POLITECNICO DI MILANO

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING

MANAGEMENT ENGINEERING MASTER DEGREE

VOICE ANALYSIS AND BUSINESS APPLICATIONS
A STATE-OF-THE-ART

Fabio Feroldi - 876701

Supervisor: Prof. Lucio Lamberti

Tutor: Ing. Elena Magri

Academic year 2019 - 2020

Table of Contents

List of Figures

List of Tables

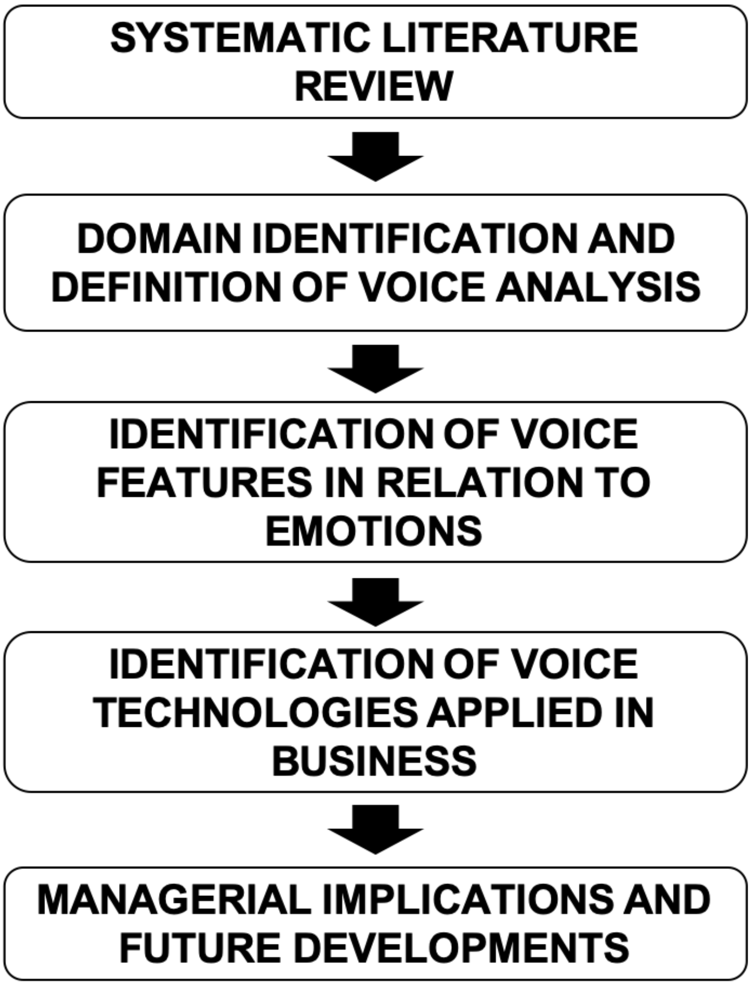
Abstract

Voice carries information not only about the lexical content, but also paraverbal aspects which describe the intentions and emotions of the speaker. The voice in different emotional states is accompanied by distinct changes in paraverbal features. In the first half of the work, we present the state-of-the-art of “Voice Analysis” methods and technologies used for classifying emotional speeches. In particular, we focus principally on studying feature extraction, selection and classification, as well on describing emotion recognition and classification algorithms and techniques. In the second part, we address the topic from the managerial point of view, investigating the existing business applications of “Voice Analysis” paired with emotion detection. Different industries have been examined to gauge how organization are integrating such innovation in their businesses, which are the future developments of this technology and which are the business opportunities that can raise in the next few years. In conclusion, the evaluation of every technology is carried out in terms of potentiality and maturity.

Executive Summary

Our thesis investigates the actual degree of knowledge about the paraverbal characteristics of voice and deepens the wide range of business applications of voice analysis. The aim of this work is to provide a state-of-the-art analysis that would help further researches to improve the existing best practices and performances of voice recognition techniques and real-world applications.

The following picture shows the work flow we followed for our dissertion:


SYSTEMATIC LITERATURE REVIEW
DOMAIN IDENTIFICATION AND DEFINITION OF VOICE ANALYSIS
IDENTIFICATION OF VOICE FEATURES IN RELATION TO EMOTIONS
IDENTIFICATION OF VOICE TECHNOLOGIES APPLIED IN BUSINESS
MANAGERIAL IMPLICATIONS AND FUTURE DEVELOPMENTS

Starting from a systematic review of the literature, in chapter 1 we identify the meaning of the word “voice analysis” and its research domain. The focus of our work is on the study of paraverbal communication and the interest that raised among researchers due to the meaningful insights that result from studying biometric characteristics of the speech. In chapter 2, we outline the research questions that we have targeted to lead our investigation. Moreover, we divide the most relevant articles into two main categories, depending on the perspective of the research. The first one is related to the study of the voice from a biomedical point of view. The second one, instead, concerns the application of voice technologies in business from a managerial point of view. In chapter 3, we describe the voice parameters that have to be captured and studied as the input of voice recognition. These parameters are classified into three main different categories, depending on the type of variables of the voice. The classes of variables are:

- Acoustic: including pitch, energy, rhythm and spectrum;
- Tone-based: describing prosody in terms of intonational phrases;
- Voice quality: referring to the characteristic auditory of the speech.

Each different phonation aspect is linked to an emotional state and it is possible to recognize and monitor them in real-time. By varying voice parameters, some important information like intentions, emotions and attitudes are likely to be detected. This task is performed by the classification algorithms which analyse the paraverbal variables collected from speech utterances and recognize the emotional intentions of the speaker. We discuss each step of the voice analysis process, providing an overview of the different approaches to address the issue. The overview is built from several theoric-empirical articles in this field, thus contains comparisons of the existing models, algorithms and best practices already applied in the experiments by researchers. By analyzing the current advancements of the recognition systems in the experiments, we discuss performance and weaknesses of such methods. Even though the most common classification methods reach acceptable levels of accuracy rate, there is still room of improvement in the effectiveness of recognition systems. In chapter 4, we shift the attention from the biometrics features of voice to the real-life implementation in the business of such technology. Exploring the different applications of the voice analysis that companies are developing and applying into their organization, we are able to understand which are the opportunities that this emerging market is generating. We classify the business applications considering a technology-driven analysis, since the same technology may find application in different industries. We choose four types of voice technology that brand are deploying in their operations and are expected to reveal their full potential in the following years. The selected ones are:

- **Voice Analytics**: it is the analysis of the customer insights which support organizations to identify and target customers for marketing programs, customer behavioural prediction, gaining customer loyalty and customer retention. Voice Analytics are already present in industries like Retail, Financial Services, Healthcare, and others, Showing tangible advantages to organizations. We illustrate the main players that are operating in the market in order to make a comparison between the solutions offered on the market. We also explain the VoC (voice of customer) concept related to marketing purposes. Generally, It intends listening to the customers' needs, in the specific context of voice analysis, it means to collect the insights from customers through the recognition and study of their voice. Afterwards, we investigate more in details the Voice Analytics applied in call centers and their importance as a source of meaningful information for organizations.

- **Biometrics Authentication**: it is a technology used to match personal voice pattern and verify the speaker's identity. Biometrics Authentication systems demonstrate numerous advantages to both companies and clients. We present a framework to evaluate the benefits gained through voice biometrics solutions in contact centres that use them to authenticate callers and verify their claimed identity in real-time. They can bring three main benefits:

    - Average Handle Time reduction (contact centre authentication time)
    - Increased self-service containment

  - Fraud prevention

They present several advantages principally applied to the Financial Service industry to ensure security and fastening online services. They are even adopted in Retail to increase efficiency and streamline the processes, and in the Hospitality industry to recognize clients and boost the service quality, as well in Healthcare to diagnose patients in a personalized way. Moreover, voice is also well suited as a biometric authentication solution across a wide range of IoT devices, including smartphones, tablets, wearables, PCs, gaming systems, smart TVs, even fixed-line telephones and automobiles.

- **Robotics**: They are meant as physical robots but also broadly as Human-Robot interaction (HRI). We introduce some characteristics of this technological innovation in business showing different examples and use cases of real-life applications. We present them divided into two groups: Social Robots and Retail Robots.

  - Social robots are autonomous machines that are designed to interact with humans and exhibit social behaviours such as recognizing, following, assisting and engaging them in conversation. We discuss the technological advancements of robots also considering the natural evolution of their social role. The social evolution is represented in two dimensions: social and physical dexterity.
  - Retail robots, by enhancing product visibility and customer experience through their attractiveness, can bring enormous advantages to brands while increasing store traffic, stimulating purchase and attract the undivided attention of shoppers.

Robots are intended to interact with people in a natural manner in different applications such as:

1. Education and Research: It is adequate for teaching Science, Technology, Engineering and Math concepts with students at all levels.
2. Health therapeutic: social robots are employed as therapy for kids with autism.
3. Service Robots: many robots worldwide perform such services as hotel check-ins, airport customer service, fast-food checkout, etc.
4. Social companion/ Caregiving: they are seen as part of the family and support for older people.

- **AI's and Virtual Assistants**: VAs are changing the way consumers interact with web interfaces. They will entirely manage the smart home environments, they can be easily integrated into many devices and be a support in our daily life. The benefits these technologies are granting to innovative firms are illustrated exploring several examples present in the market. Customer engagement and loyalty can be achieved thanks to the improvement of products/ services and driving the customer experience using Emotional AI. Moreover, "localization" concept is described as a factor which affects customer satisfaction. Localization means that communication and digital contents need to be tailored based on cultural, linguistic, functional, and other locale-specific requirements of the context. These requirements also include the personal and emotional characteristics of the consumer that are gathered through voice analysis. Voice assistants also change the way goods are purchased, driving the shopping experience toward new concepts: voice commerce. Finally, a comparison between the most famous vocal assistants is disclosed.

For each application's field, we highlight the current advancements of the real-life solutions' developments, the contingencies and the challenges that tech firms will face in the close future. A comparison between the selected applications is also carried out to highlight their different strengths and weaknesses. This analogy is made taking into account two dimensions: potentiality and maturity of a technology.

The potentiality is seen as the expected growth of the market in which the innovation can be applied. Analysts are used to apply “compound annual growth rate”, or CAGR, in order to evaluate the potential growth of a market. CAGR refers to a representational percentage that shows how much a business has grown or will grow in a time-gap of at least two years. The maturity is evaluated using the technology hype cycle model. Hype Cycle specifically focuses on the set of technologies that is showing promise in delivering a high degree of competitive advantage over the following 5 to 10 years. Gartner uses hype cycles to characterize the over-enthusiasm, or "hype," and subsequent disappointment that typically follow the introduction of new technologies. The hype cycle provides a graphical and conceptual presentation of the maturity of emerging technologies. In the last capitula, we provide a critical analysis of the business applications’ comparison, we outline our research limitations and the contribution that it can bring to future studies. In conclusion, our results are interesting from two different points of view: on one side, we investigate the capabilities and the limitations of the voice as a novel powerful way of studying human biometrics and emotions. On the other side, from a more managerial point of view, companies that are pointing at innovating their business through the adoption of disruptive technologies can find attractive incipit to create competitive advantages.

1. Introduction: verbal, non verbal and paraverbal communication

There are different ways to communicate with people. Principally, it is possible to classify three particular ways: verbal, non-verbal and paraverbal. Verbal communication is the most direct one and easy to understand because it is achieved through spoken words and their meanings. However, when we interact with someone, even the body has a language of its own. The way we sit, the gestures we make, the way we talk, how much eye contact we make, all of these are non-verbal ways of communicating that impact the messages our words are sending.

Paraverbal communication is something that is hidden in the voice when we are talking. It’s about that old saying, “*It's not what we say, it's how we say it”.*

1.1. Verbal and Non-verbal communication

Human speech is a combination of verbal and non-verbal signals. Verbal signals consist of words and linguistic units of sounds and speech organs take a prominent position among the production and transmission of signals. Verbal communication is the use of sounds and language to convey the message or for giving information. Human verbal messaging is communicated via the words that we use. It is scientifically demonstrated that people while interacting with other persons, communicate with much more than words. All the other ways of communicating are called non-verbal and sometimes are more important than verbal content because they impact on the messages sent through words. Non-verbal communication is primaeval and constitutes the earliest type of communication. Words have boundaries whereas it transcends linguistic and cultural barriers and boundaries. What people say could often be very different from what they're thinking or feeling.

Nonverbal communication plays a profound role in the messages we receive from others and gives others a wealth of information about our personalities. People can exhibit several nonverbal signals at once. Thus, when we decode a message, we usually have many more nonverbal than verbal clues about what it truly means. Decoding non-verbal communication can help to figure out how others are in fact feeling and thinking. Non-verbal communication includes the following channels [1].

Voice

The human voice through their variations, convey different meanings. Voice modulation refers to the variation of the pitch or tone while speaking. Voice has many significant features like tone (harsh, soft, whisper), pace (rapid or slow), pitch (high-low), volume, rhythm, intonation and stress placed on words and the quality of voice.

Body language

The way you move and carry yourself communicates a wealth of information to the world. This type of nonverbal communication includes facial expressions, eye contact, voice modulation, posture and gestures, attire, appearance, handshake, space, timing, behaviour and smile. But despite being the most important aspect, body language is also the most misunderstood and misinterpreted.

Facial expression

It is considered as the index of the mind and the expression of the heart. The human face is extremely expressive, able to convey countless emotions without speaking.

Eye contact

The amount of eye contact we make is an especially important type of nonverbal communication. The way a person looks at someone can communicate many things, including interest, affection, hostility, or attraction. The eyes are indeed the most expressive part of the human face. Most of us have decoded "eye language" even if we did not know about body language or nonverbal communication. This kind of techniques is very useful in legal cross-examination, in counselling sessions, in negotiations and other routine life activities.

Gestures and postures

The meaning of some gestures can be very different across cultures. Posture refers to the carriage, state and attitude of body or mind. It may be physical or mental, erect or upright. Gesture refers to any significant movement of body and deliberate use of such movement as an expression of feeling.

Attire

It plays a vital role in big organizations where there is a formal or own unwritten dress code that is well understood. Any breach of this is likely to dilute the effectiveness of the communication. It proclaims and creates the first impression.

Appearance

It is the way one looks and presents oneself indicates the importance one attaches to one’s presence. Now a day, people have the habit of judging others by their appearance. People speak not merely with their words but with their total personality. It helps in making a positive and constructive impact.

Actions and Behaviour: Actions convey messages more forcefully than words. It is said that examples set through actions are 3 far more effective in communicating intentions and concern than words. Behaviour refers to conduct, manners shown by a person towards others and is governed by thoughts and feelings.

Touch

It is possible to communicate a great deal through touch. For instance, a weak handshake, a warm bear hug can give very different and powerful messages.

Space

It is the physical distance between persons that indicates familiarity and closeness. People can use physical space to communicate many different nonverbal messages, including signals of intimacy and affection, aggression or dominance.

Silence: Silence is an important communication tool. Sometimes, an extended period of silence tells more than we mean to say. Intentional silence may contain certain feelings and attitudes that are not communicated through sounds. Additionally, silence can be an effective technique to encourage feedback.

Decoding Nonverbal signs provide insight, influences perception of individual competence, persuasiveness, power, sincerity and vulnerability. However, it would be impossible to correctly decode the nonverbal signals without being aware of cultural norms. Culture is a fundamental discriminating factor to understand the real meaning of a non-verbal message.

1.1.1. 55% 38% 7% Theory and PAD Model

From scientific researches, it is possible to assess that when we communicate, only a small percentage of our overall message comes from the words we use [2]. Albert Mehrabian is the originator of the 7-38-55 theory. The theory explains the different weights that verbal and non-verbal expressions have when people are communicating with someone. Specifically:

- 55% of our message comes from body language
- 38% of our message comes from tone of voice
- Only 7% of our message is conveyed by the words we use

These three essential elements, Mehrabian argues, account for how people convey their liking, or disliking, of another person. He especially focused on the importance of the situation in which such nonverbal ‘clues’ appear to conflict with the words used. In fact, often verbal and non-verbal messages are consistent, but they can sometimes be inconsistent, or in other terms, incongruent. If someone's words conflict with their tone of voice and/or non-verbal expressions, we often tend to believe the non-verbal clues rather than words. Mehrabian has written and researched broadly on the study of nonverbal communication. Later, he has expanded his field of interest from nonverbal communication in relation to the expression of emotions and attitudes to its application in areas such as human response, temperament and traits, and the impact of the emotional workplace environment on performance. Although his studies have brought a great contribution to the study of this issue, they present some limitations concerning both the validity of the findings and their practical application [3]. Its application is limited to cases when the communicator is expressing attitudes or emotions, and when body language and tone of voice contradict the meaning of the spoken words. Mehrabian's theoretical works extend to a series of psychometric scales used to measure different emotions. Since Nonverbal communication involves a large number of symbols (gestures, expressions) that are difficult to conceptualize, he co-created with James A. Russell the PAD (Pleasure, Arousal, Dominance) Emotional State model in order to measure a series of differing emotional conditions. The three fundamental dimensions of reference are:

- The pleasure-displeasure scale: which measures how pleasant emotion is, it is also called "valence"
- The arousal- no arousal scale: which measures the intensity of emotion
- The dominance-submissiveness scale: which measures the dominant nature of an emotion

The standard two-dimensional models with pleasure on the horizontal axe and arousal on the vertical axe is suggested to be replaced by a three dimensional model with dominance on the third axe (see **Fig. 1**).

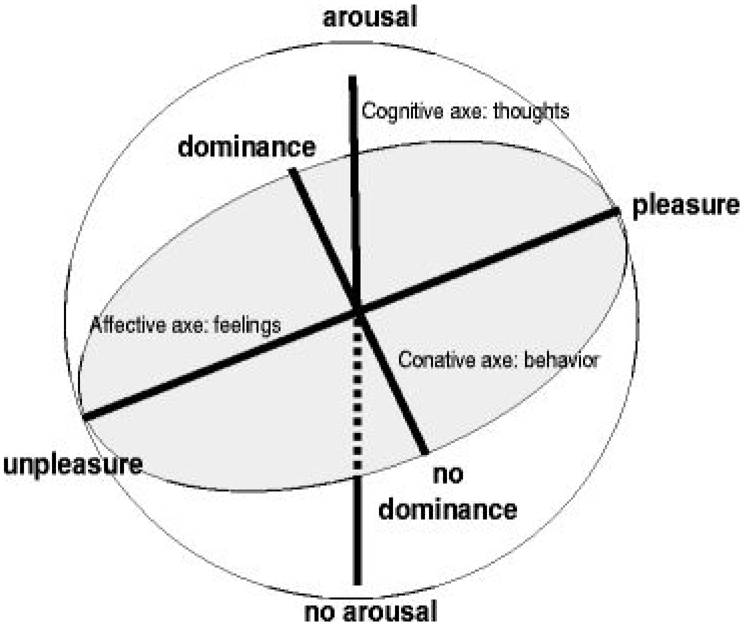


Fig. 1 Three dimensional models of pleasure, arousal and dominance

According to the PAD model published by Mehrabian and Russell in 1974, pleasure, arousal and dominance are introduced as three independent emotional dimensions to describe people's state of feeling. They conceived pleasure as a continuum ranging from extreme pain or unhappiness to extreme happiness and used adjectives such as happy-unhappy, pleased-annoyed, and satisfied-unsatisfied to define a person's level of pleasure. Arousal was conceived as a mental activity describing the emotional state along a single dimension ranging from low to high excitement and linked to adjectives such as stimulated-relaxed, excited-calm and wide awake-sleepy to define arousal. Dominance was related to feelings of control and the extent to which an individual feels restricted in his behaviour. To define the degree of dominance Mehrabian and Russell used a continuum ranging from dominance to submissiveness with adjectives such as controlling, influential and autonomous. The dominance scale represents the controlling and dominant versus controlled or submissive emotions. PAD model is also effective for measuring differences in the personality of individuals and has a widespread application [4]. Thus, this model became a useful tool for measuring consumer behaviour to assess consumer reactions to products, services, different shopping environment, marketing and advertising campaigns.

1.2. Paraverbal communication

In addition to the verbal and non-verbal plane, there is a third means of communicating: the “Paraverbal”. The term means communication through paralanguage, which can be defined as the use of manner of speaking to communicate particular meaning. Paraverbal communication is also known as voice language, it is a deeper level of communication and encoded information is sent through the prosody and vocal elements accompanying the word and speech. It is an ensemble of phenomena that are external from words and sentences but internal to voice and they are merely depending on the speaker attitude. Scholars mainly consider paraverbal features a subsystem of non-verbal ones. This includes all communication signals dealing with voice modulation and level. Inflections or emphasis are phenomena applied vocally to a message. From the point of view of form, paraverbal communication is concomitant to verbal communication (the paraverbal encoded message corresponds to the verbally coded message), but in terms of content, paraverbal communication can give another meaning to verbally coded message. Indeed, modulation or inflection can have an effect on the impact of a message, in some cases, they can completely change the meaning a person would be expected to attach to the words. Paraverbal aspects of the voice are the primary focus of our work. In order to study this topic, the research concentrates the attention on the analysis of prosodic features of speech, overlooking some principle aspects of non-vocal phenomena that do not concern the study of the voice. Prosodic features are those aspects of speech which go beyond phonemes and deal with the auditory qualities of sound. They appear when we put sounds together in connected speech. The study of paraverbal cues of the voice is still very recent and improvable, but researchers have already understood its importance and the potentiality to better understand human attitudes. Paraverbal cues are vocal cues that can be detected internally at speech behaviour and refer to the messages that we transmit through the pitch, pace and tone of our voices. Underlying these 3 key components:

1. *PITCH* is how high or low the voice goes. A person may speak in a high pitch or a low bass or baritone. For instance, if people have a squeaky high pitch it can show that they are afraid, people with a lower pitch can demonstrate calm.
2. *PACE OR CADENCE* is how fast or slow a person talks. Speaking quickly can indicate being nervous, or it can show excitement.
3. *TONE* is a combination of factors that set or convey meaning. It can be using more inflection in the voice, rather than speaking in a monotone.

1.2.1. Potentiality of paraverbal

As we previously anticipated the potentiality of paraverbal is huge in terms of personal human patterns recognition. Our voice transmits both direct and hidden messages to other people, and using paraverbal signals, we can more accurately decode the messages that other people send us. These messages of the encoder tend to reveal the degree of presence or absence of sincerity, honesty, conviction, ability, and qualifications. Paraverbal signals are clearly superior to verbal content for the contribution of particular insights [5].

They provide information about group affiliation, such as a particular dialect or accent, speaker's identity, biomedical information etc. Moreover, the most valuable knowledge they can bring, from the point of view of our work, is about the emotional state of a person. From the timbre, volume, voice modulations we can understand if the speaker is male/female, young/old, child/adult, hesitant/determined, energetic/exhausted. The pronunciation features provide information about the homeowner's environment such as urban/rural, geographical area, level of education, etc. The rhythm, the intensity, the speech flow and the use of pauses indicate which are the key elements of the speech. In fact in a speech, the less important and known elements are presented at a faster pace, instead, the main ideas are prepared with pauses in speech and less exposed. Variation in tone of the voice is useful to drive attention to important ideas, to keep a listener engaged in a subject or to calm a tense situation. Also, tone can radically change the sense and meaning of a message. It is the essential factor in the personalization of communication and in the authentic perception of the message, establishing the meaning behind the words.

From all these considerations, we can assess that the great importance of paraverbal aspects is that they affect the effectiveness of communication because it is influenced by our moods and emotions. Recently, increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. The research around the different ways of communication can be extremely wide and can open many topics. Since the paraverbal communication is the part of the voice that contains and conveys emotions, our work extends the scope only around the analysis of paraverbal. In fact, our aim is to study the voice in relation to emotion recognition.

2. Research Objectives and Methodology

In this part, we identify research's questions and objectives that can help readers in understanding the scope of our work on voice analysis topic. First, we agreed on selecting three main questions:

1. Which is the current degree of scientific knowledge about the study of voice?
Answering this question means to investigate the state-of-the-art of voice analysis, making a systematic analysis of the literature and review theoretic research, empirical experiments and conference paper, in order to realize what can be done with the study of voice.

2. Which are the most remarkable applications of voice analysis and emotion recognition in business?
To answer this point, an examination of voice analysis in the global market across several industries is needed. The aim is to figure out which are the sectors of interest and what companies are doing to implement such technologies in their business. Moreover, recognizing how broad can be the variety of applications and identifying which are the cases of success, could be helpful to categorize and analyze such potential business applications.

3. Which could be the future developments of such technologies and which opportunities can bring?
This question is functional to investigate the potentiality of voice analysis and emotion recognition in the future and the benefits they can bring to the organizations.

As concerns the methodology that has been adopted for the literature review and articles' research, we would like to point out some key steps that allow us to limit the boundaries of our issue's domain. Conducting a systematic review of the literature we have selected the most relevant articles for our research, pointing out which are the best journals that investigate the issue. We have decided to classify the articles in two main repositories, depending on the perspective of the research:

- Repository 1 contains the articles that examine voice from a biomedical point of view, thus theories and empirical experiments sharing a common purpose, to create a scientific defining knowledge of voice analysis.
- Repository 2 includes articles that explore the potentiality of the paraverbal in business applications. There is a broad variety of business sectors in which is possible to apply the study of paraverbal, so it is necessary to classify them in categories of interest with different priorities.

As a consequence of preliminary readings of specific articles about voice analysis that have been found using two main search engines Scopus and Google Scholar, it has been possible to identify some recurring keywords that can be directly associated to the thematic. The research through keywords is different between the two repositories, in particular, repository 2 is an extension of repository one in terms of quantitative and qualitative selection. In ***tab.1***, for each repository we show the names of principle journals that contain the selected articles and the keywords that we have utilized for the articles' research and selection.

REPOSITORY 1	**JOURNALS**	**KEYWORDS**
	Pattern Recognition	Voice analysis
	Neurocomputing	Speech processing
	IEEE Signal Processing Magazine	Voice recognition

	IEEE Transactions on Audio, Speech, and Language Processing	Prosodic features
	Elsevier Speech Communications Journal	Speaker recognition
	Procedia computer science	Paraverbal communication
	International Journal of Computer Applications	Emotion recognition
	International Journal of Computer Science Trends and Technology	

REPOSITORY 2	**JOURNALS**	**KEYWORDS**
	International Journal of Science and Advanced Technology	Voice analysis
	Journal of Management	Emotions
	Journal of the Academy of Marketing Science	Consumer behaviour
	Journal of Business Research	Speaker recognition
	Brain Research Bulletin	Human - Robot interaction
		Social Robot
		Artificial intelligence
		Smart speakers

Tab.1 *Articles repositories: main journals and keywords*

Through a logic combination of keywords implemented on the search engines, we have performed the research of the articles related to our scope. We present here an example of advanced research made on Scopus:

(TITLE-ABS-KEY (voice AND analysis) OR TITLE-ABS-KEY (voice) OR TITLE-ABS-KEY (paraverbal) OR TITLE-ABS-KEY (speech AND recognition) OR TITLE-ABS-KEY (voice AND recognition) OR TITLE-ABS-KEY (vocal AND prosody) AND TITLE-ABS-KEY (emotion)) AND (TITLE-ABS-KEY (purchasing AND behaviour) OR TITLE-ABS-KEY (consumer AND behaviour) OR TITLE-ABS-KEY (business) OR TITLE-ABS-KEY (marketing) OR TITLE-ABS-KEY (IoT) OR TITLE-ABS-KEY (artificial AND intelligence) OR TITLE-ABS-KEY (social AND robot) OR TITLE-ABS-KEY (HRI) OR TITLE-ABS-KEY (vocal AND assistant) OR TITLE-ABS-KEY (smart home))

The articles have been selected and inserted in a classification tab that contains relevant information. The information considered of interest and extrapolated are reported below:

- Title
- Authors
- Source (Journal)
- Publication year
- Number of citations
- Volume
- Issue
- Pages
- Affiliation
- Keywords
- Category (the topic of the articles)
- Type of the article
- The objective of the article
- Synthesis

This information has been useful to filter and manage materials and to group them into categories. Moreover, by evidencing and separating this data, the key messages of each article have been easier highlighted and it has been fruitful also later in the drafting phase.

3. Voice analysis: fundamentals and state of the art

The following capitula are supposed to provide an overview of the voice analysis process. The objective is to summarize different findings coming from recent scientific studies and to underline the current state-of-the-art of the acquaintance about the topic.

3.1. Definition of voice analysis

Voice analysis is the study of paraverbal features of the vocal communication that aim to recognize biometrics characteristics of the voice. It is the study and treatment of voice disorders and the features of speech that differ between individuals, since everyone has a unique pattern of speech stemming from their anatomy, based on the size and shape of the mouth and throat, and behavioural characteristics, such as pitch, tone, and speaking style. each individual pattern can be seen as a voiceprint of a person. The voiceprint is represented as a sound spectrogram (see ***fig.2***), which is basically a graph that shows a sound's frequency on the vertical axis and time on the horizontal axis. It displays a graphical representation of the strengths of the various component frequencies of a sound as time passes. Different speech sounds create different shapes within the graph. Spectrograms also use colours or shades of grey to represent the acoustical qualities of sound. There are two main kinds of voice analysis performed by the spectrograph, broadband (with a bandwidth of 300-500 Hz) and narrowband (with a bandwidth of 45-50 Hz). In vocal utterances, the broadband spectrogram shows vertical lines corresponding to the rapid increase in amplitude that occurs when the vocal folds clap together. These vertical lines are not visible in the narrowband spectrogram because frequency analysis is calculated over a much longer time window, too long to capture the rapid increase in amplitude that occurs at the time of vocal fold closure. The narrowband spectrogram has different strength, it is able to isolate each individual harmonic to study them separately.

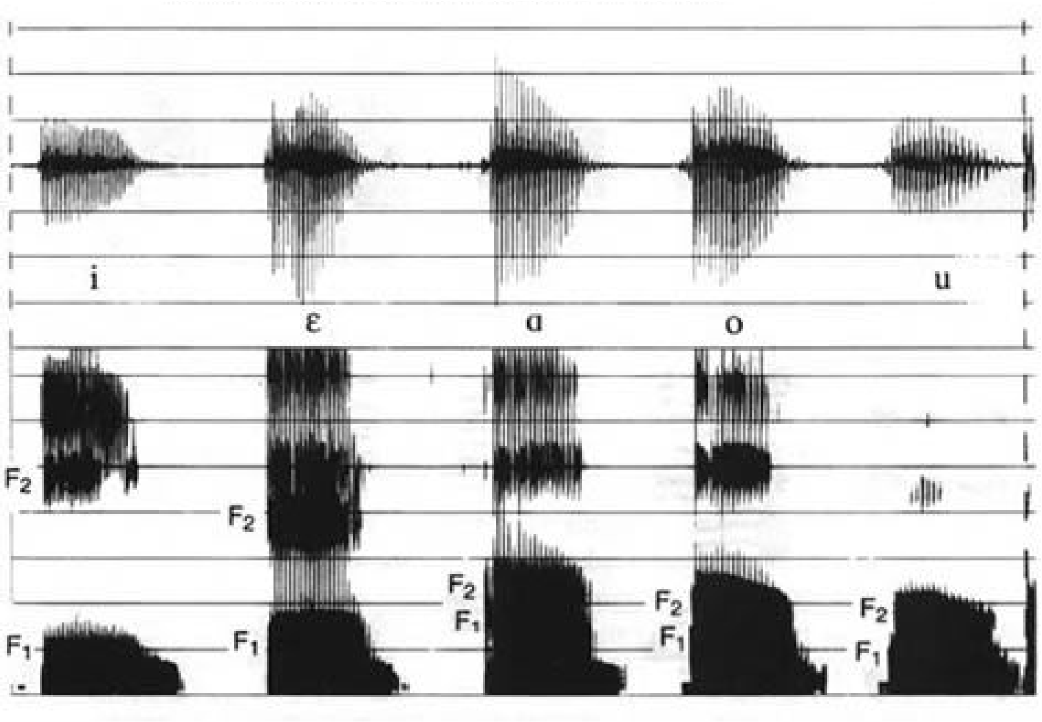


Fig.2 Sound broadband spectrogram of spoken utterances

3.2. Voice recognition vs. speech recognition

Voice recognition and speech recognition are terms that are interchangeably used. However, their meaning refers to different dimensions. The purpose of speech recognition is to arrive at the words that are being spoken. It is the process of capturing spoken words and converting them into a digitally stored set of words. The quality of a speech recognition systems is assessed according to two factors: its accuracy (error rate in converting spoken words to digital data) and speed (how well the software can keep up with a human speaker). Speech recognition technology has endless applications. Some of them that are commonly used are:

- Automatic translations (automatic speech translations in different languages)
- Dictation (automatic speech to text transcriptions)
- Hands-free computing (any computer configuration where a user can interface without the use of their hands)
- Automated customer service (automated system such as a help center or chatbots)

Voice recognition aims to recognize the person speaking the words, rather than the words themselves. Voice recognition works by scanning the aspects of speech that differ between individuals. Everyone has a way of speaking unique to them. Voice recognition technology is used to confirm the identity of the speaker or determine the identity of an unknown individual. Speaker verification and speaker identification are categories of voice recognition. The applications of voice recognition are markedly different from those of speech recognition. Most commonly, voice recognition technology is used to verify a speaker's identity or determine an unknown speaker's identity. We propose in ***tab.2*** a summary of the 4 key differences between voice recognition and speech recognition

	Voice recognition	**Speech recognition**
Recognition	Recognizes who is speaking by measuring the voice pattern, speaking style and other verbal.	Recognizes what is being said and converts them into text.
Purpose	Identify the speaker	Find out & digitally record what the speaker is saying.

Focus	Biometric aspects of the speaker to recognize them. As every voice is unique, the technology can identify the speaker by analyzing some other indicators like tempo, timbre, pitch, of their voice.	Vocabulary of what is being said by the speaker. Then, it turns the words into digital texts.
Applications	User Authentication Sentiment Analysis Human-Robot Interactions Marketing Research	Automatic translation Dictation Hands-free computing Automated customer service

Tab.2 *Voice recognition – Speech recognition key differences*

3.3. Voice Analysis and emotion recognition

Recently, increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. First, we analyze the state-of-the-art of speech signalling and processing, that exposes the current knowledge advancements about the topic from a merely technical and biomedical point of view. Starting from the illustration of different existing databases used in the experiments, we identify and group the features that are known describing voice signals. Afterwards, we study the different steps of the voice analysis process, we list and evaluate distinct types of emotions' classifiers. Indeed, we proceed to investigate the voice analysis field regarding the potential adoptions for emotions' interpretation purposes.

3.3.1. Emotional speech and emotional speech database typology

In order to design an effective voice-emotion recognition system, one of the several starting points is the proper preparation of an emotional speech database for evaluating system performance. Each database consists of a corpus of human speech pronounced under different emotional conditions. They can be recorded in one single language (English language is dominant) or multi-language. Each database is set up to contain specific emotions. The most common recordings are anger, sadness, happiness, fear, disgust, surprise, boredom and joy. Progress in applications related to emotional speech relies heavily on the availability of suitable databases. An important issue to be considered in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. According to this, they can be categorized into simulated, semi-natural and natural databases.

Simulated databases are collected from speakers by prompting them to enact emotions through specified text in a given language. The portrayed emotional expression makes use of professional or non-professional actors and actresses. Subjects are instructed to produce emotional expressions for various emotion classes, with varying degrees of intensity or arousal. Normally, speakers are supposed to emote the same text in different emotions.

Semi-natural database is created with enacted expressions, where the context is given to the speakers. The database is developed by inducing emotional expression, where users are presented with scenarios that required an emotional response. It is built by asking speakers to enact the scripted scenarios eliciting each emotion.

Natural database has the characteristic that recordings do not involve any prompting or eliciting of emotional responses, but data is collected from a real-world situation where users are not under obvious observation and are free to express emotions naturally. Sources for such natural situations could be talk shows, interviews, panel discussions and group interactions in the TV broadcast.

It is important to underline the trade-off that exists between the controllability and naturalness of the databases' recordings. Researchers believe that using spontaneous data, it is easier to incur in several difficulties, but it paints a more realistic picture of applied automatic emotion recognition, while acted data are easier to be managed controlled but may lead to no reliable results [6]. Natural datasets have the difficulty in identifying the emotion or expressive state of dialogue. In fact, there are some cases of ambiguity that researchers have faced during speech processing. One of them is the occurrence of mixed emotions in an utterance. For example, the combinations surprise-happy, frustration-anger and anger-sad, may occur frequently in the dialogue. The second reason for the ambiguity is due to the instability of emotion during the dialogue. In natural communication, emotion may not be sustainable over the entire duration of the dialogue. The emotion is expressed mostly in some segments of the dialogue, with the rest of the dialogue being neutral. Therefore, it appears difficult to define how long the emotion lasts during the occurrence of a continuous speech.

In ***tab.3*** we present a list of available databases commonly used to implement voice recognition experiments. We also report both the available language versions and emotions that are classified inside the utterances for each dataset.

Corpus	Language	Emotions
LDC Emotional Prosody Speech and Transcripts	English	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database	German	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional	Danish	Anger, joy, sadness, surprise, neutral

database		
Natural	Mandarin	Anger, neutral
ESMBS	Mandarin	Anger, joy, sadness, disgust, fear, surprise
INTERFACE	English, Slovenian, Spanish, French	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
KISMET	American English	Approval, attention, prohibition, soothing, neutral
BabyEars	English	Approval, attention, prohibition
SUSAS	English	Next to neutral speech and fear, different stress conditions are collected: medium stress, high stress, and screaming
MPEG-4	English	Joy, anger, disgust, fear, sadness, surprise, neutral
Beihang University	Mandarin	Anger, joy, sadness, disgust, surprise
FERMUS III	German, English	Anger, disgust, joy, neutral, sadness, surprise
KES	Korean	Neutral, joy, sadness, anger
CLDC	Chinese	Joy, anger, surprise, fear, neutral, sadness
Hao Hu et al.	Chinese	Anger, fear, joy, sadness, neutral
Amir et al.	Hebrew	Anger, disgust, fear, joy, neutral, sadness
Pereira	English	Hot anger, cold anger, joy, neutral, sadness

Tab.3 *Characteristics of common emotional speech databases*

Some experiments have been carried out in order to compare different kinds of database. The efforts made by researchers aim to improve the automatic perception of vocal emotion. For instance, in [7] the scope is the automatic interaction and man-man interaction evaluations in call-centres. Specifically, the focus of the research is to raise the perception of human emotion of automatic recognition system from the prosodic properties of speech, comparing two types of data acquisition methods: spontaneous and acted.

Spontaneous emotional speech

Data is collected from a real-world situation, people are free to express their emotions in a natural way. For this reason, the dataset collected using this method is the so-called NATURAL. The database for this empirical study is provided by a call-centre for an electricity company. Each conversation was segmented into phrase-level utterances. A total amount of 391 utterances were recorded from 11 different speakers. The emotion classes taken into account are only two (anger and neutral emotional states). The emotion recognition system performed the task with a classification accuracy of 81.95%.

Acted emotional speech

Data is collected from non-professional actors and actresses portraying emotional speech and gathered for a previous study on emotion recognition. The ESMBS database (Emotional Speech of Mandarin and Burmese Speakers) was composed by 720 emotional utterances recorded from speeches of 12 different actors. Six emotion classes were analyzed in this study (anger, happiness, sadness, disgust, fear, surprise). Average classification accuracy by human evaluation was found to be 65.7%. (68.3% for Burmese and 63.1% for Mandarin). These results coincide with other studies which typically describe human classification rates between 55% and 70%.

We present in ***tab 4.*** an overview of the two different types of databases with relative descriptions, examples, advantages and disadvantages.

	Acted	Natural
Advantages	Most commonly used and standardized Comparison of results is easy Number of emotions available are large	Completely natural Useful for real-world emotion systems modeling
Disadvantages	It tells how emotions should be portrayed rather than how they are portrayed Context, environment and purpose dependent information is absent Episodic in nature, not true in real-world situations	Emotions are continuous All emotions are not available Contains multiple and concurrent emotions Difficult to model Copyright and privacy issues arise

Tab.4 *Emotional speech databases: Acted vs. Natural*

3.3.2. Types of Features / Variables

Features identification is one of the most important steps for developing a comprehensive analysis of the voice. The focus is on understanding the contribution of speech prosody features towards production of emotion. Feature representation plays a key role in developing any emotion-related applications. The speech features used in many analysis studies can be broadly categorized into *Acoustic features, Tone based, Voice Quality features, and Others.*

3.3.2.1. Acoustic features

Acoustic features present various subcategories, depending on factors and parameters that are taken into account and dimensions used to monitor the chosen variables.

Pitch - Fundamental frequency

Fundamental frequency is defined as the lowest frequency at which speech signal repeats itself. The fundamental frequency of a periodic signal is the inverse of its period. It is also defined as the rate of vibration of the vocal folds. Periodic vibration at the glottis may produce speech that is less perfectly periodic because of movements of the vocal tract that filters the glottal source waveform. F0 estimation is a topic that continues to attract much effort and ingenuity, despite the many methods that have been proposed. One of the earliest and most comprehensive approach was proposed by Hermes in 1993 [8]. Afterwards, other methods have been suggested such as instantaneous frequency, statistical learning and neural networks and auditory models. Speech frequency variations contribute to prosody, and in tonal languages, they help distinguish lexical categories. Attempts to use F0 in speech recognition systems had mitigated successful results, in part because of the limited reliability of estimation algorithms. The fundamental frequency is a useful ingredient for a variety of signal processing methods too. Fundamental frequency is the main parameter to be studied when implementing voice analysis, especially if the purpose is to study the emotional aspects of the speech.

The fundamental frequency is a great source of emotional insights, for example:

- Neutral or unemotional speech has a much narrower pitch range (max-min) than that of emotional speech;
- Frequency and duration of pauses and stops normally in neutral speech are decreased;
- Neutral speech typically displays a ‘‘uniform formant structure and glottal vibration patterns,’’ contrasting the ‘‘irregular’’ formant contours of fear, sadness, and anger;
- Angry speech typically has a high median, wide range, wide mean inflection range, and a high rate of change. Vowels of angry speech have the highest F0

- Fear was discovered to have a high pitch median, wide range, medium inflection range, and a moderate rate of change;
- Sadness is shown to yield lower pitch mean and narrow range;
- Disgust generally has a low pitch median, wide range, lower inflectional range, lower rate of pitch change during inflection.

Intensity - Energy

The focus of intensity feature is to study the distribution of energy across the spectrum monitoring some parameters such as mean, median, standard deviation, maximum, minimum, range (max-min), linear regression coefficients etc. Energy usually refers to the volume or intensity of the speech. The amount of energy recorded in vocal expressions are strictly connected to specific emotional states. Therefore, the intensity contour provides information that can be used to differentiate separate sets of emotions.

For example:

- Angry and happy speech have a noticeably increased energy envelope;
- Sadness is associated with decreased intensity;
- Disgust has reduced loudness;
- Fear, joy, and anger present an increase in high-frequency energy;
- Sadness has a decrease in high-frequency energy.

Duration

Features that take part in this group are considered rhythm-based features. They are described by speech rate, the ratio of duration of voiced and unvoiced regions, and duration of the longest voiced speech. Duration is usually calculated by measuring the number of syllables per second. Properties of rhythm-based characteristics include pauses between voiced sounds, lengths of voiced segments, and rate of speech, even called articulation. Different emotions imply different durations of utterances, hence different speech rhythms describe distinct emotions.

For instance:

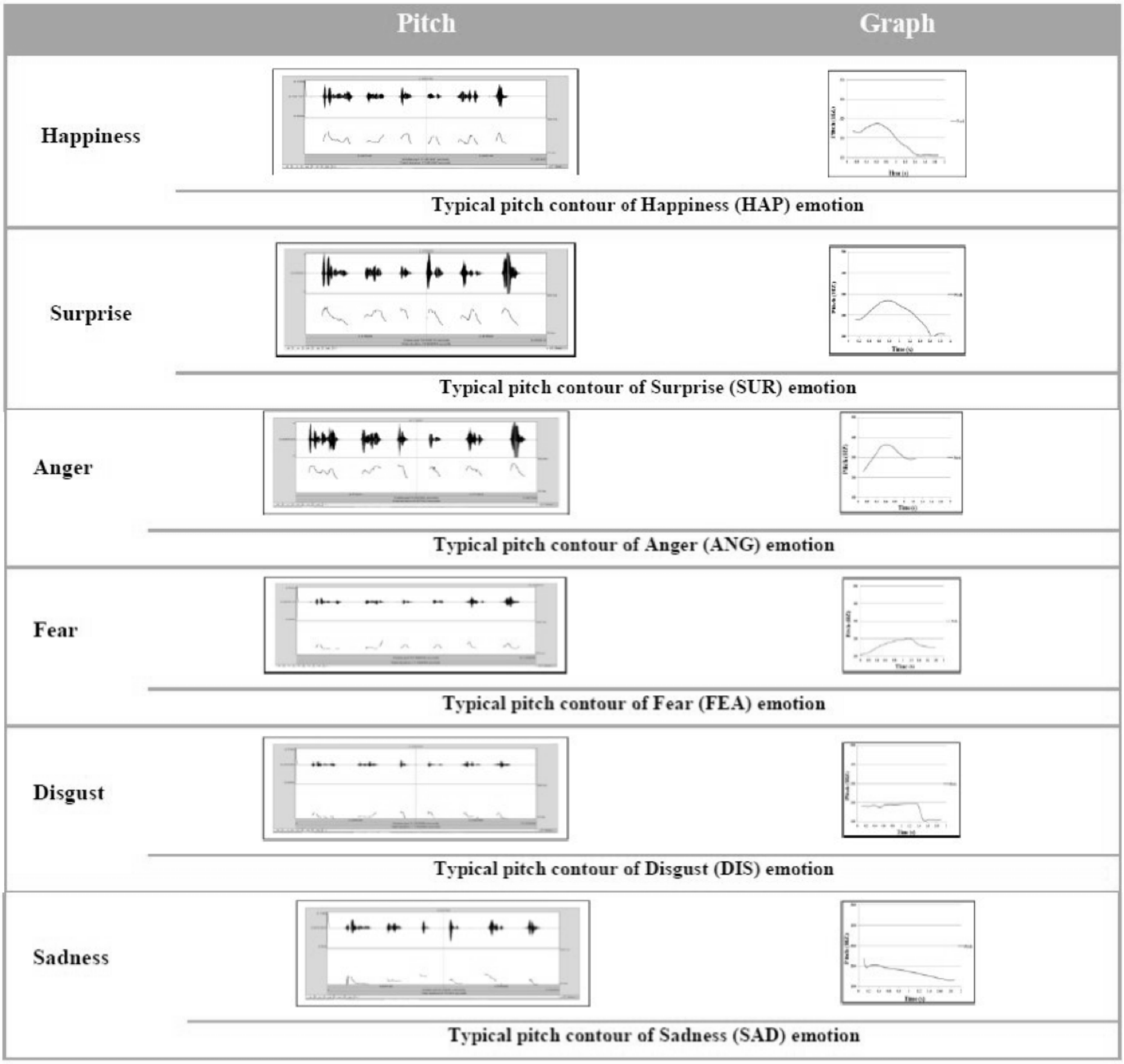
- Surprise had a normal tempo;
- Sadness has been shown to have a reduced articulation rate speech contains ‘‘irregular pauses’’;
- Anger had an increased speech rate and pauses forming 32% of the total speaking time;
- Disgust had a very low speech rate, increased pause length, with pauses typically comprising 33%;
- Divergent opinions have resulted about Fear and Happiness.

Spectral features

Vocal cord vibration generates a spectrum of harmonics, which is selectively filtered as it passes through the mouth and nose, producing the complex time-varying spectra from which words can be identified. The spectrum characterized by formant frequencies and their respective bandwidths is extensively analyzed for emotional speech. Spectral centroid indicates the centre of signals spectrum power distribution. Spectral flux is a measure of how quickly the power spectrum of a signal changing [9]. These features can be useful to identify the notes, pitch, rhythm, and melody. It is also interesting to note that there are certain changes in the spectral component which are associated with the glottal source excitation. In fact, the syllables produced with higher fundamental frequencies in angry speech tend to have weaker frequency amplitudes. A more closed phase of glottis configuration results in relatively higher amplitudes at high frequencies. A fundamental characteristic of the spectrum of a speech signal is that it is sound- specific. The deviations in spectral features are analyzed for utterances having the same lexical content. It is recognized that the emotional content of an utterance has an impact on the distribution of the spectral energy across the speech range of frequency. The most common spectral features are Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), Log-Frequency Power Coefficients (LFPC), Modulation Spectral features, Formant frequencies. From the literature review, the most successful spectral features used in speech are Log- Frequency Power Coefficients (LFPC) and Mel-Frequency cepstral coefficients (MFCC). Research has divergent opinions about which one is the most effective one. The emotional state of the speaker is unlikely to change as fast as phonemes. MFCC assigns one emotion to one short utterance of a few seconds. Hence, suprasegmental features are needed with one feature value per utterance. For these reasons, MFCC features seem to be less successful for emotion recognition since the focus is more on paralinguistic than linguistic information. From experiments' result [6], that the LFPC provided an average classification accuracy of 77.1% while the MFCC gave 59.0% identification accuracies.

3.3.2.2. Tone Based features

Tone-based level describes prosody in terms of intonational phrases or tone groups. Each tone group contains a nuclear tone, which can be rising or falling movement, combination or level tone usually on the last stressed syllable of the group. The part that leads up to this nuclear tone is called the head, the part following is the tail. Tone variation creates different geometric patterns that are usually described as pitch contours. Different types of tones and heads are usually associated with different emotions or attitudes [10]. The validity of the relation between tone shape and emotion is scientifically demonstrated, as well as relation between the phonetic realization of a tone (high rise or low rise) and emotion. Heads also can take different shapes, each shape expresses different emotions (see ***fig.3***). The emotions listed cover a wide spectrum. There are different depths of investigation about different types of emotions. Sometimes particular patterns are listed simply as emotional or neutral, sometimes very specific labels are given in order to classify the variety of emotional shades. This issue will be better analyze in the *Classification Methods* chapter.



Typical pitch contour of Happiness (HAP) emotion

Typical pitch contour of Surprise (SUR) emotion

Typical pitch contour of Anger (ANG) emotion

Typical pitch contour of Fear (FEA) emotion

Typical pitch contour of Disgust (DIS) emotion

Typical pitch contour of Sadness (SAD) emotion

Fig.3 *Typical pitch contour shapes of basic emotions*

3.3.2.3. Voice Quality (VQ) features

The term voice quality refers to the characteristic auditory which gives colour to the individual's speech. The term of voice quality has two main definitions. In a broad sense, this term describes all phonatory, articulatory and overall muscular tension of individual speech. It is also used in a narrow sense, to describe only phonation types or laryngeal qualities, such as breathiness, creakiness, harshness. By varying voice qualities, some important information like intentions, emotions and attitudes are conveyed to the listeners. Each different phonation aspect is linked to an emotional state. For instance, Breathiness is associated with angry and happy speech. Vocal fry voice is observed in sad and relaxed speech. A harsh voice, which corresponds to irregularity invoicing, was observed in fear speech. Voice quality features are usually included in order to track the phonation process and to increase the recognition performances of the system. Voice quality measures, which have been directly related to emotion, include the open-to-closed ratio of the vocal cords, jitter, harmonics-to-noise ratio, and spectral energy distribution. Experimental studies [11] demonstrate a strong relation between voice quality and the perceived emotion, however, there are still some conflicts between researchers on how to associate vocal quality terms to emotions. There is a large panel of micro-prosodic features. Among them, the most commonly used are the ones that give an indication on how voiced the speech signal is. The local jitter and shimmer evaluate the small-time variation of fundamental frequency and energy. Voice quality and prosodic contour are strictly connected and sometimes, separating them appear difficult due to their interaction. These difficulties concerned especially those affective utterances for which speakers used important modifications of articulatory reduction level, overall tension or pitch excursions. The articulation rate is one of the prosodic features which interacts strongly with voice quality parameters. The articulatory range describes the excursion size of movements for the lips, jaw and tongue. When the articulation rate increases, the speakers reduce their articulation range. As a result, vowels are less distinguished from each other, some of them are reduced or even dropped. Pitch parameters also interact with voice quality. For example, wider pitch excursions are associated with more tenseness and vowel prolongation. From the literature review, researchers used to analyze the following VQ variables, since they have some significant impact on emotion detection.

- *Mean Rd*
- *Std Rd*
- *Jitter*
- *Shimmer*

Rd means Relaxation coefficient, which is one parameter estimated from glottal models. The coefficient varies from 0 (very tense) to 2.5 (relax). The more important the Rd, the more relax the voice. Jitter and shimmer are acoustic characteristics of voice signals, and they measure the cycle-to-cycle variations of fundamental frequency and amplitude, respectively [12].

3.3.2.4. Other features

Interval features (INT)

According to music theory [13], the harmony structure of an interval or chord is mainly responsible for producing a positive or negative impression on the listener. It might be that certain frequency pairs cause a more pleasant impression on the listener than others. Interval features compose a subset of harmony features.

TEO-based features *(Teager Energy Operator)*

According to experimental studies done by Teager, the speech is produced by nonlinear airflow in the vocal system. Vortex flow varies accordingly to the emotional state of anger or stressed speech because there is fast airflow that causes vortices located near the false vocal folds that give additional excitation signals other than pitch. To measure this energy which is produced by such a non-linear process, Teager developed an energy operator which is known as Teager Energy Operator (TEO) [6].

We summarize the feature classification mentioned above in ***tab.5.*** We have agreed on considering only the basic emotions identified by Ekman. More information about Ekman's study will be presented in the next chapter. Each basic emotion is characterized by different voice behaviours, divided in categories. Results show the variation of parameters that are grasped and analysed for each emotion.

		Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**	**Surprise**
Acoustic	Pitch	Increase in mean, median, range, variability	Increase in mean, range, variability	Below normal mean F0, range F0	Increase in mean F0, range F0, perturbation, variability F0 movement	Very low range, low median, raised mean F0, slow change	Wide range, median normal or higher
	Intensity	Raised	Increased	Decreased	Normal	Slow—due to high rate of pause to phonation time, longer vowels and consonants	Tempo normal, tempo restrained
	Duration	High rate	Increased rate	Slightly slow, long pitch falls	Increased rate	Not clear	Not clear
	Spectral	High midpoint for av spectrum	Increase in high-frequency energy	Decrease in high-frequency energy	Increase in high-frequency energy	Not clear	Not clear

Tone Based	Angular frequency curve, Stressed syllables ascend frequently and rhythmically, irregular up and down inflection	Descending line, melody ascending frequently and at irregular intervals	Downward inflections	Disintegration of pattern and great number of changes in direction of pitch	Long sustained falling intonation throughout each phrase	Fall rise nuclear tone with falling head, high fall preceded by rising head, high rise tone
Voice Quality	Tense, breathy, heavy chest tone, blaring	Tense, breathy, blaring	Lax, resonant	Tense	Whisper	Breathy
Other	Clipped speech, irregular rhythm basic opening and closing, articulatory gestures for vowel / consonant alternation more extreme	Irregular stress distribution, capriciously alternating level of stressed syllables	Slurring, rhythm with irregular pauses	Precise articulation of vowel/consonant, voicing irregularity due to disturbed respiratory pattern	Voicing irregularities	Not clear

***Tab.5** Features Classification and Emotions*

3.3.3. Voice-Emotion Recognition Process

Generally, the term emotion describes the subjective feelings in short periods of time which are related to events, persons, or objects. Since the emotional state of humans is a highly subjective experience, it is hard to find objective and universal definitions. This is the reason why there are different approaches to model emotions in psychological literature. The most popular approach is the definition of discrete emotion classes, the so-called “basic emotions” formulated by Ekman in 1992. Ekman defined six basic emotions the humans are well familiar with: happiness, sadness, anger, fear, disgust, and neutral. Each of the basic emotions is not a single affective state but a family of related states. Each member of an emotion family shares certain characteristics, for example, commonalities in expression and configurational (muscular patterns) features. The usage of the term “basic” is to postulate that other non-basic emotions are combinations of basic emotions, which may be called blends or mixed emotional states. In fact, “the concept of emotion families, allows the inclusion within a family of many variations around a common theme. Thus, many different emotion terms will be found within each family” [14]. As for example, scorn is a term which describes the co-occurrence of two quite different emotions. It is considered by the author a blend of enjoyment and disgust. Emotion recognition is a statistical pattern classification problem. It consists of three major steps, feature extraction, selection and emotions classification. While the theory of classification seems to be pretty well developed, the remaining two are highly empirical issues and depend strongly on the application and database at hand.

3.3.4. Dimensionality Reduction Techniques

It is common to use dimensionality reduction techniques in speech emotion recognition applications in order to reduce the storage and computation requirements of the classifier and to have an insight into the discriminating features.

Still, in practice, dimension reduction makes classifier estimation easier, which often more than compensates for the possible loss of classification information in the transformation. Feature extraction and feature selection are two principle types of dimensionality reduction techniques [13].

3.3.4.1. Feature Extraction

It is important to efficiently characterize the emotional content of speech and at the same time making this characterization not depending on the speaker or the lexical content. While some researchers follow the ordinary framework of dividing the speech signal into small intervals, called frames, from each of which a local feature vector is extracted, other researchers prefer to extract global statics from the whole speech utterance. The majority of researchers have agreed that global features are superior to local ones in terms of classification accuracy and classification time. Global features show a main advantage over local features, their numerosity is reduced. Therefore, the implementation of feature extraction, selection and classification algorithms using global features is executed much faster than if applied to local features. Nevertheless, researchers have claimed that global features are efficient only in distinguishing between emotions characterized by high-arousal dimension. High-arousal emotions are, for instance, anger, fear, and joy. On the other hand, sadness is considered the main low-arousal emotion. Moreover, another disadvantage of global features is that temporal information present in speech signals is completely lost. Again, since the number of training vectors is much less with global features, it may be unreliable to use complex classifiers such as the Hidden Markov Model (HMM) and the Support Vector Machine (SVM) because numerosity of parameters may not be sufficient to populate the algorithms. Feature extraction techniques aims at finding a suitable linear or nonlinear mapping from the original feature space to another space with reduced dimensionality while preserving as much relevant classification information as possible. Some of the toolkits which are widely used for feature extraction are PRAAT, APARAT, OpenSMILE, OpenEAR, RAPT, ASSESS [15]. We present in ***tab.6*** a scheme of each feature extraction method associated with the most valuable types of features that are meant to extract.

METHODS	TYPES OF FEATURES
PRAAT	Acoustic - Voice quality features
APARAT	Voice quality features
OpenSMILE	Acoustic - Voice quality - Spectral features
OpenEAR	Acoustic - Spectral features
RAPT	Pitch contour
ASSESS	Acoustic - Spectral features

Tab.6 *Extraction Methods and Types of Features*

PRAAT is a prosodic feature extraction tool capable of extracting a variety of pitch features, duration features and energy features given an audio recording and its time aligned words and phones sequences. Researchers find Praat very functional because of its public domain that is supported on a variety of platforms (Windows, Macintosh, Linux, and Solaris). It also provides an existing suite of high-quality speech analysis routines, such as pitch tracking. The tool first extracts a set of basic elements representing the info of types of variables such as duration and energy information. Then a set of statistics (means and variances of pause duration, phone duration, and last rhyme duration) are calculated. Finally, it extracts the prosodic features at each word boundary. Another advantage of this tool is its flexibility, in fact, it is easy to add new features into the system. As can be observed in some experiments, using the full set of prosodic features is a virtuous choice because it improves performance considerably and also reduces overall error rate compared to using one type of feature alone [16].

APARAT is a glottal inverse filtering toolbox used for glottal flow extraction. Using this flexible software package based on iterative adaptive inverse filtering (IAIF), the estimation of the glottal flow is fairly accurate. Aparat software is freely available under an open-source license. Furthermore, some researches state that no other software incorporates such a comprehensive set of inverse filtering parameters and inverse filtering analysis tools in a package immediately usable by professionals and easily applicable in other software. TKK Aparat is useful not only in traditional speech research studies, but also in applied disciplines such as the analysis of voice fatigue in the study of occupational voice.

OpenSmile is an open-source feature extractor for incremental processing. SMILE is an acronym for Speech and Music Interpretation by Large-space Extraction. Its aim is to unite features from two worlds, speech processing and music information retrieval, enabling researchers in either domain to benefit from features from the other domain. The software supports audio low-level descriptors such as Mel-frequency cepstral coefficients, perceptual linear predictive cepstral coefficients, linear predictive coefficients, line spectral frequencies, fundamental frequency, and formant frequencies. OpenSMILE provides a simple, scriptable console application where modular feature extraction components can be freely configured and connected via a single configuration file. No feature has to be computed twice, since output from any feature extractor can be used as input to all other feature extractors internally. Unit tests are provided for developers to ensure exact numeric compatibility with future versions [17].

OpenEAR is an emotion recognition toolkit for audio and speech affect recognition. It is an open-source software freely available under the terms of the GNU General Public License. OpenEAR aims at being a stable and efficient set of tools for researchers and those developing emotional aware applications, providing the elementary functionality for emotion recognition. OpenEAR's advantages are that it combines everything from audio recording, feature extraction, and classification to evaluation of results, and pre-trained models while being very fast and highly efficient. Also, OpenEAR can be used as an out-of-the-box emotion live affect recognizer for various domains [18].

Robust Algorithm for Pitch Tracking (RAPT) uses the cross-correlation function to identify pitch candidates and then attempts to select the ‘‘best fit’’ at each frame by dynamic programming. One of the benefits of using the cross-correlation function is that it does not suffer the windowing dilemma of the autocorrelation function while maintaining resolution for high pitch values and the ability to detect low pitch values.

ASSESS by Cowie and Douglas-Cowie, is a kind of feature extraction represents a natural first stage for emotion recognition. Automatic analysis routines generate a highly simplified core representation of the speech signal based on a few landmarks. These landmarks can be defined in terms of a few measures. Those measures are then summarized in a standard set of statistics. The result is an automatically generated description of central tendency, spread and centiles for frequency, intensity, and spectral properties. However, ASSESS has not generally been used in that way. Instead the measures described above have been used to test for differences between speech styles, many of them at least indirectly related to emotion. The results indicate the kinds of discrimination that this type of representation could support.

3.3.4.2. Feature selection

Feature selection is a crucial step in the development of a system for identifying emotions in voice. The main objective of feature selection is to find the feature subset that achieves the best possible classification between classes. The classification ability of a feature subset is usually characterized by an easy-to-calculate function, called the feature selection criterion. The aim is to select what kinds of features have a greater impact on emotions. This is necessary to reduce the computational complexity in real-time applications, if the number of training patterns is too small, in order to avoid “curse of dimensionality” phenomenon. We list below some principle techniques commonly applied in empirical experiments [19].

Correlation analysis

Correlation analysis method consists of distance analysis, partial correlation analysis, and bivariate correlation analysis. Features are divided into groups through Distance Analysis, each group contains variables with similar characteristics. Partial correlation analysis of each feature in the group is carried out. Then, rank correlation analysis of the similar emotion is respectively carried out for similar features in order to calculate the correlation coefficient between the two variables. The logic is that features that may affect other features should be removed.

Principal component analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. The process is explained step by step:

Step 1: Standardization

The operation is done in order to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

Step 2: Covariance Matrix computation

The aim of this step is to understand if there is any relationship between the variables of the input data set by varying from the mean.

Step 3: Identify the principal components

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables compress most of the information within the initial variables. Hence, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set. Dimensionality is reduced without losing much information by discarding the components with low information and considering the remaining components as new variables.

Fisher criterion

It is one of the most effective methods for reducing the dimensionality of emotional speech features, thus decreasing computational calculations. The purpose is to gain the lower redundancy features and reduce the dimension of selected feature set. Fisher criterion overcomes the problems that PCA is not able to extract the discriminant embedded information from high-dimensional emotional features. The procedure of Fisher criterion is designed according to Fisher criterion and the specification of feature class.

Fisher criterion follows the following steps:

Step 1: Input the feature class S to criterion arithmetic.

Step 2: Calculate the Fisher rate (λFisher) of each dimension according to Fisher criterion arithmetic.

Step3: Calculate the sum of Fisher rate class (λFisher).

Step4: Order the feature of S according to Fisher from large to small, select the sequence of S to T according to the sequenced order

Step 5: Acquire the feature class T after selection.

Sequential Floating Forward Selection (SFFS) algorithm

It is an iterative method to find a subset of features that is near the optimal one. The algorithm proceeds in a first phase with several forward iterations. At each iteration, a new feature is added to the previous feature subset. Afterwards, the following steps are in backward directions. It means the least significant features are excluded as long as the resulting subset is better in terms of the recognition rate than the previous subset with the same number of features. This conditional exclusion step is motivated by the fact that a new included feature may carry information that was already present in other features of the previous subset. Thus, the old features can be removed without losing too much discrimination performance. This process is repeated until the desired feature vector length is reached.

3.3.5. Emotion Classification and Classifiers

Emotion recognition is a supervised learning problem. There is a large number of classifiers for supervised learning. The most popular approaches used in scientific researches are: Hidden Markov models (HMMs), Support Vector Machine (SVM) as an extension of LDA with a high-dimensional feature space, Gaussian mixture models (GMMs), Artificial Neural Networks (ANNs), Bayesian learning and k-NN (k-nearest neighbours). The classifier is estimated using data in the training set and its performance is assessed on independent test set. Several classifiers require the choice of tuning parameters or model or architecture or kernel selection. This is typically based on cross-validation. In order to keep strict separation between the design and the test set, the cross-validation then needs to be done using the training set only.

Hidden Markov Model is the most used classifier adopted for the classification of emotion in speech. It is a suite of different models adaptable to distinctive requirements. A hidden Markov model is a tool for representing probability distributions over sequences of observation. Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral vector sequences. Topology of the HMM may be a left-to-right topology as in most speech recognition applications or a fully connected topology. The assumption of left-to-right topology explicitly models advance in time. In the training phase, the HMM parameters are determined as those maximizing the likelihood function. Maximization of the function is done using the Expectation-Maximization (EM) algorithm. Next step of the HMM classifier include determining the optimal number of states, the states often have some relation to the phenomena being modeled. Each state cannot be uniquely associated with an observable event hence, the state sequence is not observable. In order to make the model more flexible, it is assumed that the outcomes or observations of the model are a probabilistic function of each state. These are known as the Hidden Markov Models. The actual state sequence is not directly observable, it is “hidden”, it can only be approximated from the sequence of observations produced by the system. Afterwards, the algorithm identifies the type of the observations, which could be discrete or continuous, and the optimal number of observation symbols in case of using discrete HMM or the optimum number of Gaussian components in case of using continuous HMM [20]. After analyzing briefly HMM based technique applied to speech recognition, we discuss advantages and disadvantages of HMMs in speech processing.

Advantages

The strengths of the HMM methods are their mathematical framework which provides straightforward solution to related problems and the efficient learning algorithms that can take place directly from raw data. Further advantages that characterized HMMs are the implementational structure which provides flexibility in dealing with various speech recognition tasks. The implementation is simple, low ratio of edges to states means that large parts of the model are simple straight-line sequences, which are easy to draw and to understand. Moreover, they have a wide variety of applications including multiple alignment, data mining and classification, structural analysis, and pattern discovery.

Disadvantages

The principle limitations of these methods are that the algorithms are expensive, both in terms of memory and compute time and HMM often has a large number of unstructured parameters . The amount of data required to train an HMM is very large. The number of parameters needed to set up an HMM is huge. As a result of the number of parameters to be estimated in a typical set of HMMs, large training data is hard to be obtained. The choice of the right model between HMM suite is also a big issue. In fact, for a given set of seed sequences, there are many possible

HMMs, and choosing one can be difficult. Smaller models are easier to understand, but larger models can fit the data better [21].

Gaussian Mixture Model (GMM) is a probabilistic model for density estimation using a convex combination of multi-variate normal densities. It can be considered as a special continuous HMM which contains only one state. Therefore, GMMs are more appropriate for speech emotion recognition when only global features are to be extracted from the training utterances [6]. Similar to many other classifiers, determining the optimum number of Gaussian components is one of the main issues. Scientific researchers consider GMM as the standard classifier because it operates on atomic levels of speech and can be effective with very small amounts of speaker specific training data.

Advantages

GMMs are very efficient in modeling multi-modal distributions and their training and testing requirements are much less than the requirements of a general continuous HMM. Hence, GMM achieves the best compromise between the classification performance and the computational requirements required for training and testing.

Disadvantages

The main limitation of the GMM algorithm is that, for computational reasons, it cannot support too high dimensionalities of the problem.

Artificial Neural Network (ANN) is another common classifier used for many pattern recognition applications. Almost all ANNs can be categorized into three main basic types: MLP, Recurrent Neural Networks (RNN), and radial basis functions (RBF) networks. While MLP neural networks are relatively common in speech emotion recognition, RBF is rarely used in speech emotion recognition. Commonly, in speech emotion recognition systems, more than one ANN is used to create an appropriate aggregation scheme, useful to combine the outputs of the individuals ANN classifiers [6].

Advantages

ANNs have some advantages over GMM and HMM. They are known to be more effective in modeling nonlinear mappings. The main advantages using this model are the ease of implementation and the well-defined training algorithm. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low.

Disadvantages

Generally, classification accuracy is fairly low compared to other classifiers.

Support Vector Machines (SVMs) are currently of great interest to theoretical and applied researcher. It is one of the most important class of machine learning models and algorithms [22]. Support Vector Machine (SVM) is widely used as a simple and efficient tool for linear and nonlinear classification as well as for regression problems. The algorithm works as a standard convex optimization problem. The resulting classifier is called the maximal margin classifier. The idea is to search the optimal separating hyperplane which has the maximal margin of separation between the training vectors from the two classes, so maximal margin classifiers estimate directly the decision boundary. Being a separating hyperplane means that the training vectors from the two classes lie on different sides of the hyperplane, and having maximal margin means that the distance from the hyperplane to the nearest training vector is maximal. The support vectors are those training vectors which lie nearest to the optimal hyperplane. In real applications, the training data is usually not linearly separable and then the maximal margin hyperplane does not exist. A solution is to seek the so-called soft-margin hyperplane. These quadratic programs can be solved either by general purpose quadratic program solvers or by techniques developed specially for SVMs.

Advantages

SVM brings some advantages that explain why researchers are more likely to use this algorithm [23]. SVMs deliver a unique solution, since the optimality problem is convex and gain flexibility since its function is non-parametric. This is an advantage compared to Neural Networks, which have multiple solutions and for this reason may not be robust over different samples. SVMs also provide a good out-of-sample generalization. This means that, by choosing an appropriate generalization grade, SVMs can be robust, even when the training sample has some bias. The global optimality of the training algorithm and the existence of excellent data-dependent generalization bounds are some advantages over GMM and HMM. Additionally, no assumptions to make data linearly separable, are needed. The transformation occurs implicitly on a robust theoretical basis and human expertise judgement beforehand is not necessary.

Disadvantages

In contrast, SVM main difficulties are mainly related to the dimensions, to the number of training instances and to the number of features. A great number of training instances leads to a great number of either variables or constraints

Bayesian Learning or Naïve Bayes (NB) classifier is based on the so-called Bayesian theorem with the naïve assumption of independence between every pair of features. This classifier in spite of the apparently oversimplified assumptions, has worked quite well in many real-world situations [24].

Advantages

Simple Bayesian classifier utilization brings some principle advantages in terms of simplicity, learning speed, classification speed, storage space, and incrementality.

Disadvantages

Bayesian classifiers have traditionally not been in a focus of research. One reason for this is that the Bayes relies on an assumption that the attributes used for deriving a prediction are independent of each other, this assumption in many cases seems to be over-simplified. Another reason is that the simple Bayes is an extremely stable learning algorithm and cannot benefit from the integration of the simple Bayesian classifiers, while many ensemble techniques are variance reduction techniques. However, from empirical experiments it has been shown that the optimality of the Bayesian classifier can be optimal even when the independence assumption is violated by a wide margin.

The k-NN (k-nearest neighbours) is a classification method that uses training data every time there is a new sample to be classified. The training samples are vectors in a multidimensional feature space and each one with a class label and the classifier will find a number of neighbours to define the class of the new sample [7]. A constant named k is defined which is going to be the number of neighbours samples.

Advantages

This classifier works very well with small and large sets of data and the cost of the learning process is zero, it is also very robust to noisy training data.

Disadvantages

The main problem is that is computationally expensive to find the k nearest neighbours when the dataset is large and the testing step is very slow because it needs to compute distance of each query instance to all training samples. This means that for live applications such as emotion recognition, it is not as useful.

Other Methods

In order to ensemble classifiers to improve over the best performing base classifier, other classifiers have been developed [20].

Extreme learning machine (ELM) is based on the confusion degree between a class of emotions and other categories of emotions. The smaller confusion degree, the greater difference between the emotion's groups is, so easier to be distinguished. The strength is that the optimal solution obtained by using ELM decision tree is unique.

Unweighted voting: the class predictions of the base-level classifiers are summed and the class with the highest number of votes determines the prediction for the ensemble.

Stacked generalization: The meta-learner, typically a series of linear models, uses the level-0 predictions and the target classes to determine which classifiers are correct or incorrect and generates a higher-level prediction based on this.

In the following ***tab.7*** we resume the most popular classifiers associated with relative advantages, disadvantages and performance achieved in empirical experiments:

Classifiers	**Advantages**	**Disadvantages**	**Accuracy rate**
HMM	Effectiveness, Flexibility, Variety of applications	High costs in terms of storage and time in complex problem	75% - 80%
GMM	Efficiency, Best compromise between performances and computational requirement	Dimensionality limitation, Cannot compute too complex classification	75% - 80%
SVM	Effectiveness, Unique solution, Flexibility	High costs in terms of storage and time in complex problem	75%

ANNs	Ease of implementation and well-defined training algorithm	Lower performance	About 60%
kNN	Flexibility, Performance	High computational costs in complex classification	> 80%
Bayesian	Simplicity, Speed, Low storage space	Cannot support complex problems, Performance	About 50%

Tab.7 *Performances of common Classifiers*

4. Voice Analysis and Business Applications

We have widely illustrated how human voice is a rich and nuanced source of emotional signaling and how paraverbal aspects of speech play a crucial role in reinforcing, amplifying, or otherwise modifying the transmission of emotional signals. The focus of this part of the work shifts to the importance of emotion detection through voice for business applications. According to [25], “an emotion can be defined as a mental state that arises spontaneously from external stimuli that is often accompanied with physical expressions, which is heavily influenced by culture to serve a purpose of a particular situation”. Humans perceive emotions, and this affects how they make judgments in particular situations. In some circumstances, emotion recognition helps people to enhance the perception process and to use such information to help make better judgments and to boost the decision-making process. A specific emotion arousal can trigger different human behaviors, but the time window in which it is expressed and exploitable is short. It is necessary both to detect an emotional state and to insert the information in the business process at the same time in order to make it valuable. Machines could help people in making right decisions by recognizing emotions faster than human capabilities. Emotional speech processing machines are able to recognize emotions of a speaker in real-time. This insight has a great potentiality especially in irrational situations where decisions have to be made fast. In the following chapters we explain how the study of voice is suitable to accomplish practical daily life needs and how companies can make business exploiting emotional speech recognition technology. Those softwares that allow companies to study consumers emotional state through voice analysis are so called Voice Analytics. Voice analytics data is an increasingly valuable asset to organizations in every industry and a powerful tool for promoting better business intelligence. According to market researches, global Emotion Analytics Market is expected to grow at USD 25 billion by 2023 at a CAGR of 17% during the forecast period 2017-2023. In ***fig.4*** the graphic of the expected market growth from 2018 to 2023.

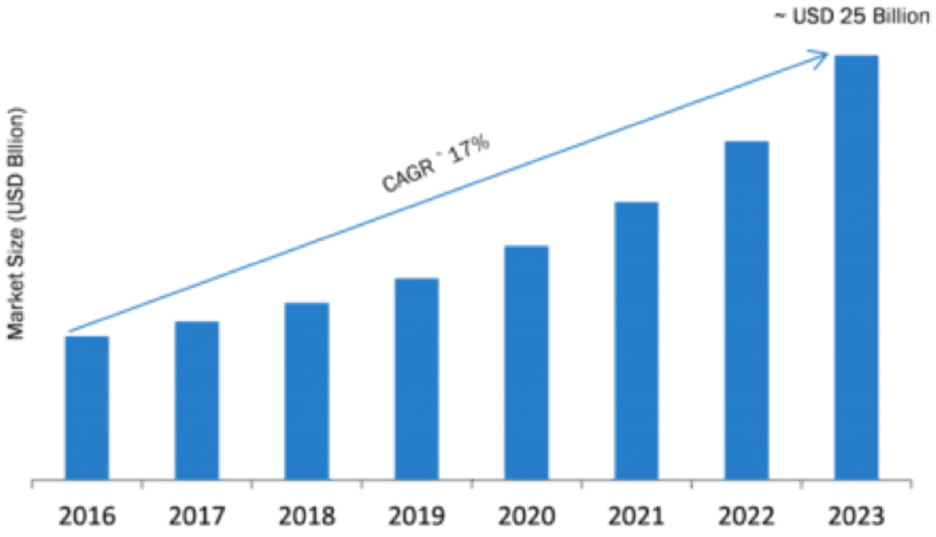


Fig.4 *Global Emotion Analytics Market*

4.1. Voice Analytics for Marketing

Today's digitally empowered customers have enormous say over what products retailer stock, how quickly lines are replenished or replaced, and over how purchasing experience is provided. Faced with the higher expectations of their customers, retailers and brands are turning shopping into a personalized and exciting multichannel experience. To ensure their transformation stays on the right track and is sustainable, brands and retailers need to know both the direction and the benefits of the latest trends in retail technology. They also need to come up with the right digital-driven strategies to keep customer engagement. For marketers, knowing how a potential customer is feeling is critical. That kind of insight can enable extreme personalization, vastly increasing the effectiveness of advertising. Marketers are interested in understanding the different emotional reactions that are generated while they experience products. The use of biometrics to capture the emotion and unconscious behaviours caused by product trial enables a deeper understanding of the consumer's experience. For marketing purposes, catching the insights coming from the interaction between consumers and products/services is fundamental, since it explains the real evaluation of the offer. This information can be used by companies to effectively promote their brands by providing the desirable advertising stimuli for the specific product they are marketing. This knowledge can help marketers in shaping product communication in order to drive purchase intention. By grasping the right emotional state of the consumer and providing him with the right stimuli, can be an effective way to trigger some unconscious reactions toward to product he is experiencing [26]. Many factors in the retail environment can influence consumers decision making. For example, researchers have demonstrated the relationship between perceptions of control in the retail environment and the customer’s judgement [27]. More findings in this experiment came out to explain the mood and attitude of the consumer. First, consumers who feel more in control in a retail environment should have a more positive attitude toward the environment. Second, consumers who feel more in control in a retail environment should approach the environment more frequently. Arousal is another fundamental factor of behaviour. “It is the basis of emotions, motivation, information processing, and behavioural reactions” [28]. In this study, researchers have investigated how arousal might be the driving force for decision-making processes and shopping behaviour at the point-of-sale. Both internal and external stimuli can cause arousal. Findings clearly confirm that a store atmosphere that evokes pleasure, can increase consumers’ spending, time spent in the store, the desire to come back and to recommend the store to others. Companies also want to provide unforgettable experiences to consumers but before they need to understand how the consumer feels. Understanding these patterns can also guide designers in controlling the emotional responses to their product designs. For instance, designers may assess which specific attributes of products stimulate arousal in individuals and design products accordingly to the emotional experience they want to provide. They may design exciting or relaxing experiences according to the image and meaning the product would convey and the specific target to which it is addressed.

4.1.1. Voice Analytics: Market and Players

According to MarketsandMarkets, the voice analytics market size is expected to grow from USD 657 million in 2019 to USD 1,597 million by 2024, at a Compound Annual Growth Rate (CAGR) of 19.4% during the forecast period. Voice analytics can help enterprises in building successful client relationships, implementing the best sales and marketing practices and understanding the changing business conditions, client insights, market trends, or service inconveniences. We provide an overview of the existing companies operating in the voice analytics market and the solutions that are offered by major players. Principal voice analytics companies are the followings.

AudEERING

AudEERING is an audio analysis company based just outside of Munich, Germany audEERING produce a number of packaged products, including:

- **Audiary**: a voice-enabled diary that allows patients with chronic diseases to record the state of their health, log their medical adherence. The technology it incorporates offers a complete analysis of the user's emotional state.
- **CallAIser**: a call centre speech analysis software that reports the parameters of telephone conversations such as duration and the relative share of the dialogue, along with the speakers' mood and the atmosphere of the conversation. This can detect and prevent escalations before they happen, allowing a more experienced call centre agent to take over and calm the situation down.

Affectiva

Affectiva was spun out of MIT Media Lab in 2009 in Boston. The company is a specialist in both face and voice emotion analytics, capable of identifying the seven basic emotions (happiness, sadness, surprise, anger, anxiety, disgust, and a neutral state). Affectiva produces emotion AI technology that is now used in gaming, automotive, robotics, education, healthcare, experiential marketing, retail, human resources, video communication, and more. The company's technology is also used by market research firms to measure consumer emotion responses to digital content. They have a range of products on offer:

- **Emotion as a Service**: a cloud-based solution that analyses images, videos and audio of humans expressing emotion. Returns facial and vocal emotion metrics on demand, with no coding or integration required.
- **Emotion SDK**: emotion-enables apps, devices and digital experiences.
- **Affectiva Automotive AI**: a multi-modal in-cabin sensing AI that identifies, from face and voice, complex emotional states of drivers and passengers.
- **In-lab Biometric Solution**: provides researchers with a holistic view of human behaviour, integrating emotion recognition technology and biometric sensors in one place.

Nemesysco

Nemesysco is a developer of advanced voice analysis technologies for emotion detection, personality and risk assessment. Founded in 2000 in Netanya, Israel.

Nemesysco's clients are typical:

- **Call centres** use Nemesysco for quality monitoring of their calls, either in real-time or immediately after, to identify the ones that are mistreated by agents. Their system also collects emotional profiles of customers and agents in the CRM system, allowing the most suitable agent for the customer's emotional profile to be matched, and the most suitable products to be offered.
- **Insurance companies** use Nemesysco's products to conduct a risk assessment and detect fraud in insurance claim calls in real-time. The technology analyses the unique vocal characteristics that may indicate a high probability of fraud or concealment of information.
- **Banks and financial institutions** use Nemesysco to perform credit risk assessment, for immediate fact verification and fraud intention detection. The voice analysis platform improves risk scoring models and reduces uncertainty for lenders, allowing them to verify past events and current information, and identify potential sensitivities.

- **Law enforcement agencies** use the Criminal Investigation Focus Tool to detect expressions of lies and measure psychophysiological reactions in suspects.

Audio Analytic

Audio Analytic was founded in 2010 in Cambridge, United Kingdom. Its sound recognition software framework has the ability to understand context through sound, allowing for detection of not only emotion but also many other specific sounds. Audio Analytic's software has been embedded into a wide range of consumer technology devices for use in the connected home, outdoors, and in the car.

Aurablue Labs

Aurablue Labs is a relatively new company, founded in 2016 in India. It leverages the power of deep learning to recognize emotions from speech signals. Call centres use it to analyse voice data and measure the quality of service delivered by agents. It can also be used by taxi firms to rate drivers based on aggression in their tone. Consumers can apparently use Aurablue technology to continuously monitor their stress levels during the day and receive alerts when it gets too high. They have also integrated their technology into the Beatz Smart Jukebox, a music player on Android that automatically adapts the playlist according to your detected mood.

VoiceSense

VoiceSense is another voice emotion analytics company based in Israel that uses Big Data predictive analytics, to predict the behaviour tendencies of individual customers. Their flagship product, Speech Enterprise Analytics Leverage (SEAL), assembles these technologies into a speech-based solution that can accurately predict future consumer behaviour.

SEAL can be applied to numerous use cases, such as:

- **Customer Analytics** for customer retention prediction
- **Fintech Analytics** for loan default prediction
- **Healthcare** for post-traumatic stress disorder (PTSD) tracking
- **Personal Assistant** with content recommendations

- **Call Center Interaction Analytics** in order to monitor customer dissatisfaction

Beyond Verbal

Beyond Verbal was founded in 2012 in Tel Aviv, Israel by Yuval Mor. Their patented voice emotion analytics technology extracts various acoustic features from a speaker's voice, in real-time, giving insights on personal health condition, wellbeing and emotional understanding. The technology does not analyze the linguistic context or content of conversations. It detects changes in a vocal range that indicate emotions like anger, anxiety, happiness, satisfaction, and cover nuances in mood, attitude, and decision-making characteristics.

4.1.2. VoC: Voice of Customers

Exploiting the voice of customers means to collect insights into customer needs, wants, perceptions, and preferences. Generally, VoC thus intends listening to the customers' needs. In this case, we take into consideration all the insights that are collected from customers through the recognition of their voice. These discoveries are translated into meaningful objectives that help in closing the gap between customer expectations and the firm's offerings. Gathered data and insights can impact on many marketing aspects across customer experience, brand, competitive analysis and product development. Understanding consumers' level of satisfaction at different touchpoints, analyzing motivations for engaging with customer service or willingness to recommend the product or brand to others are very important factors that can be used to offer better customer experiences. Central Restaurant Products is a leading wholesale distributor of foodservice equipment. They mine their calls for marketing insights. These insights from calls are used to optimize their advertising spendings. By monitoring these kinds of information, companies can boost their brand and reputation, improving satisfaction, retention and loyalty of customers.

Also, understand consumers' buying patterns and preferences, capture trends of buying behaviours, discover what customers say about competitors are meaningful insights that can be leveraged by firms to penetrate markets. Moreover, product managers are willing to gather feedback on new products' ideas, features, promotions, and understand price sensitivity. VoC programs are fast-growing segments of core business strategy for organizations. They are gaining traction in the business environment because capturing and acting on customer feedback are leading drivers of business success, since it helps companies to understand customers' complex decision-making process. The growing expectations of consumers at each stage of the customer journey are accelerating the digital transformation of retail around the world. Pre-purchase research, the shopping experience itself, the post-purchase experience is changing. The adoption of breakthrough technologies can help brands and retailers to address these changing consumer expectations in a timely and effective manner, deepening their engagement with customers. All the insights can be generated and collected using voice analytics software that use voice recognition tool to analyze a spoken conversation and to identify speaker emotions and intents. Companies in a range of industries including insurance, technology, financial services, and healthcare are leveraging this technology to study customer needs.

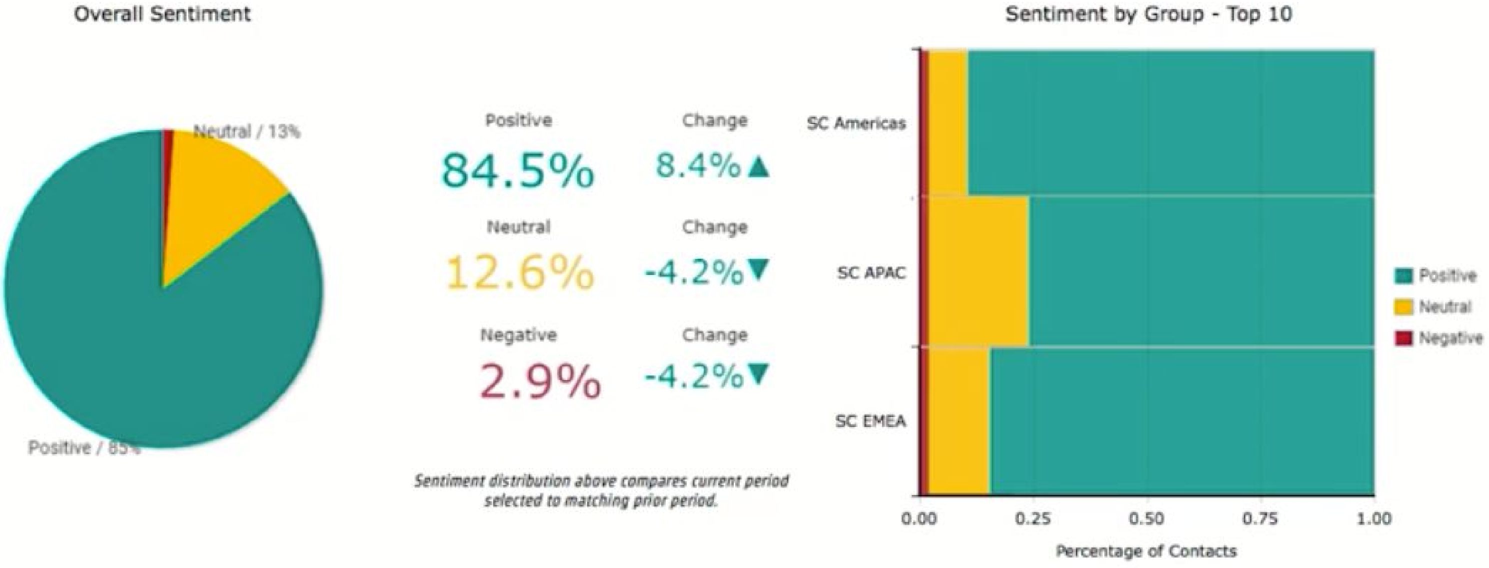
4.1.3. Voice Analytics in call centers

More and more companies are using voice analytics to gain insight into customer interactions. In this chapter, we try to deepen how do they work underlying the type of information that can be extrapolated from data analysis and how companies get benefits from it. Voice analytics program is a structured system of feedback collection, data analysis, and action planning. Data analysis is often driven by an algorithm, a scoring mechanism that monitors conversations and evaluates language and voice inflections to quantify attitudes, opinions, and emotions related to a business, product or service, or topic. One of the most widely used applications for sentiment analysis of voice is for monitoring call centre and customer support performance. It provides crucial feedback to agents and representatives, allowing them to respond appropriately to impact the outcome of the interaction, for instance, talking-down frustrated customers. For many brands, calls and contact centres are one of the richest sources of customer insights available. Analyzing call recordings and transcriptions is useful for collecting numerous insights to understand consumers needs and make smarter optimizations. Detecting emotion in real-time may lead companies to settle a series of corrective actions to improve service rating and evaluate service level agreement (SLA). For example, detecting calls could automatically be directed to someone better suited to dealing with an angry person, or sad person. Retention of customers could be evaluated against the emotional levels inside of the conversation. Also, patterns of emotions could be matched against the handling of certain policies.

Typically, the information measured by the algorithm during vocal interactions are:

- the amount of stress or frustration in a customer's voice given by the tone and the vocal inflections;
- the rate of speech indicating the speaker's emotional degree of anxiety;
- changes in the level of stress indicated by the person's speech (such as in response to a solution provided by a customer support representative).

Voice analytics leverage AI and machine learning to capture, transcribe and reveal insight from 100% of customer interactions. The platforms apply advanced Natural Language Processing (NLP) and machine learning techniques to process source audio streams in real-time. The algorithms transform the conversation between customers and agents into operational intelligence at scale with automated performance and sentiment scoring.



Sentiment distribution above compares current period selected to matching prior period.

Fig.5 Sentiment Analysis scorecard in call centres

However, the information processed is not only related to overall sentiment score, but it also correlates sentiments with KPIs like call duration, hold time, silence, evaluation score to allow the identification of behavioural trends and topic discovery. Synthesizing the operational tasks that Analytics accomplish, they have to:

- Convert audio with language patterns, acoustics and timing into categorized results for focus;
- Identify customer and agent's dialogue with sentiments to recognize optimization opportunities;
- Deliver targeted audio and transcriptions to encourage action with data-driven confidence;
- AI-driven algorithms automatically trigger notifications that alert agents to the next-best-action and may inform supervisors when the situation so requires it.

In brand reputation management applications, overall trends in sentiment analysis enable brands to identify peaks and valleys in brand's feeling or shifts in attitudes about products or services, thus enabling companies to make improvements perfectly in-line with customer demands.

4.1.3.1. Improving overall performance through Voice Analytics

When sentiment analysis scores are compared across certain segments, companies can easily identify common pain points, areas for improvement in the delivery of customer support, and overall satisfaction between product lines or services. Research by Bain & Co. has found that:

- companies that excel at customer experience grow revenues 4-8% above their competitors;
- voice of customer programs results in up to 55% greater client retention.

Insights from the contact centre are a unique resource for realizing the key drivers of satisfaction, loyalty and compliance verification. We resume below the top benefits that Voice Analytics can bring to firms.

1. *Evaluate agents' effectiveness and involvement*

   Traditional KPIs don't always tell how effective agents are. For example, a longer call can sometimes mean that an agent is adept at handling complex issues. Through voice analysis, is possible to identify whether the agents are consistently involved with calls and have a positive sentiment, or the not performing agents with negative sentiment. It is important to supplement surveys and focus groups results with Sentiment Analysis data in order to understand the impact of every interaction.

2. *Increase the performance of the call centre*

   Starting from the analysis of every call, agents can see their performance on the scorecard and realize which are the aspects to improve. Hence, companies set down targeted coaching sessions. Some case studies reveal improvements in many KPIs such as lower silence ratio and incorrect operation, higher agent productivity and lower waiting time.

3. *Test effectiveness of marketing campaigns*

   Marketers can use sentiment analysis to discover how customers view their most recent ad campaigns, to find out how customers view their brand or to understand how customer sentiment varies by product line.

4. *Quickly identify the root causes of inefficiencies*

   By pulling sentiment data into KPI reports, it is easier to identify correlations that might not be obvious, find out the weaknesses of the service delivered or why customers are displeased with some products. Revealing which are the bottlenecks can help the firms to plan and implement corrective measures to solve the issue.

4.1.3.2. ING Bank case study

(https://www.sestek.com/case-studies/ing-bank-speech-analytics-case-study).

ING Bank is an excellent example of the successful use of Voice Analytics technology in its call centre. ING uses it to analyze its calls in three ways: to shorten call length, to increase sales potential per call and to improve agent performance. ING analyzed employees performance monthly, using the following metrics.

- Seniority: it is the length of time that an individual has served in a job, using months as the unit of measurement.
- Net promoter score: it is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company’s products or services to others.
- shift order: explaining how shifts are set up.
- silence ratio: Silence costs you money and signals a problem such as agent knowledge gap, failure in processes, or systems that are too slow.

Throughout this process, ING discovered some correlations between these metrics, such as:

- agents with seniority over 24-36 months exhibited higher anger ratios, longer talk times and lower wait times;
- increases in talk times and interruption rates decreased net promoter scores by 4–5%;
- an aggressive and dominant speech tone was effective.

After getting the results, ING formulated a plan of action to increase agent performance and morale. The insights from the platform prompted ING to implement a rotation program, resulting in 65 employees transferred out of the call centre to other departments of the bank. This resulted in a 3% decrease in overall silence rates, which also provided a cost advantage by reducing call durations. With automatic evaluation reports of sales calls, ING could analyze every call with better accuracy. By identifying opportunities for improvement and acting on them with ease, ING found a 15% increase in sales quality scores. Moreover, the net promoter score increased by 10%.

4.2. Biometrics Authentication

Voice biometrics is a technology used to match personal voice pattern and verify the speaker's identity using voice as a unique identifier. Generally, a user must first create a voiceprint. The voice recognition system captures the voice print and analyzes voice’s patterns. The voiceprint is then encrypted and stored as part of the user's authentication profile, along with other authentication credentials (normally phone number validation and device authentication). The encrypted identification pattern contains physical and behavioral factors. These include pronunciation, emphasis, speed of speech, accent, as well as physical characteristics of the vocal tract, mouth and nasal passages. When a customer makes subsequent calls, voice biometrics technology captures a person's voice and compares the captured voice characteristics to the characteristics of a previously created voice pattern. If the two matches, voice biometrics software will confirm that the person speaking is the same as the person registered against the voice pattern, resulting in rapid and seamless authentication. In term of real-life applications, the most common use of this technology is for security purposes, but it has different use-cases applications:

- Entertainment: Voice recognition can be used in services such as Netflix for the access of personal contents, change TV or radio channels, open and close screens.
- Healthcare: In an industry where data security is essential, physicians can use voice biometrics to dictate and record patient's health conditions directly into the system and securely retrieve patient's personal history
- Banking: Banks can leverage the system to enable highly secure and advanced voice-based payments. With fraud on the rise, credit card companies and banks such as Citibank and ANZ use voice biometrics to proactively identify fraudsters and authenticate callers at their call centers.

4.2.1. Biometrics market and system performances

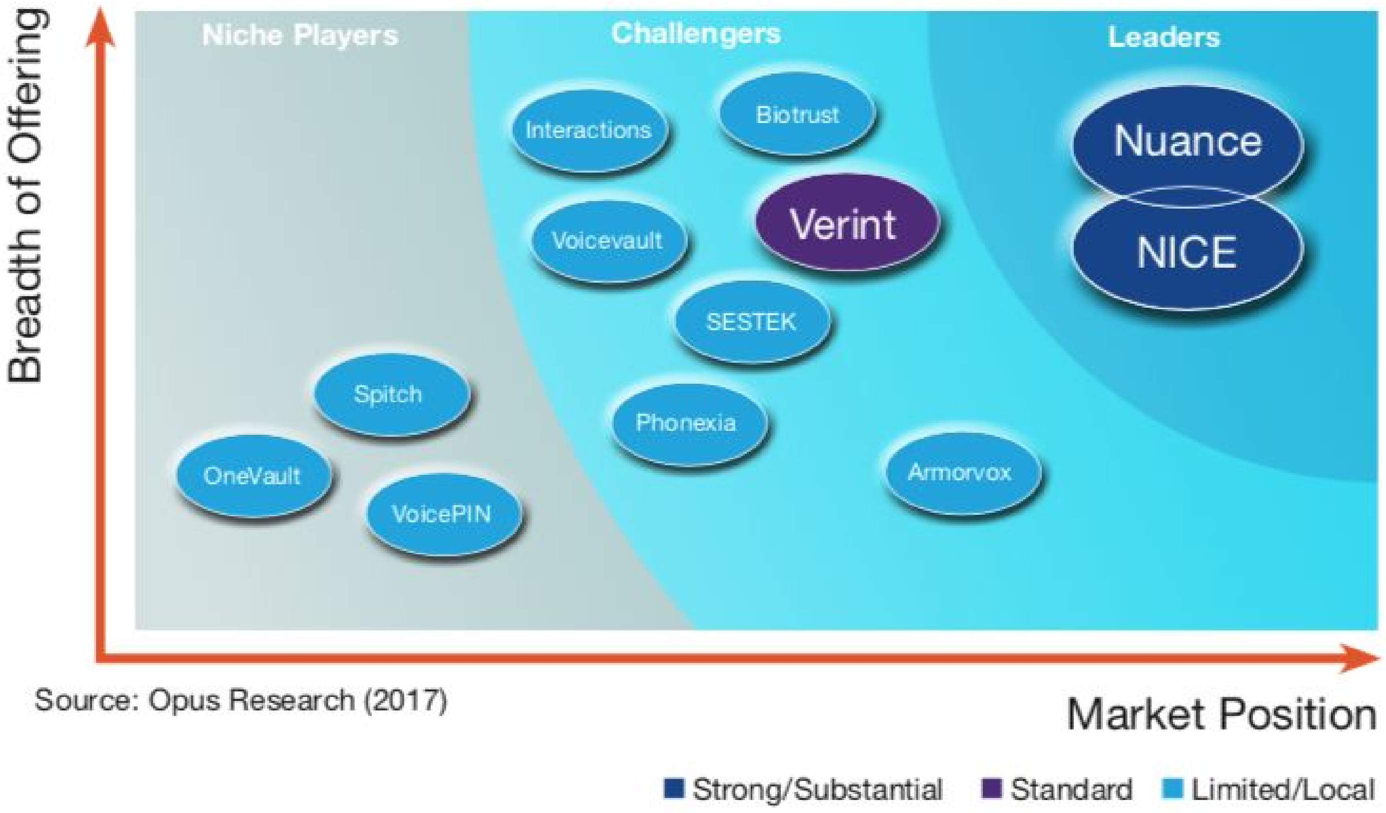
According to ResearchandMarkets report, the global Voice Biometrics Market size is expected to reach $2.7 billion by 2024, rising at a market growth of 22.7% CAGR during the forecast period. The market is segmented into Access Control & Authentication, Fraud Detection & Prevention, and Forensic Voice Analysis & Criminal Investigation. Recent experiments and researches on speaker authentication systems have reached significant technological advancements in terms of performances. In commercial applications as well, technology in voice-based biometrics seems to show a mature level [29]**.** In other experiments, the performance of voice biometrics has been studied to identify speakers and it has been observed that voice biometrics is fairly inexpensive and non-intrusive but the accuracy of voice authentication is affected by several factors and this leads to low performance in a noisy environment [30]. Some researchers have studied the performance of user recognition and tried to improve its mixing voice and face recognition. During experiments [31], it has been used audio-visual data under a multi-floor robot cooperation scenario in order to create a multi-biometric dataset comprising of face and voice modalities, namely AveRobot, tailored for evaluating people identification and verification capabilities. They collected a multi-biometric dataset of 111 participants vocalizing short sentences. The collection took place into a three-floor building by means of eight recording devices, targeting various challenging conditions. The test provides baselines for face and voice re-identification and verification tasks. The results show that the dataset they provided appears challenging due to the uncontrolled conditions. This observation could derive from the fact that the audios in AveRobot contain several noisy situations (opening doors, background speaking, alarm sounds). Another experiment was carried out by [32], who have successfully implemented a biometric system completely on the Samsung Galaxy 7 phone that fuses features from the person's face and mel-frequency cepstral coefficients (MFCCs) from the voice. The results show that fusion increased recognition accuracy by 52.45% compared to using face alone and 81.62% compared to using voice alone. The increasing results are due to the fact that different modalities, such as face and voice, provide independent sources of discriminating information that can be used for identification. Multiple modalities can also be more difficult for an attacker to bypass than a single one. Finally, high-quality identifying data from one modality can be used to compensate for low-quality data in other modalities to increase authentication accuracy.

4.2.2. Digital Payments

Technology is also radically changing the way people purchase goods, especially payments methods. With the proliferation of smartphones and wearables that have integrated payment tools, and with the availability of high-speed mobile networks, digital payments are becoming more important in the retail sector. It is evolving at a high pace since consumers find them easy-to-use, time saving and attractive as a shopping experience. Companies have to adapt to this new reality by offering seamless online and offline experiences. They can no longer afford to offer limited payment options, instead, they must offer the whole range to accomplish to clients' needs. Not managing this emerging trend is a big risk for merchants, they may lose those consumers who choose not to use certain payment methods. For instance, Google recently announced a new Google Assistant feature: the ability to make P2P payments money transfers to your friends as well as request cash using voice and Google Pay. Well-known examples include Apple Pay, Google Pay, Alipay, WeChat Pay and Venmo by PayPal. The potential benefits of voice-activated payments are being recognized by shoppers, there is a significant trust issue when it comes to using Smart Home technology for eCommerce. Voice recognition has several key advantages over other forms of identity authentication. Authentication through voice is widely accessible on mobile phones given that all phones have microphones. It has good cost-effectiveness Integrating the software into other devices such as automobiles and home appliances are considered cost-effective. The method of authentication is contactless, and therefore less invasive and more hygienic. Some disadvantages of voice recognition are that verification through voice is not as accurate as other biometric modalities, for example, facial recognition. The implementation of speaker recognition requires liveness detection to verify that a sample is from a live speaker and not a recording. Moreover, background noise can impact the quality of the sound and matching performance. *Aware* is a fitting example of a business that has demonstrated to be a trusted provider of quality biometric software and solutions for over twenty-five years. Nexa-Voice is a Mobile voice authentication SDK offered by Aware that enable multi-factor authentication on iOS and Android devices. VoicePIN is another example of a business application for mobile. VoicePIN system supports hands-free experience and eliminates the need to memorize logins and passwords. Just natural voice commands to log in or authorize purchases. Besides the security, the system has additional benefits: cost-effectiveness, unique customer experience for end-users and agents and convenience. This software has already been added to web pages or applications in banking and e-commerce to improve customer experience. Moreover, VoicePIN provides completely remote authentication.

4.2.3. CitiBank: a case study

Citibank, for example, launched the biometric authentication for institutional clients in Asia Pacific (China, India, Singapore, Thailand and Vietnam). “Around 20-30 per cent of the bank’s 1.6 million customers are expected to apply to use the voice-biometrics service”, said Vira-anong Chiranakhorn Phutrakul, consumer business manager at Citibank Thailand. Citi Voice Biometrics uses NICE software to identify roughly 130 different physical and behavioral characteristics within a person's vocal pattern and match those with a prerecorded voiceprint to verify the caller's identity. The NICE software also works to prevent fraud by identifying fraudsters within the first seconds of a call. When a caller’s voice print is identified as belonging to a fraudster, the call is classified as high risk and can be immediately transferred to a fraud specialist to prevent an unlawful transaction from taking place. According to Opus Research, in **fig.6** are represented the most important players offering biometric authentication solutions. Leadership is determined by market share and the range of services offered. Challengers category reflects both spans of product, geographic footprint and network of resellers and integrators. Niche players have confined their efforts based on product and geographic focus.



Source: Opus Research (2017)

Fig.6 *Biometrics authentication software providers*

4.2.4. Biometrics: Advantages for Companies and Clients

Biometric authentication is a promising approach to securing digital applications. It frees users from having to remember strong passwords, largely eliminates the security threats resulting from using the same password on multiple devices, and overall facilitates a more natural form of human-computer interaction. The range of authentication methods has evolved continuously over time. The evolution is accompanied by a change of authentication paradigm: from the most basic authentication factor (something you HAVE), to simple second factor like PINs and passwords (something you KNOW), through to biometrics (something you ARE). Voice biometrics solutions are used by contact centers to authenticate callers and verify their claimed identity in real-time, which produces great benefits for clients but for the company too. They can be translated into three main savings:

- Average Handle Time reduction (contact center authentication time)
- Increased self-service containment
- Fraud prevention

Average Handle Time (AHT) reduction

It eliminates the need for authentication questions asked by the agent and dismisses the customer's need to look for the answers. In **fig.7** is illustrated the formula to calculate the saving coming from the AHT reduction.

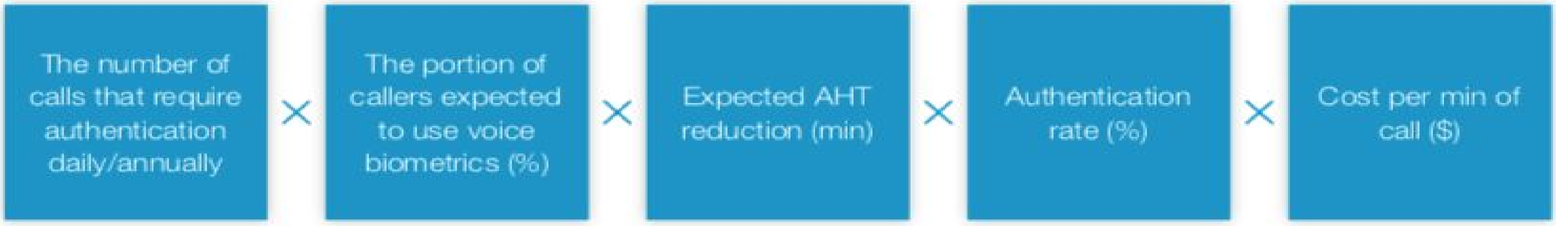

The number of calls that require authentication daily/annually
×
The portion of callers expected to use voice biometrics (%)
×
Expected AHT reduction (min)
×
Authentication rate (%)
×
Cost per min of call ($)

Fig. 7 *Average Handle Time in a contact center*

Where:

- *The number of calls that require authentication daily/annually:* based on each specific enterprise policy, a portion of the calls that reach the contact center require strict authentication, while others may require a lower level of authentication or none.
- *The portion of callers expected to use voice biometrics:* callers that are enrolled in the service. The Pareto principle, also known as "80-20 rule", suggests enrolling in the service the 20% of clients that make the 80% of calls.
- *The expected reduction in AHT: it tells* the average duration of one transaction, typically measured from the customer's initiation of the call and including any hold time, talk time and related tasks that follow the transaction..
- *The successful authentication rate: like any authentication methods, it cannot ensure 100% accuracy.* NICE's recognition system, for example, offers a real-time authentication rate of approximately 96%.

- *The cost per 1 min of call*: each contact centre has its own cost per minute of call which typically includes telecom costs, fully- loaded agent costs (including two levels of management), plus desktop technology and infrastructure costs.

Increased Self-Service Containment

When customers call the contact center, they typically start the call in the Interactive Voice Response (IVR) and transferred to an agent if needed. In **fig.8** is illustrated the calculation for the increase of Self-Service Containment.

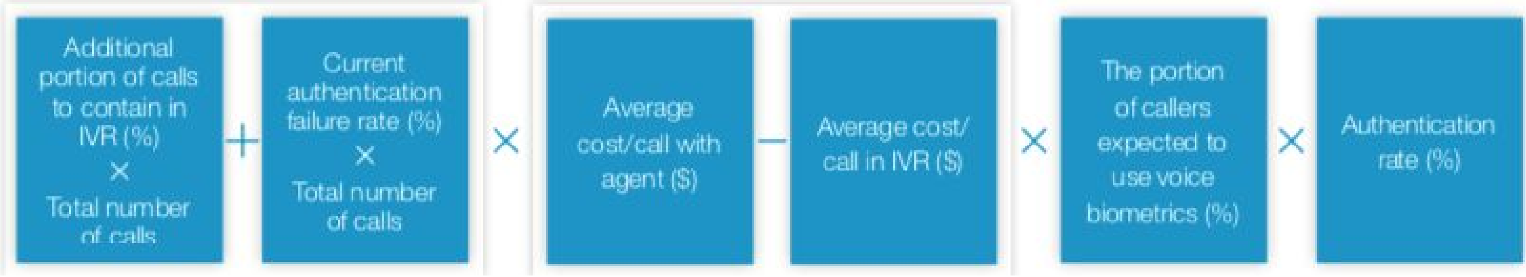

Additional portion of calls to contain in IVR (%)
×
Total number of calls
+
Current authentication failure rate (%)
×
Total number of calls
×
Average cost/call with agent ($)
–
Average cost/ call in IVR ($)
×
The portion of callers expected to use voice biometrics (%)
×
Authentication rate (%)

Fig.8 Increased Self-Service Containment calculation

Where:

- *Additional portion of calls to contain in IVR:* studies show that in most contact centers only 20% to 40% of calls are contained in self-service, all the rest reach an agent.
- *Current authentication failure rate in the IVR:* it is normally due to forgotten pin.
- *Average cost/call with agent – IVR:* since a call with an agent is 7 times more costly than a call with the IVR, containing a call in the IVR almost saves 6/7 of the entire cost of that call. Therefore, organizations prefer to contain as many calls as possible in the IVR.
- *The portion of callers expected to use voice biometrics*: making the enrollment process easy and fast may lead to daily caller enrollment rates of about 45 to 59% within one year, and 85 to 90% within three years.

- *The successful authentication rate:* NICE Real-Time authentication rate is approximately 96%.

Fraud Prevention

Enrollment rate directly affects fraud loss prevention. In other words, assuming you got an enrollment rate of 59% within a year, your fraud losses will be reduced by approximately 59% as well. The more end-customers enrolled, the more voice biometrics is in use, and the more fraudsters can be caught. According to [33], voice is considered as one of the most convenient biometric interfaces for the user. Instead of typing passwords with memorable dates and other authentication details the user's voiceprint itself is used to verify the identity of the claimed user. It is considered a powerful way to secure access to all private data of the user, who recognizes the convenience and the effectiveness of voice-authentication. The number of applications and services is growing, and different voice-based identity authentication interfaces are present in the market. The difference comes from the speech input provided by the user, that may vary. Principally, there are three kinds of input interfaces:

- Fixed pass-phrase
- Text-dependent
- Text-independent

In the fixed pass-phrase mode the user is asked to provide a pass-phrase, which has been set up and stored in his/her user profile. This mode of operation achieves high speaker verification performance but is also highly vulnerable to replay attacks, since an impostor could record the voice of the target speaker voicing the fixed pass-phrase. In the text-dependent mode of operation, the user is prompted with a short text message from a list of pre-selected utterances to read. The text-dependent mode of operation achieves less speaker verification performance but is more robust to spoofing attacks comparing to the fixed-passphrase mode. In the text-independent mode, the user is prompted a text message produced by a random text generator. In this case, the speaker models are not trained with the same test utterances, and thus achieve lower performance. However, replay attacks are practically impossible, because synthetic speech cannot be precisely fitted with the acoustic characteristics of the target speaker, and thus this mode is less vulnerable to fraud.

4.3. Robotics - HRI

The impact of robotics is expected to be cross-sectorial, with the integration of other digital elements such as AI, cloud, IoT and analytics. As technologies mature and companies gain a better understanding of how to integrate new forms of innovation into their businesses, the robotics market is supposed to accelerate. Goldstein Research analyst forecast the interactive robots market size is set to reach USD 11,890 million by 2024 from the USD 152.38 million in 2016, growing at a CAGR of 72.4% over the forecast years. By making robots smart and endowing them with robust computational skills, cloud robotics could be the promoter for the increasing of the consumer robotics marketplace. In this way, retailing could exploit technological changes in the business, improving the customers-robot interaction. Enterprises will provide more engaging and interesting customer experiences, enhancing customer loyalty. The deployment of robotics in information dissemination to augment the workforce in developed economies, such as advertising campaigns, is gaining traction. Researchers are developing more socially competent robots that are able to collaborate with people, learning by interacting with other humans. Basically, two are the main skills the robot has to learn in order to bring value to the customers. The first is the ability to acquire social input communicated by relevant clues that humans provide about their emotional state through speech interaction. In order to succeed, a machine has to learn about human emotions out of the speech, using different classifiers that discern from speech-based data from which features are extracted. When the machine learning subsystem becomes able to statistically recognize emotions from features, the system could be provided with real speech input from which features are extracted and compared with existing preferences and settings. Results could then be evaluated again and sent to re-learning or extracted for use in another system. The second is the skill to express in turn its own emotional state, so that it can affect the customer buying decision, refining in the user the sense of interacting with a human-like companion. By combining social robotics and machine learning systems the potential of robotics to assist people will increase. Such a system could be used in different real-life situations. Most common real-life applications are chatbots, retail robots, and social robots.

4.3.1. Social Robots

Social robots are autonomous machines that are designed to interact with humans and exhibit social behaviours such as recognizing, following, assisting and engaging them in conversation. With the advancement of robotic technologies, social companion robots started to take shape. They are becoming able to carry out different tasks and have interactions with humans and their environment. Social robots are intended to interact with people in a natural, interpersonal manner in order to achieve social-emotional goals in different applications such as:

- Education and Research: NAO is the world’s leading and most widely used humanoid robot for education and research. It is an ideal platform for teaching Science, Technology, Engineering and Math concepts with students at all levels.
- Health therapeutic: social robots are employed as therapy for kids with autism. Robots can tell stories on the screen and guided kids through a series of interactive games focused on improving social skills, emotional understanding, sequencing and perspective.
- Service Robots: many robots worldwide perform such services as hotel check-ins, airport customer service, fast-food checkout, etc.
- Social companion/ Caregiving: with the increasing number of solution it is possible to imagine social robots in the domestic area as a helper in the kitchen, keeping the house safe, teaching the children, being their companion

and support the elderly from reminding to take their medication until keeping them company when they feel lonely.

One of the main reasons for the increasing number of social robots as daily life companions is social isolation and loneliness that is affecting more and more individuals. Indeed, Katie Hafner of the New York Times labelled the situation as an epidemic, implying a growing number of lonely individuals. In the article, she writes that in Britain and the United States roughly one out of three people older than 65 lives alone. Moreover, in the United States, half of those older than 85 live alone. KPMG Advisory has developed a diagram that illustrates how machines are taking on an increasingly complex array of social roles in the real world. As technologies and machines' skills increase, robots follow a natural evolution of their social role. In **Fig.9** this social evolution is represented considering two dimensions: social and physical dexterity.

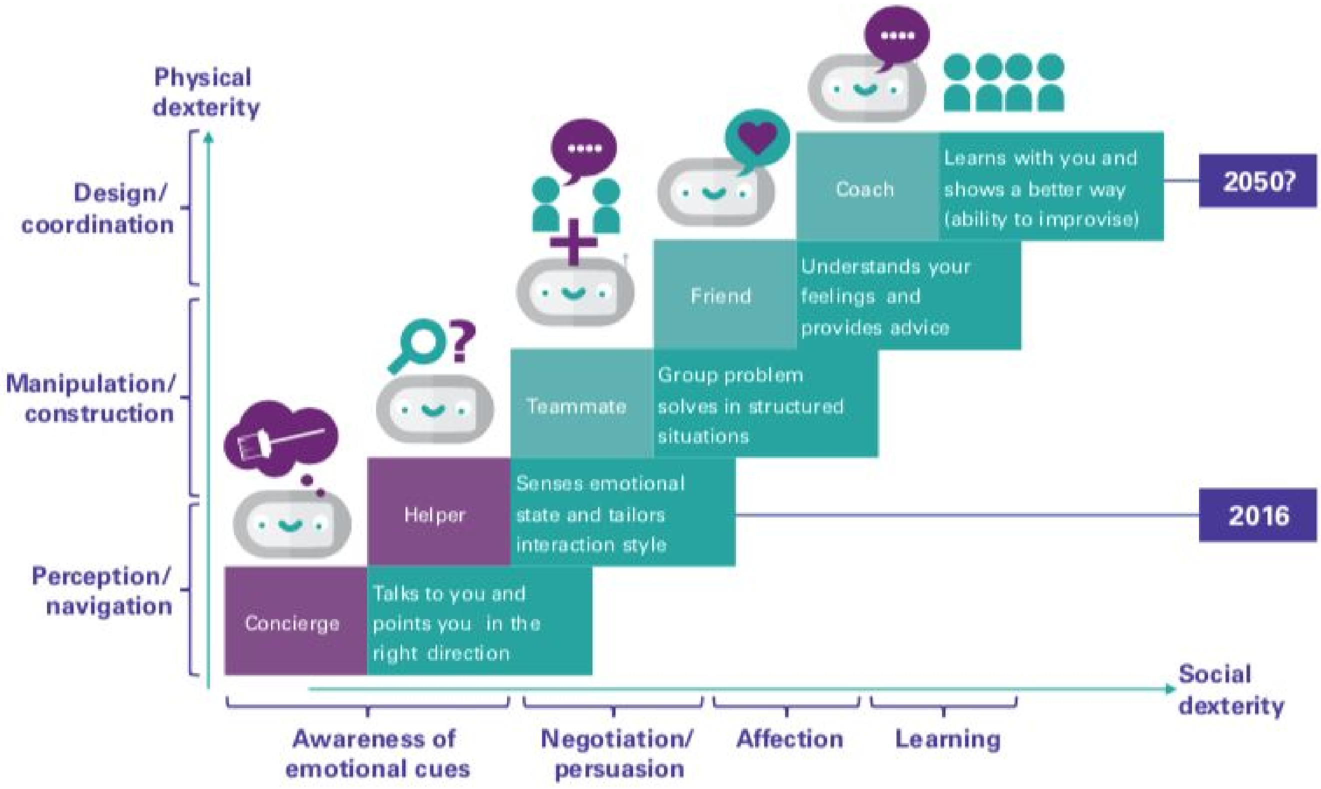

Physical dexterity
Design/ coordination
Manipulation/ construction
Perception/ navigation
Concierge
Talks to you and points you in the right direction
Helper
Senses emotional state and tailors interaction style
Teammate
Group problem solves in structured situations
Friend
Understands your feelings and provides advice
Coach
Learns with you and shows a better way (ability to improvise)
2050?
2016
Social dexterity
Awareness of emotional cues
Negotiation/ persuasion
Affection
Learning

Fig.9 *Social robot roles in human society*

Social dexterity means to have skill in performing tasks in the social sphere. At the start of the social dexterity, there is the absence of any social relationship. Indeed, the interaction is purely transactional. As social complexity increases, robots need to understand emotional cues. Again, more complex social settings would require robots to negotiate with humans. The next level of social dexterity includes robots that can build effective relationships with humans exploiting their emotional awareness and negotiation skills. In the ultimate socially complex setting, a robot would need to have the ability to learn by applying existing knowledge in a new environment and then transfer that learning to its human counterpart. Physical dexterity relates to the difficulty of fulfilling a task in the physical environment. From the most basic tasks' levels that require perception and navigation, to the most complex physical tasks that require the design of new objects, the capacity to coordinate multiple objects, and the ability to develop new solutions through improvisation and developing logic.

4.3.1.1. Kismet: a case study

In order to close the communication loop and coordinate their behaviour with people, robots have to be able to perceive, interpret, and respond appropriately to verbal and nonverbal cues from humans. Modern social robots employ both linguistic and paralinguistic information to perform different kinds of tasks. Kismet is probably the first autonomous robot explicitly designed to explore socio-emotive interactions and communication. Kismet used an expressive vocalization system to generate a wide range of emotive utterances corresponding to joy, sorrow, disgust, fear, anger, surprise, and interest. The innovation in Kismet's infrastructure is the emotional system that interacts intimately with its cognitive systems to influence behaviour and goal arbitration. The scientific literature documents the beneficial effect of emotion on creative problem solving, attention, perception, memory retrieval, decision-making, learning, and more. The emotion system is responsible for assessing the value of immediate events in order to appropriately trigger the cognitive system to help focus attention, prioritize goals, and to pursue the current goal with an appropriate degree of decision *[34].* In fact, the emotion system would help robots to improve their functionalities in complex environments and to behave appropriately in a socially acceptable manner with people. As a result, human and robot would mutually regulate others behaviour through social cues. Kismet's emotion system and cognitive system work together to implement the following capabilities:

- intelligent behaviour in a complex, unpredictable environment;
- ability to sense and recognize emotion and affect in others;
- ability to express the internal state to others;
- ability to respond to humans with social adeptness and appropriateness.

Researchers also found that simple acoustic features (such as pitch mean and energy variance) can be used by the robot to classify the affective prosody and related emotions. Using these acoustic features, Kismet could recognize the affective intent from human speech [*Breazeal C., Takanishi A., Kobayashi T. 2008. Social Robots that Interact with People. Springer Handbook of Robotics].* In another set of studies, Kismet recognizes four affective intents (praise, prohibition, attentional bids, and soothing) from a person's vocal prosody. The person who interacts with Kismet can manipulate the robot's affective state through tone of voice. The robot becomes more positive through praising tones, more aroused through alerting tones, sadder through prohibiting tones, and moderately aroused through soothing tones *[35].*

4.3.1.2. Other social robots

Pepper

The humanoid robot by SoftBank and partner French robotics firm Aldebaran. The social companion robot was introduced in two Belgian hospitals as receptionists already in 2016. More than 140 SoftBank Mobile stores in Japan are using Pepper as a new way of welcoming, informing and amusing their customers**.** The robot is able to recognize principal human emotions, respond appropriately to moods as well as questions. It can recognize the human voice in 20 languages and can detect whether it is talking to a man, woman or child.

Jibo

In 2012 MIT roboticist Cynthia Breazeal announced Jibo as “the first social robot for the home.” Actually, it is a social robot, designed to be another member of the family. The desktop robot listens in on the conversations in the family environment and assists family members with providing diary reminders, delivering messages to people, taking photos, telling stories as well as being a companion you can talk to. Software developers will be able to extend its functionality in the future and allow for Jibo to be connected with others, as well as household devices.

Miko

The tiny, wheel-powered robot is the Indian response to the robotic revolution started in the Western part of the world. It is the first companion robot developed by a Mumbai-based startup Emotix. Miko is aimed at children above the age of five years. Similar to the smart dinosaur of Cognitoys, it is also artificial intelligence-based growing and changing together with your kid. It can talk, respond, educate and entertain. It understands the specific needs and emotions of your child and reacts accordingly.

4.3.2. Retail Robots

It has been shown a growing adoption of technology and robotics in retail to attract and retain customers. Robots are becoming the first customer touch point for many physical retail environments. As the capabilities of these intelligent assistants continue to develop, robots working alongside human staff could become a common in-store best practice in the future. “Shopping centres are becoming smart improving customers satisfaction with tangible services, reliability, responding promptly to the customers' needs, providing a sense of empathy, in order to compete with the today advanced technologies” [36]. Service robots could carry out activities that are of great help for customers, such as the creation of a shopping cart or discovering where the customers' desired products are located in the shop. Personal engagement could be also useful. A shopping humanoid assistant could collect the physical information about what a customer likes or dislikes gathering data from customer behaviour. Furthermore, users' biometrics attributes could be scanned, the client can be recognized together with his mood. Thanks to Big Data, shopping companion could also access to some user’s information stored on social media and collect insights about interests and personal attitudes. Thus, the robot can recommend items according to the user’s preferences, actual emotional state, previous shopping history and personal features. Other additional services are related to the paying activity, such as the currency exchange, the gift packaging, the price comparisons, and the dispensing of coupons. In retail, automated robot platforms that embed the above-mentioned functionality are already present. Many humanoid robots have been exhibited in different shops of the world, technological enriched environments allow for a huge variety of robot services.

4.3.2.1. Using Robotics to deliver customer experience: A competitive advantage

Consumers' emotional engagement is at the core of the strategy for the use of humanoid robots in retail. Being proactive, expressive, attractive and mobile, robots create easily an empathetic link with shoppers by their humanoid behaviours. They can bring enormous advantages to the customer experience because they are connected, effective, accurate and multi-functional. Also, they are never tired of repetitive tasks, which enable staff to focus on more added-value tasks. The implementation of this strategy in retail may lead to trigger customers’ curiosity, increasing store traffic and attract the undivided attention of shoppers. Immediately, this impacts on sales. By enhancing product visibility through the attractiveness of the innovative solution, they stimulate purchase and retain loyal customers. They also allow merchants to build memorable in-store experiences, transforming the customer journey. Despite the increase in online shopping, many customers still convert sales in the physical store. So, humanoid robots guide and assist shoppers in purchasing product or improve awareness by providing services, based on their needs and requirements in continuity with their on-line journey. Instead, marketers' greatest value is the possibility to gather comprehensive data during human-robot interaction to enrich customer base and generate shopper insights.

Softbank, with their flagship product Pepper and NAO robot, are momentarily considered the most advanced robotics models offered in the non-industrial market. Pepper, rolled out in Japan and some other countries, is already reshaping the shopping experience. The robot is able to have two-way conversations with customers and understand 80% of a conversation. It is also able to read the emotions of the customer and describe the differences in products to them. More than 140 Softbank Mobile Stores in Japan are using the humanoid robot Pepper. Apple recently announced that it was adding $50 billion to Softbank's new Vision Fund, which is designed to pave the way for the next generation of tech companies. Pepper can be programmed to chat and interact with customers, give direction and answer questions. For instance, Pepper interacts in Q Square Shopping Mall (Taipei) with customers by giving directions and information, as well as greetings, and by informing shoppers about promotions and events in the mall. To brighten up the in-store experience, Pepper also plays music, dances, illuminates itself and takes selfies with customers. Thanks to their programming flexibility they can be adopted in different retail applications. Investigating the main functionalities embedded in Peeper and NAO robots, the typical use cases in retail proposed by Softbank are the following:

- *Greeter*: humanoid robots ensure a consistent and high-quality welcome to customers for lasting engagement and an increased return customer rate. Robots have a proactive engagement, they call shoppers' attention, initiate interaction by voice and animated expression, or proactively go towards people to start conversations.

- *Service Provider*: aggregating multiple services such as click and collect, payment, order & delivery and ticketing, Pepper and NAO strengthen customer engagement, raise conversion rates and improve overall customer lifetime value. They also have multi-language skills, Pepper and NAO can speak up to 21 languages and provide translation via Cloud services.
- *Sales Associate*: using their sensors, Pepper and NAO can categorize and identify the person they are interacting with, adapting their behaviours to the client’s profile. Connected, robots assist sales force throughout accurate product consultation and personalized recommendation. Their robots excel in upselling & cross-selling.
- *Loyalty management*: Create a new way in which retailers can enrol, engage and retain customers. Encourage brand loyalty and attract new business with engaging interactions and valuable services.
- *Brand Ambassador*: Constantly delivering products and brands messages. Pepper and NAO are reliable communicators in presenting corporate activities, marketing campaigns with accuracy and consistency across all locations. They provide the latest information such as product availability, special offers or location in the store.
- *Data collector*: Being perceived as 'friendly' and 'no judgement', their robots are easily accepted by shoppers, feeling more confident to answer openly to questions. They can deploy interactive shoppers’ satisfaction surveys and monitor the results in real-time to create actionable insight and improve services. They also collect information enriched by the sensors about consumers behaviors, generating precious added-value insights.
- *Retailtainment*: Using enriched interactions, taking advantage of the verbal interaction supported by the full body language, their robots are an engaging channel for in-game advertising and couponing. Those robots can have continuously evolving behaviors to interact with customers based on their real time analysis results.

In 2016 LoweBot, an autonomous retail service robot was introduced in Lowe’s Stores in the San Francisco Bay area. Customers can ask LoweBot where to find items they need inside the store. Shoppers can also ask the robot some basic customer service questions. As LoweBot helps customers with simple questions, it enables employees to spend more time offering their expertise and specialty knowledge to customers. The robot also performs real-time inventory tracking, that can be used to identify shopping patterns at the location, so that way the retailer can not only resupply its shelves but also get more understanding of which merchandise moves more quickly, and when it happens.

Neurodata Lab introduced Promobot, a robot that is able to recognize emotions and react accordingly, as well as to measure how satisfied a user is with the interaction. The robot is able to accurately recognize 7 emotions: happiness, sadness, surprise, anger, anxiety, disgust, and a neutral state. It also has two emotion recognition applications: for business and personal use. Business use means Promobot is calculating a Customer Satisfaction Index, and personal use means Promobot is calculating a Smile Index. In either scenario, the Promobot adapts its answers and reactions according to the index calculated. Neurodata Lab's Emotion AI finds many business applications in numerous fields.

Brands are testing robots even as a means of information dissemination through the application of an advertising campaign [37]. The experiments aim to understand whether live robot advertiser can better engage the audience and improve audience perception of a marketed product as compared to pre-filmed video advertising clips. The experiments were carried out in the busy CBD district of Singapore with 180 teenage passers-by. An Aldebaran Nao humanoid robot was used as the robot advertiser to merchandise robotic toys. The results suggest that based on the indicators such as valuation price and perceived liking of the product, physical robot presence will enhance information dissemination and hence improve advertising result, providing value-add to the advertised product.

4.4. AI’s and Virtual Assistants (VA)

On top of the existing transactional applications of AI-enabled technologies, conversational voice assistants are catching on in a big way. The global artificial intelligence market includes querying method, voice/speech recognition, image recognition, video analysis, context-aware processing, deep learning, language processing, gesture control, and digital personal assistant. Estimates indicate that the AI market is already growing enormously. With the advancement in Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Artificial Intelligence (AI), the business use cases for voice-based chatbots have increased over the last few years. According to MarketsandMarkets, the global voice assistant application market size is expected to grow from USD 1.3 billion in 2019 to USD 5.2 billion by 2024, at a Compound Annual Growth Rate (CAGR) of 31.9% during the forecast period. The rapid growth of this segment is due to the increased use of portable computing devices and due to the increasing accessibility of devices, coupled with cost-effectiveness and advanced features. Many retailers are now starting to experiment with voice assistant applications. Most of these projects, however, focus on transactional features of voice assistants. In these cases, consumers give voice commands to the web-interface which in turn executes them correctly. This changes the way consumers interact with web-interfaces from written to spoken but does not offer new or improved services that help the consumer forward in their shopping experience. The voice assistant can react with relevant advice or ask deepening questions to dive deeper into the individual needs. This will benefit the consumer in several ways, consumers will experience instant personalization based on the context they provide to the smart algorithms and they will save the time spent browsing through the endless variety of products an e-commerce webshop has to offer.

4.4.1. Localization

There are several factors and performance indicators to be taken into account for the evaluation of the opportunity that firms have with AI technology. Nowadays, as a consequence of business digitalization phenomena, the communication between shoppers and the company's brand happens more online than using other channels. The repeated interactions generate traffic, and this traffic is extremely important for brands in order to maintain an active engagement with customers. Customer engagement is a key success factor for companies who have brands promoted online [38]. This factor is so important for brands because it sets off an iterative process that culminates in customer loyalty. Successful companies are those who have good brand engagement but also have successful technologies that engage their consumers. Gartner projects that 85% of customer interactions in the retail sector will be managed by AI by 2020. Smart chatbots, for instance, are redefining customer service. Traditional retailers and e-commerce are now deploying AI across the whole supply chain from product development and merchandising to marketing and customer engagement, in the hope of improving operational efficiencies, reducing costs and enhancing shopping experiences. With a wide array of technologies available, companies are trying to understand how they should invest their money in incorporating technologies in their strategy to increase the number of loyal customers. Currently, in the market, companies like Microsoft, Google, Amazon, and Apple are trying ways to engage with consumers as frequently as possible via voice recognition. Using virtual assistant via voice–user interfaces presents a huge challenge but also a great opportunity for marketers. The importance of customers interaction's quality is essential. However, due to the great amount and heterogeneity of online interactions, it is not obvious how to quantify and enhance their quality. A solution could be pointing at improving another factor: *localization*.

Localization is the process of adapting the web content to the end user's cultural and locale-specific expectations [39]. Localization means that communication and digital contents need to be tailored based on cultural, linguistic, functional, and other locale-specific requirements. These local requirements may include also personal information of the user, coming from social media, life context, mood and emotional state. Understanding the customer and the full-context of the situation will help in applying the right offer, guidance, wording, security and accuracy that together will be critical in driving faster adoption of voice assistants. Localization helps firms to embrace the target culture and deliver a personalized experience, consequently gaining trust and increasing engagement of customers. Many companies are investing in artificial intelligence to improve their marketing effectiveness, service quality and provide customized experiences. For example, according to [40], there are some technologies that provide multilingual solutions for improving chatbots using AI, or software that allows brands to recognize and monitor consumption local trends or behavioural patterns of customers. This technology is considered still young, but it is a source of big opportunities to be exploited by brands. Since products or services are perceived more personalized and familiar to consumers thanks to the adaptation of contents to the user's local environment, traditional interactions with customers empowered with localization aspects can provide more value to the final client. Delivering repetitively value to clients will ensure another success factor: customer satisfaction. Companies have to make consumers living delightful experiences each time they interact with the brand in order to build a class of loyal customers.

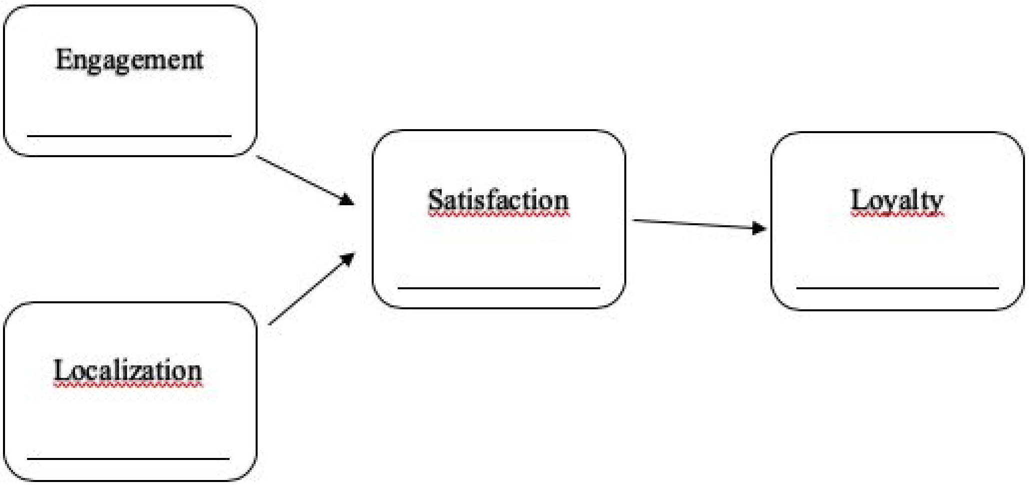


Fig. 10 *Localization affects Customer Loyalty*

For transactional activities, often habitual purchases, not much thought is required. Neither explanations of a product’s details through long conversations with the Virtual Assistant is necessary. Instead, non-transactional activities require more information to be processed and more time to investigate which is the real need to satisfy. To benefit from integrating new technologies such as a voice assistant, VA’s need to be useful, easy to use, and localized for non-transactional based activities. This will lead to a strong consumer–company engagement relationship and result in high loyalty.

4.4.2. Emotional AI: Driving Experience

Emotional artificial intelligence or Emotional AI is also known as emotion recognition or emotion detection technology. Emotion AI can detect real-time changes in mood, differentiate age, gender, and emotion. They can perform these skills by listening to changes in the voice’s paraverbal aspects. In October of 2018, Amazon filed a patent labelled “Voice-Based Determination of Physical and Emotional Characteristics of Users” related to detecting the physical and emotional wellbeing of users based on interactions captured in voice assistant data. The primary use case provided by the company is in the healthcare industry to detect illnesses. However, the patent is not limited to this unique usage and could be extended to different types of normal speech. The detection of a physical or emotional anomaly may also be used to modify how Alexa reacts to requests or what suggestions are made. It can be used to provide highly relevant content to the user at that particular time, and which may not be relevant later. Moreover, physical and emotional cues may also be combined with other characteristics such as user age, demographics, location and browsing history to further customize the interaction. Emotion sensing systems and affective computing allow smartphones to detect, analyze, process and respond to people's emotional states and moods. The proliferation of virtual personal assistants and other AI-based technology for conversational systems is driving the need to add emotional intelligence for better context and enhanced service experience.

Companies are investing in this direction to develop innovative services to provide to their clients. The number of such services is growing rapidly, and Internet-of-Things device manufacturers are also building voice control into their products. Car manufacturers, for example, are interested in understanding a driver's physical or emotional condition, or the stress and fatigue level, in order to increase safety. Any abnormalities in the driver behaviour can be identified so that an appropriate warning mechanism can be developed, to prevent traffic accidents from happening. In the experiment [41], researchers proposed a system that helps to detect abnormal driver behaviour from a dataset of driver speech recordings. They used the emotional speech data based on three different culture bases: American, European and Asian. Driver behaviour state could be normal, sleepy, talking and laughing, while the emotions that can be recognized are some basic ones (angry, sad, happy, neutral). Results show the potential of detecting sleepy driver from just the speech signals.

Advances in natural language processing (NLP) enable people to have increasingly conversational experiences with computers through voice. Consumers interact with brands through many channels, one of the most common conversational experience are call-centres. It is important for the Service Level Agreement (SLA) to take note of customer disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied. Ineffective resolution of these disputes can often lead to customer discontent and loss of business [7]. By analyzing automatic anger recognition in speech could be an important factor to improve user satisfaction. IBM, for instance, uses Watson to analyze customer service conversations. It’s worth considering that the use of voice analysis to detect emotion is already a reality in business. For instance, IBM offers its “Tone Analyzer” to improve customer satisfaction during service calls. According to TechRepublic, “the tool can pick up on seven different types of tone via conversations with customer service agents and chatbots: frustration, satisfaction, excitement, politeness, impoliteness, sadness and sympathy”.

4.4.3. Voice Commerce: a new channel and enabler for eCommerce

Voice commerce is an alternative to using a keyboard and mouse to order and purchase products online. It is not limited to finding the product itself but also ordering and buying it. According to “Voice Shopping Consumer Adoption Report” by Chatbot.AI, voice commerce is poised to become the third key online channel for shopping, joining web and mobile. From this survey made in the US, results show evidence of age differentiation between preferences for physical store shopping versus mobile and voice shopping. Indeed, 18-44-year-olds are 2.4 times more likely than 45+ adults to rate mobile shopping as their preferred method and 8.5 times more likely to rate voice shopping as their preference. The survey also reveals that one-in-five shoppers on average (21.2%) have used voice in their shopping activities at least once. In f**ig.11** data split by age are reported.

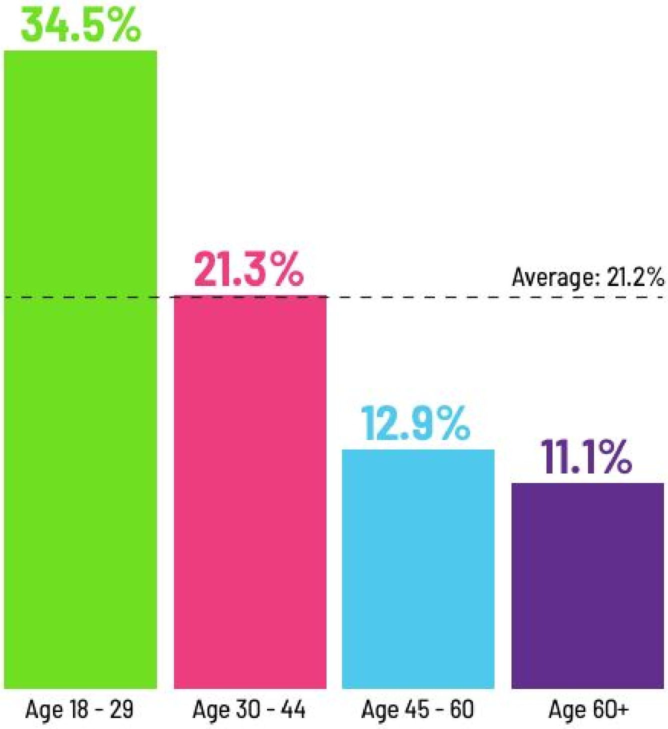


Fig.11 *Voice Commerce experience by age*

Gartner predicts that 30% of webpage visits will be by voice in 2020 and ComScore estimates voice will account for 50% of all search the same year. The increasing number of online researches and transactions executed through voice is straining many companies to look closely at their voice commerce plans and strategy. The main advantage identified by those consumers who took part in the survey are the following:

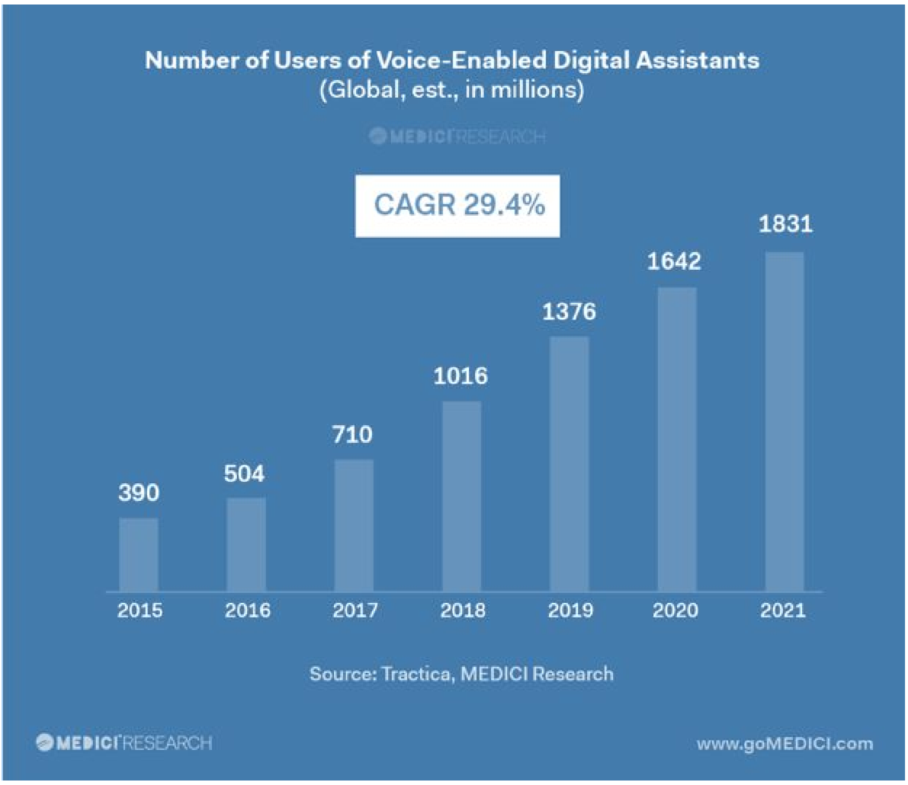
- 27.3% replies that voice commerce is hands-free;
- 20.7% reputes that it enables multitasking;
- 18.9% agrees in assessing that it is faster to get answers and results.

The principal dislikes of voice commerce are related to intangible concerns about not trusting voice for payment transactions. As a result, consumers often switch to traditional online shopping via web and mobile to purchase the products. Most customers want to see a new product before they buy it. The online challenge is to give an exhaustive and compelling product description, even with dimensions noted and photos included. Hence, voice commerce has yet to reach mass adoption levels by consumers and the technology itself may still be at a nascent stage. However, as consumers become more accustomed to voice interactions and continual innovations in technologies, voice is gaining rapidly role of a critical channel for retailers. For instance, Walmart offers voice-based shopping through Google Express making it available to both Google Home's owners and smartphone users utilizing Google Assistant. Costco, Target, and The Home Depot are other examples of Google Express retail partners. It is surprising that Google has a higher voice commerce trial rate than Amazon, but the data is consistent across three surveys in 2018. In France, Google Home devices are used also to shop at Carrefour. Amazon purchases can be made through Alexa voice assistant. In November 2016, Amazon started offering daily deals for voice shoppers. This typically provides a choice to smart speaker users of two different items on discount for that day only. In South Korea, the online store of retail conglomerate Lotte has launched a voice-activated search-and-buy service through its mobile app. Starbucks has developed voice recognition ordering in South Korea, extending its mobile order-and-pay technology by integrating with Samsung's intelligent AI chatbot, Bixby.

Voice commerce is growing, and voice assistants are changing the way people research products, compare prices and make purchases. According to the international consulting firm OC&C Strategy Consultants, the value of voice commerce is predicted to grow to over US$40 billion in 2022. Most users turn to voice commerce specifically when it comes to purchasing groceries, entertainment products and services, electronics products, and clothing. Brands and retailers are quickly adapting to the voice shopping trend. Voice is the richest source of information about how and why customers purchase or don't purchase. Leveraging this unique data asset will play a crucial role as a competitive advantage. Consumers expect retailers to know their tastes and to provide the products they want, where and when they desire them. AI makes that possible by constantly analyzing masses of customer data to provide real-time insights. Analyzing data about customer profiling and combining them with personalized context information, will enable brands to tailor the most valuable service to each client. Managers can also take advantages from AI to better manage marketing budget. Large-scale marketing campaigns through channels like TV, radio, and social media are expensive and inefficient: the message is spread to the masses, but it will only truly reach some of those people. By continually analyzing customer data, retailers will be able to provide more targeted ads, reducing the waste of marketing budgets. AI creates opportunities for the ultimate one-to-one marketing. Most innovative brands have begun using AI-enabled recommendation solutions to analyze shopping behaviour and deliver predictive insights that enable the sales team and the e-commerce site to offer relevant product ideas and recommendations. A great example is Apple retailer Humac in Denmark. They implemented LS Recommend, an AI recommendation service that suggests items the customer might need. In the first three months of use in Humac, sales of recommended items led to total margin growth of 46%, while customers' basket size grew, and return visits to buy forgotten accessories dropped down. For Amazon, the use of AI and machine learning also enabled to offer cloud-based services to users of Amazon Web Services (AWS) and to allow shoppers to take products and exit Amazon Go grocery stores without passing a checkout.

4.4.4. Alexa - Google assistant - Siri - Cortana

In this chapter we focus on the smart speakers enabled by AI, providing an overview of the big players who are controlling the actual market. Amazon Alexa, Google Assistant, Cortana and Siri are voice interfaces that provide a new way of interacting with home devices as well as business applications like conversational banking, commerce, analytics for contact centres and search, among others. Voice interface technology is relatively new and presents unlimited possibilities for organizations to explore. The market for VA-enabled wireless speakers is expanding rapidly, as we can see in ***fig.12,*** with more vendors, device types and use cases. Globally, the number of customers using voice-assistants is increasing with a CAGR of 29.4% and is estimated to reach 1.83 billion by 2021.



Source: Tractica, MEDICI Research

Fig.12 *Number of users of Voice-Enabled Digital Assistants*

Voice assistants are becoming a part of the family for many consumers, being present in their homes and on their mobile devices. This intelligent technology has changed the way people interact with their devices. In 2019, the new generation of VPA speaker products is shipping with some artificial intelligence (AI) functions running on the device rather than in the cloud. Vocal Assistants are in continuous development, and in order to be helpful at various business-related tasks, they need to be able to reason and deduce information based on what the speaker says. Also, they have to be aware of human emotions and formulate responses that are emotion-relevant. These questions aim to determine the level to which tested voice assistants are capable of handling emotion-related questions or responding in an emotion-relevant manner.

"With smartphones increasingly becoming a commodity device, vendors are looking for ways to differentiate their products, future AI capabilities will allow smartphones to learn, plan and solve problems for users. This isn't just about making the smartphone smarter, but augmenting people by reducing their cognitive load. However, AI capabilities on smartphones are still in very early stages." Said CK Lu, research director at Gartner. Indeed, Gartner predicts that by 2022:

- 10% of personal devices will have emotion AI capabilities
- 80% of smartphones shipped will have on-device AI capabilities on-device. AI is currently limited to premium devices and provides better data protection and power management than full cloud-based AI since data is processed and stored locally.

Assistants are available on most smartphone platforms [42]. Google’s assistant is integrated into Android phones and can be installed as a separate app on the iPhone. Amazon’s Alexa has Android and iOS app versions, Siri is available on all Apple devices. Looking at the past, Apple’s Siri was the first digital voice assistant to be incorporated in a smartphone (iPhone 4s), in October 2011, but the technology was later applied in smart speaker HomePod, which was launched in February 2018. Microsoft followed shortly thereafter with Cortana in 2013. Amazon launched Alexa with its Echo-connected home speaker in 2014, and Google's Assistant was announced in 2016 along with its Home speaker and is also embedded in the Google app for Android-based smartphones. These have become the most popular voice assistants in Western markets.

Amazon entered the smart speaker market in 2014, establishing as a first-mover advantage. Being first has a significant payoff since it gives a strategic advantage for enhancing the firm's image and reputation with buyers and building customer loyalty, which is hard to displace. The first launch of a home product with a large media library available out of the box, allowed Amazon to become the dominant player in the field. Amazon has a competitive advantage in this scenario for some reasons:

- It has an installed user base that is more than ten times larger than Google Home
- It has more voice applications for Alexa which makes it more attractive to many consumers
- Amazon has dozens of third-party manufacturers using Alexa Voice Service which can deliver a common experience across multiple devices that can address a wider range of consumer preferences

Innovative technology can provide a sustainable cost advantage for the first-entrant also thanks to the learning curve phenomena. When developing disruptive technology, the most important advantage to defend is know-how developed by R&D. "For first-movers, extra time over later entrants can for the capture of customers and the building of capabilities that can advance and perpetuate early gains. However, time can also benefit later entrants by helping to improve risk estimations, clear uncertainties, remove capability gaps, and facilitate learning, particularly from the costly activities (and mistakes) of earlier entrants" [43]. An optimal choice between first-mover and follower is often linked to a firm's resource-capability mix as well as market and industry conditions. In ***fig.13*** a framework of factors that influence the entry-timing is presented.

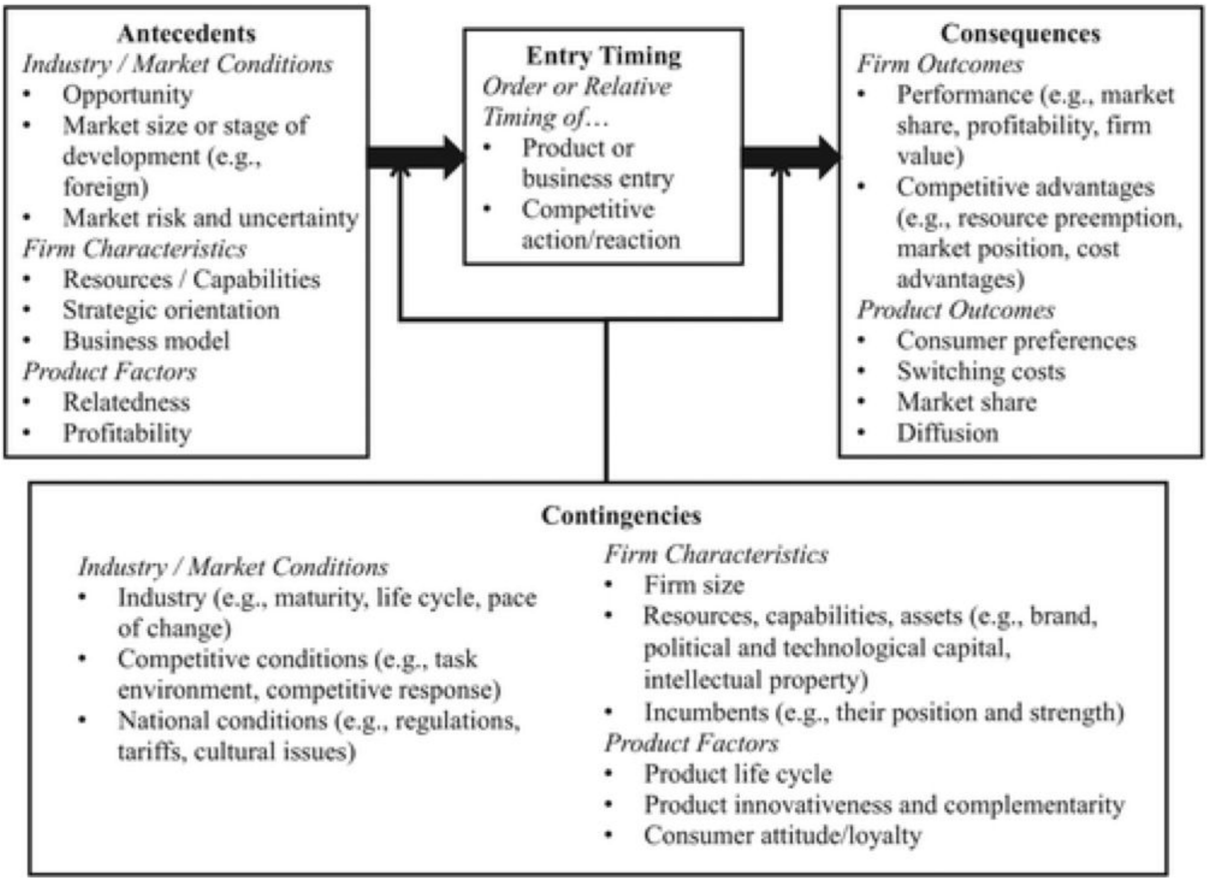

Antecedents
Industry / Market Conditions
• Opportunity
• Market size or stage of development (e.g., foreign)
• Market risk and uncertainty
Firm Characteristics
• Resources / Capabilities
• Strategic orientation
• Business model
Product Factors
• Relatedness
• Profitability
Entry Timing
Order or Relative Timing of…
• Product or business entry
• Competitive action/reaction
Consequences
Firm Outcomes
• Performance (e.g., market share, profitability, firm value)
• Competitive advantages (e.g., resource preemption, market position, cost advantages)
Product Outcomes
• Consumer preferences
• Switching costs
• Market share
• Diffusion
Contingencies
Industry / Market Conditions
• Industry (e.g., maturity, life cycle, pace of change)
• Competitive conditions (e.g., task environment, competitive response)
• National conditions (e.g., regulations, tariffs, cultural issues)
Firm Characteristics
• Firm size
• Resources, capabilities, assets (e.g., brand, political and technological capital, intellectual property)
• Incumbents (e.g., their position and strength)
Product Factors
• Product life cycle
• Product innovativeness and complementarity
• Consumer attitude/loyalty

Fig.13 *Entry Timing framework: antecedents, contingencies and consequences*

However, skills and know-how developed by first movers are easy to replicate. This opportunity opens the market to other entrants and creates competition. Google acted as a rapid second-mover exploiting spillovers from the competitor's innovation. A second-mover firm can learn from the experiences of the first mover. Also, the second firm does not face the marketing task of having to educate the public about the new product. For these reasons, Amazon's early entrance into the smart speaker race has helped it capture the greatest share of the global market in 2017, but the company's grip on the space is slipping. The current situation is predicted to change over the coming years.

Amazon may know shopping preferences, but Google through its Gmail, calendar, maps, Waze and Chrome services has a lot of context about people and their daily lives. Moreover, Android provides a key platform for Google because of Google Assistant on smartphones.

Despite Apple is the pioneer that introduced artificial intelligence, they gained a smaller market share, compared to expectations and benchmarks. Apple launched its HomePod smart speaker in February 2018. The company likely expected the device to make greater results in the market, and consequently slashed production of its HomePod and lowered sales forecasts for the device in late March. Basically, two main reasons may explain the failure. The first issue was price, the second the lack of functionalities and updates. Apple quietly dropped the price on its HomePod smart speaker by 14% from $349 to $299. Apple's HomePod only appeals to buyers in the premium segment of the space. In ***Fig.14*** is represented the expected market size of main players operating in the VPA's market.

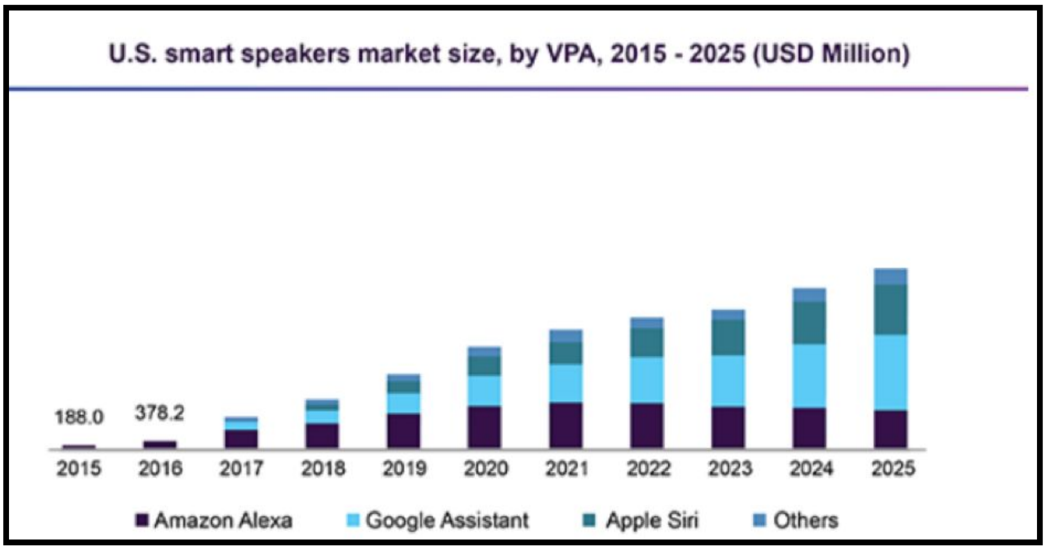


Fig.14 *Predicted voice-enabled smart speakers' market*

Most industry watchers agree the smart assistants market presents a major opportunity for companies. The opportunity is coming from the evolution of consumers' preferences and expectations. As customers preferences change, brands must re-evaluate their approach and engagement to remain in line with evolving expectations. Apple's relative failure can be seen as a signal of evolving preferences since customers when purchasing smart devices such as smart speakers are looking for products' functionalities and convenience. In this newly emerging market, customers don't feel lock-in to a single brand, indeed companies that put consumer needs first are in a position to win. Since disruptive technology is still in the early stage of its evolution, improving voice recognition and emotional analysis system could represent a competitive advantage for the next generations of voice-enabled smart devices. Gathering customers feedbacks is essential for companies to improve technology in line with the client's perspective. From a survey [44] made on the comparison between main VA's in the market: "A remarkable feature of Google Assistant was the naturalness of answering some questions. The tone and pace of the female voice used by the Google device expressed surprise, suspense and joy. These features were not always offered by Siri and Alexa."

4.5. Voice-Enabled Business Applications: A comparison

After the description of some most valuable commercial applications of voice analysis, which are being developed in the market, our work continues with their comparison based on two dimensions: potentiality and maturity.

4.5.1. Potentiality: Market Growth

One of the major benefits of innovation is its contribution to economic growth. Innovation can lead to higher productivity, meaning that the same input generates a greater output. The more productivity increases, the more goods and services are produced, the more the economy grows. Thus, the predictive market growth can explain the potential of emerging technologies. Analysts are used to applying "compound annual growth rate", or CAGR, in order to evaluate the potential growth of a market. CAGR refers to a representational percentage that shows how much a business has grown or will grow in a time-gap of at least two years. It represents one of the most accurate ways to calculate and determine returns for individual assets, investment portfolios and anything that can rise or fall in value over time. Also, CAGR can be used to compare investments of different types with one another. Since technologies can evidence short life-cycle, we consider the relative market growth of a maximum of 5 years (2019 – 2024) as a meaningful period of time to assess the potentiality of voice-enabled business applications.

4.5.2. Maturity: Hype Cycle

To evaluate the degree of technological maturity, we decided to use the Hype Cycle model, developed by the American research, advisory and information technology firm Gartner. This Hype Cycle specifically focuses on the set of technologies that is showing promise in delivering a high degree of competitive advantage over the following 5 to 10 years. Gartner uses hype cycles to characterize the over-enthusiasm, or "hype," and subsequent disappointment that typically follow the introduction of new technologies. The hype cycle provides a graphical and conceptual presentation of the maturity of emerging technologies. A hype cycle in Gartner's interpretation has five steps: Technology Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment, Plateau of Productivity **(fig. 15).**

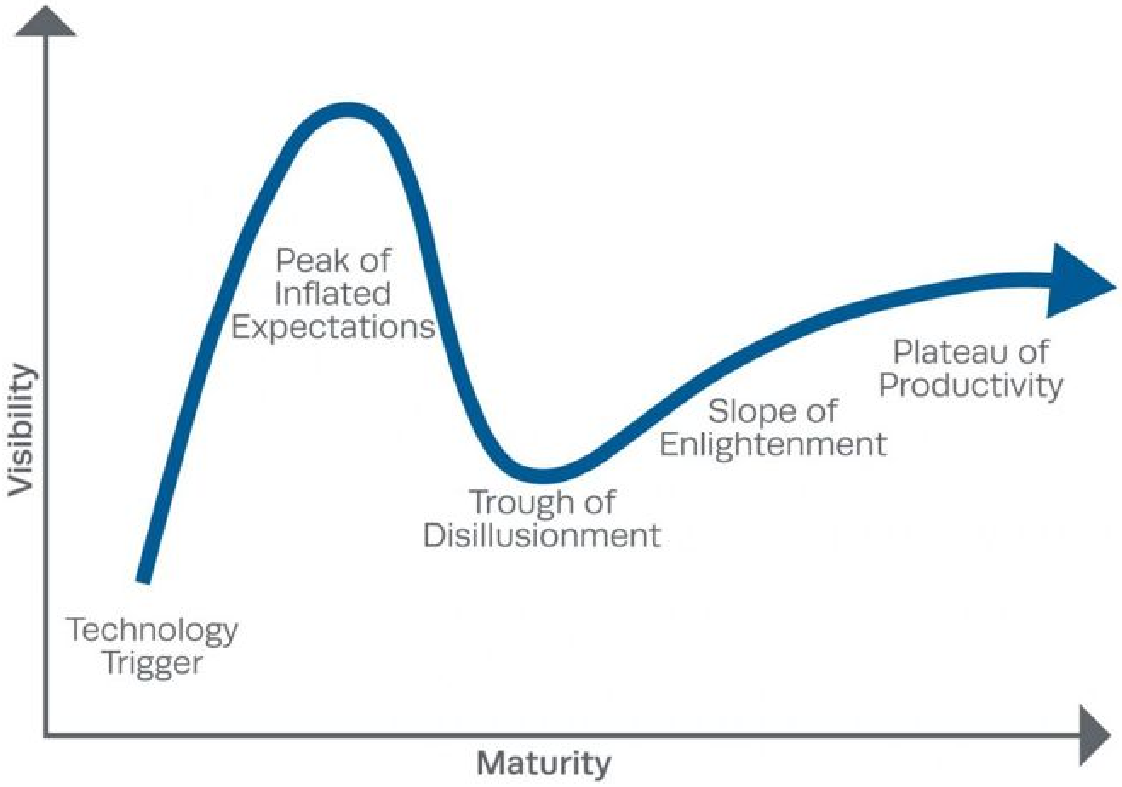


Fig.15 *Gartner's Hype Cycle for emerging technologies*

Technology Trigger

The first phase of a hype cycle is the "technology trigger" or breakthrough. Early proof-of-concept stories and media interest trigger significant publicity. Often no usable products exist, the market has still to disclosure the new technology. Normally, "innovation trigger" phase includes some main steps that are related to the birth of new tech companies, such as the first round of venture capital funding, necessary for R&D investments to build first-generation products. In this first part, the early adopters of the new technology start to investigate.

The peak of Inflated Expectations

Publicity typically generates over-enthusiasm and unrealistic expectations. It produces a number of success stories, often accompanied by scores of failures. Usually, companies begin to implement activities beyond early adopters in order to keep the hype up on the emerging innovation. The first-generation products are entering the market, often with high prices and lack of customization. Suppliers of this technology start to raise and proliferate.

Trough of Disillusionment

Technologies enter the "trough of disillusionment" because they quickly become unfashionable, the press usually abandons the topic and bring attention to other emergent technologies. It is not an end-point for those companies who continue developing commercial applications, often in this phase, they collect second/third rounds of venture capitalist funding. At this point, most valuable and strong firms consolidate their position in the market, weaker companies, by contrast, fail and exit the market. Normally, second-generation products start to be available for consumers, still considered early adopters. At the end of this phase, commonly less than 5% of the potential audience has fully adopted this technology because people lose faith in it, even while the underlying technology continues its exponential growth.

Slope of Enlightenment

Although the press may have stopped focusing on the technology, some businesses continue through the "slope of enlightenment" and experiment to understand the benefits and practical application of the technology. Continuous efforts in time and cost to improve its capabilities may lead the technology to surpass early anticipation and expectations. The market starts to see rapid advancements and the technology’s potential for further applications becomes more broadly understood, increasing number of companies. Usually, methodologies and best practices are developed together with third-generation products, which incorporates more capabilities, more personalization, even out-of-the-box.

Plateau of Productivity

The technology becomes widely implemented and mainstream adoption starts to take off. Criteria for assessing providers viability are more clearly defined. The technology's broad market applicability and relevance are clearly paying off, since technology’s place in the market and its applications are well-understood. Finally, the technology is becoming mature, abundant revenue is generated and customers start to take the product for granted. In this last phase, often adoption present high-growth rate: 20-30% of the potential audience has adopted the innovation.

A Hype Cycle can help executives in assessing relative risk and timing of emerging technologies and evaluate trade-offs between risk and innovation. Many business decisions are complex, but entry choices are particularly spiny because they often reflect a shift in strategy, operations, or even business model. Entrants must consider diverse contingencies that differ in risk exposure, resource- capability commitment, and the amount of control over entry processes and outcomes. It raises interesting questions for when businesses should adopt new technology, entering the market as second-fast-mover. As represented in ***fig.10,*** adopting a new technology during the **"**Peak of Inflated Expectations" could present many risks since the expectations are on the top, but the uncertainty of satisfying them is high too. During "Trough of Disillusionment" phase, companies have already invested a lot of resources in the innovation and giving up at this time could have negative payoffs. Otherwise, it could be an interesting moment for adopting. In "Slope of Enlightenment", knowledge advancement about the innovation makes the risk of adoption much lower. However, building the innovation's core competencies necessary to absorb the knowledge that other players already possess could be very tricky for a new-comer at this point. In the last phase, companies need to understand how to evolve the technology, how to enlarge its range of business applications, generally, how to re-invest the knowledge that has been developed.

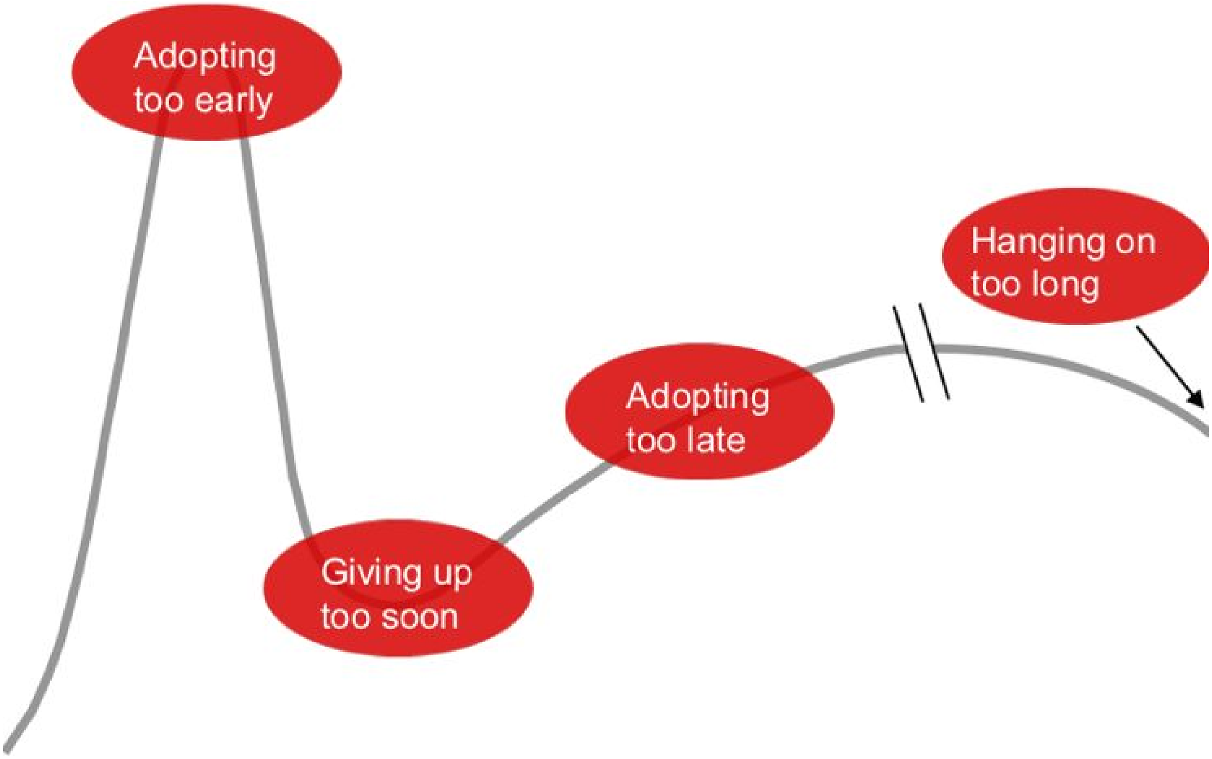


Fig.16 *Adopting new technology: timing is essential*

4.5.3. Graphical Representation: Potentiality - Maturity

Previously illustrated business applications are represented in the bi-dimensional graph (***fig.17***). The X-axis is taken from the Hype Cycle maturity dimension that explains technological maturity, while the Y-axis quantifies the predictive relative market growth (%) of the commercial application over the next five years. Afterwards, we comment on each business application explaining its position on both dimensions.

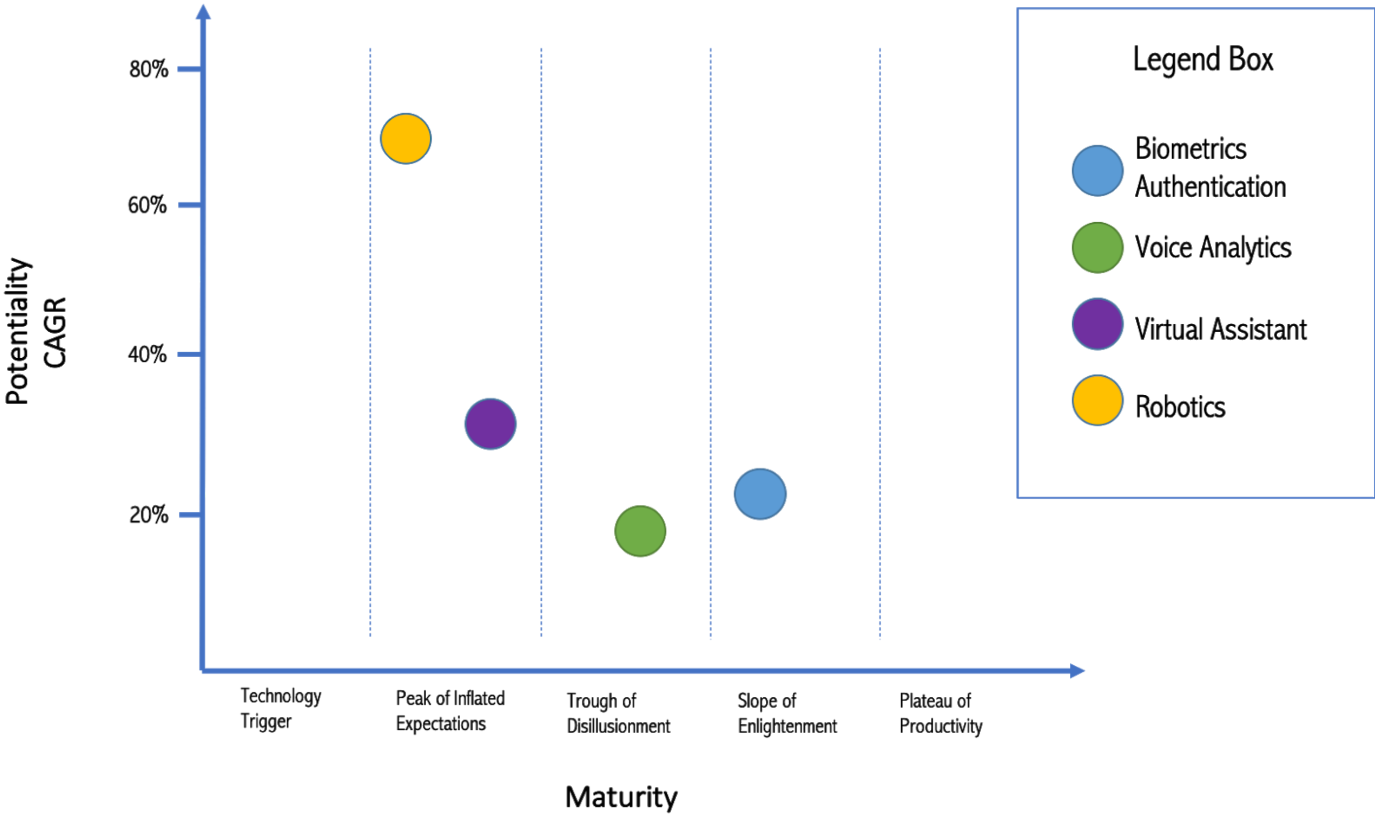


Fig.17 *Bidimensional representation of the business applications' comparison*

4.5.4. Voice Analytics

Potentiality

The voice analytics market size is expected to grow from USD 657 million in 2019 to USD 1,597 million by 2024, at a Compound Annual Growth Rate (CAGR) of 19.4% during the forecast period.

Customer analytics is the systematic analysis of the customer insights which support organizations to identify and target customers for marketing programs, customer behavioural prediction, gaining customer loyalty and customer retention. Factors such as growing need to extract insights from customer interactions, the rising demand to monitor consumers behaviour, need to extremely personalize products and service, increasing attention on a customer-centric approach, are expected to drive the market growth. To achieve such result, data coming from customers interactions have to be processed. Data optimization is done by the combination of process and technologies such as behavioural analysis, sentiment analysis, predictive modelling, segmentation, and data visualization. The voice analytics solution empowers users to analyze recordings of conversations to identify the emotional state and intent of the speakers. Every enterprise has the necessity to understand the changing business conditions, client insights, market trends, or service inconveniences. This information makes companies able to create corporate branding, marketing campaigns and to build successful client relationships by continuously supporting them. Moreover, analytics solutions allow organizations to monitor performances with lower efforts, in more effective and rapid ways. Enterprises focus on their core capabilities to maintain their competitive advantage over other players, but they need better insights to drive more revenue and value for their shareholders.

Maturity

The move from more traditional styles of research to finding the right use of technology, in order to use insights to automate actions in businesses is still maturing. The technical innovation of voice analytics is evolving, and there are substantial differences between current and older applications. From the market's beginning, these solutions held great potential for companies, but their contributions will increase as their capabilities are incorporated into every process of the organization. Analytics are necessary to understand many aspects of customers and are a great way to get started in building a VoC program. They have different denominations, basing on their specific purposes, and they include customer journey analytics, emotion detection, customer engagement center (CEC) interaction analytics, voice analytics, interaction analytics, and analytics for customer intelligence. Gartner forecasts that by 2020, 50% of analytical queries either will be generated via search, natural language processing or voice, or will be automatically generated. By 2021, natural language processing and conversational analytics will boost analytics' adoption from 35% of employees to over 50%, including new classes of users, particularly front-office workers. From a recent study on vendor market for Voice of the Customer (VoC) products and services (State of Voice of the Customer Programs, 2017), nearly three-quarters of large companies rate their voice of the customer (VoC) programs as being successful. However, companies aren't close to reach the "plateau of productivity" of this technology. Indeed, only 14% of companies have reached the two highest levels, out of five, of Temkin Group's VoC Maturity Model:

1. Novices: Early stages of development
2. Collectors: The VOC team spends most of its time focused on discussions
3. Analyzers: The VOC team is doing data crunching, but is not well integrated with rest of the company
4. Collaborators: There is now a strong relationship between the VOC team and rest of the organization
5. Transformers: Customer insight data is now integrated into processes throughout the company

Moreover, in last reports, Gartner positions this kind of business applications on the down-sloping part of its hype cycle curve. For these considerations, we can consider voice analytics technologies' maturity in the "Trough of Disillusionment" phase.

4.5.5. Biometrics Authentication

Potentiality

Global Voice Biometrics Market size is expected to reach $2.7 billion by 2024, rising at a market growth of 22.7% CAGR during the forecast period.

Based on Application, the market is segmented into Access Control & Authentication, Fraud Detection & Prevention, and Forensic Voice Analysis & Criminal Investigation. Rapid advancements in technologies used for authentication and payment purposes have helped the biometrics industry in expanding its client base across the world. The requirement for easy, faster, and convenient user authentication is projected to impact voice biometrics software. Moreover, since the number of online transactions and fraudulent activities is growing, the financial services sector is expected to boost the overall market. This is supposed to create a significant demand for voice biometric software over the coming years. According to Marketing reports, in Global Voice Biometrics Market, the Access Control & Authentication market dominated the Global Voice Biometrics Market by Application 2017. The Fraud Detection & Prevention market is expected to witness a CAGR of 23.4% during (2018 - 2024). Additionally, The Forensic Voice Analysis & Criminal Investigation market is expected to witness the highest CAGR of 24.2% during (2018 - 2024).

Maturity

Mobile banking is a real growth area and banks are moving away from standard authentication methods, including passwords and hardware tokens, to more agile and convenient ones. Biometrics is one technology that is seeing rapid growth as a password and token replacement. Although fingerprint modes are commonly integrated into mobile apps, a voiceprint is emerging as the modes of choice for identity verification, authentication and fraud management in banks and FinTech suppliers. Banks are increasingly adopting biometric technology and deploying these systems across a wide variety of banking channels, from the traditional (ATMs and branches) to the new banking channels of mobile and IoT. Biometric technologies help them to better identify new customers, securely authenticate their existing ones, verify identity for high-value transactions and combat fraud. Indeed, Goode Intelligence forecasts that by 2020 biometrics will be in use by 1.9 billion bank customers around the world. Biometrics technologies seem to be more mature compared to voice analytics, and according to Gartner’s Hype Cycle for Risk Management report, user Authentication Technologies through Biometrics are climbing the “Slope of Enlightenment”.

4.5.6. Robotics

Potentiality

Goldstein Research analyst forecast the interactive robots market size is set to reach USD 11,890 million by 2024 from the USD 152.38 million in 2016, growing at a CAGR of 72.4% over the forecast years.

Advancements in technology, such as cloud, IoT, machine learning, and artificial intelligence, are making robotics more attractive because of their capabilities Manufacturers are trying to incorporate such technologies into robots to increase performances and improved human-robot communication to transfer the real-time information. The rising demand for robots with advanced capabilities to assist humans in everyday lives is increasing their penetration in various sectors, such as retail, enabling interactive experiences and enhancing the engagement with in-store customers. The main issue of robotic solutions is the requirement of high capital investments needed to adopt them in business. Once the lack of awareness about the advantages of robots in the emerging economies and the high cost of these machines will be reduced, the adoption rates of service robotics will grow exponentially in many industries. Indeed, the emergence of low-price robotic solutions will be a key driver boosting the market growth, and robotics manufacturers are overcoming these constraints to expand the industry.

Maturity

Smart robots are gaining great hype in the marketplace, as providers execute on their plans to expand their offerings and deliver solutions across the wider spectrum of industry-specific use cases and enterprise sizes. The market is becoming more dynamic, opening to new technologies. Advancements in natural language processing (NLP) are increasing dramatically the sophisticated skills with which robots can sense, plan, act and learn. Enterprise NLP usage is increasing as capabilities improve, along with new use cases based on conversational agents and automatic voice recognition. Although existing syntactic and semantic based methods are increasingly augmented and displaced with deep neural networks (DNNs) approaches, dialogue capabilities appear still weak. DNNs are experimental and fragile, and understanding inferences, context and synthesis pose significant challenges. Capabilities of emotion recognition and emotion computing have to be deep studied. Emotion computing refers to establishing a computer system for perceiving and recognizing human emotions based on the information that arouse and influence human emotions. By combining leading technologies like human-robot interface technology, artificial intelligence reasoning and cloud computing, the emotional recognition and interaction technologies will be applied to even wider fields and play a more important role in the future. Thus, making them be able to express, recognize and understand emotions as well as imitate, extend and expand human′s emotion. As a result, a harmonious human-robot environment could be built, and the robot would gain higher intelligence [45]. As a consequence, it is reasonable that Gartner positioned Smart Robots technologies climbing the “Peak of Inflated Expectations”.

4.5.7. Virtual Assistants

Potentiality

The global voice assistant application market size is expected to grow from USD 1.3 billion in 2019 to USD 5.2 billion by 2024, at a Compound Annual Growth Rate (CAGR) of 31.9% during the forecast period.

According to Market Research, the adoption rate of voice assistant application solutions is expected to grow thanks to the advancement in Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) ecosystems. Improving customer experience and purchase process are the key factors driving the voice assistant application market. The voice assistant applications enhance online communication, enable sophisticated interaction with intuitive response times, improve customer retention, and understand people’s natural language with voice recognition. The users are guided vocally by the assistant with the help of a human-like interaction platform. Based on Application, the market is segmented into Mobile Application, Web application and Devices. Virtual assistants are available across various smartphones and tablets. Voice assistant applications deployed over the website can be useful for precise navigation, instant answering to customer queries and customer insights. Smart speakers are among virtual assistant embedded in devices (not mobile). The growing trend of voice assistants' popularity and increasing businesses' digital innovation is contributing notably to this market in the upcoming years. From a report of 'MarketsandMarkets' archive: "The cloud delivery model offers a range of benefits to enterprises, such as scalability, flexibility, faster route to market, and lower cost. Hence, small enterprises are expected to adopt cloud-based voice assistant solutions heavily during the forecast period".

Maturity

Conversational user interfaces have seen explosive growth in interest with chatbots, messaging platforms and virtual assistants. The emerging pattern of voice assistants acting as a guide or concierge in front of these conversational interfaces is likely to gain a lot of traction. Most CUI implementations are still primitive and thus are not able to respond to complex queries. Advancements in capabilities are largely coming from improvements in natural-language understanding, speech and voice recognition, which will bring CUIs closer to the promises. Businesses that haven't begun deploying AI to interact with customers and employees should start now, because customers and employees are increasingly expecting conversational interfaces to be available for delivering services. Still, only 4% of enterprises have a conversational interface solution in production, while a further 38% is experimenting or planning with the technology, according to Gartner CIO Survey 2018. Additional capabilities around context and intent handling, are still quiet immature and probably will be improved within next year. Today, media and press are pushing this technology on the top of the expectations, and this sparks a huge amount of interest and speculation. "The effects of speech recognition can be seen on a daily basis. Consumers and workers increasingly interact with applications without touching a keyboard," said Matthew Cain, VP and Distinguished Analyst of Gartner.

Conversational user interfaces have proliferated due to the adoption of chatbots and virtual personal assistants (VPAs) by businesses, and consumer adoption of devices with speech interactions including smartphones. Virtual Assistants' hype may show a slight drop when some initial VA platforms disappear or are adapted to new achievements, pressured by technological improvements or the entry of new players. The technology hopes to achieve the "plateau of productivity" in 2 to 5 years. By the way, at the moment we can assume VA's maturity in the "Peak of Inflated Expectations" phase. We also report some events of interest that happened last few years in AI-enabled voice assistant industry:

- In March 2018, Google added the text-to-speech synthesis technology to its Cloud Platform. The enhancement would enable developers to select from 32 different voices in 12 languages.
- In March 2018, AWS enhanced its service, Amazon Polly, wherein the company added a new Speech Synthesis Markup Language (SSML) Breath feature. The feature is designed for developers so that they can add appropriate pauses in speeches to sound more natural.
- In December 2017, Baidu partnered with Huawei, wherein both the companies would together develop the open AI mobile ecosystem. According to this partnership, both companies would work on enhancing image and voice recognition on smart devices.

5. Conclusions

In conclusion, we would return to our initial “research objectives” that we have targeted to lead the research in order to answer them in a timely manner.

1. *Which is the current degree of scientific knowledge about the study of voice?*

The main findings raised from the literature evidence that it is possible to recognize the emotional state of a speaker studying his voice. The emotions that can be effectively detected in a speech are principally the Ekman’s basic ones. More complex sentiment analysis lead to higher computational costs due to the dimensionality limitations that characterized current technologies. Most of the experiments made to test the accuracy of algorithms are carried out in labs or sites protected from external sound disturbs. As a consequence, the performance of the recognition systems visibly decreases in high dimensional problems and in noisy environments.

Deepening: More considerations that are present in the literature related to the voice analysis help us in understanding the effort made by scientific research to study the paraverbal features of the voice and the relationship with the sentiment analysis. From the empirical results, we could identify the most important characteristics of the voice in order to detect the variations of the speaker’s emotional state. Current advancements of the recognition systems for paraverbal features allow us to pinpoint their performance and weaknesses. Even though the most common classification methods reach acceptable levels of accuracy rate, there is still room of improvement in the effectiveness of recognition systems.

2. *Which are the most remarkable applications of voice analysis and emotion recognition in business?*

Four meaningful categories of business applications have been indicated. Each category includes multiple successful use cases of enterprises that apply such innovation in their business to improve the overall efficiency and effectiveness in their service and products. Each group of technology-driven business applications can be adopted in different industries as an advantage against competition. The sectors that look more likely to embrace this innovation are: Retail & Marketing, Banking and Insurance, Healthcare, Hospitality and others.

Deepening: As concerns the categories of voice analysis business applications, we have selected the most valuable markets in which such technology are taking place and we have classified the applications adopting a technology-driven evaluation. We approached the analysis of each market benchmarking the innovative solutions already offered and implemented in business. We highlighted the wide range of applicability in different fields and the benefits this technology could bring to the enterprises through the analysis of selected case studies and companies' reports.

3. *Which could be the future developments of such technologies and which opportunities can bring?*

The materials collected in our research demonstrate that most innovative companies are investing in voice technologies due to the opportunities they can carry to organizations. The evolution of the interactions between customer and brands is reshaping the way organizations deliver their service. Indeed, voice is gaining ground as convenient and avant-garde means through which search on the web, perform transactions, interact with products and services. New markets such as voice analytics, voice biometrics, vocal assistants, service and social robots have raised on the wave of voice's growing value. The researches which have been made show that these are fast-growing markets since during the next five years are expected to increase their value roughly from 20% up to 70% Moreover, thanks to the advancements in the research and technology performance, there is every likelihood that companies would gain greater benefits from integrating voice technologies in their business.

Deepening: The investigation of the business applications available on the market permits us to identify which factors are boosting the adoption of voice analysis in different industries. For each application, we evaluated its potential growth in relation to its maturity on a bi-dimensional graph in order to emphasize the opportunities that the innovation can create. The implementation of emotion recognition technologies through voice showed a positive effect on economic performances too, evidencing an increase in revenues in different case studies. Although the findings of our research suggest that investing in voice technologies can turn in a competitive advantage for companies in separate sectors, our analysis does not demonstrate the effectiveness of such innovation in detecting emotions compared to other biometrics methods such as facial recognition.

In the following chapters, we would provide some critical remarks about the research objectives and the bidimensional model proposed in the previous section. We would also underline the potentialities of the illustrated voice technologies, our research's limitations and the contribution that our work could provide to the scientific knowledge.

5.1. Graphical Evaluation: Discussions

The considerations around the previously mentioned model aim to critically analyzed the graphical representation of the business applications which have been investigated. From a first look at the graphical comparison, the technologies draw a downward sloping curve. We expected to find these results because of the dimensions taken into account. Generally, the greater the maturity the less the growth potential because the advancements in technology's maturity lead to explore different potentialities of the innovation, increase successful business cases and consumer adoption. Hence, the potential market growth logically decreases stage by stage. During the "technology trigger" stage, the innovation is still young, and a lot of work has to be done in order to comprehend its breadth of capabilities and real potentiality. Technologies may present very different evolutions in terms of rapidity and market penetration. However, it is possible to assume that the greatest expected CAGR of the market is reached in the "peak of inflated expectation" stage since the innovation start to realize its real potential.

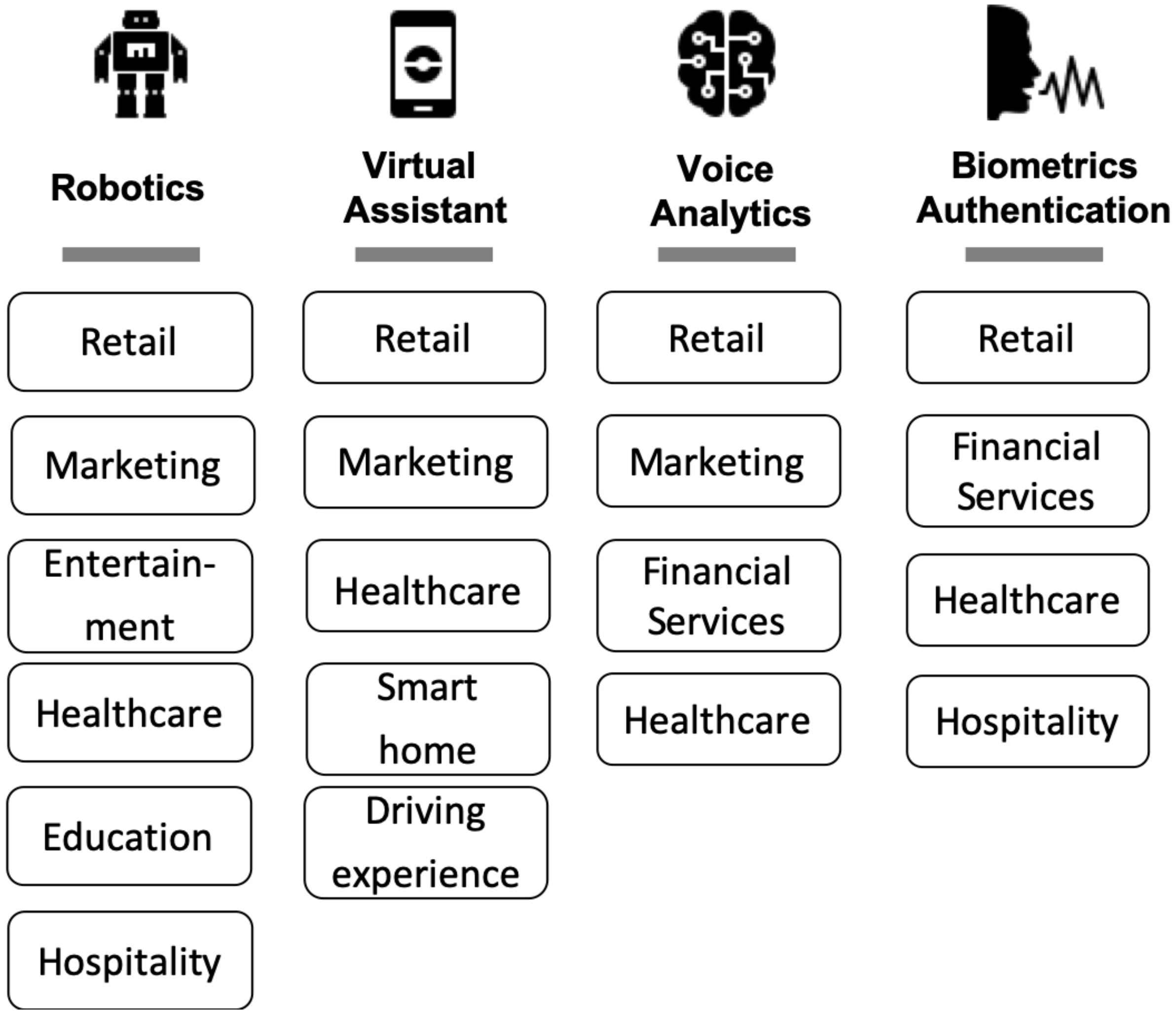
Robotic applications are living this phase thus manifest the greatest potential growth in the next five years compared with the others, especially with Voice Analytics which are more than three times lower as potentiality. The market of robots has such great potential also thanks to the increasing number of industries in which they can be applied, such as retail, entertainment, education, healthcare marketing purposes and hospitality. Companies are now starting to pilot robots in their business in order to fully evaluate the competitive advantage they can bring. The more the technology is novel, so standing in the initial phases of its maturity, the greater the number of investments that are necessary to reach a mature level of development. In fact, a great bulk of funding is coming toward the robotic sector. SoftBank Group Corp has come to an agreement with different big investors, including Microsoft and Apple, of around $108 billion for a second Vision Fund aimed at investing in robots. The Japanese conglomerate itself plans to invest $38 billion in the fund.

Voice analytics instead present the lowest potential growth compared with the other technologies. Their expectations are declining because most innovative companies already have been applying them to their business and the "hype" around these new solutions is slowing down. In fact, they are positioned in the "trough of disillusionment" stage. According to a study by Adobe Analytics, 91% of 401 business decision-makers surveyed said they already are making significant investments in voice, and 94% said they plan to increase their investments in the next year. 66% of the interviewed strongly agree that voice can help drive conversion and increase revenue. At the same time, 71% said they strongly agree that it can help improve the user experience, increase consumer engagement (65%), and increase customer loyalty (64%). Naturally, to understand the customers better, data from service calls should be able to help organizations add a new dimension to enterprise analytics. Using the additional component of voice, industries like Retail, Financial Services, Healthcare, and others can add depth to their analytics. According to Marketsandmarkets report, among industry verticals, Financial Services are expected to continue holding the largest market share, but Retail is expected to grow at the highest CAGR during the forecast period. The regional analysis of global Voice Analytics is considered for the key regions such as Asia Pacific (APAC), North America, Europe, Latin America and Rest of the World. North America is the leading region across the world. Asia-Pacific is expected to grow at the highest CAGR among other regions during the forecast period.

Voice Biometrics is positioned as medium-low potential growth. Since they are crossing the "slope of enlightenment" phase, the graphical representation shows a higher potentiality than Voice Analytics technologies. The reason for this gap is that Voice Biometrics have reached a higher awareness of risks, benefits as well as the application of the technology. This is often the result of hard work as well as focused experimentation by some organizations. Voice biometrics has a number of distinct advantages as a method for user authentication. It comes very naturally for people to produce for mobile authentication and can follow on from the success of fingerprint biometrics being easily integrated into smartphones. Technology's providers can launch other variants of the product at this stage as it is the right time to capitalize on the building understanding. In fact, voice is also well suited as a biometric authentication solution across a wide range of IoT devices, including tablets, wearables, PCs, gaming systems, smart TVs, even fixed line telephones and automobiles. The efforts made to develop successful voice recognition systems, allow the technology to reach great level of accuracy. VoiceVault, one of the major players in the market, claims their technology that can be optimized to deliver a false accept rate of 0.01% with a false reject rate of less than 5% for high-security applications. It can also be optimized to deliver a false reject rate of 0.05% with a false accept rate of less than 1% for cost reduction applications. In Nuance's November 2012 press release, it claims to be delivering a 99.6% successful authentication rate while surpassing industry security requirements [46].

Virtual Assistants are characterized by a medium potentiality since they exhibit the half market growth rate compared to robotics' one, but still higher than both voice analytics and voice biometrics technologies. The great potentiality of AI-enabled technologies derives from the variety of applications that can be offered and the easiness to embed them into smart devices. These virtual assistants often need to be connected to other systems such as payment gateways to facilitate transactions however they are a powerful way to streamline purchases and capture decision points about customers. So far, there are two types of devices that support voice assistants: smartphones and smart speakers. The number and the usage of smartphones is indisputable, while smart speakers' market has to reach its potential yet. The technology is demonstrating its value in many sectors of application. In Smart Home, voice-controlled devices prove to be easier and more comfortable to use. In the Healthcare industry, patients can use voice to request assistance and access information including appointment schedules or directions, with answers and information sent to their phone. In E-commerce there are several examples of business applications such as ordering a taxi, booking a restaurant, flights and hotels, as well as buying cinema tickets, can easily be done through a voice command. Thanks to the great hype surrounding virtual assistants, we can see from the graph them standing on the "peak of inflated expectations" like robotic applications. However, according to Gartner's report, we can notice them moving towards the "Trough of Disillusionment," with a plateau expected to be reached in only 2-5 years. The technology is still in the beginning steps of its maturity, many investments are necessary to increase systems' performance. In order to work efficiently, voice assistants have to be smart to understand what people say and interpret it the right way.

Different business applications have different impacts in different industries. In **fig.18** we outline the sectors in which each technology is applied in the current state-of-the-art. The variety of industry in which a business application can penetrate may impact on its potentiality since the demand for those solution increases with the number of industries served. In fact, Robotics shows the highest number of sectors paired with the highest potentiality. As we evidenced from the literature review, different use cases of robotics applications can be found in marketing & retail sector, where robots are applied to enhance the purchasing experience. In the hospitality sector, service robots have the task of welcoming clients during events and engage them in a conversation. Instead, social robots are gaining traction in healthcare thanks to their therapeutic applications and in the entertainment playground as social companions. In line with the considerations, Virtual Assistants can potentially penetrate more industries than Voice Analytics and Biometrics Authentication. VAs are changing the way consumers will purchase goods, driving the purchasing experience toward new concepts. They will entirely manage the smart home environments, they can be easily integrated into many devices and be a support in our daily life. The most valuable business applications of voice analytics are first for retail & marketing purposes in order to study consumers' behaviour, secondly for every contact centre monitoring and optimization. Finally, Biometrics Authentication through voice presents several advantages principally applied to financial service to ensure security, preventing frauds, and fastening the online services. They can increase efficiency and streamline the processes even when practised in retail and hospitality services to recognize clients, as well in healthcare to diagnose patients.



Robotics	Virtual Assistant	Voice Analytics	Biometrics Authentication
Retail	Retail	Retail	Retail
Marketing	Marketing	Marketing	Financial Services
Entertain-ment	Healthcare	Financial Services	Healthcare
Healthcare	Smart home	Healthcare	Hospitality
Education	Driving experience		
Hospitality			

Fig.18 *Technologies and related industries*

5.2. Research Limitations and Future Developments

The findings of this study have to be seen in the light of some limitations. There are two major issues in this study that could be addressed in future research. First, limitations can exist due to constraints on research design or methodology. Second, the novelty of voice analysis researches and relative business applications reveals some difficulties in terms of lack of available and reliable data.

The primary limitation is due to the fact that the study focuses on the state-of-the-art exploration of voice analysis topic, so principle sources of information consist in the systematic review of scientific articles. Empirical experiments could be integrated into the work to assess the validity of previous studies and to improve the consistency of current findings. Presenting alternative methodologies is a way to fill in the gaps of this academic research and to overcome limitations in future studies. Moreover, the majority of studies that have been selected from literature and examined in this work, underlines some specific issues concerning voice recognition experiments carried out in laboratories. One important issue is related to the speech databases available for researches. Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. Principal reasons can be attributed to the fact that the available recorded utterances are few and they have low quality audio. Since commonly only one emotional speech database is investigated in each study, it is likely that some of the conclusions established in that studies cannot be generalized to other databases. Inconsistencies in the different types of audio and their quality is an important challenge in speech processing. Usually, problems are related to channel's mismatch in the handset or in the recording apparatus. They also may arise because of the speaker related conditions like illness and stress. More complications transpire in noisy conditions, when experiments are performed outside lab environment. Other discrepancies come out in overlapping speech cases, when multiple speakers talk on a single microphone and each speaker is producing similar level of audio at the same time. It is difficult to handle these because single audio is very hard to recognize. Also, handling of whispered speech brings several problems, since it is very hard to collect as it is not available under natural speech scenarios. To address this problem, more cooperation across research institutes in developing emotional speech databases is necessary. As far as the lack of available or reliable data, the issue concerns more the second section of the study since businesses are still exploring the great potential of voice-enabled systems. Therefore, most of the considerations are made based on data collected from secondary sources. The scarcity of reliable data is an obstacle that limits the scope of the analysis and makes difficult to identify significant relationships from the data. In order to support the business application discussion with meaningful information, it may be worthwhile conducting interviews among companies that are implementing commercial applications discussed in previous chapters. It would be interesting to interview at least one business figure for each application industry in order to gather consistent info about capabilities, opportunities, recent developments of current voice-driven technologies. This work provides an overview of most potential voice recognition business applications that can be used to investigate different innovative solutions among distinct sectors. Being aware of the potentiality of emergent technology in relation to its maturity level in the global market could be advantageous for companies to evaluate strategic decisions such as entry-timing choice of adopting technology innovations. The approach proposed is intended to be an exploratory study in order to lay the groundwork for a more complete research study in the future.

6. References

[1] K Vipin, Vipin Sharma. 2011. Conference: “Decoding Non-Verbal Communication” in Second International Seminar on Teaching and learning of ESL in Technical Education.

[2] Mehrabian, A. & Wiener, M. 1967. Decoding of Inconsistent Communications. Journal of Personality and Social Psychology. 6, 108-114.

[3] D.Lapakko. 2007. Communication is 93% Nonverbal: An Urban Legend Proliferates. Communication and Theater Association of Minnesota Journal, 34, 7-19.

[4] Iris Bakker, Theo van der Voordt, Peter Vink, Jan de Boon. 2014. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. Current Psychology 33. 405-421

[5] Oana Rusu, Maria Chiriţă. 2017. Verbal, non-verbal and paraverbal skills in the patient-kinetotherapist relationship. Timisoara Physical Education and Rehabilitation Journal 10(19).

[6] Moataz El Ayadi,, Mohamed S. Kamel, Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44 572–587.

[7] Donn Morrison, Ruili Wang, Liyanage C. De Silva. 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 49, 98–112.

[8] Alain de Cheveigne, Hideki Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America 111(4). 1917-30.

[9] Dipti Patil, Shamla Mantri, Ria Agrawal, Shraddha Bhattad, Ankit Padiya, Rakshit Rathi. 2014. A Survey: Pre-processing and Feature Extraction Techniques for Depression Analysis Using Speech Signal. International Journal of Computer Science Trends and Technology (IJCST). Volume 2, Issue 2.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor. 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine. Volume 18 , Issue 1.

[11] Marie Tahon, Gilles Degottex, Laurence Devillers. 2012. Usual voice quality features and glottal features for emotional valence detection. International Conference on Speech Prosody.

[12] Mireia Farrús, Javier Hernando, Pascual Ejarque. 2007. Jitter and shimmer measurements for speaker recognition. 8th Annual Conference of the International Speech Communication Association.

[13] B. Yang, M. Lugger. 2010. Emotion recognition from speech signals using new harmony features. Signal Processing 90, 1415–1423

[14] Paul Ekman. 1992. An argument for basic emotions. Cognition & Emotion. Volume 6, Issue 3-4. 169-200.

[15] P. Gangamohan, Sudarsana Reddy Kadiri and B. Yegnanarayana. 2016. Analysis of Emotional Speech—A Review. Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions, Chapter: 11.

[16] Zhongqiang Huang, Lei Chen, Mary Harper. 2006. An Open Source Prosodic Feature Extraction Tool. 5th International Conference on Language Resources and Evaluation.

[17] Florian Eyben, Martin Wöllmer, Björn Schuller. 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. 9th ACM International Conference on Multimedia.

[18] Florian Eyben, Martin Wo ̈llmer, and Bjo ̈rn Schuller. 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops.

[19] Zhen-Tao Liua,b, Min Wua,b, Wei-Hua Caoa,b,∗, Jun-Wei Maoa,b, Jian-Ping Xua,b, Guan-Zheng Tan. 2018. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273, 271–280.

[20] Tin Lay Nwe, Say Wei Foo , Liyanage C. De Silva. 2003. Speech emotion recognition using hidden Markov models. Speech Communication 41, 603–623.

[21] Chandralika Chakraborty, P.H. Talukdar. 2016. Issues and Limitations of HMM in Speech Processing: A Survey. International Journal of Computer Applications 141, 7.

[22] Veronica Piccialli, Marco Sciandrone. 2018. Nonlinear Optimization and Support Vector Machines. 4OR 16, 2, 111–149.

[23] Laura Auria, Rouslan A. Moro. 2008. Support Vector Machines (SVM) as a Technique for Solvency Analysis. SSRN Electronic Journal. Volume 1, Issue 1.

[24] Alexey Tsymbal , Seppo Puuronen , David W. Patterson. 2003. Ensemble Feature Selection with the Simple Bayesian Classification. Information Fusion Volume 4, Issue 2, Pages 87-100.

[25] S. Lugović, I. Dunđer, M. Horvat. 2016. Techniques and Applications of Emotion Recognition in Speech. 39th International Convention on Information and Communication Technology.

[26] Richard P. Bagozzi, Mahesh Gopinath, Prashanth U. Nyer. 1999. The Role of Emotions in Marketing. Journal of the Academy of Marketing Science. Volume 27, Issue 2. 184-206.

[27] James C. Warda, John W. Barnes. 2001. Control and affect: the influence of feeling in control of the retail environment on affect, involvement, attitude, and behavior. Journal of Business Research 54, 139–144.

[28] Andrea Groeppel-Klein. 2005. Arousal and consumer in-store behaviour. Brain Research Bulletin 67, 428–437.

[29] Hagai Aronowitz. 2012. Voice Biometrics for User Authentication. Speech Processing Conference.

[30] Adewole Kayode S., Abdulsalam Sulaiman Olaniyi, Jimoh R. G. 2011. Application of voice biometrics as an ecological and inexpensive method of authentication. International Journal of Science and Advanced Technology.

[31] Mirko Marras, Pedro A. Marìn-Reyes, Javier Lorenzo-Navarro, Modesto Castrillon-Santana, Gianni Fenu. 2019. AveroBot- An audio-visual dataset for people re-identification and verification in human-robot interaction. Conference paper.

[32] M. Gofman, N. Sandico, S. Mitra, E. Suo, S. Muhi, and T. Vu. 2018. Multimodal biometrics via discriminant correlation analysis on mobile devices. Int'l Conf. Security and Management (SAM). 174-181.

[33] Saeid Safavi, Hock Gan, Iosif Mporas, Reza Sotudeh. 2016. Fraud Detection in Voice-based Identity Authentication Applications and Services. The IEEE International Conference on Data Mining series (ICDM).

[34] Cynthia Breazeal. 2004. Function Meets Style: Insights from emotion theory applied to HRI. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). Volume 34, Issue 2.

[35] C. Breazeal. 2003. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies, Volume 59. 119–155.

[36] Francesca Bertacchini, Eleonora Bilotta, Pietro Pantano. 2017. Shopping with a robotic companion. Computers in Human Behavior 77, 382-395.

[37] Clarice J. Wong, Yong Ling Tay, Lincoln W.C. Lew, Hui Fang Koh, Yijing Xiong, Yan Wu. 2018. Advbot: Towards Understanding Human Preference in a Human-Robot Interaction Scenario. 15th International Conference on Control, Automation, Robotics and Vision.

[38] Emi Moriuchi. 2019. Okay, Google!: An empirical study on voice assistants on consumer engagement and loyalty. Psychology and Marketing.

[39] Anthony Pym. 2012. Website Localization.

[40] Nimdzi Insights. 2019. Artificial intelligence, localization, winners, losers, heroes, spectators, and you.

[41] Norhaslinda Kamaruddin, Abdul Wahab.. 2011. Heterogeneous Driver Behavior State Recognition Using Speech Signal. Recent Researches in Power Systems and Systems Science.

[42] Matthew B. Hoy. 2018. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, Medical Reference Services Quarterly 37, 81-88.

[43] Miles A. Zachary, Peter T. Gianiodis, G. Tyge Payne, Gideon D. Markman. 2015. Entry Timing: Enduring Lessons and Future Directions. Journal of Management Vol. 41 No. 5, 1388–1415.

[44] Gustavo López, Luis Quesada, Luis A. Guerrero. 2018. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. International Conference on Applied Human Factors and Ergonomics.

[45] Tian-Miao Wang, Yong Tao, Hui Liu. 2018. Current Researches and Future Development Trend of Intelligent Robot: A Review. International Journal of Automation and Computing. Volume 15, Issue 5. 525-546.

[46] John Gibbons, Anna Lo, Aditya Chohan, A. Taleb Damaree, Jonathan Leet, and Vinnie Monaco. 2014. Voiceprint Biometric Authentication System. Proceedings of Student-Faculty Research Day, CSIS, Pace University.