or to predict a person’s reactions (including, but not limited to internal states) when he or she is exposed to a given set of signals (e.g., facial expression, voice, gesture, graphics, music, etc.).” (Xiao et al., 2013, p. 244).

[Insert Figure 1 about here]

Figure 1 describes the framework of AE and its relations to human empathy (HE). At the center of the framework is the human brain, which processes input signals (bottom-up information processing), and regulates cognitive, emotional and behavioural responses (i.e., output signals) based on the information it inferred from the signals (top- down information processing) (Decety and Jackson, 2004; Reik, 1949). According to Ding (2007), the human brain often needs to balance the needs of different intrapersonal agents (called “mini-minds”) and the interactions of the mini-minds result in a person’s true internal states, which then determine his/her responses. The objective of HE or AE models is to predict or infer the true internal states of an individual either based on the input signals (i.e., ex-ante prediction) received by the individual or based on the output signals (i.e., ex-post inference) emitted by the individual.

It is worth noting that the proposed framework employs a broader view of “empathy”, compared to the “empathy” commonly studied in psychology and neuroscience. Specifically, it incorporates three distinctive types of empathy: the widely used notion of empathy emphasizing on the inference and perception of others’ transient internal states (e.g., feelings and thoughts), the empathy emphasizing on the inferences of others’ non-transient states (e.g., traits), and the empathy emphasizing on the inferences of others’ responses (Decety and Ickes, 2011). It’s also worth noting that a key characteristic of empathy (human or artificial) is to understand the internal states of others without any explicit explanations from them. This is the main reason why biological audio/visual signals are often used in empathy tasks. Considering the integral role of A/V data in empathy tasks, AE provides a framework for understanding the value of A/V data and incorporating A/V data analytics into marketing research.

Under the proposed framework, the empathy tasks can be classified along two dimensions: first, the focal information processing mechanism. The ex-ante prediction focuses on the top-down information processing and aims to predict how the focal person responds to given signals; The ex-post inference focuses on the bottom-up information processing and aims to infer the focal person’s internal states from the signals emitted by him/her. An example of ex-ante prediction is to predict how a consumer would respond to the voice in an advertisement (e.g., Chattopadhyay et al., 2003). An example of ex-post inference is to infer whether a consumer is happy or not from his/her voice on the phone (e.g., Hall et al., 2014). For researchers, understanding the focal information processing mechanism provides a basis for developing appropriate AE models in different application contexts. For example, if the focus of an AE model is to make predictions on an individual’s responses or internal states, understanding the individual’s decision rules related to the input signals would be important. The analytical methods used in this case should be able to measure the effects of the input signal features on an individual’s responses or internal states. If the focus of an AE model is to make inferences on an individual’s internal states, understanding the meanings of the emitted signals would be important. The analytical methods used in this case should be able to analyze and map the emitted signals to the specific internal state.

Second, the focal internal states. The dynamic empathy tasks aim to make inferences/predictions about a focal person's transient internal states or responses at a specific time, usually from transient signals; The static empathy tasks aim to make inferences/predictions about a focal person's non-transient internal states such as personal or social traits or inherent preferences, usually from non-transient signals. An example of dynamic empathy is when a person or computer model attempts to infer a consumer's emotional states from his/her transient voice or facial expressions (e.g., Hall et al., 2014; Lu et al., 2016). An example of static empathy is when a person or computer model attempts to infer how trustworthy a person is from his/her non-transient facial appearance (e.g., Oosterhof and Todorov, 2009). The distinction between static and dynamic internal states is important in the contexts of AE because it determines the sources of data and external validity of AE models. The dynamic AE models focus on the short-term cognitive process and can only be generalized to other similar settings for the individual. The static AE models on the other hand focus on the long-term cognitive process and can be generalized over time.

Building on the conceptual framework, an empathy model can be defined with four key elements: (1) a human or non-human model A; (2) a focal person B who was exposed to signals S_b and whose internal states I_b or responses R_b are unknown to A; (3) signals observed by or known to A, which could be S_b or $R\ _b$; and (4) the intra-person decision rules f_b used by B in regulating responses to external signals S_b; Depending on the model inputs and outputs, four types of artificial empathy models are often used for business decision making, which are summarized in Table 2.

[Insert Table 2 about here]

Static-Ex-post. Artificial empathy models that belong to this group observe the signals emitted from the focal person/party (e.g., voice, images, facial appearance, physical conditions, etc.) and attempt to make inferences on the non-transient internal states (e.g., stable/inherent preference, traits, physical states, etc.) they experienced. A large body of literature on voice- and face-based trait perception belongs to this group (e.g., Belin et al., 2011; McAleer et al., 2014; Todorov, 2017; Vernon Richard et al., 2014). For example, it's been shown that the face- based perceptions such as competence, likeability, and trustworthiness are predictive of election outcomes and leadership success (Rule and Ambady, 2008; Todorov et al., 2005). Similarly, people make inferences on the speaker's personality or social traits based on the acoustic features of voices (Frühholz and Belin, 2018; Klofstad et al., 2012; Smith et al., 1975; Wang et al., 2021; Zoghaib, 2019). For example, Wang et al. (2021) found that vocal tones in video pitches are used by receivers to make inferences about a persuader's competence. In business contexts, studies have shown that the brand-related images consumers post to the social media websites can be used to make inferences about consumers' perceptions of the brand (Klostermann et al., 2018; Liu et al., 2020) and that images posted by customers reveal their consumption experiences and are predictive of the restaurants' business performance (Zhang and Luo, 2022). Another popular application of the static-Ex-post models is the recruitment and allocation of employees. For example, it has been shown that inferences from the audio and visual cues displayed by job candidates have impacts on the interviewer's judgements in recruitment and correlate with job performance (DeGroot and Motowidlo, 1999; Hickman et al., 2021).

Static-Ex-ante. Artificial empathy models that belong to this group observe the signals received by the focal person/party (e.g., visual contents, products, packages, brand logos, websites etc.) and attempt to make predictions about their impacts on the focal person/party's non-transient internal states as well as the resulting responses (e.g., stable/inherent preference, traits, physical responses, etc.). The static-Ex-ante models are often used by marketing researchers and practitioners to understand consumers' perceptions and preferences for the design of marketing mix, including products, packages, brand logos, and content used in marketing communications, etc. For example, Orth and Malkewitz (2008) and Orth and Crouch (2014) studied how the product package design features such as complexity and prototypical design type, affected consumers' brand impressions. Landwehr et al. (2013) and Liu et al. (2017) developed models to determine the optimum level of complexity and prototypicality in a product. The use of static-Ex-ante models can provide insights on how to design and optimize the content to maximize its effectiveness. For example, Xiao and Ding (2014) built a model to estimate consumers' preferences for the faces displayed in print ads and their impacts on consumers' responses toward the brand. Similarly, Kim et al. (2021) show that consumers' preferences for voice features used in audio ads drive their preferences toward the product/brand. The static-Ex- ante models can also be used to make personalized product/service recommendations. For example, Zhou et al. (2020) modeled users' preference on facial features in the context of dating and networking and matched users based on their preferences.

Dynamic-Ex-post. Artificial empathy models that belong to this group observe the signals emitted from the focal person/party (e.g., voice, images, facial expressions, body movements, etc.) and attempt to make inferences on the focal person/party's transient internal states (e.g., emotional state, contextual preference, physiological states, etc.). The inference of emotions from various signals such as voices, images and videos has been studied extensively in multiple disciplines including psychology, neuroscience, marketing, and computer science (e.g., Bagozzi et al., 1999; Li and Deng, 2020; Scherer, 1995; Young et al., 2020). In marketing, Customers' emotional responses toward the product or content have been used to infer their preferences for the product or content features (Lu et al., 2016; Liu et al., 2018). Studies have shown that the emotional states inferred from consumers' voice and visual signals during the service interactions are predictive of their satisfaction levels and product/service evaluations (Hall et al., 2014; Mattila and Enz, 2002). Singh et al. (2018) show that the visual cues displayed by customers and salespeople in sales interactions can be used to infer the salesperson's effectiveness in handling customer queries. In addition to emotions, one can also infer other types of transient internal states such as deception, uncertainty, or sarcasm from the audio/visual signals emitted by others (Cheang and Pell, 2008; Ekman et al., 1991). For example, Ekman et al. (1991) examined two types of smiles and pitch and found that the pitch increases significantly in deceptive speech. Pon-Barry and Shieber (2011) developed a model to infer speakers' uncertainty from their spoken responses to questions.

Dynamic-Ex-ante. Artificial empathy models that belong to this group observe the signals received by the focal person/party and attempt to make predictions about their impacts on the focal person/party's transient internal states and the resulting responses (e.g., emotional states, contextual preference, physiological responses etc.). In business contexts, the dynamic-Ex-ante model focuses on the selection or design of stimuli (e.g., visual contents, products, packages, brand logos, web- sites etc.) in order to maximize favorable non-transient responses. For example, Liu et al. (2018) developed a model to optimize short movie trailer clips in order to maximize positive customer responses. Li and Xie (2020) and Shin et al. (2020) developed models based on image posts on social media to understand how image features such as image quality, face presence, and visual content affect consumers' responses to the post.

Businesses can benefit from AE models in several ways. First, an individual's non-transient audio and visual signals reflect important information about the individual, and are often used by others to make inferences about the individual's invariant states such as biological (e.g., identity, gender, age, ethnicity), physical (e.g., height, weight), and social traits (e.g., trustworthiness, attractiveness, dominance). Research has shown that people can make inferences on gender, identity, personality (McAleer et al., 2014), and a variety of social traits such as attractiveness, trustworthiness, competence, dominance, and warmth (Hodges-Simeon et al., 2010; Oleszkiewicz et al., 2017) from voice and face (Bruce and Young, 2013). The vocal and facial appearances of communicators, job candidates, product/service providers, and potential friends or partners have significant impacts on individuals' decisions in the contexts of voting, communication, hiring, personal selling, service, social networking and dating (Cunningham et al., 1990; DeGroot and Motowidlo, 1999; Keh et al., 2013; Mayew and Venkatachalam, 2012; McAleer et al., 2014; Schroeder and Epley, 2015; Small and Verrochi, 2009; Xiao and Ding, 2014). For example, a trustworthy-looking face on a website increases the likelihood of attracting investments (Duarte et al., 2012; Rezlescu et al., 2012) and a competent and intelligent-looking face increases the likelihood of being selected as leaders (Todorov et al., 2005; Zebrowitz et al., 2002). The voices of individuals have been found to serve as an important cue for perceptions on traits such as personality, competence, and credibility (Miller et al., 1976; Pon-Barry and Shieber, 2011; Smith et al., 1975; Wang et al., 2021; Weninger et al., 2012). Although the majority of existing literature focuses on facial cues, other types of visual cues emitted by individuals such as hand movement, body language, gaze, and gait may also convey important information about their invariant internal states (Whittle, 2014). In addition, same audio/visual cues may be perceived differently in different contexts, for example, babyfaceness may be associated with negative traits such as immature in one context or positive traits such as trustworthy in another (Berry and McArthur, 1985; Gorn et al., 2008). Thus, it's important for practitioners to understand how these cues are used by their target customers in the specific business contexts they are operating in.

Second, the audio and visual signals emitted by an individual pro- vide important insights into their transient states such as emotional, psychological, and cognitive status and facilitate interpersonal interac- tions and communications (Fischer and Manstead, 2008; Levenson and Ruef, 1992). For example, customer's audio and visual cues displayed during service interaction have been shown to have predictive power of their satisfaction levels (Hall et al., 2014; Mattila and Enz, 2002; Ma and Dubé, 2011). Viewers' audio and visual signals have been used to measure their emotional responses to advertisements (Brown et al., 1998; Nelson and Schwartz, 1979; Nighswonger and Martin, 1981). Un- derstanding viewer's facial expressions while watching video clips has been shown to provide important insights on the design of movie trailers to achieve the best viewer responses (Liu et al., 2018). Lin et al. (2021) found that positive facial expressions displayed by broadcasters in a live stream have positive effects on viewers' emotional states and responses during the live stream. In addition to emotional states, researchers and practitioners may also look at other types of transient internal states such as interestedness, uncertainty, curiosity, level of anxiety, etc.

Third, understanding how the audio and visual stimuli affect customers' internal states and decision making would help the firm in designing marketing mix to maximize the marketing effectiveness. Re- searchers in marketing have looked at a variety of design elements of the marketing mix, including the design of products (Burnap et al., 2021; Liu et al., 2017) and packages (Orth and Malkewitz, 2008; Sundar and Nose- worthy, 2014), brand logos (Dew et al., 2021; Luffarelli et al., 2019a,b), print, audio and video advertisements (Kim et al., 2021; Pieters et al., 2010; Poor et al., 2013; Texeira et al., 2012; Xiao and Ding, 2014), social media marketing content (Jalali and Papatla, 2016), websites (Hauser et al., 2009; Urban et al., 2014), personal selling (Marinova et al., 2018; Pennington, 1968), and in-store marketing (Caldwell and Hibbert, 2002; Chandon et al., 2009; Zhang et al., 2014). It's been proven that changing the audio/visual elements of marketing mix has significant effects on customers' responses to the marketing mix (e.g., Caldwell and Hibbert, 2002; Pieters et al., 2010; Xiao and Ding, 2014). The same logic can also be applied when designing audio/visual-based products/services. For example, using viewers' real-time emotional responses toward movie trailers, Liu et al. (2018) proposed an optimization procedure for producing short clips to promote movies. In addition, it's been shown that the audio and visual signals emitted by the frontline employee have significant effects on customer’s evaluation of the shopping experience and appropriate display of audio and visual signals during service inter- action can improve customers’ satisfaction level (Marinova et al., 2018; Pugh, 2001). With the rise of video-focused channels such as TikTok and livestreaming (Giertz et al., 2021; Lin et al., 2021), practitioners and researchers are increasingly interested in understanding how to design and optimize the audio and visual elements of the video communication process to elicit optimal customer responses in various online business contexts.

Lastly, different people may interpret the same audio and visual stimuli differently (Kim et al., 2021; Schweinberger et al., 2014; Xiao and Ding, 2014). Understanding the differences in the preferences for or responses to audio and visual stimuli among different customers or segments provides important insights for firms to customize the audio/visual elements in their marketing activities or product/service offerings. For example, Xiao and Ding (2014) showed significant variations among different segments of customers on their preferences for faces used in print ads. Kim et al. (2021) identified significant differences among different customer segments in their preferences for the voices used in radio ads. Zhou et al. (2020) proposed an approach to match dating site users by modeling individual users’ preferences for facial features. In all three cases, firms can improve the marketing effective- ness or business performance significantly by choosing the right audio and/or visual stimuli for each customer or customer segment (Kim et al., 2021; Xiao and Ding, 2014; Zhou et al., 2020). Researchers and practitioners may explore ways to customize or optimize the audio and visual elements of marketing mix based on individual (non-transient or transient) customer preferences, which will likely lead to product or service innovation.

3. *METHODS AND TOOLS USED IN A/V ANALYTICS*

It’s much easier for the computer models to process structured data compared to the unstructured audio/visual data. Therefore, the main objective of A/V data analytics is to convert the audio or visual data into a structured form and extract useful information. In this section, we discuss the analytical procedures and review the methods and tools that are commonly used in A/V analytics. The objective of this section is not to provide a detailed guideline but to give an overview of the key techniques and tools that are available and discuss how to choose appropriate methods/tools in different application contexts.

3.1 Audio analytics using voice data

This section aims to examine the human voice as a main source of audio data in the business contexts and provide an overview of various methods for voice analysis. The human voice is one of the primary means to express meaning and feeling, both through linguistic and non-linguistic aspects. The voice can be varied not only between different people but also between circumstances within a person to convey various information, for example, the person's appearance and biological information such as body size, age and gender, paralinguistic information such as personality, emotional state and mood, and even social characteristics of the speaker such as his/her social status (Zhang, 2016). On the listener's side, people are able to make judgments about these aspects based on the voice, regardless of whether they are correct about this inference (Kreiman and Sidtis, 2011). Advances in technology have led to easier data collection, data storage, and analysis. Various fields such as physiology, psychology, and computer science have examined the voice regarding artificial and human empathy, which calls for an inter-disciplinary approach.

Understanding Voice and Audio Signals. Voice is defined by Merriam-Webster Dictionary as "sound produced by vertebrates by means of lungs, larynx, or syrinx, especially sound so produced by human beings." Although voice is generally produced in the form of speech, there are also other forms of voice expression; for example, people may sing, laugh, and cry, and babies make babbling and cooing sounds. Generally, voice and speech are used interchangeably, but it's useful to distinguish the two of them from the researcher's perspective. Speech is defined in Merriam-Webster as "the communication or expression of thoughts in spoken words." In other words, voice is the production of sounds as a result of the movement of the vocal folds, whereas speech is more articulated production of "phonemes," which is the smallest meaningful unit of sound. Humans can produce sounds with their voice from birth, but they need to be trained over time to produce speech that is understandable by other people, which is also related to the community that person is a part of and the language that is spoken within the culture. In the computer science literature, voice recognition usually refers to the method for detecting the source, or the speaker who is talking, whereas speech recognition refers to detecting the content that is being said. In other words, voice recognition answers the question of “who” is speaking, whereas speech recognition answers the question of “what” is being said. Per the purpose of this review, we focus on the non-linguistic aspects of voice instead of the content. The language, or the content aspect of the voice is often discussed in natural language processing or text analytics (for a detailed review on text analytics, please see Berger et al. 2020).

We start with how voice is produced in the human body. We do not go into details here, for a detailed explanation, the reader may refer to the work done by Kreiman and Sidtis (2011, Chap. 2). The human vocal system consists of three parts: the lungs that create air pressure and airflow, the vocal folds that are located in the larynx whose vibration constricts the airflow to form the voice source, and the vocal tract that modifies the voice source and outputs specific sounds. When the lungs and its diaphragm muscles pump out air (i.e., breathing), the vocal tract, which is basically a tube with two flaps (i.e., the vocal folds) located just above the lungs, vibrate to produce sound. When the person is just breathing without intending to speak, the vocal folds are held apart and no sound is produced. To produce sound, the vocal folds vibrate faster or slower, which results in higher or lower pitches. Loudness depends on how fast the air moves past the folds. Then the sound is articulated by a combination of nose, mouth, tongue, and lips to control the flow of air and sound coming from the vocal folds, which produces phonemes.

After the voice comes out of the mouth, it is transmitted in space in the form of longitudinal waves where the air pressures are varied and particles in the surrounding air have vibrational motion which transmits the wave further through the air. This wave can be picked up by human ears to get processed further to be perceived in the brain as sounds. However, for analysis purposes they have to be stored in an analog signal form to be processed by researchers. The analog form of the signal is stored in phonograph or magnetic storage such as old-style LP records or audiotapes, and it can be converted into a digital signal to be stored in digital form that can be processed by computers and digital devices. Because digital computers store and process information in numbers, the digital signal is represented as a series of numbers.

Digital conversion of the signal mainly consists of two simultaneous processes. Because the analog signal is continuous, the idea is to map input values in the analog signals, which is an extremely large set (i.e., a continuous set which basically can have an infinite number of values) to output values in a discrete, smaller, and countable set. First, the continuous signal has to be discretized in time, which is called the sampling procedure. Here, the waveform is sampled at certain time points, usually spaced periodically, and the result is a sequence of real numbers. In the sampling process, the sampling rate or the sampling frequency reflects the periodic spacing, and is defined as the number of samples per second. Usually, frequencies are measured in cycles per second, or hertz (Hz), and a 10kHz sampling rate means that the analog signal is sampled 10,000 times per second. The second operation is quantization, which is the discretization of signal amplitude. The continuous amplitude variation in the analog signal is represented as a series of levels or steps. In other words, this process is to assign a digital amplitude value that is closest to the original amplitude. Each level, or step is called a quantum, thus the name quantization.

Generally, given that audio signals are non-stationary in nature, that is, they change properties very rapidly, they are analyzed in short-terms, where the signal is divided into short-term intervals (i.e., frames), with the premise that the signal is relatively stationary within the frame. The duration of the frame is called the frame length, and can be as short as 5 ms to as long as several hundred ms (Kazama et al. 2010). To account for the rapid change of the signal in the edges of the frames, the frames are selected to overlap and this degree of overlap is called the frame interval. Then, the particular frame is weighted according to a “window,” where the signal is multiplied by a window function that minimizes the signal amplitude outside of the window. Examples of window functions include the rectangular, Hamming, and Hanning windows (Podder et al. 2014).

Audio Signal Processing. Signals, regardless of whether they are in analog or digital form, can be analyzed in the time domain and frequency domain. The two domains are interchangeable by mathematical equations, such as the Fourier transform (continuous Fourier transform and discrete Fourier transform for continuous and discrete time signals, respectively). The two domains are generally used for different purposes of analysis due to differences in characteristics, and each having certain advantages and disadvantages. Figure 2 shows a representation of the two domains.

[Insert Figure 2 about here]

Time-domain audio features are extracted directly from the samples of the audio signal in general. The time-domain is a good representation of the original sound, and because waveforms are continuous, it reflects temporal variations in the signal and is particularly useful when temporal aspects are important in the analysis. However, due to the continuous nature of the signal, waveforms in the time domain are generally difficult to interpret and analyze (Kent and Read 1992). Thus, time-domain features are usually combined with more sophisticated frequency domain analysis.

On the other hand, the spectrum, which is obtained from the time-domain representation using the Fourier transform, allows relatively easy and economical characterization of many important features (e.g., formant frequencies of vowels, energy regions of aperiodic sounds). The spectrum can be used to characterize steady-state events or, with proper sampling, dynamic events such as transitions. On the other hand, analysis may not be straightforward about some properties of interest, such as temporal aspects. Also, spectral analyses can require quite a lot of time and computing resources, but the advancement of technology has helped run over this hurdle (Kent and Read 1992). The Discrete Fourier Transform is one of the most important methods in digital signal processing. Simply put, Fourier showed that periodic waveforms, no matter how complex, can be expressed as the sum of an infinite series of sinusoidal components, which vary in amplitude and phase. Each component is an integer multiple of the fundamental. Essentially, the Fourier transform transforms a periodic amplitude by time waveform into a frequency waveform, i.e., a spectrum, which is a graph of the amplitude of the various frequency components (Gerhard 2003).

Acoustic feature extraction. From the time domain and the frequency domain, various acoustic features, such as the pitch and intensity are extracted, and these acoustic features are used as inputs in various models of voice analysis. To effectively perform different types of analysis, such as speech recognition, personal trait or emotion detection, it is important to determine and extract the features that are appropriate for carrying out the purpose of the method. It has been a withstanding challenge in the voice analysis field to extract relevant and efficient voice features, and numerous researches have proposed and identified a large set of acoustic features for different purposes. Here we introduce some of the key features used in audio analysis.

- *Pitch*, which refers to humans' perception of how high or low the voice is, is one of the most important features of voice. For example, actor Sam Elliot has been known for having a deep, resonant, and "western"-appeal in his voice and has been performing voice-overs for various brands such as Dodge, American Beef Council, Coors, and Doritos. Numerous researches have examined how to extract pitch features. Pitch results from the vibration of the vocal folds during human voice production. As the vocal folds vibrate faster, the speech signal repeats itself more frequently, and humans perceive a higher pitch. For analysis purposes, the problem is to detect the periodicity in the waveform, but because the speech signal is very noisy and changes quality frequently (in other words, it's aperiodic), the task is not easy (Kent and Read 1992). Generally, extracting pitch features involve extracting the fundamental frequency, which is the lowest frequency component, or partial, of the sound signal. Because the partials are harmonically related in a periodic waveform, as shown by Fourier, knowing the lowest frequency will give information about the pitch of the waveform. Pitch features can be extracted in both the time and the frequency domain.
- The *zero-crossing rate* of an audio signal in the time domain is the rate of how the sign changes in the signal from positive to negative, and vice versa. of the signal during the frame. The intuition here is that the more frequent the sign changes in the signal, the more frequent the signal repeats itself. ZCR is known to reflect in a rather coarse way, the pitch characteristics of the signal, but can also be interpreted as a measure of noisiness of a signal.
- *Formants* represent resonances of the vocal tract and estimating their location and frequencies at that location is important in various voice analysis tasks. One generally used method to extract formants is linear predictive coding (LPC), which brings its concept from time series analysis in statistics. Specifically, as we mentioned earlier, the vocal tract responds to modify the voice source and outputs various sounds. Here the vocal tract is analogous to "filters," which change in time, and the parameters of the filter will characterize the voice. LPC is based upon the basic idea that speech does not vary substantially from sample to sample. Thus, a

sample in digitized signals is partly predictable from the immediately preceding samples, and any sample can be closely approximated by a linear function of samples that precede it. The resulting weighted terms in the equation represent the resonances of the vocal tract, and using these values, the spectral envelope can be represented in a compressed form, with a low bit rate.

- *Cepstrum* analysis is a spectral analysis method where the output is the inverse Fourier transform (or sometimes the original Fourier transform) of the log of the magnitude spectrum of the input waveform. The name cepstrum comes from reversing the beginning four letters in the word "spectrum", implying that the spectrum is modified. The variable corresponding to the cepstrum transform is called "quefrency". The intuition for this method comes from the fact that the harmonic spectrum of the signal which results from the Fourier transform of a pitched signal usually consists of regularly spaced peaks. When the log magnitude of a spectrum is taken, these peaks are compressed, and their amplitude is scaled down. The result is a periodic waveform in the frequency domain, and the period of this waveform, i.e., the distance between the peaks, corresponds to the fundamental frequency of the original signal. The peak in the resulting Fourier transform of this waveform represents the period of the original waveform. Thus, the cepstrum analysis allows the researcher to better examine the periodicity (or the rate of change) between multiple frequency bands.
- *Aggregate pitch contour statistics* such as mean, maximum, minimum, and range have been shown to outperform features that describe the pitch shape in explaining human voice emotions. Also, analyzing global statistics for the whole utterance is reported to be more accurate and robust than statistics for shorter speech regions (e.g., voiced segments) (Busso et al. 2009). However, no definite conclusion has been made.
- The *intensity of the voice* refers to how loud or quiet the voice is, and is represented by the intensity of the waveform amplitude. The power of the signal is a common approach to compute the intensity (Giannakopoulos and Pikrakis 2014).
- The *time properties* of the audio signal can be considered. These include the duration of the speech, the relative portion of the voice frames and the silenced frames, and the articulation rate, in other words, how fast the person is talking.

Table 3 provides a list of useful tools for audio data analysis. Given that there are a variety of features that can be adopted for voice analytics, recent research efforts have been focused on selecting an optimal set of voice features (Schuller et al., 2010). Despite the effort, little success has been achieved in obtaining a set of features that performs consistently over different contexts, conditions, and multiple data sets (Eyben et al., 2015). Thus, researchers have applied high-dimensional feature sets that consist of an exhaustive number of acoustic parameters to capture all variances (Tahon and Devillers, 2015). For most machine learning algorithms, such high-dimensional feature sets complicate the learning process, by increasing the likelihood of overfitting and hindering generalization. Moreover, dealing with many acoustic parameters is computationally burdensome and may be difficult to apply with limited resources on a large scale (Eyben et al., 2015). In the meanwhile, some recent research has introduced various methods to process the original, raw speech time signal and obtain results instead of extracting features (Ma et al. 2018). Also, deep learning methods have relaxed constraints in the number of inputs that can be applied in the model.

[Insert Table 3 about here]

3.2 Visual data analysis using image data

Human experiences in the real world are largely visual. Using digital cameras, we can capture information from the real world in digital forms such as images and videos. We begin this section with the key steps in image data analysis, followed by an overview of video data analysis.

Image acquisition. The digital form of visual data is usually represented as a 2-D (e.g., gray images), 3-D (e.g., color images) or 4-D (e.g., color videos) array. It’s worth noting that many factors affect the digitalization process of visual information, such as lighting, camera optics and sensors. For researchers, understanding how digital images and videos are captured before analyzing them can provide useful insights for image processing in later stages. Marketing researchers can obtain images from publicly available sources, including social media sites (e.g., Facebook, LinkedIn, Instagram, etc.), retailing and e-commerce websites (e.g., Amazon, Taobao, etc.), image posts in customer reviews (e.g., Tripadvisor, Yelp, etc.), and product listing platforms (e.g., Rightmove, Airbnb etc.). Companies may have access to more detailed image information about their customers, e.g., customer profile images, photos of customer service encounters or the shopping process, etc.

A digital image is represented as a 2-D or 3-D array in a computer, where each element is called a pixel (short for "picture element"). For a grayscale image, each pixel value represents the brightness/intensity at a point (x, y) in the *m(height)×n(width)* image plane. A color image is usually represented as a *m(height)×n(width)×3* array, whose elements specify the intensity values of each color channel in a color space. Take the widely used RGB color space as an example, each element represents the intensity values of the red, green, and blue color channels. The HSV (Hue, Saturation, Value) color space is also commonly used as it reflects how humans experience color. Other standard color spaces include the XYZ and LAB color space developed by CIE (the International Commission on Illumination), CMY (Cyan, Magenta, Yellow), and YCbCr, where the luminance information is stored as a single component (Y) and chrominance information is stored as two color-difference components (Cb and Cr). Each color space represents color in different ways, which makes them more convenient for certain types of calculation or more intuitive to identify colors. One should choose the color space that's most suitable for the extraction or understanding of image features.

The process of visual data analysis generally involves three steps: (1) image preprocessing. Pre-processing the image (or image frames from a video clip) and converting it into a suitable form for further analysis; (2) feature detection and description. Selecting and extracting useful features from the data; (3) machine learning. Building machine learning models for specific tasks, such as recognition, detection, and classification. Table 3 provides an overview of some useful tools for visual data analysis.

Image preprocessing. Images obtained from the internet usually vary a lot in terms of image quality, size, orientation, lighting conditions, etc. A critical first step in many widely used computer vision techniques (e.g., face recognition, facial expression recognition, object recognition, etc.) is to preprocess the images and make it easier for the machine learning models to analyze. The main objective of the image preprocessing step is to improve the quality of image, enhance the details that are important for later analysis or remove the noises in the images. For example, in order to recognize faces from facial images, the faces need to be aligned by head orientation and normalized to the same size (Xiao and Ding, 2014). Image preprocessing is also widely used in training deep convolutional neural networks (CNNs) with limited sample size as a way to augment image data (Xu and Ding, 2021; Zhang *et al.*, 2022). It's also used in marketing research for creating image stimuli. Ert *et al.* (2016) blurred the faces in Airbnb hosts' profile images and used them as stimuli to study the impact of hosts' profile photos on guests' decisions. Xiao and Ding (2014) replaced faces from print advertisements with faces of models to investigate the effects of facial features on the performance of print advertisements.

The preprocessing of an image usually involves operations to map pixel values from the original image to an improved image, such as geometric transformation (e.g., scaling, cropping, rotation, affine), image filtering and enhancement (e.g., contract adjustment, color enhancement), and noise reduction (e.g., deblurring). Table 4 summarizes the commonly used preprocessing operations and their main applications.

[Insert Table 4 about here]

Commonly used preprocessing techniques include point-based and area-based operators. The point-based operators mainly operate on the input pixel value and/or global image parameters. Examples include 2-D geometric transformations (e.g., resizing, cropping, rotation), brightness adjustments, color correction and transformations, and histogram equalization. Point-based operators are usually the first steps in image preprocessing to improve image quality or normalize images for batch processing, which are important steps in recognition tasks and image classification tasks. The area-based operators use a set of pixel values within the neighborhood of the input pixel to determine its output pixel value. Examples include linear filtering (e.g., box filter, gaussian filter), non-linear filtering (e.g., median filter, bilateral filter), morphological operation (for binary images), and Fourier transform. The area-based operators are often used in preprocessing to remove noises in the image, adding soft blur, enhancing certain features such as edges and corners, and sharpening details of the image. These operations help improve the performance of feature extraction and significantly improve the performance of the image analysis algorithms, especially when dealing with images collected from the internet or in natural environments (Chaki and Dey 2018).

Feature detection and description. This involves detecting and representing the focal image or regions of interest with a set of features, which then can be used in building machine learning models. The visual content can be represented as a hierarchy of features. At the lowest level are the pixels, which carry brightness and/or color information. The next level are local features that are derived from the pixel-level information such as edges, lines, corners, blobs (regions), etc. The next higher level are global features such as texture, shape, color distribution, etc., which are then combined to detect objects and their attributes. At the highest level are the meanings and relationships as perceived by the human. Feature detection and description are essential for many image and video analytics applications, including object recognition, image classification and segmentation/image retrieval, object tracking and motion estimation, etc.

Images often contain a large amount of information, many of which may not be relevant to the focal task. The key to feature detection is to detect the features that can represent the input image and are most relevant to the focal task. For example, in image matching and classification tasks, researchers usually use localized features such as key points and patches, edges, blobs, corners, and lines (Hossein-Nejad *et al.*, 2021). In human detection tasks, researchers have mainly looked at the key body parts, such as torso, head, and limbs (Moeslund *et al.*, 2006). Since the main focus of this monograph is empathy, it’s also worth considering how features would affect viewer’s internal states and drive responses.

Some commonly used image features in marketing research include local features (e.g., key points or small image patches, edges, lines, corners, blobs) and global features (e.g., color pattern, texture, and shape), etc. The main difference between local and global features is that global features capture key aspects of the entire image such as color, texture or shape, whereas local features represent the image through some local structures of the image (e.g., key points, edges, corners, blobs). Global features are easy to compute, suitable for differentiating images/objects that are more distant from each other, e.g., when one wants to classify images of faces from images of trees, in which case, color histograms of the two categories would be very different from each other. Local features generally require more techniques, more computing resources and are more distinctive, thus are suitable for tasks such as image matching, and object recognition (Bianco *et al.*, 2015).

From the perspective of human perception, some image features are more relevant than others (Adaval *et al.*, 2019). For example, it's been shown that color affects customers' perception of product features, feelings, and attention (e.g., Gorn *et al.*, 1997, 2004; Hagtvedt and Brasel, 2017). Shapes have been found to have significant effects on the perception of facial appearance (e.g., Zhou *et al.*, 2020) as well as product and brand attributes (e.g., Folkes and Matta, 2004; Jiang *et al.*, 2016; Krider *et al.*, 2001; Luffarelli *et al.*, 2019b; Raghubir and Krishna, 1999). For example, Zhou *et al.* (2020) developed a method using Fourier transformation to extract the shape features of key facial parts and measured their effects on perceptions and dating choices. Luffarelli *et al.* (2019b) studied the effect of the degree of (a)symmetry in brand logo designs on brand perceptions and brand equity. Studies also show that sizes of brand, text and picture elements can affect customer's attention and responses to print ads (see Peschel and Orquin, 2013; Pieters and Wedel, 2004; Wedel and Pieters, 2008). In addition, higher level image features such as complexity and aesthetics have also been found to affect customers' perceptions of brands and responses to brand's marketing efforts (Henderson *et al.*, 2003; Hoegg *et al.*, 2010; Pieters *et al.*, 2010; Townsend and Shu, 2010). For example, Pieters *et al.* (2010) measured the impact of visual complexity of poster ads on customers' responses along two dimensions: feature complexity and design complexity. Feature complexity was measured using the size of the image file, and design complexity was measured via six proven principles related to the number, shape, color and size of the objects in the ad. Townsend and Shu (2010) studied the impact of product aesthetics in the context of financial decisions. The aesthetics was manipulated through the design of financial documents along aspects such as number of images and colors. Table 5 summarizes the commonly used image features in marketing and corresponding detection and representation techniques.

[Insert Table 5 about here]

3.3 Visual data analysis using video data

Video has become a rich, rapidly growing source of high value, high volume data (Kang *et al.*, 2019). Depending on the creator and purpose of video data, videos can be broadly categorized into four groups: video content created by firms to assist marketing communication; video content created by content providers to be consumed by others; video content created and shared by customers; as well as video recordings of natural lives (e.g., surveillance videos) (Xiao *et al.*, 2013).

Videos carry rich information about the people, environments and incidences that are depicted in them. With the abundance and accessibility of video data and the increase in computation power, video analytics have found their value in a wide range of applications, including but not limited to enhancing customer experience, human computer interaction, access control, anomaly detection, autonomous driving, etc., in numerous industries, such as retailing, transportation, security, healthcare, military, etc. (Connell *et al.*, 2013; Liu *et al.*, 2013). More and more researchers have used video data in business decision making such as inferring customer preference (Lu *et al.*, 2016), designing trailers (Liu *et al.*, 2018), improving in-store marketing effectiveness (Singh *et al.*, 2018; Zhang *et al.*, 2014), monitoring customer responses during the consumption of online content (Zhang *et al.*, 2020), and understanding the effectiveness of marketing efforts (Li *et al.*, 2019b).

Video data analysis in the marketing domain can be generally grouped into three categories according to the focal unit of analysis. The first group focuses on analyzing the video at the frame level and aggregating the frame-level features in the analysis. For example, Li *et al.* (2019b) measured the variation among frames in crowdfunding videos and analyzed its impact on the crowdfunding success rate. Rajaram and Manchanda (2020) analyzed the content of frames extracted at different time point of influencers' videos and measured the impact of the content design on the effectiveness of the influencers' videos. The second group focuses on analyzing the temporal changes in video frames in order to understand the temporal video features such as motion and transitions in scenes. For example, Liu *et al.* (2018) and Zhang *et al.* (2020) used temporal segmentation tools to detect the transitions between scenes and shots in movies or movie trailers and extract video features at the scene or shot level. A third group of video data analysis focuses on the dynamics of objects or events depicted in the video. For example, Liu *et al.* (2018) used a moment-to-moment emotion identification tool to detect facial expressions based on facial movements. Lu *et al.* (2016) tracked the movement of hands in video recordings of customers' garment evaluation process. The image analytical methods that we previously discussed also apply to video data analysis in the first two groups as they mainly focus on the global or local features extracted at the frame level.

Here we'd like to focus on video-based tracking as it's one of the fundamental video analytical methods that are relevant to marketing and has yet to realize its potential in marketing research. Video-based tracking has many applications in the marketing domain, such as video surveillance, product/service evaluation, augmented reality and human–computer interface. For example, a retailing store may want to track individual customers' shopping path, shopping behaviors and interactions with salespeople or other customers (Zhang *et al.*, 2014). One may also want to track a customer's product or service evaluation process (Lu *et al.*, 2016). In the context of augmented reality and human–computer interfaces such as smart fitness mirrors, one may want to track the body movements of customers in order to provide feedback.

Generally speaking, video-based tracking involves monitoring an object's spatial and temporal changes in a video sequence. Tracking generally involves the following steps：object detection, representation, and tracking. This process could apply to the tracking of a single object or multiple objects in a video stream. Most algorithms proposed in the literature focus on the tracking of a single object, and in practice tracking of multiple objects can be processed as running multiple single-object trackers in a simultaneous and independent way.

There are two ways to track an object, one is tracking by detection, which develops a model of the object and identifies it in each frame. Another is tracking by matching, which first predicts how the object moves, and searches for the domains in the next frame that matches with it. The main difference between the two lies in the tracking stage.

Object detection. The object detection can be done at either the frame level or the video level. One key difference between the two is whether the detection utilizes information of the adjacent frames. Video-based object tracking involves matching detected objects between consecutive frames using different features of the target/object, such as motion, velocity, color, texture, with the purpose to generate the trajectory of the moving object by locating its position in each frame. Some of the commonly used video-based object detection techniques include background subtraction, temporal differencing, and optical flow (Forsyth and Ponce 2011). Background subtraction aims to distinguish the moving object by subtracting its background scene image. It starts with building a representation of the background scene, called the background model, then detecting any significant change in an image region from the background model, which signifies a moving object and will be identified for each incoming frame. Background subtraction is the most frequently used among these three in practice because of its computational effectiveness and reasonable accuracy, although it is sensitive to dynamic scene changes such as illumination variations and extraneous events (Forsyth and Ponce 2011; Joshi and Thakore 2012). Temporal differencing methods utilize the pixel-wise difference between several consecutive frames to detect moving objects (Joshi and Thakore 2012). Optical flow methods track the object by estimating the distributions of apparent velocities of movement of objects between two video frames. It's often used to characterize and quantify the motion of objects in videos (Fortun, Bouthemy, and Kervrann 2015).

Object representation. Object representation focuses on describing the detected focal object or behaviors using visual features. Objects detected from videos can be represented using shape-based (e.g., points, boxes, blobs, silhouettes, etc.), color-based (e.g., color histogram), motion-based (e.g., optical flow), or texture-based descriptors (Forsyth and Ponce 2011; Joshi and Thakore 2012).

Object tracking. The following tracking stage focuses on segmenting the focal object from the video frames either by building a discriminative classifier in order to classify the target candidates as the tracked object or background (i.e., tracking by detection) or by matching visual features between the tracked object and a candidate region in consecutive frames (i.e., tracking by matching). The tracking methods can be classified into three types: point, kernel, and silhouette-based tracking (Yilmaz et al., 2006). Point-based tracking segments the focal object according to its feature points. Kernel-based tracking segments the focal object according to the appearance/shape of the object. Silhouette-based tracking segments the focal object based on the information encoded within the region of the object. Compared to the other two, kernel-based tracking method has been used more often due to its high accuracy and low computational cost (Forsyth and Ponce 2011).

3.4 Machine learning models in A/V data analytics

Machine learning models can be classified into three groups: supervised, unsupervised and reinforcement learning. Typical uses of machine learning models in A/V data analytics focus on pattern recognition, segmentation, and classification tasks. The vast literature on machine learning techniques is beyond the scope of this review (for detailed reviews on classic machine learning and deep learning models, please see Brei 2020; Voulodimos et al. 2018). In this review we focus on discussing when and how to choose the right machine learning approaches for different research contexts using A/V data. Table 6 summarizes the three main machine learning approaches, commonly used techniques, and main applications in A/V data analytics.

[Insert Table 6 about here]

Generally speaking, supervised learning is the most commonly used model in A/V analytics. It usually involves paired input-output training dataset, learning the best mapping between the inputs and output values and using that to predict the output values for new inputs. Classic models in supervised learning require hand-crafted features as well as paired input-output training dataset and have been mainly used for classification, object detection and recognition tasks such as voice or face recognition, emotion or facial expression recognition, image classification, etc. Some of the popular labeled datasets used in A/V data analytics include: ImageNet (https://www.image-net.org/), COCO (https://cocodataset.org/), LVIS (https://www.lvisdataset.org/), and Speech Commands dataset (https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html).

In recent years, deep learning models have significantly advanced the field of A/V data analytics, especially in the domain of supervised learning (Pouyanfar et al. 2018; Voulodimos et al. 2018). Deep learning models have been shown to achieve superior performance than the classic machine learning models in many image analytics tasks such as image classification, image segmentation, object recognition and tracking (Voulodimos et al. 2018), as well as in voice recognition and inferences tasks (Shen et al. 2021). The deep learning models that are frequently used for visual data analysis include AlexNet, Visual Geometry Group network (VGG-Net), residual neural network (ResNet), high-resolution net (HRNet), recurrent neural networks (RNNs), etc. (Yan et al. 2020). In marketing contexts, supervised deep learning models have been used in understanding brand image (Liu, Dzyabura, and Mizik 2020) and the impact of video content (Li, Shi and Wang 2019), predicting product return rates (Dzyabura et al. 2019), predicting social perceptions from faces (Messer and Fausser 2019), and optimizing product design (Burnap et al. 2021).

Unsupervised learning in the domain of A/V data analytics mainly aims to reduce the dimensionality of the input data, or to characterize a given dataset by identifying the underlying patterns or relations. It has been mainly used for segmentation and clustering tasks. Compared to supervised deep learning models, unsupervised deep learning models received less attention from the marketing researchers. Dzyabura and Peres (2021) used unsupervised learning in analyzing 4,743 collages for 303 brands in order to understand brand associations. One main area of application in marketing is the generation of A/V data and stimuli using the deep generative models such as variational autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAE and GAN models can generate synthetic A/V data without labels associated with the dataset. A VAE model takes input data and compresses or reduces its dimension to a latent space (i.e., encoding) and later tries to reproduce the original data as accurately as possible (i.e., decoding) (Kingma and Welling 2013). A GAN model automatically discovers and learns the patterns and regularities in input data so that the model can generate new output that resembles one that has been drawn from the original dataset (Goodfellow et al. 2014). One can take a preferable voice or image that they want to benchmark and generate new A/V data using these models. For example, Dew et al. (2021) applied VAEs to generate logos that incorporate desirable aspects of a brand. Burnap et al. (2021) used VAEs and GANs to generate product designs that appeal to consumers.

Reinforcement learning involves learning a sequence of actions in an interactive environment by maximizing the expected reward (Sutton and Barto 2018). It has mainly been used for solving feature/model selection problems for image classification, object detection, visual tracking, and navigation (Bernstein and Burnaev 2018). Deep reinforcement learning has been used to improve the performance of object detection and tracking, image segmentation, and video analysis (for a detailed review of deep reinforcement learning models using visual data, see Le et al. 2021). Marketing applications of reinforcement learning include robotics, games, and self-driving cars.

Two approaches that have been increasingly applied in A/V analytics are transfer learning and representation learning. Both techniques aim to improve the efficiency of learning tasks. Transfer learning is a technique where a model that is trained and stored for one task is applied as is or fine-tuned for another task at hand. It has become a popular approach in A/V data analytics to utilize the knowledge obtained from existing deep learning models and save resources and improve efficiency in training new models. Commonly used models in transfer learning tasks are object detection models such as VGGNet (Simonyan and Zisserman 2014) and ResNet (He et al. 2016). For example, Zhang et al. (2022) fine-tuned VGG16 to detect objects to label their Airbnb images. Shen et al. (2021) tested various transfer learning models such as VGG16 and ResNet50 to infer uncertainty from human voice. Representation learning refers to a set of techniques used to learn representations of the data in order to extract useful features (as oppose to hand-crafted features) for specific machine learning tasks such as speech recognition, object detection or classification (Bengio et al. 2013). Representations can be learned from either labeled input data or unlabeled input data. For example, Dew et al. (2021) used unsupervised representation learning methods to understand the relations between visual features of brand logos and customers' perceptions.

A number of open-source platforms are available for researchers who are interested in A/V data analytics. TensorFlow, Keras, Caffe, Microsoft Cognitive Toolkit, and PyTorch are the popular platforms among others (Pouyanfar et al. 2018). Table 7 gives an overview of the deep learning models applied in the marketing contexts. The majority of these studies use output features either from pre-trained models or application programming interfaces (APIs). Some of the popular APIs include Google Cloud Vision API (Klostermann et al. 2018; Li and Xie 2020; Nanne et al. 2020), Clarifai (Dzyabura and Peres 2021; Nanne et al. 2020; Zhang and Luo 2022), YOLOV2 (Nanne et al. 2020), and Amazon Rekognition (Wang et al 2020). Some studies use fine-tuned pre-trained models or develop a new deep learning model for the focal task, for example, Messer and Fausser (2019) developed a CNN model to predict the social perception from faces. Nanne et al. (2020) compared the performances of Google Cloud Vision, Clarifai, and YOLOV2 in the context of understanding user generated image content. By using 21,738 user generated Instagram images related to 24 different brands, they tested the three APIs on a dataset of 21,738 Instagram pictures related to 24 different brands and found that Google Cloud Vision outperforms the other APIs in detecting objects, whereas Clarifai provides more useful image tags for the understanding of the users' portrayal of a brand, and YOLOV2 is outperformed by both in analysing brand-related user generated images.

[Insert Table 7 about here]

4. TRENDS AND FUTURE DEVELOPMENTS OF A/V ANALYTICS IN PRACTICE

The use of audio and visual data analytics is reshaping many aspects and many sectors of the business world. The use of A/V analytics benefits businesses through at least four ways. First, A/V analytics provide new tools for creation of new customer values. For example, in the insurance industry, Lapetus Solutions (https://www.lapetussolutions.com/solutions/chronos/) develops a machine-learning model using face analytics to predict people's age and estimate longevity from facial images. This allows the insurance companies to provide efficient and cost-effective life insurance underwriting within 10minutes. In the fashion and beauty industry, companies such as ModiFace (http://modiface.com/) uses 3-D facial micro-feature tracking technology to track facial movements and expressions in real-time through 68 non-identifying parameters, including key facial features such as lip, eye edges, iris size and location, and skin texture features such as spots, texture, and wrinkles. Using the face and video analytics, it also builds Augmented Reality (AR) applications such as Sephora Virtual Artist, E-commerce AR, and in-store AR mirrors, which allow customers to virtually try on beauty products. A similar app called 'Makeup Genius' launched by L'Oreal brought the brand 20 million app users and 60,000,000 virtual product trials in one year (Fortune 2021). In the smart home and security sector, A/V analytics have been used by companies such as Google (https://store.google.com/#alert-types-more), Honeywell (https://www.resideo.com/us/en/honeywell-home-app/), Ring (http://ring.com/) and

IntelliVision (https://www.intelli-vision.com/) to provide intelligent surveillance and control service to users by automatically detecting sounds and movements, and recognizing objects.

Second, A/V analytics can improve the effectiveness of value communication. With the audio and visual content stimuli, firms can develop ex-ante artificial empathy models to predict and optimize customers' responses to the marketing content. With the audio and visual information of customer responses, firms can develop ex-post artificial empathy models to infer customers' internal states. For example, using speech analysis and computer vision techniques, Affectiva (https://www.affectiva.com/) helps brands optimize the video content by inferring the emotional states of a viewer as they watch the videos. It has been used by more than 1400 brands, including MARs, Kellogg's and CBS, in optimizing the marketing effectiveness.

Third, A/V analytics can improve the efficiency and effectiveness of value delivery. One key advantage of applying automatic audio and video analytics is to extract useful information more efficiently and more effectively to support business decision-making. This is increasingly important for both online and offline businesses amid the rising labor cost and growing data volume. Allgovision (https://www.allgovision.com/business-intelligence.php) uses video analytics to detect the number of customers and monitor the queue in order to improve the service efficiency. Gorilla (https://www.gorilla-technology.com/Video-Analytics) uses deep learning-based video analytics and IoT to analyze real-time video feeds for in-store people tracking and activity detection. A speech synthesis system called WinkTalk adapts the synthetic voice styles according to listener's facial expressions (Székely et al. 2014), which can be potentially useful for computer-assisted customer service.

Finally, A/V analytics can improve the efficiency and effectiveness of employee recruitment and allocation. Hirevue (https://www.hirevue.com/), for example, developed machine learning models to assess applicants' skills and match with job positions by analyzing the audio and visual information from their interviews.

The trend of using A/V analytics to better create, communicate and deliver value to customers has been further accelerated in the last two years as much of the world was forced to go online due to the COVID-19 pandemic. With the increasing use of online channels in the business sector, more and more customer interactions happen in an environment where firms have less control of. The increasing use of mobile channels and developments in technologies such as augmented reality (Yaoyuneyong et al. 2016) poses new challenges as well as opportunities for marketers. The access to A/V data and analytical tools not only gives the firms "eyes" and "ears" but also "keys" to unlock the benefits of analytics-based decision making. For both online and offline businesses, the A/V analytics provide innovative ways to gain better customer insights, enhance customer experience, understand customer preference, discover unmet needs and optimize marketing effectiveness.

REFERENCES

Adaval, R., G. Saluja, and Y. Jiang (2019). Seeing and thinking in pictures: A review of visual information processing. *Consumer Psychology Review*. 2(1): 50–69.

Bagozzi, R. P., M. Gopinath, and P. U. Nyer (1999). The role of emotions in marketing. *Journal of the Academy of Marketing Science*. 27(2): 184–206.

Balducci, B. and D. Marinova (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*. 46(4): 557–590.

Belin, P., P. E. Bestelmeyer, M. Latinus, and R. Watson (2011). Understanding voice perception. *British Journal of Psychology*. 102(4): 711–725.

Bengio, Y., A. Courville, and P. Vincent (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 35(8): 1798–1828.

Berger, J., A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*. 84(1): 1–25.

Bernstein, A. V. and E. V. Burnaev (2018). Reinforcement learning in computer vision. *Tenth International Conference on Machine Vision (ICMV 2017)*. 10696: 458–464.

Berry, D. S. and L. Z. McArthur (1985). Some components and consequences of a babyface. Journal of Personality and Social Psychology. 48(2): 312.

Bharadwaj, N., M. Ballings, P. A. Naik, M. Moore, and M. M. Arat (2022). A new livestream retail analytics framework to assess the sales impact of emotional displays. Journal of Marketing. 86(1): 27–47.

Bianco, S., D. Mazzini, D. Pau, and R. Schettini (2015). Local detector and compact descriptors for visual search: A quantitative comparison. Digital Signal Process. 44: 1–13.

Brei, V. A. (2020). Machine learning in marketing: Overview, learning strategies, applications, and future developments. Foundations and Trends® in Marketing. 14(3): 173–236.

Brown, S. P., P. M. Homer, and J. J. Inman (1998). A meta-analysis of relationships between ad-evoked feelings and advertising responses. Journal of Marketing Research. 35(1): 114–126.

Bruce, V. and A. Young (2013). Face Perception. Psychology Press.

Bruner, G. C. (1990). Music, mood, and marketing. Journal of Marketing. 54(4): 94–104.

Buie, D. H. (1981). Empathy: Its nature and limitations. Journal of the American Psychoanalytic Association. 29: 281–307.

Burnap, A., J. R. Hauser, and A. Timoshenko (2021). Design and evaluation of product aesthetics: A human-machine hybrid approach. Available at SSRN 3421771.

Busso, C., M. Bulut, S. Lee, and S. Narayanan (2009). Fundamental Frequency Analysis for Speech Emotion Processing. The Role of Prosody in Affective Speech. Berlin, Germany: Peter Lang Publishing Group. 309–337.

Caldwell, C. and S. A. Hibbert (2002). The influence of music tempo and musical preference on restaurant patrons’ behavior. Psychology and Marketing. 19(11): 895–917.

Chaki, J. and N. Dey (2018). A Beginner’s Guide to Image Preprocessing Techniques. CRC Press.

Chandon, P., J. W. Hutchinson, E. T. Bradlow, and S. H. Young (2009). Does in-store marketing work? Effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase. *Journal of Marketing*. 73(6): 1–17.

Chattopadhyay, A., D. W. Dahl, R. J. B. Ritchie, and K. N. Shahin (2003). Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. *Journal of Consumer Psychology*. 13(3): 198–204.

Cheang, H. S. and M. D. Pell (2008). The sound of Sarcasm. *Speech Communication*. 50(5): 366–381.

Clarke, J. S., J. P. Cornelissen, and M. P. Healey (2019). Actions speak louder than words: How figurative language and gesturing in entrepreneurial pitches influences investment judgments. *Academy of Management Journal*. 62(2): 335–360.

Connell, J., Q. Fan, P. Gabbur, N. Haas, S. Pankanti, and H. Trinh (2013). Retail video analytics: An overview and survey. In: *Video Surveillance and Transportation Imaging Applications*. Vol. 8663.

Cunningham, M. R., A. P. Barbee, and C. L. Pike (1990). What do women want? Facial metric assessment of multiple motives in the perception of male facial physical attractiveness. *Journal of Personality and Social Psychology*. 59(1): 61–72.

Decety, J. and P. L. Jackson (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*. 3(2): 71–100.

Decety, J. and C. Lamm (2006). Human empathy through the lens of social neuroscience. *The Scientific World Journal*. 6: 1146–1163.

Decety, J. and W. Ickes (eds.) (2011). *The Social Neuroscience of Empathy*. MIT Press.

DeGroot, T. and S. J. Motowidlo (1999). Why visual and vocal interview cues can affect interviewers' judgments and predict job performance. *Journal of Applied Psychology*. 84(6): 986–993.

DeShields, J., W. Oscar, K. Ali, and K. Erdener (1996). Source effects in purchase decisions: The impact of physical attractiveness and accent of salesperson. *International Journal of Research in Marketing*. 13(1): 89–101.

Dew, R., A. Ansari, and O. Toubia (2021). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science.* Articles in Advance.

Ding, M. (2007). A theory of intraperson games. *Journal of Marketing*. 71(2): 1–11.

Duan, C. and C. E. Hill (1996). The current state of empathy research. *Journal of Counseling Psychology*. 43(3): 261.

Duarte, J., S. Siegel, and L. Young (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*. 25(8): 2455–2484.

Dzyabura, D., S. El Kihal, J. R. Hauser, and M. Ibragimov (2019). Leveraging the power of images in managing product return rates. Available at SSRN 3209307.

Dzyabura, D. and R. Peres (2021). Visual elicitation of brand perception. *Journal of Marketing*. 85(4): 44–66.

Ekman, P., M. O’Sullivan, W. V. Friesen, and K. R. Scherer (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*. 15(2): 125–135.

Ert, E., A. Fleischer, and N. Magen (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*. 55: 62–73.

Eslami, M., K. Vaccaro, K. Karahalios, and K. Hamilton (2017). Be careful; things can be worse than they appear: Understanding biased algorithms and users’ behavior around them in rating platforms. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11.

Eyben, F., K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, and C. Busso, . . . and K. P. Truong (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*. 7(2): 190–202.

Fischer, A. H. and A. S. R. Manstead (2008). Social functions of emotion. In: *Handbook of Emotions*. Ed. by M. Lewis, J. Haviland-Jones, and L. F. Barrett. 3rd edition. New York, NY: Guilford Press. 456–468.

Folkes, V. and S. Matta (2004). The effect of package shape on consumers.’ judgments of product volume: Attention as a mental contaminant. *Journal of Consumer Research*. 31: 390–401.

Forsyth, D. and J. Ponce (2011). *Computer Vision: A Modern Approach*. Prentice Hall.

Fortun, D., P. Bouthemy, and C. Kervrann (2015). Optical flow modelling and computation: A survey. *Computer Vision and Image Understanding*. 134: 1–21.

Fortune (2021). A.I. in the beauty industry: How the pandemic finally made consumers care about it. url: https://fortune.com/2021/01/11/ai-artificial-intelligence-personalized-beauty-cosmetics-brainstorm-reinvent/ Retrieved on 8 Oct. 2021.

Frühholz, S. and P. Belin (eds.) (2018). *The Oxford Handbook of Voice Perception*. Oxford University Press.

Gerhard, D. (2003). *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Regina, SK, Canada: Department of Computer Science, University of Regina. 0–22.

Giannakopoulos, T. and A. Pikrakis (2014). *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press.

Giertz, J. N., W. H. Weiger, M. Törhönen, and J. Hamari (2021). “Content versus community focus in live streaming services: How to drive engagement in synchronous social media”. *Journal of Service Management*. 33(1): 33–58.

Goldman, A. (1993). “Ethics and cognitive science”. *Ethics*. 103: 337–360.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair, . . . and Y. Bengio (2014). “Generative adversarial nets”. *Advances in Neural Information Processing Systems*. 27.

Gorn, G. J., A. Chattopadhyay, J. Sengupta, and S. Tripathi (2004). “Waiting for the web: How screen color affects time perception”.*Journal of Marketing Research*. 41(2): 215–225.

Gorn, G. J., A. Chattopadhyay, T. Yi, and D. W. Dahl (1997). “Effects of color as an executional cue in advertising: They’re in the shade”. *Management Science*. 43(10): 1387–1400.

Gorn, G. J., Y. Jiang, and G. V. Johar (2008). “Babyfaces, trait inferences, and company evaluations in a public relations crisis”. *Journal of Consumer Research*. 35(1): 36–49.

Guan, Y., Y. Tan, Q. Wei, and G. Chen (2020). “Information or distortion? The effect of customer generated images on product rating dynamics”. *Working Paper*.

Hagtvedt, H. and S. A. Brasel (2017). “Color saturation increases perceived product size”. *Journal of Consumer Research*. 44(2): 396–413.

Hall, J. A., P. Verghis, W. Stockton, and J. X. Goh (2014). “It takes just 120 seconds: Predicting satisfaction in technical support calls”. *Psychology and Marketing*. 31(7): 500–508.

Hartmann, J., M. Heitmann, C. Schamp, and O. Netzer (2021). “The power of brand selfies”. *Journal of Marketing Research*. 58(6): 1159–1177.

Hauser, J. R., G. L. Urban, G. Liberali, and M. Braun (2009). “Website morphing”. *Marketing Science*. 28(2): 202–223.

He, K., X. Zhang, S. Ren, and J. Sun (2016). “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

Henderson, P. W., J. A. Cote, S. M. Leong, and B. Schmitt (2003). “Building strong brands in Asia: Selecting the visual components of image to maximize brand strength”. *International Journal of Research in Marketing*. 20: 297–313.

Hennig-Thurau, T., M. Groth, M. Paul, and D. D. Gremler (2006). “Are all smiles created equal? How emotional contagion and emotional labor affect service relationships”. *Journal of Marketing*. 70(3): 58–73.

Hickman, L., N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo (2021). “Automated video interview personality assessments: Reliability, validity, and generalizability investigations”. *Journal of Applied Psychology.* Articles in Advance.

Hodges-Simeon, C. R., S. J. Gaulin, and D. A. Puts (2010). “Different vocal parameters predict perceptions of dominance and attractiveness”. *Human Nature*. 21(4): 406–427.

Hoegg, J., Alba, J. W., and Dahl, D. W. (2010). The good, the bad, and the ugly: Influence of aesthetics on product feature judgments. *Journal of Consumer Psychology*, 20, 419–430.

Hossein-Nejad, Z., Agahi, H., and Mahmoodzadeh, A. (2021). Image matching based on the adaptive redundant keypoint elimination method in the SIFT algorithm. *Pattern Analysis and Applications*, 24(2), 669-683.

Hughes, C., Swaminathan, V., and Brooks, G. (2019). Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns. *Journal of Marketing*, *83*(5), 78-96.

Hui, S. K., Fader, P. S., and Bradlow, E. T. (2009). Path data in marketing: An integrative framework and prospectus for model building. *Marketing Science*, 28(2), 320-335.

Hui, Sam K., Yanliu Huang, Jacob Suher, and Jeffrey Inman (2013). Deconstructing the 'First Moment of Truth': Understanding Unplanned Consideration and Purchase Conversion Using In-Store Video Tracking. *Journal of Marketing Research*, 50 (4), 445–62.

Hul, M. K., Dube, L., and Chebat, J. C. (1997). The impact of music on consumers' reactions to waiting for services. *Journal of retailing*, 73(1), 87-104.

Ickes, W. (1997). *Empathic Accuracy*. The Guilford Press, New York.

Jalali, N. Y., and Papatla, P. (2016). The palette that stands out: Color compositions of online curated visual UGC that attracts higher consumer interaction. *Quantitative Marketing and Economics*, 14(4), 353-384.

Jia, H., Kim, B. K., and Ge, L. (2020). Speed Up, Size Down: How Animated Movement Speed in Product Videos Influences Size Assessment and Product Evaluation. *Journal of Marketing*, *84*(5), 100-116.

Jiang, L., Yin, D., and Liu, D. (2019). Can joy buy you money? The impact of the strength, duration, and phases of an entrepreneur's peak displayed joy on funding performance. *Academy of Management Journal*, 62(6), 1848-1871.

Jiang, Y., Gorn, G. J., Galli, M., and Chattopadhyay, A. (2016). Does your company have the right logo? How and why circular-and angular-logo shapes influence brand attribute judgments. *Journal of Consumer Research*, 42, 709–726.

Joshi, Kinjal A., and D. G. Thakore (2012). A survey on moving object detection and tracking in video surveillance system. *International Journal of Soft Computing and Engineering*, 2(3), 44-48.

Kachur, A., Osin, E., Davydov, D., Shutilov, K., and Novokshonov, A. (2020). Assessing the Big Five personality traits using real-life static facial images. *Scientific reports*, 10(1), 1-11.

Kahle, Lynn R., and Pamela M. Homer (1985). Physical Attractiveness of the Celebrity Endorser: A Social Adaptation Perspective, *Journal of Consumer Research*, 11 (4), 954-961.

Kang, Daniel, Peter Bailis, and Matei Zaharia (2019). Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine. *In CIDR.*

Kazama, M., S. Gotoh, M. Tohyama and T. Houtgast (2010). On the significance of phase in the short term Fourier spectrum for speech intelligibility. *The Journal of the Acoustical Society of America* 127(3), 1432-1439.

Keh, H. T., Ren, R., Hill, S. R., and Li, X. (2013). The beautiful, the cheerful, and the helpful: The effects of service employee attributes on customer satisfaction. *Psychology & Marketing*, *30*(3), 211-226.

Kent, R. D., and Read, C. (1992). *The Acoustic Analysis of Speech*. London: Whurr.

Kim, H. J., Wang, Y., and Ding, M. (2021). Brand Voiceprint. *Customer Needs and Solutions*, 1-14.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Klofstad, C. A., Anderson, R. C., and 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.

Klostermann, J., Plumeyer, A., Böger, D., and Decker, R. (2018). Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, 35(4), 538-556.

Kreiman, J. and D. Sidtis (2011). Voices and listeners: Toward a model of voice perception. *Acoustics Today*, 7(4), 7-15.

Krider, R. E., Raghubir, P., and Krishna, A. (2001). Pizzas: π or square? Psychophysical biases in area comparisons. *Marketing Science*, 20(4), 405-425.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., and Kannan, P. K. (2016). From social to sale: The effects of firm-generated content in social media on customer behavior. *Journal of marketing*, *80*(1), 7-25.

Landwehr, J. R., Wentzel, D., and Herrmann, A. (2013). Product design for the long run: Consumer responses to typical and atypical designs at different stages of exposure. *Journal of Marketing*, 77(5), 92–107.

Le, N., Rathour, V. S., Yamazaki, K., Luu, K., and Savvides, M. (2021). Deep reinforcement learning in computer vision: a comprehensive survey. *Artificial Intelligence Review*, 1-87.

Levenson, R. W., and Ruef, A. M. (1992). Empathy: a physiological substrate. *Journal of Personality and Social Psychology*. 63, 234–246.

Li, H., Simchi-Levi, D., Wu, M. X., and Zhu, W. (2019a). Estimating and Exploiting the Impact of Photo Layout in the Sharing Economy. Available at *SSRN 3470877*.

Li, S., and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing.1-20.*

Li, X., Shi, M., and Wang, X. S. (2019b). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, 36(2), 216-231.

Li, Y., and Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1), 1-19.

Liaukonyte, J., Teixeira, T., and Wilbur, K. C. (2015). Television advertising and online shopping. *Marketing Science*, 34(3), 311-330.

Lin, Y., Yao, D., and Chen, X. (2021). Happiness begets money: Emotion and engagement in live streaming. *Journal of Marketing Research*, *58*(3), 417-438.

Liu, H., S. Chen, and N. Kubota (2013). Intelligent video systems and analytics: A survey. *IEEE Transactions on Industrial Informatics*, 9(3), 1222-1233.

Liu, L., Dzyabura, D., and Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4), 669-686.

Liu, X., S. W. Shi, T. Teixeira, and M. Wedel (2018). “Video content marketing: The making of clips”. *Journal of Marketing*. 82(4): 86–101.

Liu, Y., Li, K. J., Chen, H., and Balachander, S. (2017). The effects of products' aesthetic design on demand and marketing-mix effective- ness: the role of segment prototypicality and brand consistency. *Journal of Marketing*, 81(1), 83–102.

Lu, S., L. Xiao, and M. Ding (2016), A Video-Based Automated Recommender (VAR) System for Garments, *Marketing Science*, 35 (3), 484–510.

Lu, S., Yao, D., Chen, X., and Grewal, R. (2021). Do larger audiences generate greater revenues under pay what you want? evidence from a live streaming platform. *Marketing Science*, *40*(5), 964-984.

Luffarelli, J., Stamatogiannakis, A., and Yang, H. (2019). The visual asymmetry effect: An interplay of logo design and brand personality on brand equity. *Journal of Marketing Research*, 56(1), 89-103.

Luffarelli, J., Mukesh, M., and Mahmood, A. (2019). Let the logo do the talking: The influence of logo descriptiveness on brand equity. *Journal of Marketing Research*, 56(5), 862-878.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., and Cai, L. (2018). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. In *Interspeech.* 3683-3687.

Ma, Z., and Dubé, L. (2011). Process and outcome interdependency in frontline service encounters. *Journal of Marketing*, 75(3), 83–98.

Malik, N., Singh, P. V., Lee, D. D., and Srinivasan, K. (2019). A Dynamic Analysis of Beauty Premium. Available at *SSRN 3208162.*

Marinova, D., Singh, S. K., and Singh, J. (2018). Frontline problem-solving effectiveness: A dynamic analysis of verbal and nonverbal cues. *Journal of Marketing Research*, 55(2), 178-192.

Mattila, A. S., and Enz, C. A. (2002). The role of emotions in service encounters. *Journal of Service research*, 4(4), 268-277.

Mayew WJ, Venkatachalam M (2012). The Power of Voice: Managerial Affective States and Future Firm Performance. *Journal of Finance*, 67(1),1–43.

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one* 9(3), e90779.

Mei, T., Hua, X. S., Yang, L., and Li, S. (2007). VideoSense: towards effective online video advertising. *In Proceedings of the 15th ACM international conference on Multimedia*, 1075-1084.

Messer, U., and Fausser, S. (2019). Predicting Social Perception from Faces: A Deep Learning Approach. *Working paper.*

Miller, N., Maruyama, G., Beaber, R. J., and Valone, K. (1976). Speed of speech and persuasion. *Journal of personality and social psychology*, 34(4), 615.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, *104*(2-3), 90-126.

Nanne, A. J., Antheunis, M. L., van der Lee, C. G., Postma, E. O., Wubben, S., and van Noort, G. (2020). The use of computer vision to analyze brand-related user generated image content. *Journal of Interactive Marketing*, 50, 156-167.

Nelson, Ronald G. and David Schwartz (1979). Voice-Pitch Analysis. *Journal of Advertising Research*, 19(5), 55-59.

Nighswonger, N. J., and Martin Jr, C. R. (1981). On using voice analysis in marketing research. *Journal of Marketing Research*, 18(3), 350-355.

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., and Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin and Review*, 24(3), 856-862.

Oosterhof, N. N., and Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, 9(1), 128.

Orth, U. R., and Crouch, R. C. (2014). Is beauty in the aisles of the retailer? Package processing in visually complex contexts. *Journal of Retailing*, 90(4), 524-537.

Orth, U. R., and Malkewitz, K. (2008). Holistic package design and consumer brand impressions. *Journal of Marketing*, 72(3), 64-81.

Pennington, A. L. (1968). Customer-salesman bargaining behavior in retail transactions. *Journal of Marketing Research*, 5(3), 255–262.

Peschel, A. O., and Orquin, J. L. (2013). A review of the findings and theories on surface size effects on visual attention. *Frontiers in Psychology*, 4, 902.

Pieters, R., and Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing*, 68(2), 36–50.

Pieters, R., Wedel, M., and Batra, R. (2010). The stopping power of advertising: Measures and effects of visual complexity. *Journal of Marketing*, 74(5), 48-60.

Pon-Barry, H., and Shieber, S. M. (2011). Recognizing uncertainty in speech. EURASIP Journal on Advances in Signal Processing. *Special Issue on Emotion and Mental State Recognition from Speech.*

Poor, M., Duhachek, A., and Krishnan, H. S. (2013). How images of other consumers influence subsequent taste perceptions. *Journal of Marketing*, *77*(6), 124-139.

Pouyanfar, S., Yang, Y., Chen, S. C., Shyu, M. L., and Iyengar, S. S. (2018). Multimedia big data analytics: A survey. *ACM computing surveys (CSUR)*, *51*(1), 1-34.

Pugh, S. Douglas (2001). Service with a Smile: Emotional Contagion in a Service Encounter, *Academy of Management Journal*, 44 (5), 1018–27.

Raghubir, P., and Krishna, A. (1999). Vital dimensions in volume perception: Can the eye fool the stomach? *Journal of Marketing Research*, 36, 313–326.

Rajaram, P., and Manchanda, P. (2020). Video Influencers: Unboxing the Mystique. arXiv preprint arXiv:2012.12311.

Reik, T. (1949). *Character analysis*. New-York: Farrar, Strauss and Giroux.

Rezlescu, C., Duchaine, B., Olivola, C. Y., and Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one* 7(3), e34293.

Rooderkerk, R. P., and Lehmann, D. R. (2021). Incorporating Consumer Product Categorizations into Shelf Layout Design. *Journal of Marketing Research*, 58(1), 50-73.

Rule, N. O., and Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science*, *19*(2), 109-111.

Scherer, K. R. (1995). Expression of emotion in voice and music. Journal of voice, 9(3), 235-248.

Schroeder J and Epley N. (2015). The Sound of Intellect Speech Reveals a Thoughtful Mind, Increasing a Job Candidate's Appeal. *Psychological Science*, 26(6) 877–891.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, *1*(2), 119-131.

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., and Zäske, R. (2014). Speaker perception. Wiley Interdisciplinary Reviews: *Cognitive Science*, 5(1), 15-25.

Shen, T., Kim, H., Ding, M. (2021). Voice-based Empathetic Method for Enhancing Voice Technology Applications in Marketing with Certainty/Uncertainty. *Working paper.*

Shin, D., He, S., Lee, G. M., WHINSTON, A. B., Centintas, S., and Lee, K. C. (2020). Enhancing Social Media Analysis with Visual Data Analytics: A Deep Learning Approach. *Management Information Systems Quarterly*, 44(4), 1459-1492.

Simonyan K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Singh, S., Marinova, D., Singh, J., and Evans, K. R. (2018). Customer query handling in sales interactions. *Journal of the Academy of Marketing Science,* 46(5), 837-856.

Small, D. A., and Verrochi, N. M. (2009). The face of need: Facial emotion expression on charity advertisements. *Journal of Marketing Research*, *46*(6), 777-787.

Smith, A. N., Fischer, E., and Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?. *Journal of Interactive Marketing*, *26*(2), 102-113.

Smith BL, Brown BL, Strong WJ, and Rencher AC (1975). Effects of speech rate on personality perception. *Lang Speech,* 18(2),145–152

Sundar, A., and Noseworthy, T. J. (2014). Place the logo high or low? Using conceptual metaphors of power in packaging design. *Journal of Marketing*, 78(5), 138-151.

Sutton, R. and A. G. Barto (2018). *Reinforcement learning: An introduction.* MIT press, Székely, E., Ahmed, Z., Hennig, S., Cabral, J. P., and Carson-Berndsen, J. (2014). Predicting synthetic voice style from facial expressions. An application for augmented conversations. *Speech Communication*, *57*, 63-75.

Tahon, M., and Devillers, L. (2015). Towards a small set of robust acoustic features for emotion recognition: challenges. *IEEE/ACM transactions on audio, speech, and language processing*, *24*(1), 16-28.

Teixeira, T., Wedel, M., and Pieters, R. (2012). Emotion-induced engagement in internet video ads. *Journal of Marketing Research*, 49(2), 144-159.

Teixeira, T., Picard, R., and El Kaliouby, R. (2014). Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Science*, 33(6), 809-827.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., and Zhang, Y. (2019). What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *Journal of Marketing*, *83*(4), 1-20.

Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, 308(5728), 1623-1626.

Todorov, A. (2017). *Face value: The irresistible influence of first impressions.* Princeton University Press.

Townsend, C., and Shu, S. B. (2010). When and how aesthetics influences financial decisions. *Journal of Consumer Psychology*, 20, 452–458.

Urban, G. L., Liberali, G., MacDonald, E., Bordley, R., and Hauser, J. R. (2014). Morphing banner advertising. *Marketing Science*, 33(1), 27-46.

Vernon Richard J. W., Clare A. M. Sutherland, Andrew W. Young, and Tom Hartley (2014), Modeling First Impressions from Highly Variable Facial Images. *Proceedings of the National Academy of Sciences*, 111 (32), 3353-3361.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*.

Wang, X. S., He, J., and Grewal, R. (2020). Image Features and Demand in the Sharing Economy: A Study of Airbnb. *Working paper.*

Wang, X. S., Lu, S., Li, X. I., Khamitov, M., and Bendle, N. (2021). Audio Mining: The Role of Vocal Tone in Persuasion. *Journal of Consumer Research*. 48(2), 189-211.

Wang, Z., S. N. Singh, Y. J. Li, S. Mishra, M. Ambrose, and M. Biernat (2017). Effects of Employees' Positive Affective Displays on Customer Loyalty Intentions: An Emotions-as-Social-Information Perspective. *Academy of Management Journal*, 60 (1), 109-129.

Wedel, M., and Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6), 97-121.

Wedel, M., and Pieters, R. (2008). A review of eye-tracking research in marketing. In N. K. Malhotra (Ed.) *Review of Marketing Research*, 4, 123-147.

Weninger, F., J. Krajewski, A. Batliner and B. Schuller (2012). The voice of leadership: Models and performances of automatic analysis in online speeches. *IEEE Transactions on Affective Computing*, 3(4), 496-508.

Whittle, M. W. (2014). *Gait analysis: an introduction*. Butterworth-Heinemann.

Xiao, Li, and Min Ding (2014). Just the Faces: Exploring the Effects of Facial Features in Print Advertising. *Marketing Science*, 33 (3), 338–52.

Xiao, Li, Hye-jin Kim, and Min Ding (2013). An Introduction to Audio and Visual Research and Applications in Marketing. in *Review of Marketing Research*, Malhotra, N., ed. Bingley: Emerald Group Publishing Limited, 213–253.

Xu, J., and Ding, M. (2021). Transparent Model of Unabridged Data (TMUD). *Available at SSRN 3849871*.

Yan, Xiyu, Huihui Gong, Yong Jiang, Shu-Tao Xia, Feng Zheng, Xinge You, and Ling Shao (2020). Video scene parsing: An overview of deep learning methods and datasets. *Computer Vision and Image Understanding*, 201, 103077.

Yaoyuneyong, G., Foster, J., Johnson, E., and Johnson, D. (2016). Augmented reality marketing: Consumer preferences and attitudes toward hypermedia print ads. *Journal of Interactive Advertising*, 16(1), 16-30.

Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking: A survey. *Acm computing surveys (CSUR)*, *38*(4), 13-es.

Young, A. W., Frühholz, S., and Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences,* 24(5), 398-410.

Zebrowitz, Leslie A., Judith A. Hall, Nora A. Murphy, and Gillian Rhodes (2002). Looking Smart and Looking Good: Facial Cues to Intelligence and Their Origins, *Personality and Social Psychology Bulletin*, 28 (2), 238-249.

Zhang M. and L. Luo (2022), Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp. Forthcoming at *Management Science***.**

Zhang, Q., Wang, W., and Chen, Y. (2020). Frontiers: In-Consumption Social Listening with Moment-to-Moment Unstructured Data: The Case of Movie Appreciation and Live Comments. *Marketing Science*, 39(2), 285-295.

Zhang, S., Lee, D., Singh, P. and Srinivasan, K. (2022). What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features. *Management Science*. Forthcoming.

Zhang, W., Chintagunta, P. K., and Kalwani, M. U. (2021). Social Media, Influencers, and Adoption of an Eco-Friendly Product: Field Experiment Evidence from Rural China. *Journal of Marketing*, *85*(3), 10-27.

Zhang, X., Li, S., Burke, R. R., and Leykin, A. (2014). An examination of social influence on shopper behavior using video tracking data. *Journal of Marketing,* 78(5), 24–41.

Zhang, Z. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4), 2614-2635.

Zhou, Y., S. Lu, and M. Ding (2020). Contour-as-Face Framework: A Method to Preserve Privacy and Perception. *Journal of Marketing Research,* 57(4), 617-39.

Zoghaib, A. (2019). Persuasion of voices: The effects of a speaker’s voice characteristics and gender on consumers’ responses. *Recherche et Applications en Marketing (English Edition)*, *34*(3), 83-110.

Figure 1. A Framework of Artificial Empathy

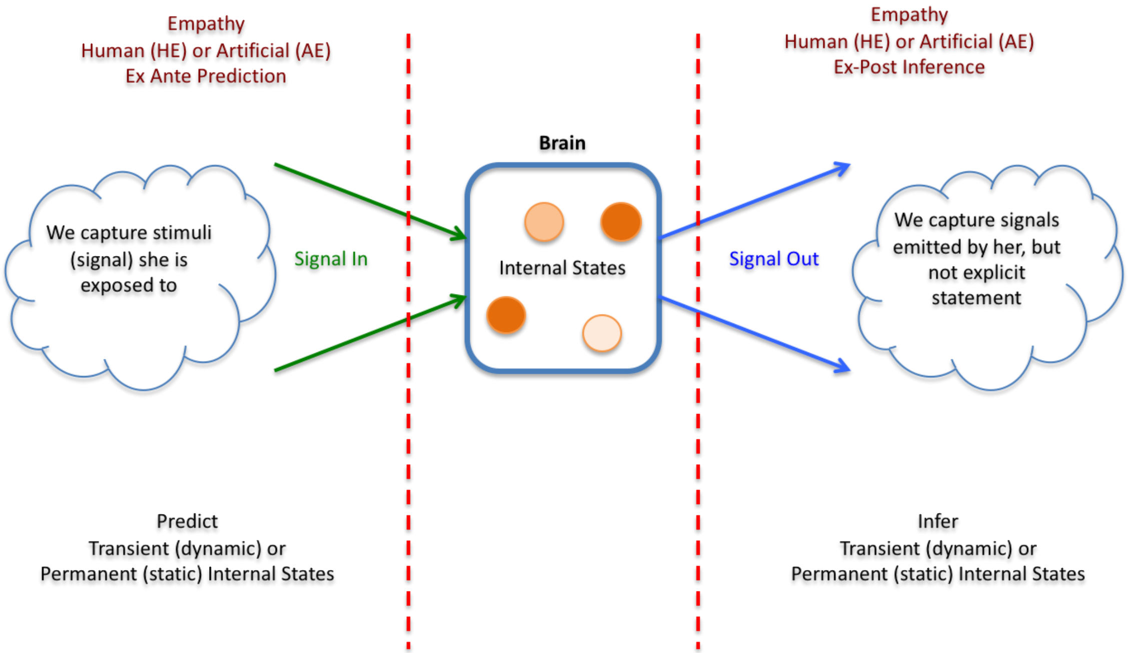
Empathy
Human (HE) or Artificial (AE)
Ex Ante Prediction
Empathy
Human (HE) or Artificial (AE)
Ex-Post Inference
Brain
We capture stimuli (signal) she is exposed to
Signal In
Internal States
Signal Out
We capture signals emitted by her, but not explicit statement
Predict
Transient (dynamic) or
Permanent (static) Internal States
Infer
Transient (dynamic) or
Permanent (static) Internal States

Figure 2. Time Domain and Frequency Domain (Xiao et al. 2013)

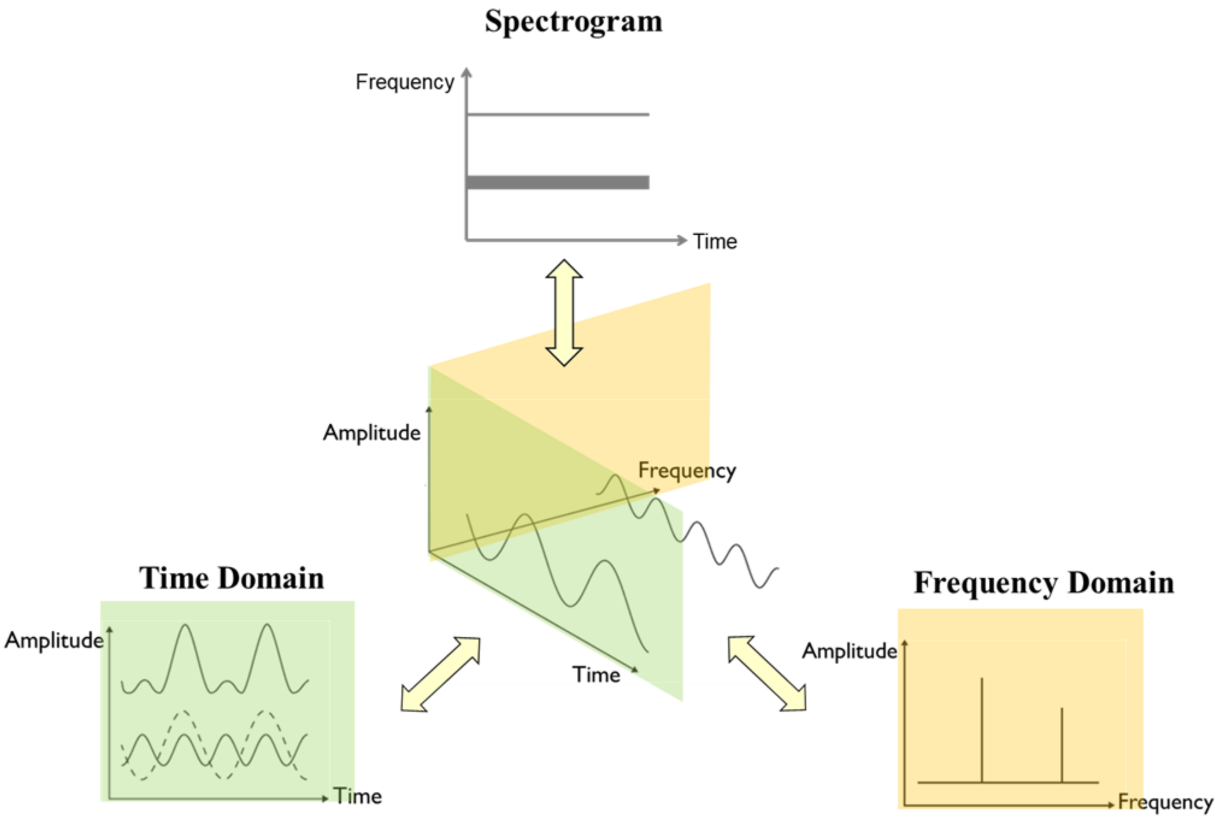

Table 1. Types of Audio/Visual Data in Business Contexts

	A/V data	**Sender**	**Interpreter**	**Business Contexts**	**Representative References**

Social Media Posts (e.g., Instagram, YouTube, Facebook)	Visual	Consumers	Consumers	Influencer marketing;
				User generated content;
					Klostermann, Plumeyer, Böger and Decker (2018); Li and Xie (2020); Liu, Dzyabura, and Mizik (2020);

Online Profiles (e.g., LinkedIn, Tinder)	Visual	Consumers	Consumers/Firms	Customer relationship management;
				Customer Privacy;
					Zhou, Lu, and Ding (2020); Malik et al. (2019)

Online Reviews (e.g., Amazon, TripAdvisor, Yelp)	Visual	Consumers	Consumers/Firms	Product review;
				Consumer satisfaction;
					Zhang and Luo (2022); Guan et al. (2020)

Recordings of Service Encounters (face to face and phone call)	Audio
	Visual
		Consumers	Firms	Customer relationship management;
				Consumer satisfaction;
				Personal Selling;
					Hall et al. (2014); Singh et al. (2018); Zhang et al. (2014)

In-Store Shopping Behavior (e.g., product trails)	Visual	Consumers	Firms	Product/service recommendation; In-store marketing;	Lu, Xiao and Ding (2016); Hui, Fader, and Bradlow (2009)
Research Data (e.g., customer interviews)	Visual Audio	Consumers	Firms	Marketing research	Nighswonger and Martin (1981);
Product, Package, Brand Logo Design	Visual	Firms	Consumers	Marketing mix design and optimization	Liu et al. (2017);
Design of Promotional Content (e.g., Advertisements)	Visual Audio	Firms	Consumers	Marketing mix design and optimization	Xiao and Ding (2014); Kim, Wang, and Ding (2021);
Offline Store and Shelf Design	Visual	Firms	Consumers	Marketing mix design and optimization	Rooderkerk and Lehmann (2021)
E-commerce Channel (e.g., website, app) Design	Visual	Firms	Consumers	Marketing mix design and optimization	Hauser et al. (2009);
Design of the Interpersonal Selling/Shopping Process	Visual Audio	Firms	Consumers	Sales management; Live streaming	Lu et al. (2021); Bruner (1990); Hul, Dube and Chebat (1997)
Media Products (news, broadcasts, magazines, movies, TVs, podcasts)	Visual Audio	Third-party	Consumers	Entertainment marketing; Ad planning	Liu, et al. (2018);
Over-the-top Content (e.g., Netflix, YouTube etc.)	Visual Audio	Third-party	Consumers	Entertainment marketing; Ad planning	Mei et al. (2007)
Board Meetings, Funding Pitches	Visual Audio	Firms	Investors	Funding	Jiang,Yin, and Liu (2019); Tegtmeier et al. (2021)

Table 2. Types of Artificial Empathy Models in Marketing Contexts

	Static – Ex-post	**Static – Ex-ante**	**Dynamic – Ex-post**	**Dynamic – Ex-ante**
Input	Responses	External signals	Responses	External signals
Output	Internal states	Responses	Internal states	Responses
AV data used	Non-transient audio and visual signals, e.g., • design of the interpersonal selling/shopping process • user generated content • job interviews	Non-transient audio and visual signals, e.g., • design of promotional content (e.g., face and voice in advertisements); • customer online profiles • design of marketing mix (e.g., product, package, brand logo)	Transient audio and visual signals, e.g., • recordings of service encounters • recordings of product trail	Transient audio and visual signals, e.g., • design of A/V content (e.g., trailer) • design of promotional content (e.g., social media post, print, audio, and video ad)

Application context	Interpersonal selling/shopping;
	Brand image management;
	Employee recruitment and allocation;
		Marketing mix design and optimization;	Service management
			Personalized product/service recommendation;
				Marketing mix design and optimization;

Representative References	DeGroot and Motowidlo (1999);
	Klostermann et al. (2018);
	Zhang and Luo (2022);
	Liu et al. (2020);
	Wang et al. (2021).
		Landwehr et al. (20113);
		Xiao and Ding (2014);
		Liu et al. (2017);
		Kim, Wang, and Ding (2021);
		Zhou, Lu, and Ding (2020).
			Mattila and Enz (2002); Hall et al. (2014);
			Lu, Xiao, and Ding (2016);
			Singh et al. (2018).
				Liu et al. (2018);
				Li and Xie (2020);
				Shin et al. (2020).

Future research opportunities	• Inferring internal states from new types of non-transient audio/visual cues, e.g., hand movement, body language, gaze and gait, etc. • Exploring the use of audio/visual cues in the inferences of new types of non-transient internal states in both online and offline contexts.	• Understanding how the audio and visual elements of marketing mix interact with each other and with other elements of marketing mix in influencing customers' perceptions and attitudes. • Optimizing the audio and visual elements of marketing mix based on customer preferences. • Exploring how the design of audio and visual elements of marketing mix vary across	• Inferring internal states from new types of transient audio/visual cues, e.g., hand gestures, head and body movements, etc. • Exploring the use of audio/visual cues in the inferences of new types of transient internal states in online and offline contexts.	• Understanding how the audio and visual elements of marketing mix interact with each other and with other elements of marketing mix in influencing customers' cognitive, emotional and physiological states. • Optimizing the audio and visual elements of marketing communication or service delivery, especially in the online contexts (e.g., livestreaming).

	• Understanding how the inferences of non-transient internal states change or vary across different business contexts.	different types of products or brand images.	• Product/service innovation based on dynamic inferences of internal states.	• Exploring how the design of audio and visual elements of marketing communication or service delivery process vary across different types of products or brand images.

Table 3. Useful Tools for Audio/Visual Data Analysis

Tools/Platforms	**Useful Toolboxes/Libraries**	**Type of Data**	**Description**	**Source**
Matlab	Image Acquisition Toolbox Image Processing Toolbox Computer Vision Toolbox Pattern recognition and machine learning Toolbox; Deep learning Toolbox Lidar Toolbox Vision HDL Toolbox (mainly for video) Audio Toolbox	Audio/Image/ Video	For image and audio processing, analysis, visualization and algorithm development; For audio processing, speech analysis, acoustic measurement;	https://www.mathworks.com/
OpenCV	N.A.	Image/Video	OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning library. It has more than 2500 classic and state-of-the-art computer vision and machine learning algorithms. It has C++, Python, Java and MATLAB interfaces.	https://opencv.org
ImageJ/Fiji	N.A.	Image/Video	ImageJ/Fiji is an open source image processing program for multidimensional image data with a focus on scientific imaging.	https://imagej.net/ImageJ
Octave	Octave-forge image package	Image	The Octave-forge Image package provides functions for processing images, feature extraction, image statistics, spatial and geometric	https://octave.sourceforge.io/image/index.html

			transformations, morphological operations, linear filtering, etc.	
Python	Python Imaging Library (Pillow 8.2) scikit-image pyAudioAnalysis pyannote-audio	Audio/Image/ Video	The Python libraries of algorithms for image and audio data processing.	https://pypi.org/project/Pillow/ https://scikit-image.org https://github.com/tyiannak/pyAudioAnalysis/wiki https://github.com/pyannote/pyannote-audio
PyTorch	Torchaudio Torchvision	Audio/Image/ Video	PyTorch is an open-source machine learning framework for developing deep learning models.	
TensorFlow	Pre-trained deep learning models and APIs	Audio/Image	TensorFlow is an open-source machine learning framework for developing deep learning models.	https://github.com/tensorflow/tfjs-models https://github.com/tensorflow/models/tree/master/official#computer-vision
Keras	Pre-trained deep learning models	Image	Keras is a Python deep learning API.	https://keras.io/api/applications/
Caffe	Models, and worked examples for deep learning	Image/Video	Caffe is a Deep learning framework.	https://caffe.berkeleyvision.org/
Dlib	N.A.	Image/Video	Dlib is a modern C++ toolkit containing machine learning and computer vision algorithms. It has C++ and Python interfaces.	http://dlib.net
Praat	N.A.	Audio	Praat is a general-purpose speech tool which has a standalone user interface and works with scripts: it enables editing, segmentation and labeling, and prosodic manipulation	www.praat.org
SpeechBrain	N.A.	Audio	SpeechBrain an all-in-one speech toolkit based on PyTorch. It provides speech recognition, speaker recognition, speech enhancement, multi-microphone signal processing	https://speechbrain.github.io/

Table 4. Commonly Used Image Pre-processing Techniques

	Techniques	**Description**	**Applications**
Point-based	2D geometric transformations	Resizing, cropping, rotation, affine and projective transformations, etc.	Image normalization, correct for deformation or distortion in images;
	Histogram equalization	Transforming the intensity values in an image so that the histogram of the output image matches a specified histogram (usually a flat histogram).	Enhancing contrast of images;
	Color transformation	Transforming the image from one color space to another.	Color correction and color image filtering;
Area-based	Linear filtering (e.g., gaussian filter, box filter)	An output pixel's value is a weighted sum of pixel values within a small neighbourhood of the input pixel.	Image smoothing, sharpening, noise removal, and edge enhancement;
	Non-linear filtering (e.g., median filter, bilateral filter)	An output pixel's value is a non-linear combination of neighboring pixels.	Image smoothing and noise removal while preserving edges;
	Morphological operation	Adjusting pixel value of binary images based on the value of other pixels in its neighbourhood.	Noise removal, image segmentation, region of interest selection;
	Fourier transforms	Analysing the frequency characteristics of an image.	Image sharpening, blur, and noise removal;

Table 5. Commonly Used Image Features and Analytical Methods

	Features	**Feature detection methods**	**Feature representation methods**	**Applications**

Local features-low level	Key points	Harris detector;
		Gabor-Wavelet detector;
			Histogram of gradient orientations (HOG);
			Scale invariant feature transform (SIFT);
			Speeded-Up Robust Features Descriptor (SURF);
			Local Binary Pattern (LBP);
			Multi-scale-oriented patches (MOPS)
				Image matching, Image classification;

	Edges and contours	Harris detector;
		Gradient operators (Robert, Prewit, Sobel, and Laplacian);
		Canny edge detectors;
			Edge histogram descriptor;
			Fourier descriptors
				Image matching;
				Image classification;
				Image segmentation

	Corners	Harris detector;
		Förstner corner detector;
		Features from Accelerated Segment Test (FAST);
		Laplacian-of-Gaussian (LoG);
		Difference of Gaussian (DoG);
		Determinant of Hessian (DoH);
			Histogram of gradient orientations (HOG);
			Scale invariant feature transform (SIFT);
			Speeded-Up Robust Features Descriptor (SURF);
				Image matching; Image classification;

	Blobs/regions	Laplacian-of-Gaussian (LoG);
		Difference of Gaussian (DoG);
		Determinant of Hessian (DoH);
		Features from Accelerated Segment Test (FAST)
			Scale invariant feature transform (SIFT);
			Speeded-Up Robust Features descriptor (SURF);
			Maximally stable extremal regions (MSER) descriptor
				Image classification;
				Object detection and tracking;

Global features-low level	Color	N.A.	Color histogram descriptor;
			Color structure descriptor;
			Color layout descriptor;
			Scalable color descriptor (HSV color space);
			Dominant color descriptor;
				Image classification;
				Image segmentation;

	Texture (repeated elements in an image)	Sub-elements (e.g., edges, spots) detectors; Vector quantization	Local Binary Patterns;
			Autocorrelation;
				Object recognition;
				Image segmentation;

			Wavelet transform;
			Gabor transform;	

	Shapes	Centroid Distance; Contour Curvature;
		Area Function;
			Fourier descriptor;
			Convex Hull;
			Chain codes;
			Shape context;
				Image retrieval;
				Image classification;
				Object recognition;

Global features-high level	Visual complexity/clutter	Image file size;
		Image texture;
		Edge density;
		Number of objects;
		Symmetry;
		Feature Congestion measure;
		Subband Entropy;
		Unsupervised Activation Energy (UAE)
			Bag-Of-Visual-Words	Visual content design;
				Brand logo design;
				Product design;
				Package design;
				Shelf/store design;
				Webpage design;

	Visual aesthetics	Colorfulness;
		Color distribution;
		Color harmony;
		Luminance and exposure;
		Edge distribution;
		Size and aspect ratio;
		Symmetry;
		Image composition;
			Bag-Of-Visual-Words	Visual content design;
				Brand logo design;
				Product design;
				Package design;
				Shelf/store design;
				Webpage design;

Table 6. Machine Learning Models, Techniques and Applications

	Supervised learning	**Unsupervised learning**	**Reinforcement learning**

Inputs	Training data paired with outputs (discrete labels (Classification)/Continuous values (regression));
	Hand-crafted Features;
		Training data	Environmental states;
			Agent's action set;
			Transition dynamics;

Outputs	Model parameters;
	Error rates;
		Patterns in data	Optimization Strategy

Learning structure	Paired inputs-outputs are given to a learning algorithm, which adjusts the model parameters to maximize the agreement between model predictions and target outputs.	Identifying the patterns in training data.	Adapting the model parameters and behaviors in an environment in order to maximize total reward.

Commonly used techniques	Support vector machine (SVM);
	Nearest neighbors;
	Bayesian classification;
	Logistic regression;
	Decision trees and forests;
		Clustering;
		K-means and Gaussian mixture modeling;
		Principle component analysis;
		Manifold learning;
			Q learning;
			Monte Carlo methods;
			State–action–reward–state–action (SARSA);
			Temporal Difference (TD) methods;

Application context	Semantic image classification (label the content of an image)/ object detection (e.g., face, body)	Image segmentation;
		Face recognition;
			Semantic image classification/object detection;

Table 7. Image-based Deep Learning Models in Marketing Contexts

Types of visual data	**Deep learning models**	**Features extracted/studied**	**Marketing context**	**Type of AE model**	**Reference**
Instagram (10,375 user generated images)	Google Cloud Vision API	1250 semantic labels	Brand image/perceptions	Static-ex-post	Klostermann et al. (2018)
Professional social network (43,533 MBA graduate profiles with profile pictures)	Pre-trained CNN by the Open Face project	Facial features	Attractiveness	Static-ex-ante	Malik et al. (2019)
The 10K Adults Face database	CNN model	Facial features	Warmth and competence traits	Static-ex-ante	Messer and Fausser (2019)
Self-developed platform (91, 856 brand-related user generated images)	Clarifai (API)	5,426 semantic labels	brand perceptions	Static-ex-post	Dzyabura and Peres (2021)
Yelp (755,758 user generated photos)	Clarifai (API)	5,080 semantic labels	Restaurant survival	Static-ex-post	Zhang and Luo (2022)
Airbnb (11,496 host-generated property images)	VGG16 model (fine-tuned)	Image quality	Guests' preference	Static-ex-ante	Zhang et al. (2022)
Tweeter, Instagram (492,860 consumer generated images)	VGG-16 CNN model (fine-tuned)	3 image classes (brand selfies, consumer selfies, packshots)-consumer response	Consumer response	Static- ex-ante	Hartmann et al. (2021)
E-commerce platform (15,006 consumer generated images)	MobileNet model (fine-tuned)	Aesthetic score	Consumer rating	Static- ex-ante	Guan et al. (2020)
31,367 real-life facial images	ResNet model (fine-tuned)	Facial features	Big five personality traits	Static- ex-post	Kachur et al. (2020)
Airbnb (220,000 host-generated property images)	ResNet50 model (fine-tuned)	Image quality and image room types	Room demand	Static- ex-ante	Li et al. (2019a)
Flickr, Instagram (16,360 user generated images)	CaffeNet model (fine-tuned)	56 brands	Brand attributes (e.g., rugged, fun, glamorous, healthy)	Static-ex-post	Liu et al. (2020)
Tweeter (4,537 tweets with images)	Google Cloud Vision API	Image quality, face presence, happy emotion	User responses	Dynamic-ex-ante	Li and Xie (2020)
Instagram (21,738 brand-related user generated images)	Google Cloud Vision API, Clarifai, YOLOV2	Semantic labels	Brand-related UGC	N.A.	Nanne et al. (2020)
Tumblr (53,417 image posts)	Yahoo! CNN model	Visual content features (complexity, aesthetics, texture, celeberity endorsement, etc)	Consumer responses	Dynamic-ex-ante	Shin et al. (2020)

Airbnb (host-generated property images)	Amazon Rekognition	Interior design score, room type (e.g., living room and bedroom)	Room demand	Static- ex-ante	Wang, He, and Grewal (2020)
706 brand logos	VAE and GAN	Logo design features (e.g., color, shape, font, complexity, etc.)	Brand personality	Static- ex-ante	Dew et al. (2021)
7,000 rated car images and 180,000 unrated car images	VAE and GAN	Car design features	Customer rating	Static- ex-ante	Burnap et al. (2021)
1650 influencer videos	YAMNet EfficientNet-B7	Video content features (texts, audios, images)	Viewer responses	Static- ex-ante	Rajaram and Manchanda (2020)