Emotion-Aware Voice Interfaces Based on Speech Signal Processing

Yong Ma

München den 14. Oktober, 2022

Emotion-Aware Voice Interfaces Based on Speech Signal Processing

Yong Ma

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Yong Ma

München, den 14. Oktober, 2022

Erstgutachter: Prof. Dr. Andreas Butz
Zweitgutachter: Prof. Dr. Yaxi Chen
Tag der mündlichen Prüfung: 28. November, 2022

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, 17.10.2022

Yong Ma

Name (+ Unterschrift)

Abstract

Voice interfaces (VIs) will become increasingly widespread in current daily lives as AI techniques progress. VIs can be incorporated into smart devices like smartphones, as well as integrated into autos, home automation systems, computer operating systems, and home appliances, among other things. Current speech interfaces, however, are unaware of users' emotional states and hence cannot support real communication. To overcome these limitations, it is necessary to implement emotional awareness in future VIs.

This thesis focuses on how speech signal processing (SSP) and speech emotion recognition (SER) can enable VIs to gain emotional awareness. Following an explanation of what emotion is and how neural networks are implemented, this thesis presents the results of several user studies and surveys.

Emotions are complicated, and they are typically characterized using category and dimensional models. They can be expressed verbally or nonverbally. Although existing voice interfaces are unaware of users' emotional states and cannot support natural conversations, it is possible to perceive users' emotions by the speech based on SSP in future VIs.

One section of this thesis, based on SSP, investigates mental restorative effects on humans and their measures from speech signals. SSP is less intrusive and more accessible than traditional measures such as attention scales or response tests, and it can provide a reliable assessment for attention and mental restoration. SSP can be implemented into future VIs and utilized in future HCI user research.

The thesis then moves on to present a novel attention neural network based on sparse correlation features. The detection accuracy of emotions in the continuous speech was demonstrated in a user study utilizing recordings from a real classroom. In this section, a promising result will be shown.

In SER research, it is unknown if existing emotion detection methods detect acted emotions or the genuine emotion of the speaker. Another section of this thesis is concerned with humans' ability to act on their emotions. In a user study, participants were instructed to imitate five fundamental emotions. The results revealed that they struggled with this task; nevertheless, certain emotions were easier to replicate than others.

A further study concern is how VIs should respond to users' emotions if SER techniques are implemented in VIs and can recognize users' emotions. The thesis includes research on ways for dealing with the emotions of users. In a user study, users were instructed to make sad, angry, and terrified VI avatars happy and were asked if they would like to be treated the same way if the situation were reversed. According to the results, the majority of participants tended to respond to these unpleasant emotions with neutral emotion, but there is a difference among genders in emotion selection.

For a human-centered design approach, it is important to understand what the users' preferences for future VIs are. In three distinct cultures, a questionnaire-based survey on users' attitudes and preferences for emotion-aware VIs was conducted. It was discovered that there are almost no gender differences. Cluster analysis found that there are three fundamental user types that exist in all cultures: Enthusiasts, Pragmatists, and Sceptics. As a result, future VI development should consider diverse sorts of consumers.

In conclusion, future VIs systems should be designed for various sorts of users as well as be able to detect the users' disguised or actual emotions using SER and SSP technologies. Furthermore, many other applications, such as restorative effects assessments, can be included in the VIs system.

Zusammenfassung

Mit den Fortschritten der KI-Techniken werden Sprachschnittstellen (VIs) in unserem täglichen Leben immer weiter verbreitet sein. VIs können in intelligente Geräte wie Smartphones integriert werden, aber auch in Autos, Hausautomatisierungssysteme, Computerbetriebssysteme und Haushaltsgeräte. Die derzeitigen Sprachschnittstellen kennen jedoch nicht die emotionalen Zustände der Benutzer und können daher keine echte Kommunikation unterstützen. Um diese Einschränkungen zu überwinden, ist es notwendig, Emotionsbewusstsein in zukünftige VIs zu implementieren.

Diese Arbeit befasst sich mit der Frage, wie Sprachsignalverarbeitung (SSP) und Sprach-Emotionserkennung (SER) es VIs ermöglichen können, Emotionen zu erkennen. Nach einer Erklärung, was Emotionen sind und wie neuronale Netze implementiert werden, werden in dieser Arbeit die Ergebnisse mehrerer Nutzerstudien und Umfragen vorgestellt.

Emotionen sind kompliziert und werden in der Regel anhand von Kategorien- und Dimensionsmodellen beschrieben. Sie können verbal oder nonverbal ausgedrückt werden. Obwohl bestehende Sprachschnittstellen die emotionalen Zustände der Benutzer nicht kennen und keine natürlichen Unterhaltungen unterstützen können, ist es möglich, die Emotionen der Benutzer durch die auf SSP basierende Sprache in zukünftigen VIs zu erkennen.

Ein Teil dieser Arbeit, der auf SSP basiert, untersucht mentale Wiederherstellungseffekte beim Menschen und deren Messung anhand von Sprachsignalen. SSP ist weniger aufdringlich und zugänglicher als herkömmliche Messungen wie Aufmerksamkeitsskalen oder Reaktionstests und kann eine zuverlässige Bewertung der Aufmerksamkeit und der mentalen Erholung liefern. SSP kann in zukünftige VIs implementiert und in der zukünftigen HCI-Benutzerforschung eingesetzt werden.

Anschließend wird ein neuartiges neuronales Aufmerksamkeitsnetz vorgestellt, das auf spärlichen Korrelationsmerkmalen basiert. Die Erkennungsgenauigkeit von Emotionen in kontinuierlicher Sprache wurde in einer Nutzerstudie anhand von Aufnahmen aus einem echten Klassenzimmer demonstriert. In diesem Abschnitt wird ein vielversprechendes Ergebnis gezeigt.

In der SER-Forschung ist nicht bekannt, ob die vorhandenen Methoden zur Erkennung von Emotionen gespielte Emotionen oder die echten Emotionen des Sprechers erkennen. Ein weiterer Teil dieser Arbeit befasst sich mit der Fähigkeit des Menschen, seine Emotionen zu handeln. In einer Benutzerstudie wurden die Teilnehmer aufgefordert, fünf grundlegende Emotionen zu imitieren. Die Ergebnisse zeigten, dass sie sich mit dieser Aufgabe schwer taten; dennoch waren bestimmte Emotionen leichter zu imitieren als andere.

Ein weiteres Thema der Studie ist die Frage, wie VIs auf die Emotionen der Benutzer reagieren sollten, wenn SER-Techniken in VIs implementiert sind und die Emotionen der Benutzer erkennen können. Im Rahmen dieser Arbeit wurde untersucht, wie man mit den Emotionen der Benutzer umgehen kann. In einer Benutzerstudie wurden Benutzer angewiesen, traurige, wütende und verängstigte VI-Avatare glücklich zu machen und wurden gefragt, ob sie in einer umgekehrten Situation genauso behandelt werden möchten. Die Ergebnisse zeigen, dass die Mehrheit der Teilnehmer auf diese unangenehmen Emotionen eher mit einer neutralen Emotion reagierte, aber es gibt einen Unterschied zwischen den Geschlechtern bei der Emotionswahl.

Für einen auf den Menschen ausgerichteten Designansatz ist es wichtig zu verstehen, welche Präferenzen die Benutzer für zukünftige VIs haben. In drei verschiedenen Kulturen wurde eine fragebogenbasierte Umfrage zu den Einstellungen und Präferenzen der Benutzer für emotionsbewusste VIs durchgeführt. Es wurde festgestellt, dass es fast keine geschlechtsspezifischen Unterschiede gibt. Eine Clusteranalyse ergab, dass es drei grundlegende Benutzertypen gibt, die in allen Kulturen existieren: Enthusiasten, Pragmatiker und Skeptiker. Folglich sollte die zukünftige VI-Entwicklung verschiedene Arten von Verbrauchern berücksichtigen.

Zusammenfassend lässt sich sagen, dass künftige VIs-Systeme für verschiedene Arten von Benutzern konzipiert werden sollten und in der Lage sein sollten, die verdeckten oder tatsächlichen Emotionen der Benutzer mithilfe von SER- und SSP-Technologien zu erkennen. Darüber hinaus können viele andere Anwendungen, wie z. B. die Bewertung von Wiederherstellungseffekten, in das VIs-System aufgenommen werden.

Contents

Chapter 1

Introduction

Over recent decades, voice interfaces (VIs) play an increasing role in our daily lives. VIs can build the communication bridge between humans and machines. VIs exist in both mobile and stationary devices and allow users to easily interact with them through voice or speech commands. More specifically, with the advancement of artificial intelligence (AI), speech recognition and synthesis, as well as natural language processing (NLP), VIs can make speech interaction accessible for more and more users, including older adults and people with disabilities [44]. In comparison to conventional interfaces such as keyboards, mice, and touch displays, VIs are hygienic, which is significant during a pandemic since users do not need to touch anything. Furthermore, VIs keep users' hands and eyes free for other tasks.

Generally, users can interact with VIs by voice or speech and VIs can recognize the user's spoken commands. Then VIs will give their answers based on their understanding and they will not say anything afterward like humans' communication. In other words, current VIs do not initiate talking with users and they only respond to users after perceiving users' voices and contexts. It means that VIs are not able to communicate with users naturally. They cannot perceive users' emotions via voice and express their own emotions when they talk with users, although it is feasible to perceive users' emotions by natural language processing and semantic sentiment analysis. With the advancement of speech signal processing and speech emotion recognition, it becomes possible to apply these techniques in future VIs. For instance, VIs can detect users' emotions in the communication between them by speech emotion recognition. However, it is quite challenging for the VIs to distinguish between real and disguised emotions. Most of the current speech emotion datasets were based on acting voice emotion [97, 8] and thus it is not easy to distinguish between the users' real and acted moods, which means that users may trick emotion-aware VIs. Even if the VIs can understand users' emotions, how to deal with users' emotions is another challenging issue.

1.1 Motivation

Nowadays VIs, such as Alexa, Siri, Google Assistant, and Cortana, make communication between humans and computers possible, using speech recognition and natural language processing to understand spoken contexts, and speech synthesis and dialog management to answer their questions. Nevertheless, the present functionalities of VIs are insufficient to suit people's needs. Current VIs can perform various tasks, including playing music, taking notes, creating shopping lists, etc., but they cannot perceive users' personalities and emotions, particularly unpleasant feelings. More precisely, despite the fact that speech emotion detection and speech emotion synthesis techniques are currently available, users find it difficult to converse emotionally with the VIs. Additionally, from the human perspective, emotions facilitate human-to-human communication by allowing individuals to express themselves emotionally while also comprehending and recognizing the feelings of others and determining the appropriate emotion to reply to them based on their emotions. Due to the user's demand for emotions in VIs, emotion-aware techniques will be quite important in future VIs.

Assuming VIs have the ability to understand users' emotions, they can not only carry on a conversation with users more naturally but also help users regulate their emotions and even monitor their mental health. As previously mentioned, emotional awareness can enhance human-to-human communication, and hence emotion-aware VIs can improve the users' experience throughout the conversation between VIs and users. Apart from emotional dialogue, emotion-aware VIs can help users regulate their emotions. Emotion regulation refers to the ability to exert control over one's emotions [215], and it involves behaviors such as hiding visible signs of sadness or fear, focusing on reasons to feel happy or calm, etc [119]. Innovative HCI intervention techniques can be combined with fundamental emotion regulation models from psychology [300], and emotional awareness can help emotion regulation more effectively, such as e-learning [105]. Moreover, emotion regulation is crucial to mental health, and a lack of effective emotion regulation is a potential transdiagnostic factor for mental illness [147]. For instance, negative and positive emotion regulation are associated with experiencing anxiety and depression in adolescence [346], and efficient emotion regulation can be beneficial for mental health. In addition, emotion-aware VIs can provide mental health support and assistance for physical activities [221].

With the advancement of speech signal processing (SSP) and speech emotion recognition (SER), it is possible to use knowledge of the users' emotional states in VI systems. Measuring restorative effects from speech signals, for example, can be implemented in automated driving [210]. However, numerous challenges remain in emotion-aware VIs. For one thing, the majority of SER databases are acted speech emotions [171, 310], which means that training datasets are also imitative emotional speech. The present SER system is not capable of detecting the emotion of speech accurately. For another, it is uncertain whether users can mimic fundamental emotions [208]. In addition, assuming that future VIs can reliably understand users' emotions, the difficult issue is how to respond to users' feelings, particularly negative emotions. All of these issues have to be solved before building emotion-aware voice interfaces that satisfy the users.

1.2 Background

Speech recognition is the basis of intelligent VIs and this technique can be traced back to the 1950s. In 1952, Davis and his colleagues in Bell Lab built an automatic single-speaker digital identification machine called " Audrey " which could recognize the sound of numbers from 0 to 9 with 90% accuracy. The researchers in the RCA laboratory at Princeton University developed a monosyllabic word recognition system in 1956 that can recognize the different syllables contained in a given person's ten monosyllabic words. Another researcher from the Lincoln laboratory at MIT, 1959, made a non-person-specific speech recognition system for ten vowels. From the 1960s to the 1970s, research on VIs mainly focused on the recognition devices that can recognize vowels (Radio Research Lab, Tokyo), phonemes (Kyoto University) and spoken digits (NEC Laboratory) [4] and isolated-words speech recognition, such as IBM Shoebox [201]. Researchers from Carnegie-Mellon University, IBM, and Stanford University built the Harpy speech recognition system [206] that could understand entire sentences.

With further breakthroughs in speech recognition technology in the 1980s, rigorous statistical modeling frameworks. gradually replaced template-based matching methods. Hidden Markov Models (HMM) became the main foundation for automatic speech recognition and understanding systems [155]. During these periods, researchers from IBM developed the experimental transcription system based on HMM that could recognize spoken words and type them on paper [22]. This speech recognition system was capable of processing approximately 20,000 words and it could achieve a 95% recognition rate at 5000 words [145]. In 1989, Lee et al. [190] developed the SPHINX speech recognition system using HMM method in combination with the Vector Quantization (VQ), capable of recognizing 4,200 consecutive utterances including 997 words. During the 1990s and the early 2000s, speaker-independent systems came into being with a number of innovative pattern recognition methods, such as Support Vector Machine (SVM) [67].

With the rise of deep neural network techniques, Mahamed et al. [224, 223] proposed a more efficient speech recognition method based on deep belief networks, which can significantly improve the recognition accuracy, even approaching 98% in the standard environment. During this period, the VI system started to grow by leaps and bounds, many VIs products came out such as Siri, Google Assistant, Cortana, and Alexa. These products can not only communicate with humans normally but also perform tasks or services for them based on their commands or questions.

Current VIs can not only comprehend and respond to users' speech but also communicate more organically with their human counterparts in a variety of contexts [20], with the improvement of natural language processing (semantic comprehension) and artificial intelligence technology. More specifically, VIs can recognize users' voices using speech recognition techniques, interpret voice information using natural language understanding techniques, and respond to users using dialog management and speech synthesis techniques [273, 291] as shown in Figure 1.1. Researchers used the VIs model in Figure 1.1 to apply it in a variety of device scenarios and undertake some novel applications [65]. Reddy et al. [276] are looking at what might happen if ordinary objects among people had conversational

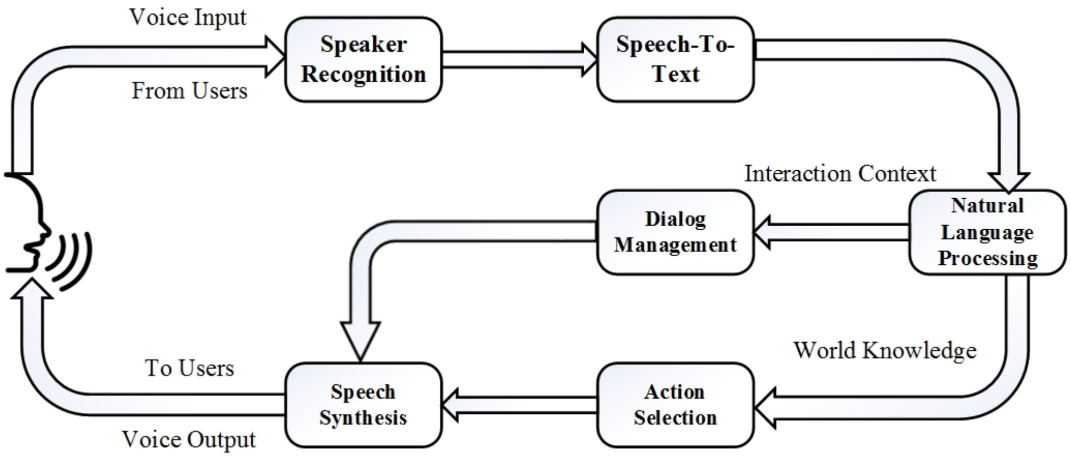


Figure 1.1: The structure of current intelligent voice assistant model, which is inspired by the Rakotomalala et al. model [273]

capabilities. This assumption also serves as a springboard for future study toward the development of ordinary conversational VIs at home. Additionally, researchers investigate the impact of user characteristics and preferences on how users interact with unfamiliar VIs [228]. Nevertheless, there are still certain challenges in interacting with VIs, such as improving measure reliability, validity, and consistency, evaluating speech interfaces in real-world contexts and reducing barriers to building speech interfaces [65].

On the technical side, current VIs are unable to communicate with users in the same way that humans do. Namely, according to existing methodologies and the VIs paradigm, VIs can sense human emotions by sentiment analysis and yet are unable to detect users' emotions by their voice signal processing. For example, the VIs can perceive users' emotions during dialogues by achieving word embedding with rich semantic and emotional knowledge and incorporating this emotional information into the deep learning architecture [142, 79]. It is also possible to understand users' emotions by acquiring emotional audio features and integrating these features into the speech emotion engine [289]. However, due to limits in the emotional speech database and algorithm flaws, the present SER approach is still challenging to implement in the current VIs.

1.3 Speech Signal Processing in VIs

Speech signal processing (SSP) is a feasible tool that can be implemented in future VI systems, although current research about VIs mainly focuses on user experience, user context, user trust, etc [232, 267, 273]. In general, users' voices can be regarded as input signals, and speech features analysis is performed after preprocessing (see figure 1.2). Many speech techniques, such as speech recognition, speech augmentation, etc, can be integrated into the VIs after the initial processing of SSP (preprocessing and speech characteristics analysis).

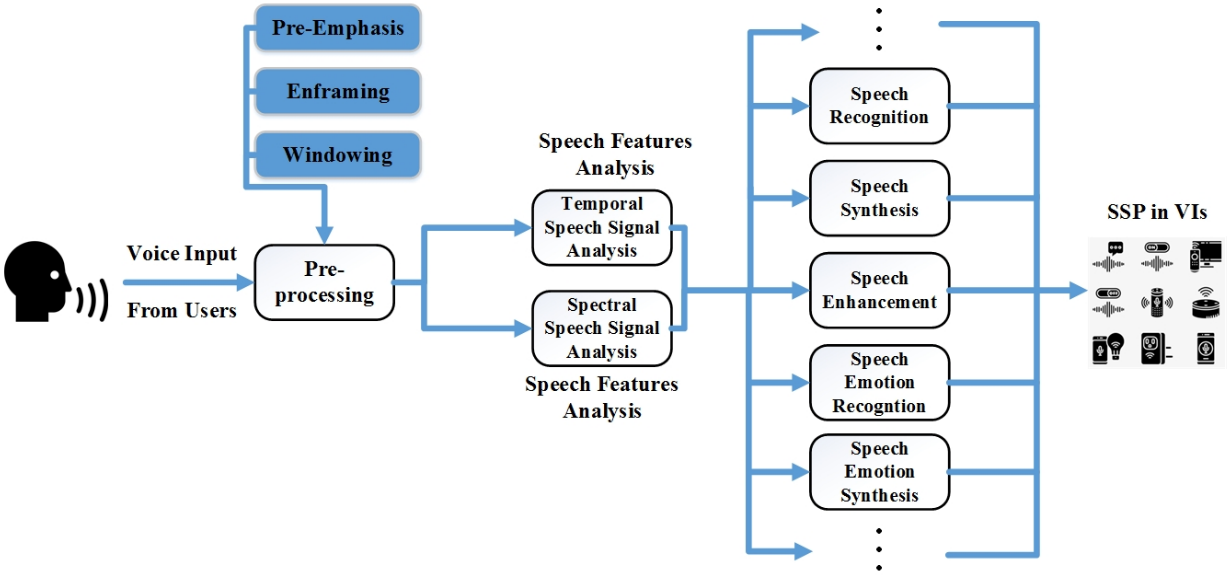


Figure 1.2: The Structure of SSP in VIs

Human social behaviors are secret signals [45, 253] and speaking is the most significant way to express themselves. In general, speech, which is the primary analog form of the message, can be translated to an electrical waveform, which can then be altered by both analog and digital signal processing, and then converted back to acoustic form as required [272]. SSP typically included speech recognition, speech synthesis, speech enhancement, speech storage, etc [110, 272]. Many SSP applications, such as dynamic family interactions in voice assistants [32], have been achieved in VIs by combining SSP with deep-learning methods [121]. For instance, researchers designed four voice assistant personalities (Friend, Admirer, Aunt, and Butler) to improve automotive user interfaces [40]. Furthermore, VIs can be able to build a bridge between patients and physicians by utilizing an interactive home healthcare system with an integrated voice assistant [89]. Patients can also use these VIs to get medical information, monitor and evaluate health conditions, and interact with ease based on handy applications and services. As previously stated, speech enables the creation of a link between humans and robots. As speech technology advances, the relationship between humans and VIs will become more intimate, such as consumer–voice assistant interaction [315] and future children's VIs [116].

1.3.1 SSP Techniques

SSP, namely speech analysis and processing consists of spectral and temporal speech signal analysis, as well as the necessary machine learning methods [82, 270]. From Figure 1.2, speech recognition, speech synthesis, and speech enhancement are the main speech techniques that have been implemented in the VIs. With the development of speech techniques and artificial intelligence, SSP plays a critically important role in VIs, although some of the latest advanced speech processing methods are not implemented in the current VIs. For instance, researchers presented health monitoring based on speech techniques [12], and Chang et al. and Arevian et al. [56, 16] proposed mental health monitoring through computational analysis of speech.

Speech-based health detection is introduced with the growth of SSP and deep learning [71]. Currently, the COVID-19 pandemic is the most prevalent disease [64] which has already seriously affected human health. Although it is possible to detect the coronavirus by some medical measurements, they are highly costly and uncomfortable for humans. Detecting COVID-19 from breathing and coughing sounds [64, 202] offers an alternative. In these study, the audio recordings of breathing or coughing were obtained via mobile devices or the web, and the convolutional neural networks (CNNs) was used to determine whether the speaker is infected with COVID-19 or not using raw breathing and coughing audio and spectrograms [290]. Additionally, it is feasible to develop speech interfaces for doctors and patients and improve healthcare services based on the state of art speech techniques [183].

Mental health detection using SSP is another application. Researchers proposed affective and mental health monitor using mobile phones [57]. Smartphones can capture users' voices and SSP based on glottal timing factors can assisted to assess mood, stress, and mental health. Krajewski et al. [176] presented speech-based fatigue detection using SSP and Greeley et al. [118] developed a fatigue detection system based on speech recognition and speech features analysis. Life stress is not uncommon for psychological and physical health problems and speech signals can be assisted to measure and assess humans' stress by analyzing certain features of speech and speech acoustics [319, 299]. Furthermore, SSP can assist in the diagnosis and analysis of early autism disorder states [238, 263]. Liu et al. [203] conducted depression detection from speech using acoustic features analysis. With the advancement of deep learning, researchers proposed a novel depression detection method using convolutional neural networks (CNNs) [91, 144]. In addition, SSP can be utilized to detect users' personalities based on prosodic features [225, 115, 13]. Likewise, SSP can be applied in social signal processing [328].

1.3.2 Applications of SSP in VIs

Speech recognition is the most common application in VIs. Normally, VIs can recognize users' voices via SSP and the corresponding machine learning algorithms and the spoken dialog system can be developed by speech recognition and synthesis techniques [186]. With the upgrading of SSP and deep learning, the techniques of speech recognition can be embedded in a variety of VIs [121]. For example, researchers proposed VIs in games using speech recognition [11]. In this study, VIs identify users' speech as input, and these voices are used in games for selection, navigation, control, and performance activities. Song et al. [303] presented research on frustration detection in game playing from voice signals using supervised contrastive learning. Many other VIs applications based on speech recognition are also proposed in some research, such as voice-controlled smart homes [53].

However, it is extremely difficult for VIs to recognize users' voices in noisy environments, especially in outdoor surroundings. Speech enhancement [76] has become quite vital for VIs in speech recognition and can be regarded as background noise processing. During the speech enhancement, traditional machine learning and deep learning were used to process and remove the additive noise from a speech signal. Researchers applied the speech enhancement to speech-based intelligent driver assistant systems using wavelet analysis and blind source separation [191]. Commercial speaker verification systems with speech enhancement can obtain better performance [100]. Moreover, speech enhancement can be used in wake-up word detection in VIs and these techniques can improve the accuracy of identification [38].

There are many other SSP techniques that can be implemented in VIs. Sleepiness and sleeping quality detection based on SSP in VIs are incredibly typical use. Researchers presented sleepiness detection using related-speech features and traditional machine learning algorithms and the self-report on the Karolinska Sleepiness Scale can be applied to derive sleepiness reference values [177, 175]. According to the correlation between the speech features and clinically validated questionnaire scores, it is also possible to use SSP techniques to estimate sleep quality, anxiety, and mood [165]. Another form of sleep detection is snore sound classification which can be also applied to sleeping quality detection. The Naive Bayes model based on speech wavelet features from snore sound [271] and data augmentation approach based on semi-supervised conditional generative adversarial networks (GANs) [354] can obtain high accuracy in snore sound classification. These speech techniques can be integrated into VIs such as smartphones and applied in snore detection [326]. In addition, SSP can be employed in silent speech interaction [166], making this interactive system useful in noisy surroundings. An ultrasonic image sensor is connected to the bottom of the jaw in this study, and it scans the interior situation as the user speaks without really emitting a voice. The user's voice is resynthesized and can be utilized to operate existing speech interaction systems such as smart speakers by detecting ultrasound images utilizing deep convolutional neural networks.

In summary, it is not impossible to integrate the SSP techniques into the VIs although not too many techniques about SSP were applied in VIs. VIs can easily record users' voices, which are then uploaded to the remote server. SSP and artificial intelligence algorithms can be used in the server to process speech data.

1.4 Emotion-Aware VIs

The expression of emotions in speech is the fundamental part of human communications [19] and thus emotions can enhance the interaction between users and VIs. Emotion-aware VIs are those that can recognize users' emotions and respond emotionally to them. Specifically as shown in Figure 1.3, emotion-aware VIs can record and detect users' voices first, and then recognize users' voice emotions using speech emotion recognition and natural language understanding methods. Subsequently, VIs will understand the users' speech and give the corresponding replies based on dialog management. Moreover, VIs will also respond to users with suitable emotions. Speech synthesis and speech-emotion synthesis plays a

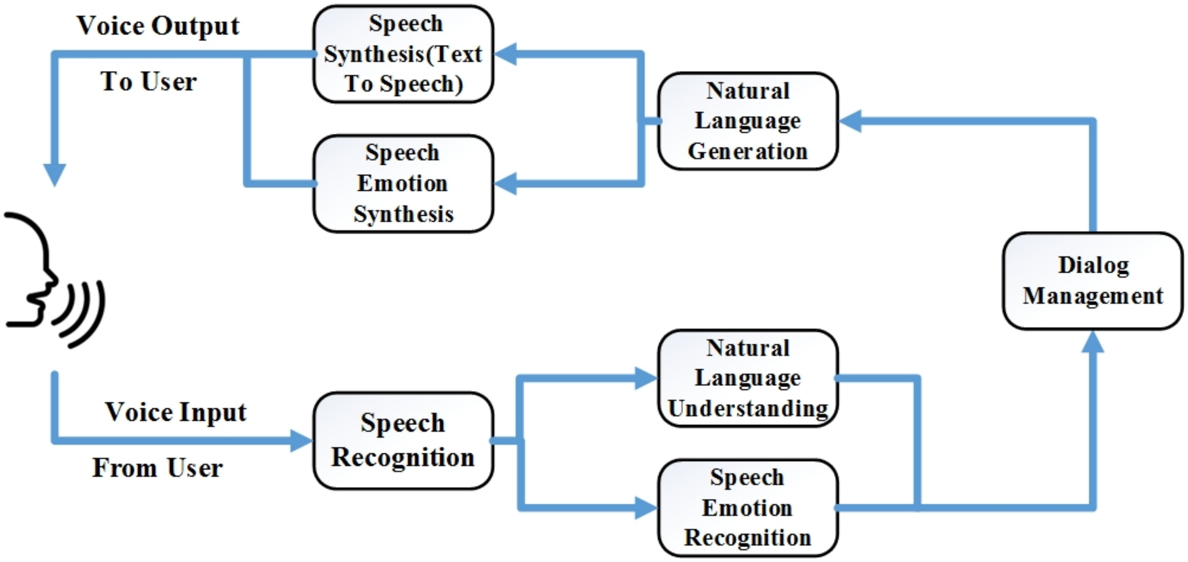

Voice Output
To User
Speech Synthesis(Text To Speech)
Speech Emotion Synthesis
Natural Language Generation
Dialog Management
Natural Language Understanding
Speech Emotion Recognition
Speech Recognition
Voice Input
From User

Figure 1.3: The Emotion-Aware VIs Structure

significant role during this period.

Currently, although emotion-aware in speech are unavailable in most VIs, VIs can sense users' emotion using sentiment analysis [351] and emotional context understanding [27]. However, it is not precise enough to perceive users' emotions merely using natural language processing. Languages and text can be leveraged to express emotions and the semantics in double meaning can lead to the impression of false feelings. Furthermore, differences across cultures [199] exist, and text-based emotion perception in different languages is challenging. With respect to speech, it is not difficult for humans to recognize emotion even if they do not comprehend the text language. Neumann et al. [234] presented cross-lingual and multilingual emotion recognition on English and French SER database with comparable interaction features and Liang et al. Liang et al. [197] proposed adversarial learning for speech emotion recognition in the datasets CHEAVD 2.0 (Chinese SER database) and AFEW (English SER database). Therefore, it is feasible to embed the SER techniques into VIs that can more precisely detect users' emotions.

1.4.1 Speech Emotion-Aware Techniques

Speech emotion recognition (SER) is the detection, analysis, and processing of human emotional states. More concretely, SER is crucial in speech emotion-aware techniques and SER algorithms can be leveraged to detect input voice emotion by extracting emotion-related speech characteristics from the input voice signal. As artificial intelligence advances, more deep learning techniques are increasingly used in SER [162]. For instance, some researchers proposed the SER method using DNN-HMMs with restricted Boltzmann Machine (RBM) and achieved relatively good results in the eNTERFACE'05 database [195]. Han et al. [125] presented SER approach utilizing deep neural networks and extreme learning machines.

With the development of Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN), more and more research about SER focus on CNN and RNN method, such as [143, 21, 356, 352]. The attention mechanism is another crucial technique in deep learning which also plays a critical function in SER. Since the emotionally salient frame in a speech signal exists, researchers proposed the SER approach utilizing RNN with local attention [222]. Self-attention can be also used in SER which is presented by Tarantino et al. [314]. A transformer network is the state-of-the-art deep learning technique that can be also utilized in SER, such as CTNet [196].

Generally, emotions are fundamental in human communication and they also play an essential role in VIs, especially emotion-aware VIs. With the rapid development of deep learning technology in speech emotion identification, speech emotion recognition technology is being used in an increasing number of human-computer interface and voice interaction technologies. For one thing, SER can enable rather natural communication between humans and computers, and many fascinating HCI applications, such as emotion recognition in contact centers, in-car board systems, and so on, are built on SER approaches [275]. For another, SER can enhance user experience in the HCI applications such as e-learning [23], emotional voice assistants on shopping [157] and emotional robot assistant [278].

However, due to a variety of constraints, emotion-aware VIs presents many difficulties. First of all, it is not easy to detect real emotions with high accuracy, even though current SER approaches can yield encouraging outcomes as deep learning develops. On the one hand, emotions are mental states connected with varied ideas, sensations, behavioral reactions, and a level of pleasure or unhappiness [185, 50, 51] and it is quite difficult to recognize certain emotions accurately, especially emotions from speech [93]. On the other hand, most of the speech emotion databases are acting or imitating emotions, rather than the spontaneously generated or induced real emotions [8, 162, 1]. Secondly, despite the existence of various approaches for speech emotion synthesis, such as [341] and [323], it is still a massive challenge for emotion-aware VIs to respond to users' emotions. It is feasible to incorporate human emotion reaction strategies, but it is still difficult to truly comprehend human tragedies and apply them to VIs.

1.4.2 Emotion-Aware VIs

Emotion-Aware techniques are becoming increasingly significant in human-robot interaction and social robots as artificial intelligence (AI) advances [317]. Despite the fact that there is little study on speech emotion-aware interaction and that most emotion-aware interaction research focuses on facial emotion or physiological sensors, such as [335] and [105], emotion-aware VIs play a vital role in future HCI research. On the one hand, emotions can be functioned as an important role in HCI [31, 257] and also be used as a tool for evaluating user experience [6]. Due to many limits of emotional measures, such as inconvenience, privacy concerns, and high cost, speech signals that can bypass these constraints and be gathered simply can provide an alternative. Moreover, speech emotions can be simply articulated, and humans can easily sense speech emotions. On the other hand, as aforementioned at the beginning, emotion-aware techniques can facilitate the interaction between humans and machines. In comparison to conventional emotion-aware interaction not based on voice, emotion-aware voice interaction has the capacity to remove constraints such as inconvenience and unpleasantness without the need for sensors or video recording, and voice signals are simply acquired.

As previously stated, there are certain concerns with emotion-aware VIs. To begin with, it is difficult to distinguish between real and false emotions or imitated emotions, despite the fact that there is already a range of spoken emotion detection systems as deep learning techniques progress. On the one hand, most speech emotion databases include acted emotions, and gathering true speech emotion data is challenging. On the other hand, individuals find it difficult to discriminate between genuine and fabricated emotions. Secondly, how to respond to human emotions becomes another challenge if emotion-aware VIs can perceive human emotions with high accuracy. Thirdly, it is also critical to investigate future emotion-aware VIs with respect to ethical and privacy concerns. Furthermore, it's also worth looking at potential emotion-aware VI-specific domains, such as automotive emotion-aware VIs.

1.5 Research Questions and Contexts

This thesis focuses mostly on emotion-aware VIs, such as speech emotion recognition and associated applications. More specifically, the thesis studies the prospect of future VIs that can sense the user's mood and respond to the user by deploying responsive emotions. In addition, the thesis investigates users' attitudes regarding the potential functions of the future emotion-aware visual interfaces.

1.5.1 Research Questions

Despite the fact that several speech technologies are now being employed in current VIs, there are still numerous research issues that need to be addressed. The main research questions in this thesis are presented in the following:

- **TRQ 0:** How to build an emotion-aware VIs?

This is a big question that will not be answered completely in this thesis. This thesis investigates only some aspects of this top-level research question. As the provided studies also involve research questions, the research questions in this thesis are referred to as TRQs (thesis research questions). The first two research questions are primarily technical. The one question is whether it is possible to measure users' restoration effects using speech signal processing. Another question is whether a speech emotion detection system can be employed in a real-world situation such as the teaching environment. In the other word, which approach can effectively detect users' emotions in the wild? In general, the speech-emotional databases are from an experimental context, and it is unknown if the algorithms based on these databases can be deployed in real-world circumstances.

- **TRQ 1**: Is it possible to measure users' restoration effects using speech signal processing?
- **TRQ 2**: Which approach can effectively detect users' emotions in the wild?

The next two research issues concern how voice interfaces should interact with users. One question concerns the ability to imitate emotions. More specifically, it is questionable whether existing emotion recognition systems correctly detect such performed emotions, or rather the speaker's actual feelings. The second question concerns how the voice interface should respond to emotions. Although it is feasible for VIs to sense users' emotions through semantic analysis or voice emotion recognition, responding to users' emotions remains a significant challenge.

- **TRQ 3**: To what extent can users pretend emotional states?
- **TRQ 4**: How should a voice assistant react to users' emotions?

Additionally, how to solve ethical and privacy problems in future VIs design is a big obstacle. The last research question is to ascertain users' perspectives regarding emotion-aware VIs.

- **TRQ 5**: What is the users' attitude towards emotion-aware voice interfaces and the resulting ethical and moral questions?

1.5.2 Research Contexts

Along with resolving these difficulties, the thesis exploited the relevant research using speech techniques. As aforementioned, emotion-aware VIs can enhance human-machine communication if the computer can understand users' emotions and respond to them with the appropriate emotion. In the other words, the emotion-aware VIs can recognize the users' emotions through speech signals and converse with them using proper emotion. Speech emotion detection and response are the most difficult aspects of this procedure.

A prototype was constructed to respond to TRQ 1. This was a software prototype rather than a hardware device. A user study was conducted using this prototype. With regard to the issue of speech emotion detection (TRQ 2), this thesis provided the subject of speech emotion detection by doing research on emotion identification from continuous speech in classroom teaching. The traditional speech emotion databases are not used in machine learning methods in this study, and the replacement dataset is obtained from classroom teaching contexts.

Additionally, another difficulty is detecting real and fake emotions (TRQ 3). The thesis conducted a user study in which a limited number of participants were requested to imitate five fundamental emotions, and an open-source emotion-in-voice detector (Vokaturi[1]) was

[1] https://vokaturi.com/

utilized to offer feedback on whether their performed emotion was detected as intended. If the emotion-aware VIs can detect users' emotions approximately using speech emotion recognition methods, the next challenge is determining how to deal with the detected emotion emotions (TRQ 4). This thesis discusses ideas for how humans react to other people's emotions, specifically how humans deal with an avatar's negative emotions. It presents user research in which participants were faced with an angry, sad, or terrified avatar.

TRQ5 was the subject of a questionnaire-based survey that was assessed using cluster analysis (k-medoids). Participants from several countries, including Germany, Egypt, and China, take part in our survey.

1.6 Thesis Contribution

This thesis provides several contributions:

- It presents novel approaches for measuring users' restoration effects through analysis of voice signals [210]. In comparison to existing methods for measuring restoration effects, this thesis proposes a novel method based on speech signal analysis. The virtual restorative environments user study will offer speech data from users. Time and frequency domain features will be extracted in speech signal processing and these features are correlated with conventional attention ratings. They also have been proven to be valuable for evaluating restoration effects using speech feature analysis.

- It exploits speech emotion recognition in classroom teaching circumstances using attention neural networks. In general, most of the speech emotion recognition experiments are carried out in the acted emotional speech database. The experiments in this thesis were conducted in classroom teaching contexts. As speech data is time series data, LSTM can help resolve pre- and post-association. Furthermore, in speech emotion recognition, the attention network can contribute to the resolution of major emotional expression issues. Thus, the main contribution in this part is speech emotion recognition in classroom teaching situations utilizing LSTM-attention neural networks.

- It created a brief user study to explore whether users could trick an emotion-aware VoiceBot [208]. More specifically, the main contributions in this section are the establishment of an emo-voice website and the implementation of a user study to investigate whether users can imitate basic emotions and cheat the existing speech emotion detector with acted emotions.

- It analyzes how to handle users' emotions by responding to three different unpleasant emotions [209]. Another emo-voice website with three bad emojis was built in this section. The thesis contributes to the tactics for dealing with users' emotions by drawing on strategies from users' reacting emotions in communication between users and emojis.

- Based on a research of users' preferences and views toward emotion- and personality-aware VIs across different cultures, it proposes three different user types (Enthusiasts, Pragmatists, and Skeptics) [207]. More concretely, contributions in this section include questionnaire design and analysis of questionnaire data. The questionnaire included 20 questions based on technological, social, and context dimensions, as well as 19 Likert-Scale questions. Another contribution in this section is the construction of three different user groups based on the correlation coefficients for all Likert-Scale questions and cluster analysis.

1.7 Thesis Outline

This thesis presents an early concept of future emotion-aware VIs as depicted in Figure 1.3. More precisely, future emotion-aware VIs will be able to recognize human emotions and respond to them with appropriate emotions. Moreover, these VIs can identify human emotions in real-time and in real environments and they can also recognize both real and simulated human emotions. The thesis is separated into six chapters that go through these studies in depth.

Chapter 1 - Introduction The first chapter motivates the topic of this thesis and explains a vision for emotion-aware vice interfaces. It also discusses the origins of emotion-aware VIs as well as the background of VIs and speech-emotion-aware techniques.

Chapter 2 - Emotions in Speech This chapter focuses on the many ideas of emotions, such as dimensional and category models. It also gives the notion of speech emotion.

Chapter 3 - Speech Features Analysis The third chapter presents temporal and spectral speech features including short-time energy, zero-crossing rate, Mel-Frequency Cepstrum Coefficient (MFCC), and so on. Additionally, it introduces a user study on measuring restoration effects using speech signal analysis.

Chapter 4 - Speech Emotion Recognition The fourth chapter is divided into two sections: speaker segmentation and speech emotion recognition. It provides an overview of existing emotion speech databases and speech emotion recognition algorithms in this section. Furthermore, it presents a speech emotion recognition approach based on an attention neural network in a classroom teaching circumstance.

Chapter 5 - Emotion-Aware VIs Exploration This chapter investigates the possibilities of future emotion-aware VIs, including the users' capacity to mimic emotion and how VIs deal with users' emotions.

Chapter 6 - User Preference for Emotion-aware Voice Assistants The sixth chapter investigated users' attitudes toward and preferences for emotion-aware voice assistants in three different cultures. An online questionnaire was employed in this study to investigate differences and similarities in attitudes in Germany, China, and Egypt. Three user types (Enthusiasts, Pragmatists, and Skeptics) were discovered that occur across all cultures using cluster analysis.

Chapter 7 - Conclusion and Outlook The last chapter focuses on the benefits and drawbacks of these five studies of future emotion-aware VIs. Furthermore, it suggests other possible applications for future emotion-aware VIs.

1.8 Own Previous Publications

Some of the content of this thesis has been previously published, in [193, 194, 207, 208, 209, 210, 309]. Below, I will clarify, how this material was used in my thesis and what my contribution to the respective paper was:

"A journey through nature: Exploring virtual restorative environments as a means to relax in confined spaces" [193] This publication provides the basis for chapter 3, and some figures from this publication are used in this thesis. In the publication, I contributed some ideas of the study and collected experimental data.

"Cultivation and incentivization of HCI research and community in China: Taxonomy and social endorsements." [194] This publication just provides a citation in the thesis. I contributed the some ideas of the study and wrote some parts of the paper text.

"Enthusiasts, pragmatists, and skeptics: Investigating users' attitudes towards emotion- and personality-aware voice assistants across cultures." [207] This publication provides the basis for chapter 6, and many diagrams from the publication are used in this thesis. To the publication, I contributed the general idea of the study, I supervised the student who designed the questionnaire and ran the study, I coordinated translation of the questionnaire into Arabic, did the Chinese translation myself, evaluated the results of the study, wrote major parts of the paper text, and coordinated the entire writing and publication process of the paper as its first author.

"Fake moods: Can users trick an emotion-aware voicebot?" [208] This publication provides the basis for chapter 5, many figures and tables from the publication are used in this thesis. Among my contributions to the publication is the design of the user study and the development of the website. Furthermore, I conducted the user study and analyzed the experimental data. In addition, I evaluated the results, drafted the paper text, and coordinated its publication as first author of the paper.

”How should voice assistants deal with users’ emotions?” [**209**] This publication provides the basis for chapter 5, many figures and tables from the publication are used in this thesis. To the publication, I contributed the general idea of the study, I supervised the student who designed the website and ran the study. Additionally, I evaluated the results, drafted the majority of the paper text, and coordinated its writing and publication as first author of the paper.

”You sound relaxed now - Measuring restorative effects from speech signals.” [**210**] This publication provides the basis for chapter 3, many figures and tables from the publication are used in this thesis. In the publication, I contributed the idea of the study and collected the experimental data. Moreover, I evaluated the results, drafted the paper text, and coordinated its publication as first author of the paper.

”Unsupervised speaker segmentation framework based on sparse correlation feature.” [**309**] This paper provides the citation in the thesis. My contribution to this publication was the idea for the algorithms and the data collection. Additionally, I evaluated the algorithm and results, drafted the major part of the paper text, and coordinated the entire writing and publication process.

Chapter 2

Emotions in Speech

Emotion is an essential part of human beings. It is frequently characterized in psychology as a complicated state of feeling that results in bodily and psychological changes that impact thinking and behavior [96, 174]. Emotion can be linked to a range of psychological characteristics such as temperament, personality, mood, and motivation, all of which entail physiological arousal, expressive actions, and conscious experience [229]. With the advancement of affective computing and AI, emotion now plays a key role in HCI, particularly VIs. The following chapter discusses the various emotion models as well as speech emotions.

2.1 Emotion

Emotion, according to a definition of Neurobiology of Emotion [75], can be divided into at least two categories: primary and secondary emotions. Primary emotions are considered to be natural and recognized as the usual form of feeling for a one-year-old child, whereas secondary emotions are assumed to arise from higher cognitive processes. From the opinion of Russel [283], emotion can be constructed in two ways: core emotions, which represent neurological states such as tiredness, and mental constructions, which denote acts such as facial expressions. The latter represents activities, such as facial emotions and tones, as well as correlations between actions. Scholars who have researched emotions have not yet been able to agree on them due to their complexity. Yet, emotions are mainly classified into three types: physiological, neurological, and cognitive. Emotions, from a physiological definition, are caused by responses within the body [286]. According to neurological theories, brain activity causes emotional reactions [111]. The cognitive theories propose that ideas and other mental activities play an important part in the formation of emotions [236].

Emotions have an impact on our attitudes and other feelings, as well as our cognitive performance, behavior, and psychology. Simultaneously, feelings are readily realized as a result of repeated emotional experiences, and when individuals feel glad to do a task many times, they fall in love with it. Emotions are mainly composed of three components [5, 305, 168]:

- Subjective experience refers to an individual's self-perception of various emotional states.
- External expressions, or expressions measured as movements of body parts at the start of an emotional state. Facial expressions (patterns of changes in facial muscles), gestural expressions (expressive motions of other areas of the body), and intonation expressions (aspects of speech variations in tone, rhythm, speed, and so on) are examples of expressions.
- Physiological arousal, or the physiological reaction caused by emotion, is a type of arousal.

2.1.1 Emotion AI

Emotion AI, namely affective computing, is computing that is capable of measuring, analyzing, and influencing emotions in response to human outward manifestations [258]. More specially, these techniques enable computers and systems to recognize, interpret, and imitate human feelings and emotions. As mentioned in chapter 1, emotion AI enables human-computer interaction more naturally when computers have emotions. In other words, affective computing can enhance computers' capacity to sense circumstances, comprehend human emotions and intentions, and respond accordingly. Further, affective computing techniques can bridge the interaction between humans and computers or robots. The difficult challenge, however, is precisely recognizing human emotions.

Generally, psychologists utilize questionnaires to measure human emotions, such as the state-trait anxiety inventory (STAI), the profile of mood states (POMS), and the positive and negative affect schedule (PANAS) [218]. Researchers, for example, can use the achievement emotions questionnaire (AEQ) to measure emotions in students' learning and performance [248]. However, due to the limitations of subjective aspects, collecting emotional questionnaires is inconvenient, and the final result may be inaccurate. Physiologic sensors can offer another alternative. Intuitively, human emotional states can be identified using a number of sensors that can monitor a wide range of data [92] such as heartbeat, respiration, pulse rate, and so on. For instance, it is feasible to detect humans' anxiety state via leveraging questionnaires to gauge subjective sensations, record and analyze facial muscle activity, monitor blood pressure with a sphygmomanometer, test adrenaline levels in blood samples, and so on [114]. Many sorts of emotions can be detected by combining information such as bodily movements and gestures, facial expressions, psychological signals, and speech [55].

Text and voice play a crucial part in emotion-aware VIs, and these information can assist to perceive human emotions as previously stated. Researchers in [345] presented a deep dual recurrent encoder model that uses text data and audio signals concurrently to gain a better comprehension of speech data and then integrates these knowledge to predict emotional states. More specifically, emotion recognition in a text document includes natural language processing (NLP) and deep learning techniques, and deep learning assisted semantic text analysis has been offered for human emotion detection utilizing large data [120]. Speech emotion recognition, on the other hand, is primarily concerned with speech feature analysis and classification methods [97]. As deep learning algorithms and different speech-based emotional databases progress, emotion in voice can be detected with increasing accuracy. In this thesis, speech signals are the main inputs, and speech emotion recognition is the main method for detecting users' emotions.

2.1.2 Emotion in HCI

Emotion is an integral element of human beings, and emotional computing plays a crucial role in HCI, with numerous HCI studies focusing on emotion [42, 257]. Emotions can be viewed as one form of assistive tool that can help bridge the connection between humans and computers. Emotional computing promotes more intuitive, natural computer interfaces by taking the user's emotional state into consideration, and it has significant potential for enhancing human-computer interaction. As emotion in AI advances, increasingly affective computing techniques can be employed in the HCI domain, e.g. user experience and affective interaction [255]. Generally, user experience questionnaire (UEQ) [287] can be leveraged to assess the quality of the user experience, and usability has traditionally been seen as the most important indication of user experience. Some researchers proposed that emotional response evaluation can be regarded as another integral part of the user experience [6]. The emotional responses effectively added to our understanding of the user experience.

Affective interaction is another application of emotion in HCI which involves emotion in education and learning, consumption, health monitoring, entertainment and gaming, and so forth [337]. In education and learning research, educational technologies based on user-centered design, such as mixed reality or AI-based systems, can successfully contribute to learning [242]. If emotions are included in instructional design and technology, these emotional designs can boost not just the learner's positive engagement but also facilitate comprehension and knowledge transfer [245, 214]. Emotions can also play a critically important role in the research of consumption behaviors. Affective information, such as affective recommender systems [318], can be implemented to identify consumers' moods and efficiently up-regulate their positive emotions during consumption. Furthermore, emotional information can be beneficial during customer service calls [126, 321] as well as physical and emotional health monitoring [227, 84]. In the other affective interaction, emotions can be regarded as an important component in the study of entertainment and gaming since they establish a link between the distal and proximal factors of participating in entertainment [313]. In addition, effective interaction can be implemented in voice interfaces as discussed in the previous section 1.4.

2.2 Emotion Models

In the realm of psychology, there are two primary types of theoretical perspectives on emotional models - category viewpoint and dimensional viewpoint. Emotions are separate and essentially diverse constructs according to the category approach. More specially, emotions can be classified into separate categories, and each of these categories differs somewhat in their external manifestations and physiological arousal patterns [149]. This emotional model is generally represented by five or six basic emotions. The same can be said of emotions when they are looked at from a dimensional perspective. More concretely, emotion models can be categorized as multi-dimensional models, which include bi-dimensional models, tri-dimensional models, and 4-dimensional models. A detailed description of emotion models is provided in the following section.

2.2.1 Discrete Emotion Models

Generally, discrete emotion models are composed of several basic emotions that are easy for individuals to grasp and comprehend. According to Robert [261] from the early 1980s, there were eight core emotions, which he grouped into four pairs (joy-sadness, anger-fear, trust-distrust, surprise-anticipation). Ekman defines discrete emotions as a small set of universal and inherent emotions that contain six basic emotions: fear, anger, disgust, sorrow, happiness, and supervise [94, 96, 96]. Currently, researchers use different basic emotion models depending on their needs. For instance, the six emotions model has been used in facial emotion recognition for HCI applications [62] and the majority of the speech emotional database is likewise based on the six emotions model [311].

Normally, there are two advantages regarding the discrete emotions utilization in HCI research. First of all, it is easy for individuals to grasp emotions when the discrete emotion model is used to describe emotions and sentiments. It also corresponds to people's intuition and common sense. Consequently, it is advantageous for use in real-world applications. Subsequently, the discrete emotion model can facilitate emotion awareness and emotion response in HCI. However, this paradigm seems incapable of precisely characterizing the nature of emotions or evaluating emotional states from a computational standpoint. It is difficult to define which emotions are required for HCI research, and there is no consensus among researchers on this. Furthermore, emotion categories are qualitative descriptions of undefined experiences, and the subjective emotional experience cannot be stated statistically.

2.2.2 Dimensional Emotion Models

Dimensional models of emotion visualize emotional states as points in space, and the distance between those points indicates how emotionally different they are [269]. Different emotional states can be seen scattered in different locations in space according to their dimensions, and the distance between them reflects the intensity of the emotions. Different emotions in the dimensional model are not independent of one another, but are continuous, allowing for progressive, seamless transitions. As opposed to discrete emotion models, dimensional emotion models are continuous and have the benefit of capturing a wide variety of emotions and documenting the progression of feelings.

American psychologist Johnston utilized a one-dimensional axis to depict emotions, with the positive half-axis indicating happiness and the negative half-axis representing sadness [154]. In this model, when a person is stimulated by negative sentiments, the emotion advances toward the negative axis; when the stimulus is withdrawn, the negative emotion decreases and moves closer to the origin. When positive emotions stimulate the emotional state, it moves toward the positive semi-axis and finally returns to the origin when the stimulus fades.

Based on the two-dimensional emotion model, emotions can be categorized according to their polarity and intensity [70]. Polarity is the contrast between positive and negative emotions, whereas intensity is the distinction between strong and weak emotions. Currently, the most extensively used model today is Russell's validity-arousal two-dimensional model [282], which was created in 1980. As shown in Figure 2.1, this dimensional emotion paradigm is made up of the valence dimension and the arousal dimension. The valence dimension's negative half-axis represents unpleasant emotions, while the positive half-axis indicates pleasant emotions. The arousal dimension's negative half-axis reflects peaceful emotions, while the positive half-axis depicts intense emotions.

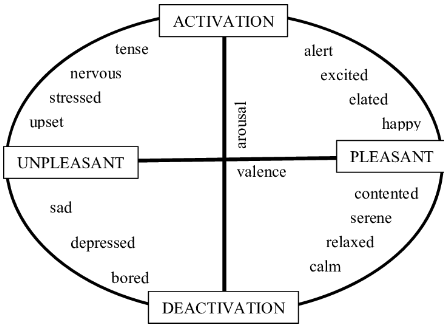


Figure 2.1: The Two Dimension Emotion Model [172, 282]

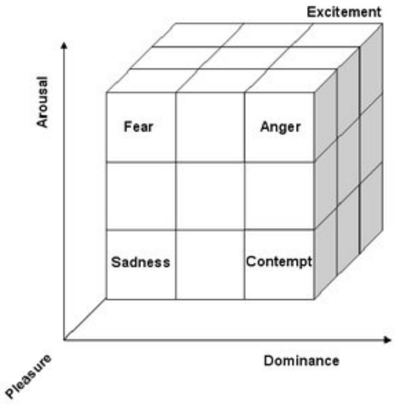


Figure 2.2: The Three Dimension Emotion Model [179]

In addition to the polarity and strength dimensions, extra characteristics are incorporated in the emotion description in the 3-dimensional emotion model. Mehrabian proposed a **PAD** emotion model with aspects of pleasure, arousal, and dominance dimensions [217]. From Figure 2.2, **P** represents pleasantness, which indicates the positive and negative aspects of the individual's emotional state; **A** denotes arousal,which signifies the individual's level of neurophysiological activation; and **D** stands for dominance, which designates the individual's state of control over the situation and others. This 3-dimensional model not only provides a theoretical framework to describe the emotional space, it also employs a quantitative approach to establish the position and relationship of various emotional categories in that space, allowing it to be widely used in emotional psychology, affective computing, human behavior, marketing, and product satisfaction [39]. Another 3-dimensional emotion model based on monoamine neurotransmitters was presented by Lövheim [205]. In addition to these three-dimensional emotion models, psychologists offered a 4-dimensional emotion model (psychologically-motivated emotion categorization model) that incorporates the dimensions of sensitivity, aptitude, pleasantness, and attention [52]. Although the dimensional emotion models provide a theoretical and methodological basis for quantitative emotional computing, it remains challenging to determine how many dimensions are required to accurately represent human emotions, and which dimensions are necessary.

2.2.3 Other Emotion Models

Aside from the more commonly employed dimensional and discrete emotion models, some psychologists have proposed other emotion models that are based on different theories, including cognitive-based emotion models [240], probabilistic emotion models based on emotional energy [258], and event-based emotion models [324]. These emotional models are useful in analyzing human emotions from a variety of perspectives and enriching the mathematical description of emotions.

2.3 Emotion Models in HCI

As previously said, there are too many emotion models to pick from, making it difficult to choose which emotion model should be employed in HCI research. Discrete emotional models, such as basic emotions, are easily recognizable and have the benefit of cultural commonality and simple discriminability. Accordingly, implementing this emotion model in an HCI context is not unfeasible. However, due to the general flawed amount of this paradigm, it appears incapable of accurately representing real-world emotions. Some emotions, in fact, cannot be explained using merely a discrete emotion model. Conversely, the dimensional emotion model can depict a wide range of emotions as points in space, such as the bi-dimension model and tri-dimensional model. Although dimensional emotion models have been employed in affective computing, it remains problematic to identify how many dimensions are required to effectively describe human emotions.

Despite the fact that it is generally recognized that various emotion models have certain benefits and limitations, as shown in the Table 2.1, it is still difficult to select an appropriate emotion model and utilize it in HCI research, particularly in VIs studies. Since the discrete emotion model is intelligible for users, it is applied in emotion-aware VIs in this thesis. SER techniques can be applied to VIs, and many speech emotion databases in SER are discrete emotion databases, such as SAVEE emotion datasets [128, 130] and Emo-DB. Furthermore, unlike dimensional emotion models, discrete emotion models are easily identified and understood by consumers.

Table 2.1: The Advantages and Disadvantages in Discrete and Dimensional Models

Emotion Models	Advantages	Disadvantages
Discrete Emotion Model	Easy Identification and Comprehension	Insufficient Amount
Dimensional Emotion Model	Large Number and Continuation	Difficult Selecting

The discrete emotion model is used in the majority of the user studies in this thesis. Regardless of the fact that the purpose of teacher emotion recognition study is to distinguish positive emotions, the database is still a discrete emotion model. The Vokaturi[1] instruments can detect the participants' emotions, making them vital tools for studying mimic emotions and emotional reactions. The discrete emotion model continues to be used in this speech emotion database.

2.4 Beyond Basic Emotions

The absence of emotions in voice assistants is undoubtedly one of the reasons why existing voice assistants do not seem genuine. However, emotions alone will not suffice to achieve the aim of lifelike voice assistants. In addition to the fundamental emotions, there are non-basic emotions such as engagement, boredom, perplexity, and irritation [87]. VIs can respond to the user's needs more properly. There have been few studies on detecting non-basic emotions. This study makes use of facial expressions [350] or facial expressions combined with speech analysis [2]. It is unclear how reliable the identification of non-basic emotions by speech alone is.

Appropriate datasets are necessary for neural networks to recognize non-basic emotions. There is a database that contains frustration and enthusiasm [49], but there is no database that includes all non-basic emotions because it is unclear what non-basic emotions are. It is unclear, for example, whether love is a fundamental emotion, a non-basic emotion, or no emotion at all [178]. Aside from non-basic emotions, users can be in other mental states, which may be crucial for voice assistant awareness. If the user is fatigued or inebriated, the voice assistant should adjust its behavior. If the purpose of a voice assistant is to behave like a person, the voice assistant must be aware of the circumstances. A voice assistant should not begin lengthy talks if the user is in a hurry. Future studies should focus on context awareness as well as emotion and personality awareness.

[1] https://vokaturi.com/

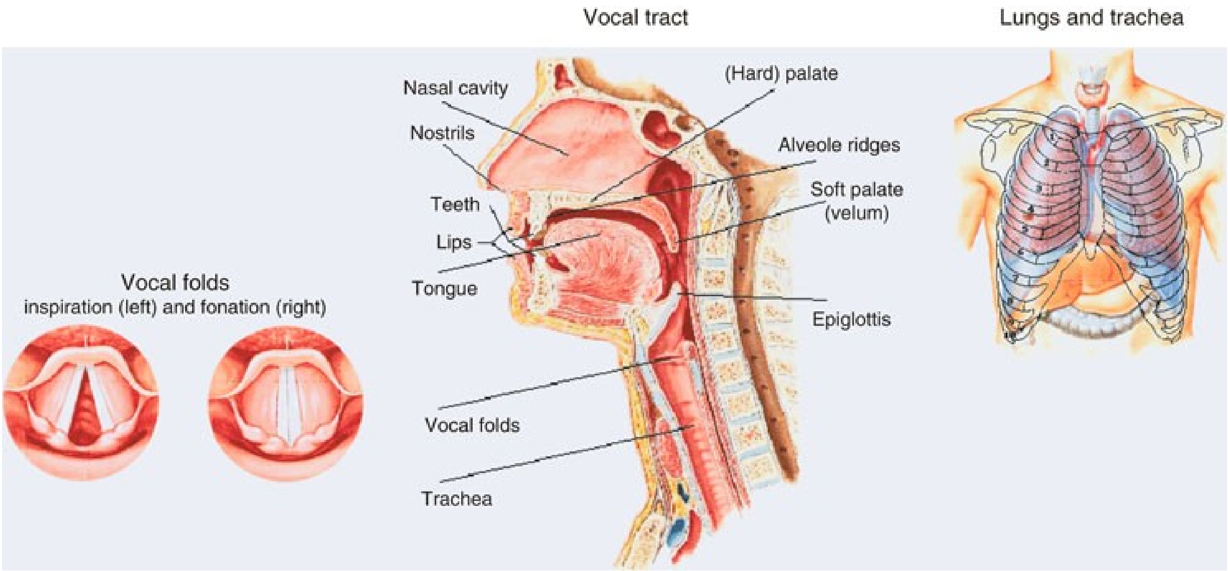


Figure 2.3: Speech Production [88]

2.5 Speech Emotion

Speech plays an essential role in VIs since speech emotions foster more natural communication between humans and VIs. The discrete emotion model is applied in VIs research throughout this procedure. This section contains information regarding speech production and speech emotions.

2.5.1 Speech Production

As described in Figure 2.3, the human voice is generated by the vocal organs, which include the lungs, trachea, vocal tract, pharynx, nose, and mouth and are controlled by the brain [4, 88, 272]. These organs create a complicated channel in which the larynx is referred to as the vocal tract and the area between the vocal tract and the lips is referred to as the vocal tract. The lungs and trachea give energy to the vocal system, while the larynx creates sound, which is subsequently regulated by the vocal tract. More specially, when no sound is created, air enters the lungs by breathing. Air is released from the lungs via the trachea, forming a stream that ultimately flows via the vocal cords. When the vocal cords are stretched, they vibrate and the airflow from the vocal tract to the lips produces a muted tone. If the vocal cords are relaxed, the passage through the vocal tract has little effect on the lung air. If the vocal tract is constricted, the air is accelerated and propelled out, resulting in a fricative or clear tone; if the vocal tract is directly closed, the air is compressed here and makes a bursting sound when it opens again.

According to the process of creating speech, it is possible to analyze speech signals in time and frequency domains. Time-frequency domain speech features can be employed in affective computing and HCI research, such as speech emotion recognition [97] and voice interfaces [210].

2.5.2 Emotions in Speech

Recognizing emotional information in speech is an important part of authentic human-computer interaction. When conveyed with different emotions, the same verbal content has dramatically different interpretations. Only when the computer accurately detects both the content and the emotion of the speech can it correctly perceive its semantic meaning; hence, understanding the emotion of the speech may make the human-computer interaction more natural and fluid.

Humans can detect changes in the emotional state of another person by listening to speech since the human brain has the ability to perceive and interpret. The human brain is capable of seeing and comprehending information in a speech signal that reflects the speaker's emotional state (e.g., special intonation, change of intonation, etc.). The computer's modeling of the above-mentioned human emotion perception and understanding process is known as automatic speech emotion recognition. Its purpose is to extract emotional acoustic components from collected speech data and to determine the mapping connection between these acoustic elements and human emotions.

As depicted in Table 2.2, a user who is sad or disgusted will typically speak slower and lower-pitched, with resonant or grumbling chest tone; whereas a user who is fearful, angry, or pleased will speak quicker and louder, with high pitched, powerful high-frequency energy, as well as more explicit enunciation [258]. It is not difficult for humans to recognize emotion states depending on the characteristics of speech emotion, and it is also advantageous to use these speech characters in VIs research.

Table 2.2: Speech Emotion [42]

	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	Much faster	Slightly faster	Slightly Slower	Faster or slower	Very much slower
Pitch average	Very much higher	Very much higher	Slightly lower	Much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Normal	Higher	Lower	Higher	Lower
Voice Quality	Irregular voicing	Breathy chest tone	Resonant	Breathy blaring	Grumbled chest tone
Pitch change	Normal	Abrupt on stressed syllables	Downward inflections	Smooth upward inflections	Wide downward terminal inflections
Articulation	Precise	Tense	Slurring	Normal	Normal

2.6 Summary

This chapter mainly introduced the concept of emotion, emotion models, emotion models in HCI, and speech emotions. Especially, it is difficult to offer explicit definitions of emotion, yet emotion can be defined from physiological, neurological, and cognitive perspectives. Furthermore, emotions are mainly composed of three components: subjective experience, external expressions, and physiological arousal.

In general, there are two primary types of theoretical perspectives on emotion models - discrete and dimensional emotion models. The dimensional emotion model can depict a wide range of emotions as points in space, such as the bi-dimension model and tri-dimensional model. Yet, it seems that it is not easy to determine how many dimensions are required to effectively describe human emotions. Since discrete emotion models are easily identified, they are well suited to use in HCI research, especially in VIs. In addition to the discrete emotions, non-basic emotions such as perplexity, fatigue, irritation, etc., may be also employed in future emotion-aware VIs.

Speech emotion is one of the most significant expressions of emotion, and it can facilitate more natural communication between humans and VIs. Emotion-related speech features, such as time-frequency domain speech features, can be utilized in speech emotion recognition, which can also be employed in emotion-aware VIs. It is possible to calculate these speech features based on speech production, although producing speech is a complex procedure.

Chapter 3

Speech Features Analysis

As aforementioned, speech plays a critical role in VIs, and speech features analysis is an integral part of speech processing. A well-known fact is that speech signals are commonly recognized as non-stationary and time-varying [123]. The human voice is formed through the vocal tract as a result of intense impulses contained within the vocal chambers. Due to the slow movement of the oral muscles of the voice box, the speech signal can be regarded as stable and time-invariant in the "short time" range. This trait dictates that the "short-time analysis" approach can be employed in the analysis and processing of speech data. The term "short time" usually refers to a period of 10-30 ms. To examine its distinctive properties, the speech signal can be separated into segments, each of which is termed a frame, and the frame length is generally this brief duration. Furthermore, each frame can be used to generate feature parameters in time and spectral domains. This chapter will go over additional specifics concerning speech features and their applications, as well as a user study on the use of speech signal processing for measuring the effectiveness of restoration [210, 193].

3.1 Speech Features in the Time and Frequency Domain

Human's voice can be regarded as a kind of waveform that can be captured by the microphone. The microphone usually transforms the acquired sound vibrations in the air into an electrical signal, which can then be sampled to produce a discrete-time waveform file. That is what it refers to as the digital speech signal. The speech signal is a time-varying, non-smooth signal that contains a variety of information. As it aforementioned, short-time speech signals are steady, therefore it is not complicated to analyze the speech signal in a short time period. Pre-processing plays a critical role in speech signal analysis [160, 272]. For one thing, it can eliminate the influence on the quality of the speech signal caused by the human vocal cords itself, as well as variables caused by the equipment used to record the speech signal, such as aliasing, high harmonic distortion, high frequencies, and so on. On the other hand, it can ensure that the produced speech signals are uniform and smooth after pre-processing, which can subsequently be used to improve the quality of speech signal analysis, such as speech feature extraction.

In general, the study and processing of phonetic characteristics, which include features in both the temporal and spectral domains, is referred to as speech signal analysis. As previously stated, temporal features (time domain features) [281] can be extracted in the "short time" range, making them easy to extract and interpret physically; spectral features (frequency domain features) [17] can be obtained by converting the speech signal from the time domain into the frequency domain using the Fourier Transform.

3.1.1 Pre-processing in Speech Features Analysis

Pre-processing typically consists of pre-emphasis, framing, and windowing. **Pre-emphasis** is used to weight the high-frequency section of the speech, eliminate the influence of mouth-lip radiation, and improve the high-frequency resolution of the speech [325]. A transfer function, such as a first-order FIR high-pass digital filter, is commonly used for it. The pre-emphasis filter is frequently configured as follows [285].

$$H(z) = 1 - az^{-1} \tag{3.1}$$

In the equation 3.1, a is specified as a constant that is commonly referred to as the pre-emphasis coefficient. The value of a typically ranges between 0.9 and 1.0, and this value is commonly taken as 0.97.

Typically, **framing** is configured to include a partial overlap between two adjacent frames [272]. The main reason for this is that the speech signal is time-varying, and the feature changes are minor in the short time range, so it is considered as a steady-state signal, and each frame duration is between 10ms and 30ms as it was aforementioned. Yet, the speech signal contains changes beyond this short time range [272, 7]. The fundamental tone changes between two adjacent frames and its characteristic parameters may change a lot. Nevertheless, in order to ensure that the characteristic parameters change smoothly, two non-overlapping frames are inserted between the overlapping frames to extract the feature parameters, forming an overlapping part between adjacent frames.

Windowing is the process of multiplying a speech signal waveform by a time window function as a way to emphasize its predetermined characteristics [146, 124]. As a time-varying signal with short-time smoothness, speech is sampled as $x(n)$, a time-varying signal with a nearly unlimited length, but the processing frame is similar to multiplication by a finite-length window function [268, 272].

$$y(n) = \sum_{n=-\propto}^{\propto} x(m) * \omega(n-m) \tag{3.2}$$

The output of the equation 3.2 can be viewed as the convolution form which corresponds to the discrete signal $x(m)$ flowing through a unit impulse corresponding to the $\omega(m)$ FIR filter. The window function has a low-pass characteristic in general, and different window function selections will have varying bandwidth and spectral leakage. In speech signal analysis, the primary window functions include rectangular window (3.3), Hamming window (3.4), and Hanning window (3.5), among others [262, 262]. In the following three equations (3.3, 3.4, 3.5), **L** represents the window length.

$$Rectangular\ Window:\ \ \omega(n)=\begin{cases}1 & 0\leq n\leq L-1\\ 0 & others\end{cases} \tag{3.3}$$

$$Hamming\ Window:\ \ \omega(n)=\begin{cases}0.5*(1-\cos(2\pi n/(L-1))) & 0\leqslant n\leqslant L-1\\ 0 & others\end{cases} \tag{3.4}$$

$$Hanning\ Window:\ \ \omega(n)=\begin{cases}0.54-0.46*\cos(2\pi n/(L-1)) & 0\leqslant n\leqslant L-1\\ 0 & others\end{cases} \tag{3.5}$$

3.1.2 The Time Domain Features

The time domain features (temporal features) are easier to extract and analyze in the "short-time" range, such as short-time energy, zero crossing rate, short-time auto-correlation, and so on.

The phrase "**short-term energy**" refers to the magnitude of the data energy in each frame of a speech signal, which is used to differentiate between initials and finals, voiced and unvoiced, and so on [152]. Assume a time domain signal for a speech waveform is set to $x(n)$ and a window function is set to $\omega(n)$. If $y_i(n)$ is the i-th frame speech signal produced after frame processing, then $y_i(n)$ is as follows:

$$y_i(n)=\omega(n)*x((i-1)*inc+n),\ \ 1\leqslant n\leqslant L-1,\ \ 1\leqslant i\leqslant fn \tag{3.6}$$

In the equation 3.6, $\omega(n)$ is the window function, generally a rectangular or Hamming window. $y_i(n)$ represents one frame value, where $n=1,2,...,L$, $i=1,2,...,fn$, L is the frame length; inc is the frame shift length; and fn represents the numbers of frames after framing.

The following formula can be used to represent the short-time energy in the i-th frame voice signal $y_i(n)$.

$$E(i)=\sum_{n=0}^{n}y_i^2(n),\ \ 1\leqslant i\leqslant fn \tag{3.7}$$

Short-time zero-crossing rate indicates the number of times the speech signal waveform crosses the horizontal axis in one speech frame [152]. A zero-crossing rate occurs when the time-domain waveform crosses over the time axis in continuous speech signals, whereas it occurs when the neighboring sample values change signs in discrete signals. The number of times a sample value changes sign is the short-term average over-zero rate. It can discriminate between turbid and consonant sounds, as well as detect speech signals from background noise and determine the wordless section. It can also establish the beginning and ending positions of unvoiced and voiced segments.

Assuming $y_i(n)$ is the speech signal $x(n)$ after framing and the frame length is L, the short-time zero-cross rate is as follows.

$$Z(i) = 1/2 * \sum_{n=0}^{L-1} \left| sgn[y_i(n)] - sgn[y_i(\leqslant n-1)] \right|, \ \ 1 \leqslant i \leqslant fn \tag{3.8}$$

In the equation 3.8, $sgn[x]$ is a symbolic function; if $x \geqslant 0$, the value of $sgn[x]$ is 1; if $x < 0$, the value of $sgn[x]$ is -1.

The short-time auto-correlation is the result of intercepting a signal segment at the N-th sample point with a short-time window and calculating auto-correlation [152]. The short-time auto-correlation of a speech signal can be utilized primarily for endpoint identification and fundamental extraction, as well as to determine the fundamental period of turbid tones and feature parameters in speech recognition. Assuming $y_i(n)$ is the speech signal $x(n)$ after framing, the short-time auto-correlation of each data frame is defined as follows.

$$R_i(k) = \sum_{n=0}^{L-k-1} y_i(n) * y_i(n+k) \tag{3.9}$$

In the equation 3.9, L denotes speech frame length and k is delayed amount.

3.1.3 The Frequency Domain Features

The frequency domain features (spectrum features) are produced by transforming the time-based signal to the frequency domain with the Fourier Transform, such as Mel-Frequency Cepstrum Coefficients(MFCCs), Linear Prediction Coefficient (LPC), Linear Prediction Cepstrum Coefficient (LPCC), fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. These features can be utilized to discern notes, pitch, rhythm, and melody, which can then be employed in speech recognition and emotion identification. In section 3.1.3, it focuses mostly on the introduction of MFCCs.

MFCC is a depiction of a sound's short-term power spectrum that is based on the human peripheral auditory system [137]. More specifically, this spectrum feature is the study of the frequency spectrum of speech based on the findings of human auditory experiments. There are two auditory mechanisms involved in MFCC. First, according to the theory of the mel scale [306], the frequency domain of human subjective perception is nonlinear. Human perceptual frequency can be expressed as follows:

$$F_{mel} = 1125 * \log(1 + \frac{f}{700}) \tag{3.10}$$

In the equation 3.10, F_{mel} represents perceived frequency in Mel and f denotes the physical frequency in Hz [78].

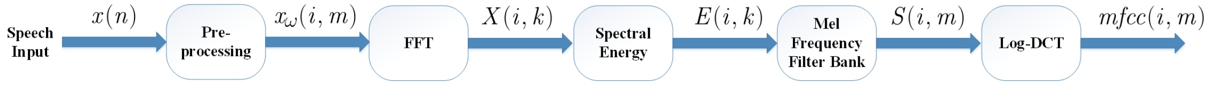


Figure 3.1: The Flow Procedure of MFCC Feature Extraction

The second mechanism is the crucial band, which is made up of several frequency groups and generally corresponds to human auditory neuron tuning curves [294, 353]. The frequency group refers to the segmentation of the human ear's basilar membrane into many tiny portions, each of which corresponds to a frequency group. For the same frequency group of sounds, the brain superimposes them together for processing. The partition of speech into a sequence of frequency groups in the frequency domain based on the critical band division generates a filter bank known as the Mel filter bank. Assuming the band-pass filter sets $H_m(k), 0 \leqslant m \leqslant M$, M denotes the number of filters (each filter with triangular filtering properties), and the center frequency is $f(m)$. The band-pass filter has the following transfer function:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leqslant k \leqslant f(m) \\ \frac{f(m+1))-k}{f(m+1)-f(m)} & f(m) < k < f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{3.11}$$

In the equation 3.11, $0 \leqslant m \leqslant M$. In this formula (3.1), $f(m)$ can be defined as the following equation:

$$f(m) = (\frac{N}{f_s}) * F_{mel}^{-1}(F_{mel}(f_l) + m * \frac{F_{mel}(f_h) - F_{mel}(f_l)}{M+1}) \tag{3.12}$$

In the equation 3.12, f_l is the filter frequency range's lowest frequency; f_h is the filter frequency range's highest frequency; N is the length at fast Fourier transform (FFT); f_s is the sampling frequency, and F_{mel} is the inverse function of F_{mel}^{-1}.

MFCC extraction follows a flowchart in Figure 3.1. As it shown in this figure, assign the signal obtained after pre-processing $x(n)$ as $x_i(m)$, where the subscript i represents the i-th frame after framing. As a result of performing the FFT for each frame, the time domain signal is converted to the frequency domain signal $X(i,k)$. The spectral energy can be calculated using the function of $[X(i,k)^2]$. The Mel filter bank energy can be calculated by multiplying the energy spectrum $E(i,k)$ of each frame by the frequency domain response $H_m(k)$ of the Mel filter and summing them up. MFCC can be finally obtained using the logarithm of the Mel filter bank energies, followed by a discrete cosine transform (DCT).

$$mfcc(i,n) = \sqrt{\frac{2}{N}} * \sum_{m=0}^{M-1} \log[S(i,m)] * \cos(\frac{\pi n(2m-1)}{2M}) \tag{3.13}$$

In the equation 3.13, $S(i, m)$ denotes the Mel filter bank energy, m indicates the m-th Mel filter (there are M in total), i represents the i-th frame, and n stands for the spectral line after DCT.

3.2 General Rhythm and Social Speech Features

In addition to general speech features, rhythmic and social speech features can also contribute to speech analysis. Unlike the general speech features, rhythmic speech features can sense the movement in speech, which is characterized by stress, tempo, and syllable amount [235]; social speech features can detect nonverbal verbal cues and predict behavioral outcomes in various social situations [302]. Rhythmic speech features include tempo, brightness, key, and so on, whereas social speech features include activity, engagement, emphasis, and mirroring. In section 3.2, it focuses mostly on partial rhythmic and social speech features.

3.2.1 General Rhythm Speech Features

As aforementioned, rhythmic speech features can reflect the movement in speech as well as the tendency of consonants and vowels of languages, such as sentiment, utterance form (statement, inquiry, or command), irony or sarcasm, emphasis, and so on [163]. Consequently, these features can be applied to speech analysis, multilingual speech recognition, speech emotion recognition, etc.

Tempo is a common rhythmic feature that can estimate the length of musical tones, and it can also be employed in music to investigate a section of a specific change in tempo speed or change in rhythm speed [73, 181]. The beat is often used to identify the number of strikes per minute, and it is also employed in speech features to compute the tempo of the speech rhythm feature by recognizing the periodicity of the speech signal.

The **brightness** of voice is the degree of purity and appeal to the ear, which typically refers to the change in the short-time energy of the speech signal at frequencies over 1500 HZ [181]. It indicates the clarity of the voice diction and represents the degree of penetration of the speaker's voice in speech characteristics. Thus, it is also possible to analyze voice signals using brightness.

The **key** of a voice is the frequency level of sound, which also reflects the degree of human hearing to differentiate the tone of a voice [181]. In rhythmic speech features, Key can be classified into C, A, and G tones that represent the speaker's gene frequency variance.

The **inharmonicity** is another rhythmic speech feature that indicates the amount of energy outside the ideal harmonic series [180, 181]. There are also numerous more rhythmic speech elements that can be exploited in speech analysis. Due to the thesis's limitations and relevance to the user context, it only introduces these four rhythm speech features.

3.2.2 Social Speech Features

Social speech features generally can be highly successful in predicting behavioral outcomes in various social interactions, such as negotiations, dating, etc [72, 302]. According to the idea of Pentland [252], social speech features often include activity, engagement, emphasis, and mirroring.

Activity is known as the percentage of a person's speaking time that can be calculated by z-scoring the fraction of speaking time plus the frequency of voiced segments [302, 72]. The ratio of the duration of speaking frames in the conversation to the entire duration of the conversation was used to calculate the fraction of speaking time. The rate of voicing segments obtained in the conversation's speaking region was used to calculate voicing frequency.

$$Fraction\ speaking\ time = s/n, \tag{3.14}$$

where $s = speaking\ frames$, $n = total\ frames\ in\ speech\ segment$ [302].

$$Voice\ rate = v/(v+u), \tag{3.15}$$

where $v = voiced\ frames$, $u = unvoiced\ frames$ [302].

Engagement can be described as the impact that one person has on the other's conversational turn-taking [72, 302]. Individual turn-taking dynamics impact each other in a two-person dialogue, which can be described as a Markov process [250]. Engagement can be calculated by assessing one person's effect on the other. To calculate the engagement, it is possible to use a measure of the coupling between the two Hidden Markov Models (HMM) that describes the influence each individual has on the other. The calculation is as follows.

$$P(S_t^i|S_{t-1}^i......S_{t-1}^N) = \sum \alpha_{ij} * P(S_t^i|S_{t-1}^i) \tag{3.16}$$

where $\alpha_{ij} = influence\ coupling\ parameter\ between\ interacting\ chains\ i\ and\ j$, $P(S_t^i) = probability\ that\ chain\ i\ is\ in\ state\ S\ at\ time\ t)$, $i = chains\ from\ 1\ to\ N$, $t =$ discrete time steps.

Emphasis is the variation in speech prosody, especially variation in pitch and loudness [72, 302]. In general, emphasis is determined by extracting the mean energy, frequency of the fundamental format, and spectral entropy in each voiced segment. More specifically, the speaker's emphasis can be obtained by summing up the mean-scaled standard deviation of the energy, formant frequency, and spectral entropy. The following equation 3.17 can be used to calculate emphasis.

$$Emphasis\ measure = \sum std(\varepsilon) + std(\mu) + std(\rho) \tag{3.17}$$

where ε = formant frequency, μ = spectral entropy, ρ = energy in frame, and std is standard deviation [302].

Mirroring is another social speech trait that is found in back-and-forth exchanges in dialogue, often consisting of single words like (' OK?', 'Yes!', 'done') and short interjections

(1 sec) like ‘uhhuh’, ‘hummh’ [72, 302]. The following formula can be used to describe the mirroring calculation.

$$Mirroring\ measure = \{(S1(i) - S2(i)) \leqslant 1sec\}/n \tag{3.18}$$

where $S1(i) = time\ of\ occurrence\ of\ short\ speaking\ frame(< 1s)\ for\ speaker\ 1$,
$S2(i) = time\ of\ occurrence\ of\ short\ speaking\ frame(< 1s)\ for\ speaker\ 2$,
$n = length\ of\ speech\ segment\ in\ seconds$,
$\{\} = total\ number\ of\ such\ pairs\ in\ time\ n$ [302].

3.3 User Study: Measuring Restoration Effects From Speech Signal

Speech features analysis can be utilized in a variety of applications, including HCI applications [65], VIs applications [121], social behavior processing [327, 328, 231], physiological and psychological health [37, 80], etc. In general, temporal and frequency domain features can be used for speech recognition, emotion recognition, and personality recognition. These recognizing techniques can well be incorporated into VIs systems to improve user-VI interaction. The rhythmic and social speech features can be assisted to detect mental health and social behavior. These functions are typically embedded in VIs, and hence have a significant impact on VIs.

IVs are commonly used in automated driving, and this system can improve the user experience in automated vehicles, such as voice control in autonomous driving [81, 211]. Another possibility is to use speech signal analysis to evaluate restorative experiences in automated vehicles. This part contains a use case for measuring restoration effects using speech features analysis. The main work from this part consists of an unsupervised speaker segmentation method and attention restorative metrics based on speech features analysis. The user study was conducted in in-car Virtual Restorative Environments (VREs) [193] and the results of the speech features analysis are reported in this part. The majority of the figures, tables, and results in this section are drawn from the citation paper [210].

3.3.1 Research Background

Physical and mental recovery has become a critical topic in human-centered computer research due to the rising stress for people in current metropolitan contexts. Some of this study looks toward the use of (virtual) natural surroundings to efficiently recover attention [237] and alleviate weariness. One potential use for such a restorative environment is autonomous driving, which can provide travelers with a means to reconnect with nature while psychologically recharging in automated automobiles. Currently, existing research often employs self-report questionnaires or reaction tests as primary indicators to assess the effects of attention restoration. Traditional procedures, on the other hand, disrupt the experiment’s flow, alter the attention being assessed, and may suffer from subjectivity and memory effects. Subjectivity and memory effects may potentially be present in the study.

Other techniques of detecting attentiveness using sensors or devices were proposed. For example, analyzing attention recovery can be done by utilizing eye movement to evaluate attention [107]. Restorative attention can also be detected using EEG-based [36, 230] and ECG-based [54] features. However, it exists certain concerns, such as uncomfortable human supervision, to which individuals may be constrained in a particular environment. An alternative is an attention restoration detecting system based on SSP. Dhupati et al. [83] presented the concept of employing SSP in the investigation of sleepiness or tiredness detection. It enables the automatic detection and assessment of attention restoration, as well as the analysis and evaluation of the mental state. Speech cues, in particular, are crucial in emotion recognition and can aid in the automated evaluation of mental recovery stages. The extraction of time and frequency domain speech features, in conjunction with an efficient speech segmentation algorithm [357], can therefore legitimately detect and evaluate the extent of attention restoration in this context.

In this user study, it presents a unique assessment approach for attention restoration and applies it to a virtual reality-simulated in-car restorative environment. The effects of attention restoration are measured via speech signal analysis. It is also possible to distinguish different restoration levels based on speech features using classic machine learning approaches such as SVM and KNN algorithms.

3.3.2 User Study Design

The design of this use case is based on user-centered design (UCD) [3] and semi-structured interviews. Ten users were polled about their favorite non-driving-related activities (NDRAs), daily relaxing environments, and design requirements for relaxing environments in self-driving cars. The most often stated NDRAs were watching movies, sleeping, online chatting, internet surfing, and playing games. Some individuals like to unwind in a quiet room with soft lighting and music, or by going for a stroll in nature or reclining on the beach. Based on the findings of the interviews, an initial prototype for in-car Virtual Restorative Environments (VREs) featuring water, beach, seagulls, and trees was proposed. In order to experience in-car VREs that are as near to reality as possible, speed and navigation information can be provided naturally in the form of animation in an immersive setting.

To completely immerse travelers in the offered in-car restorative environment, this user study employs HMD-based VR using an Oculus Rift VR headset. In this user study, two modes were presented: autonomous driving and relaxing. Users can switch between the modes by using an Oculus Touch Controller. participants can feel the vehicle's velocity and the navigation map while driving in automatic mode as is shown in figure 3.2. On the windscreen, it displays the vehicle's primary status as well as a progress bar showing the present position to the passengers. In the relax model, participants can experience the in-car VRE with a 360° video of live-captured natural scenery[1]. Birds warble and wave in the background, accompanied by light flute music[2] in VR. The video brightness

[1] 360 Degree Cinemagraphs, *accessed: October 2019*

[2] Relaxing Pan Flute Music, *accessed: October 2019*

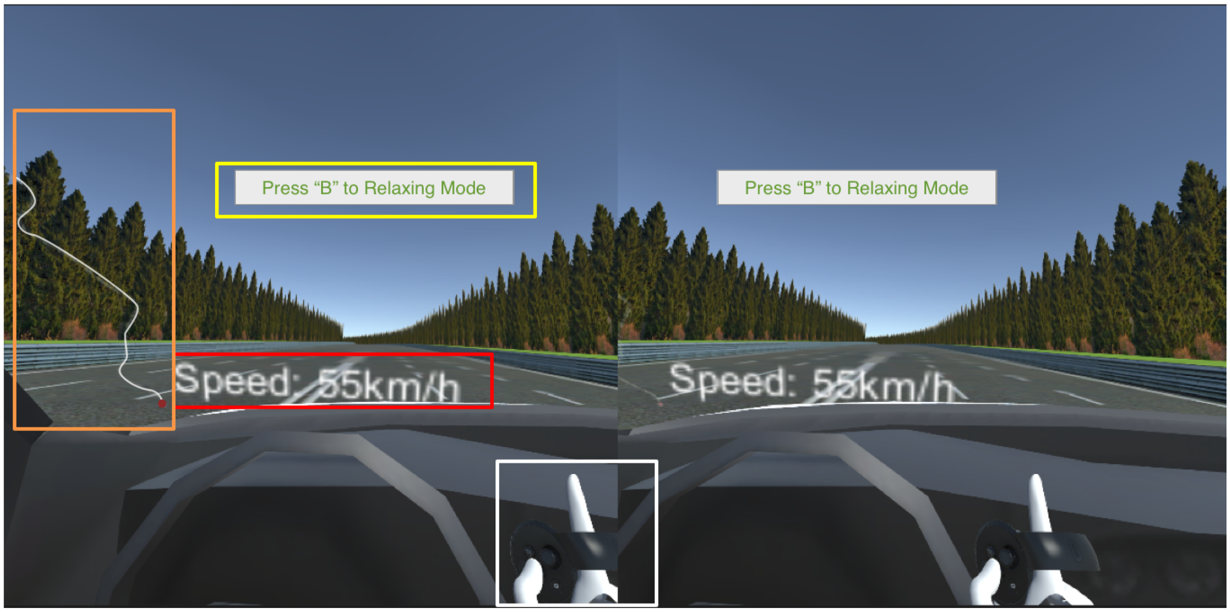


Figure 3.2: Automated driving mode: The UI displays car speed (red box), navigation map (orange box), hand model (white box), and study instructions (yellow box) [193]

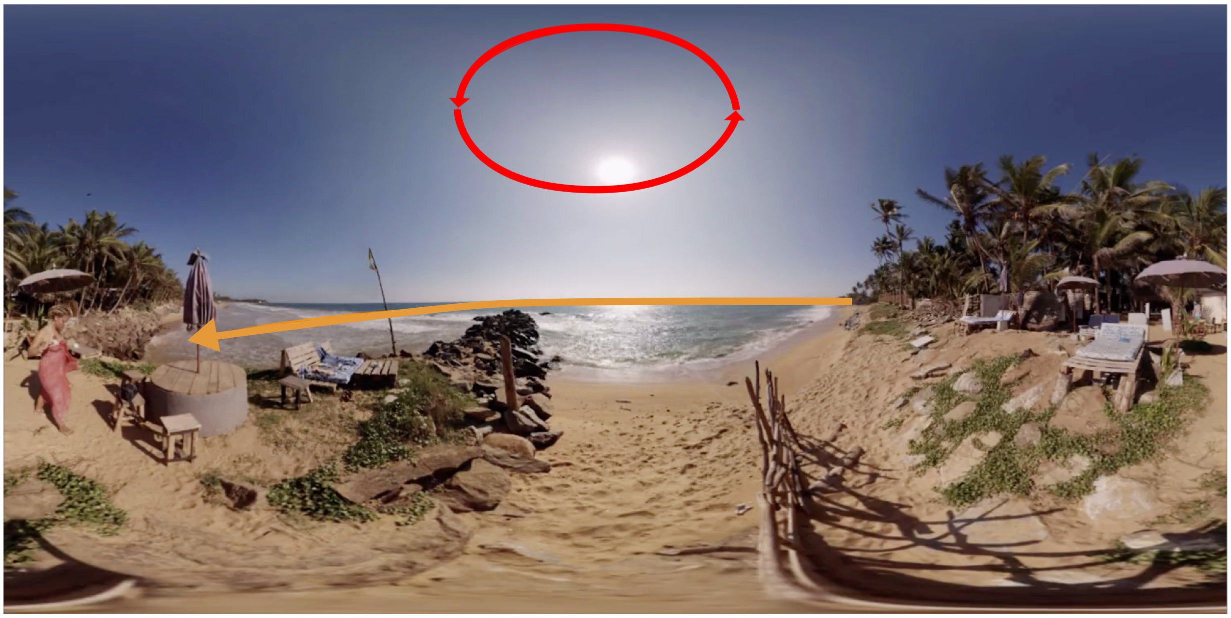


Figure 3.3: Relaxing mode: The moving path of seagulls (in red) indicates speed and the ship (in orange) shows the progress of the entire journey [193]

can be changed to improve the participants' experience in the simulated environment. In addition to this natural peaceful scenery, it shows 3D representations of seagulls and a ship as metaphors for vehicle velocity and travel progress, respectively (see Fig. 3.3).

3.3.3 User Study Procedure

The within-subject design was utilized in the user research, and each participant experienced the in-car VRE. The data obtained included a comparison of subjective and objective attention measurements before and after the restorative experience, as well as an analysis of the correlation between the post-experience measures of attention and the properties of the speech signals. It was hypothesized in the experiment that the in-car restorative environment in VR would have strong restorative effects on attention, and that specific aspects of the voice signal would be correlated with attention metrics.

Apparatus, Participants, and Measurement

The study procedure was approved by the local Ethics review board of LMU of Munich. Unity 3D was installed on a normal PC linked to an Oculus Rift for this user study. Speech signals were recorded at 48 kHz using a digital audio recorder (Aigo voice recorder R6611). The microphone was around 15cm away from the participants' lips. For voice signal analysis, Matlab 2017a and MIRtoolbox 1.7.1 were utilized. In the experiment, it included a total of 12 volunteers (5 male) aged 25 to 28 years (M=26, SD=0.9). Moreover, half of them had previously driven and used virtual reality. Furthermore, they preferred resting in a quiet environment with gentle music, watching movies, or going for a walk in natural settings. It uses objective and subjective attention metrics, as well as the suggested speech signal analysis approach, to determine attention restoration.

Subjective attention can be measured using the attention function index. The 13-item Attention Function Index (AFI) [63] is intended to assess perceived efficacy in daily tasks supported by attention and working memory. It is frequently used as a self-assessment of cognitive performance. It utilizes an 11-point Likert scale, with 0 representing *Not at all* at all and 10 representing *Extremely well*. Object attention can be measured using the digit span backward test. This measurement utilized the Inquisit program, which displays the numerals 0-9 in a random sequence. Participants had to repeat them in the exact opposite sequence, and the experimenter used the mouse to pick the number they uttered from a circle of digits. Their voice was recorded in the object attention measurement.

Speech Signal Analysis

Unsupervised speech segmentation can be used to segment speech components from the Digit Span Backward test (DSB) voice recordings. Short-time energy, pitch, and MFCC will be extracted as time- and frequency-domain speech features. Furthermore, prosodic speech features such as tempo, key, and harmonics will be obtained from the speech parts of the participants. These features can be assisted to study and evaluate the degree of attention restoration. The entire processing pipeline is shown in Figure 3.4 and explained below.

Denoising and enframing the speech data, as well as selecting a timing window, are all part of the preprocessing procedure. In this procedure, the frame size was set as 25ms and a frame step was set as 10ms. As a requirement for extracting time-frequency domain speech

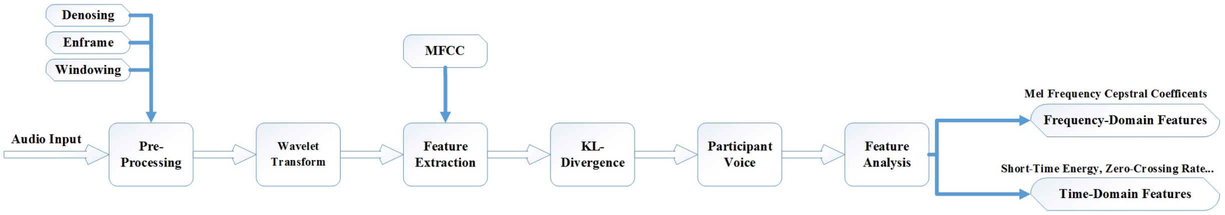


Figure 3.4: The signal analysis pipeline tightly integrates speech feature extraction and speaker segmentation

features, the data is then processed using a Daubechies-4 Wavelet Transform [74]. At this point, the KL-Divergence [342] can accurately split the data into speech and non-speech segments when combined with the previously recovered features. Features are only used from speech data segments At the end of the processing, both time- and frequency-domain features are forwarded to further processing.

Experiment Procedure

Participants were invited to the experimental room and given a brief overview of the study. Figure 3.5 displays the total timeline of this study method. After collecting personal information, participants were instructed to complete the AFI surveys as a subjective assessment of attention and the DSB test on a computer as an objective measure of attention. The baseline included both subjective and objective measures of attentiveness. During the DSB test, the recorder was also used to collect audio data from the subjects. Then, as a warm-up for the restorative experience, participants were invited to watch a video clip of a traffic jam[3] so as to induce context-specific stress. At this stage, the same measurements (AFI and DSB) were performed again.

The experimenter showed an instructional film of the system's functionality prior to the restorative experience. Following a training period, participants either experienced the in-car restorative environment or closed their eyes in VR for 10 minutes (as in the other condition). They conducted the same attention tests after the restorative experience, and the results were recorded and compared to the previous baseline. Finally, the experimenter conducted an interview in which individuals were asked to discuss their experiences throughout the restorative experience. The entire experimental flow is shown in figure 3.5.

3.3.4 Study Results

To fully comprehend the impacts on attention restoration, objective and subjective attention assessments, as well as speech signal analysis, were applied to analyze. The Pearson correlation coefficient was utilized to assess whether there was a significant correlation and difference between the two sets of attention measurements before and after the restorative experience, which were based on time-frequency domain speech components in Elan [184].

[3] Incredible Traffic, *accessed: October 2019*

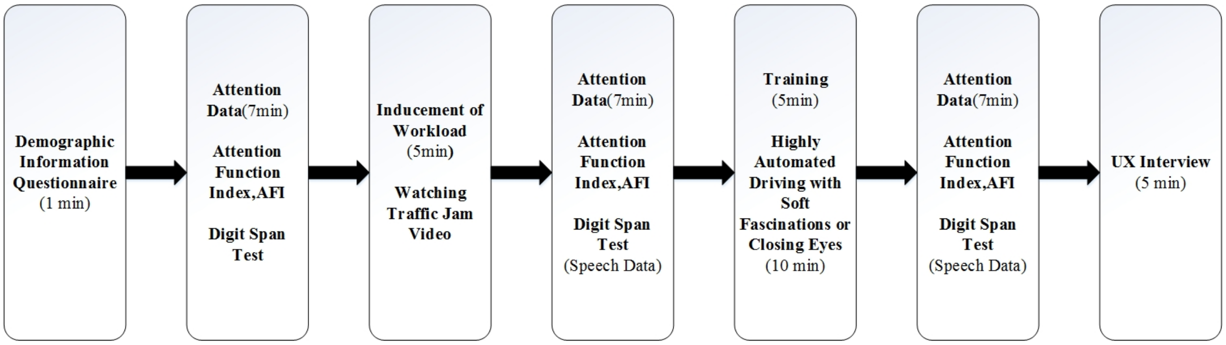

Demographic Information Questionnaire (1 min)
Attention Data(7min)
Attention Function Index,AFI
Digit Span Test
Inducement of Workload (5min)
Watching Traffic Jam Video
Attention Data(7min)
Attention Function Index,AFI
Digit Span Test (Speech Data)
Training (5min)
Highly Automated Driving with Soft Fascinations or Closing Eyes (10 min)
Attention Data(7min)
Attention Function Index,AFI
Digit Span Test (Speech Data)
UX Interview (5 min)

Figure 3.5: Timeline of the study procedure [210]

Furthermore, classical machine learning can be utilized to verify if speech signal analysis can accurately evaluate the effect of attention restoration. A high correlation impact is seen when $r \geq 0.5$ and the accuracy of the machine learning approach is more than 0.85 in two categories and 0.7 in three classifications.

Traditional Attention Measures

Attention Score The AFI's attention score represents the participants' subjective assessment of their current attentional state. The VRE condition had a somewhat higher average rise than the shutting eyes condition. Participants gained 2.9 (SD=7.6) points on average following the VRE restorative experience, compared to 1.4 (SD=8.4) points after closing their eyes. The absolute AFI score was marginally higher after the restorative experience in the VRE condition (Mdn=61) as compared to the pre-VRE measurement (Mdn=59), W=65, p=0.31.

Working Memory The DSB examines the digit span to determine the participant's working memory capacity as a measure of their direct attention ability at the moment. While the two error maximal length (TEML) measurement did not indicate any change in both situations before and after the restorative experience, it investigated the mean span (MS) measurement further.

Across conditions, the VRE condition resulted in a somewhat greater MS increase than closing the eyes: participants increased their short-term memory by 0.16 (SD=0.7) digits in the VRE, compared to 0.11 (SD=0.5) digits after closing the eyes. Furthermore, after being exposed to the VRE, subjects obtained their greatest performance in terms of working memory span (M=5.11, SD=1.0) compared to the previous DSB tests.

Attention Evaluation Based on Speech Signal Analysis

Speaker Segmentation Result Table 3.1 depicts the outcomes of speaker segmentation and figure 3.6 shows one of them. A total of 21 people completed the attention scales and made usable voice recordings. The average accuracy in speaker segmentation was around 94%, while the f1 score was around 89%. As a result, MFCC may be viewed as an appropriate speech characteristic for detecting the participant's voice. A 4-Daubechies Wavelet transform may also effectively reduce noise, and KL-Divergence based on MFCC can segment the participants' voices.

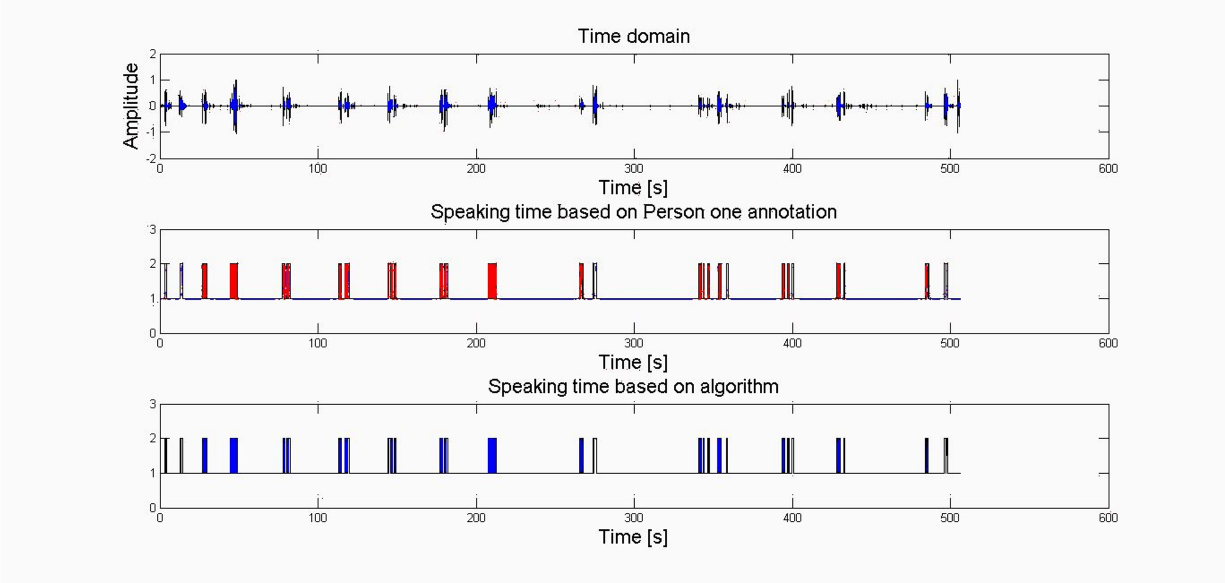


Figure 3.6: Result of speaker segmentation based on WT-KLD, compared to manual annotation of the data.

The decreased accuracy and f1 score in certain participants are due to the wavelet transform failing to properly reduce noise, such as Gaussian white noise, and therefore the KLD approach failing to isolate the participant's speech. Furthermore, certain low soliloquies or other people's voices (experimenter's voice) were captured in the recorder, and these voices may have an impact on the participants' speech detection during the whole speaker segmentation procedure.

General Time-Frequency Speech Features As shown in table 3.2, the time domain features, such as short-time energy, zero-crossing rate, max peak in autocorrelation, and 3 formants exhibit a substantial positive connection with the AFI score. The first three frequency peaks were selected in the spectrum with a high degree of energy as formants.

These six speech features can be utilized to properly investigate and measure attention restoration effects in this experiment. Furthermore, the short-term energy and zero-crossing rate were lower before and after the in-car VRE session. In other words, these two features appear to increase when participants go from a state of fatigue to a state of relaxation. More study on this secondary discovery, however, is required to develop a stable metric.

Additionally, the frequency-domain feature MFCC, as shown in Figure 3.7, was linked to the various Attention Restoration Levels (ARL). The top two photographs in figure 3.7

File No.	*Accuracy*	*F1-score*
01	0.7906	0.8831
02	0.8672	0.9289
03	0.7642	0.8664
04	0.9366	0.9673
05	0.8680	0.9293
06	0.9976	0.9988
07	0.7210	0.8379
08	0.9215	0.9591
09	0.9389	0.9685
10	0.6482	0.7098
11	0.9389	0.9685

Table 3.1: Accuracy and F1-score for Speaker Segmentation in the different data files

show spectrograms before and after experiencing an in-car restorative natural environment in VR, whereas the bottom images show before and after closing the eyes. The right spectrograms are substantially brighter than the left ones, implying that time-frequency speech properties are also connected with the attention restoration level and can help identify such levels.

Attention Restoration Level Classification Based on these seven effective speech features that are related to attention restoration phases, two classic machine learning techniques can be employed. The low-level state is defined as happening prior to the VRE experience, while the high-level state is defined as occurring after the VRE experience. The SVM and KNN algorithms were employed to provide the encouraging preliminary findings shown in table 3.3. Even if three levels are specified (closed eyes, VRE experience, and before), the SVM and KNN methods shown in table 3.3 can be used to reliably identify these three states. This demonstrates the link between speech qualities and the extent of attention restoration. More specially, it is feasible to evaluate attention restoration based

Table 3.2: Pearson Correlation Coefficients between Different Features and the Attention Restoration Level [210]

Short-Time Energy	0.8994
Zero-crossing Rate	0.9192
Max peak in autocorrelation	0.9209
Formant 1	0.9209
Formant 2	0.5798
Formant 3	0.7515

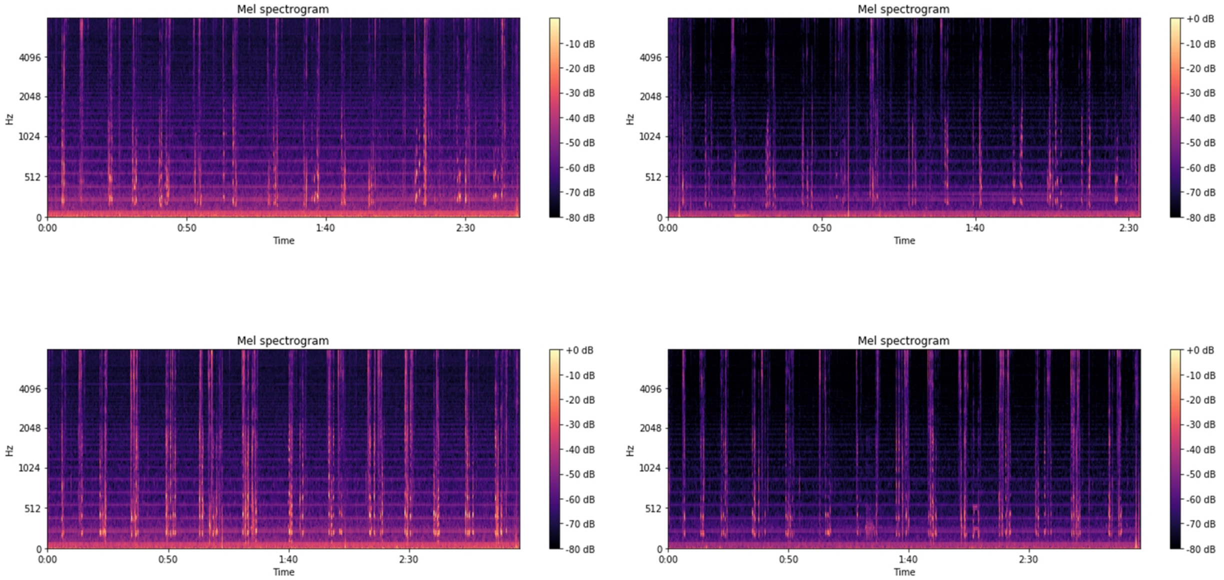


Figure 3.7: Spectrograms of the MFCC feature in different attention-restoration states

on speech signal analysis.

Table 3.3: Recognition Accuracy of SVM and KNN using 2 and 3 attention restoration levels

	KNN Method	*SVM Method*
2 level classification	0.8775	0.9228
3 level classification	0.6051	0.7162

Rhythmic Speech Features

Table 3.4 indicates how the rhythmic speech features of tempo, key, and inharmonicity correlate with the level of attention restoration. The tempo is a measure of the number of speech units of a specific kind generated in a particular length of time [180, 181] and this acoustic feature can be obtained by detecting periodicities. The tempo has a substantial correlation with the attention restoration level in table 3.4. The key feature is the tonality speech feature, which can evaluate tonal center positions and their relative clarity [180, 181]. According to table 3.4, the key has a high association with the amount of attention restoration. Inharmonicity [180, 181] is a prosodic speech feature that reflects the amount of energy that is not in the ideal harmonic series. In addition, the inharmonicity value is connected to the amount of attention restoration. Thus, these three rhythmic speech features can also assist to analyze and assess attention restoration.

Table 3.4: Correlation coefficients between the attention restoration level and three prosodic phonetic features

	Tempo	*Key*	*Inharmonicity*
ARL	0.8765	0.8679	0.7869

Discussion of the User Study

It developed an evaluation procedure for the restorative effects of an in-car VRE in this user study. As a baseline for evaluating restoration effects, traditional methods based on attention ratings and reaction tests were utilized. According to the results of the attention surveys, all participants considerably increased their attention capacity following the restorative experience. In the response test, they obtained moderate to poor restoration in terms of overall accuracy and reaction time on average. The objective assessment of attention shows no substantial gains over the subjective evaluation.

It offered a fresh assessment approach based on speech signal analysis in addition to these known metrics. The study did a correlation analysis after extracting short-time energy and zero-crossing rate in the time domain and MFCC in the frequency domain following speaker segmentation and discovered that these acoustic features were substantially connected with the aforementioned attention measurements. Additionally, classic machine learning techniques such as KNN and SVM can help to illustrate the link between generic speech characteristics and restorative attention metrics. Using these two classical methods and the aforementioned relative speech characteristics, encouraging early results in 2-level classification before and after VRE experience may be obtained. Based on these speech features, the KNN and SVM techniques can recognize distinct restorative attention stages in three-level categorization. Furthermore, prosodic phonetic elements including pace, key, and inharmonicity are linked to subjective judgments of the restorative impact. As a result, these speech features, like subjective assessments, can be used to identify and evaluate the restorative impact.

In terms of speaker segmentation, which is vital in the provided assessment system, it offered an artificial speech segmentation algorithm aimed at detecting the participants' voices during the restorative experience. However, human speech labeling and microphone placement can have an impact on speaker segmentation precision. On the one hand, correctly labeling participants' start and finish voices is difficult. The distance between the participants and the microphone, on the other hand, may have an effect on the manual label. This approach's long-term goal is to build a fully autonomous processing chain for evaluating attention levels based on audio data from interviews, which are always done as part of the research, or even audio from user interactions.

3.4 Summary

This chapter mainly focused on speech features analysis which includes time and frequency domain features, rhythmic speech features, and social speech features. More specifically, the time domain features (temporal features) are easier to extract and analyze in the “short-time” range, such as short-time energy, zero crossing rate, short-time auto-correlation, and so on. The frequency domain features (spectrum features) are produced by transforming the time-based signal to the frequency domain with the Fourier Transform, such as Mel-Frequency Cepstrum Coefficients(MFCCs), etc. Rhythmic speech features can sense the movement in speech, which is characterized by stress, tempo, and syllable amount. Social speech features can detect nonverbal verbal cues and predict behavioral outcomes in various social situations. These features can be used for speech recognition, speech emotion recognition, and speech personality recognition and these recognizing techniques can be embedded in the VIs system to improve user-VIs interaction.

Then, it offers a user study on measuring restoration effects in in-car VRE using voice features analysis. Although not a VIs example, it might be used in future autonomous driving VIs. In this user study, It proposes a unique attention restoration assessment approach that analyzes the speech signal of study participants (n=21) before and after they encounter an in-car restorative environment in VR. This is less intrusive and more accessible than other metrics such as attention scales or reaction tests. The results of this study reveal that specific temporal and frequency domain speech features, such as short-time energy and Mel Frequency Cepstral Coefficients (MFCC), are correlated with subjective and objective measures of attention. As a result, speech signal analysis can give a reliable indicator of attention and its restoration. An unsupervised speech segmentation technique based on a Wavelet Transform and Kullback-Leibler Divergence was also suggested to make speech signal analysis more practical.

Chapter 4

Speech Emotion Recognition

Speech emotion recognition plays a critically important role in VIs. As previously stated, emotion-aware VIs are able to provide a communication bridge between users and VIs while also improving the user's interaction experience through advancements in speech emotion recognition and synthesis. In general, speech emotion recognition includes speech signal pre-processing, speech feature extraction, speaker segmentation, and speech emotion recognition (see figure 4.1). Pre-processing comprises pre-emphasis, enframing, and windowing with frame sizes ranging from 10 to 30ms. In a subsequent stage, time- and frequency-domain information can be retrieved. Noise is removed from the speaker before a classification algorithm is employed to recognize the various emotional states. In this section, it mainly focuses on speech segmentation and speech emotion recognition. Furthermore, it also presents a user study about emotion recognition from continuous speech in classroom teaching.

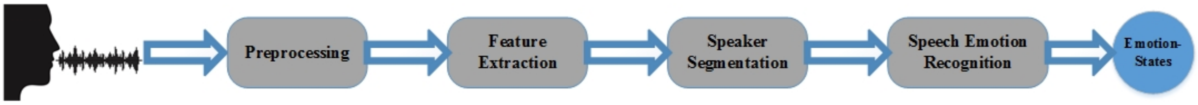


Figure 4.1: Processing steps needed for a speech emotion recognition system for continuous speech

4.1 speaker segmentation

Speaker segmentation is extremely crucial in VIs. In general, user-VI communication can be improved by integrating speaker segmentation techniques into the VIs and using high-efficiency speaker segmentation algorithms. Traditionally, speaker segmentation algorithms are divided into two types: supervised and unsupervised learning methods. An unsupervised method can recognize a speaker's voice without prior knowledge, for example using cepstral analysis for tracing unique voice segments [26]. Speech features, such as short-time energy, magnitude, zero crossing rate, and autocorrelation [152], can also support unsupervised speaker segmentation methods. Whereas the supervised method can be used when the speaker's speech information is known in advance and can be used for training. For instance, GMM [344], SVM [131], deep learning techniques such as DNN, CNN, and RNN/LSTM [24], can be used to recognize the speaker's voice during the conversation. The greater the training sample set used in the supervised learning method, the higher accuracy will be achieved. Yet, the most difficult aspect of this approach is gathering this large dataset. Improving the accuracy of unsupervised speech recognition offers another alternative. This section provides various instances of unsupervised speaker segmentation methods.

4.1.1 Unsupervised speaker segmentation

Traditional unsupervised approaches, such as HMM method [309], can segment a speaker's voice based on the difference in speech energy. This method can serve the purpose of separating the speaker's voice from other noise or silence, thereby identifying the segments of data, in which emotion will be recognized. As the limitation of the unsupervised method and the inconvenient training procedure, it presents an unsupervised speaker segmentation method as shown in figure 4.2. The method uses two stages: a Wavelet Transform and detection based on Kullback-Leibler-Divergence (KLD).

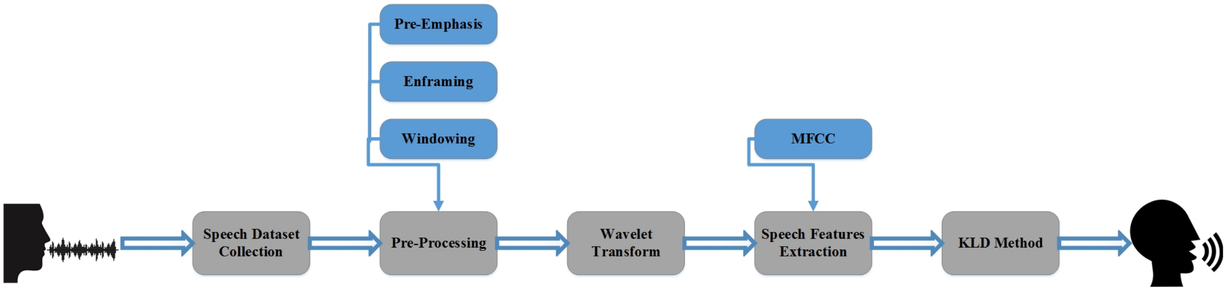


Figure 4.2: Processing steps for Speaker Segmentation in detail

For noise reduction, it implemented some pre-processing steps and a wavelet transform. The pre-processing consists of pre-emphasis, enframing, and windowing. In the pre-emphasis, it used a filter to emphasize higher frequencies in order to equalize the effect of the glottal source and energy radiation from the lips. Moreover, as speech signals are non-stationary, their characteristics change in very short time intervals. Enframing and windowing help to maintain a more steady state of the signal. In this system, it chose a frame size of 25ms and a frame step of 10ms. It then successfully used Daubechies (D4) wavelets [265] and its inverse transform to reduce the noise in the voice data. The Kullback-Leibler Divergence (KL-Divergence) can be utilized to reliably segment the speaker's speech from the whole voice data.

KL-divergence [342] is then used to find the change points between speakers and silence: The MFCC allows us to detect these change points since the symmetric Kullback-Leibler distance between two adjacent frames of MFCCs detects whether this frame is a change point between speech and non-speech. When all of the change frames have been identified, the K-means approach is used to categorize them and thereby segment the speaker's voice data.

4.1.2 Study Result and Evaluation

The unsupervised learning technique presented in the preceding section was used in two different scenarios. One is an interview environment, while the other is a classroom teaching context. The Elan[1] is a ground truth annotation software that can be utilized to manually label the speakers' voice segments. This section contains the results and analyses for these two scenarios.

The interview scenario in this user case was generated from the last stage of the user study. In general, the last stage of user research focuses on communication between participants and experimenters and they mostly poke at the experimental design and user experience. Participants' voices can be gathered during this procedure, and an unsupervised approach can capture participants' voices without training their vice before. Furthermore, as previously stated, participants' speeches can be used to recognize their emotions and study engagement. The following table shows the result of speaker segmentation in an interview scenario.

Participant ID number	*Accuracy*	*F1-score*	*Participant ID number*	*Accuracy*	*F1-score*
01	0.6323	0.7747	**06**	0.9976	0.9988
02	0.8672	0.9289	**07**	0.6222	0.6937
03	0.6562	0.7265	**08**	0.7210	0.8379
04	0.9366	0.9673	**09**	0.9215	0.9591
05	0.8680	0.9293	**10**	0.6482	0.7098

Table 4.1: Results for speaker segmentation in the interview context.

As shown in the figure 4.1, the average accuracy of speaker segmentation was about 79% and the average F1-score[2] [117] was above 0.77. Therefore, the unsupervised method based on wavelet transform and Kullback-Leibler divergence (WT-KLD) can be used to detect the speaker's voice in the absence of any training data. More specifically, wavelet transform can efficiently remove noise, while KL-Divergence can assist to segment the participants' voices. However, not all of the results are satisfactory, and some of them can attain incredible precision. On the one hand, the wavelet transform approach cannot eliminate all noise from the voice. On the other hand, MFCC may not be the sole element

[1] https://archive.mpi.nl/tla/elan

[2] https://deepai.org/machine-learning-glossary-and-terms/f-score

associated with human voice difference, and there may be additional aspects that can be utilized to discern various people's voices.

In this user study, the classroom teaching scenario is primarily focused on the teacher's voice. Traditionally, the majority of speech in the classroom teaching environment comes from instructors' voices, and hence teachers' voices play a crucial role in classroom teaching. A video camera can capture instructors' appearances as well as record their voices. Students occasionally engaged in debates with the speaker during the lecture. The teacher's voice dominates the audio data, but there is additional noise such as whispering and the sound of the recording computer. The following table shows the result of speaker segmentation in a classroom teaching scenario.

Participant ID number	*Time*	*Accuracy*	*F1-score*	*Participant ID number*	*Time*	*Accuracy*	*F1-score*
A01	45min	0.7642	0.8664	**C01**	45min	0.9598	0.9795
A02	45min	0.6968	0.8213	**C02**	45min	0.7906	0.8831
B01	45min	0.8782	0.9352	**D01**	45min	0.9565	0.9778
B02	45min	0.8801	0.9362	**D02**	45min	0.9389	0.9685

Table 4.2: Results for speaker segmentation in the classroom teaching context.

As depicted in table 4.2, A-D refers to the 4 teachers, whereas 01 and 02 indicate the various lectures. The average accuracy of the speaker segmentation was about 86% and the average F1-score was above 0.92. Table 4.2 shows the results for each recording, which exhibit a certain variation between teachers. However, in general, this demonstrates that MFCC was an appropriate speech feature for recognizing the instructors' voices and that KL-Divergence based on MFCC could effectively distinguish and segment the speech signal, with noise reduction using D4 Daubechies wavelets. The accuracy (0.9598) and F1-score (0.9795) are the greatest in the audio data of recording C01. Despite the fact that the students' voices were dominant at the end of the lesson, this recording delivered an excellent outcome. For recording A02, the accuracy (0.6968) and F1-score (0.8213) are the lowest. An ear assessment of this audio indicated that it featured portions in which the students' and teachers' voices blended, making noise reduction and segmentation nearly difficult.

4.2 Speech Emotion Recognition

In human-human interactions, various modalities, such as facial expressions, body language, and speech, are used to exchange information [161]. However, recognizing human emotional states is considered an extremely difficult task in most dynamic scenarios, such as driving, and thus continues to face challenges [164]. With the advancement of sensors and their widespread use, multiple sensors can be utilized to capture and measure emotion. The two most commonly used data in the field of research [69] are audio data (speech) and video data (facial expressions). As a result, speech emotion recognition has received a lot of attention in recent years, and it had a big impact on the field of Human-Computer Interaction (HCI) research as previously mentioned. Speech emotion recognition has applications ranging from academia to industry, including psychiatric diagnosis, smart productions, lie detection, intelligent call centers, educational software, and so on [98]. In order to achieve high accuracy in speech emotion recognition, research is involved in several areas, including denoising speech signals in real-life scenarios, as well as finding the most effective feature set and classifiers.

In general, numerous approaches, such as conventional machine learning methods, can be applied to recognize speech emotions. Hidden Markov models (HMM) and short-term spectral characteristics were utilized by Lee et al. [187] and Schuller et al. [288] to identify speech emotions. They discovered that a model trained using spectral aspects of vowel sounds outperformed a model built with prosodic features. Other methods that were also employed for emotion recognition include Support Vector Machines (SVM), the Kernel Extreme Learning Machine (KELM), Convolutional Neural Networks (CNNS) as well as Recurrent Neural Networks (RNNs) [8]. Currently, deep learning approaches are the primary focus of researchers in this field. Su et al. [307] imposed a graph attention mechanism on a gated recurrent unit network to improve the performance of SER. Dissanayake et al. [86] proposed another SER method using an autoencoder-based CNN and a Long-Short-Term Memory(LSTM) network structure. Moreover, SER based on self-attention [314], interaction-aware attention networks [343], and augmenting generative adversarial networks (AGAN) [182] can also effectively improve the recognition performance. This section focuses on speech emotion recognition in considerable depth.

4.2.1 Emotional Speech Databases

Emotional speech datasets are significant in speech emotion recognition, since building and choosing a good database for speech emotion recognition tasks is vital for assessing speech recognition performance. In general, emotional speech datasets are divided into three distinct types[97, 274] - natural database, acted database, and elicited database.

Natural databases are built on spontaneous speech from real-life conversations such as call centers and television video clips. This is the most genuine database, with the highest level of naturalness. However, one significant shortcoming of this strategy is the issue of copyright and privacy. Another issue is subjectivity in emotion labeling; various human evaluators may judge different aspects of the same statement, making it difficult to describe the audio emotion.

Although natural databases are the most optimal for use in speech emotion recognition, due to data collecting limitations, several studies in the field of SER have asked professional actors or expressive people to replicate predetermined emotion states. Because it is easier operable with simple equipment, this is the most widely utilized strategy. Furthermore, there is no need for emotion labeling effort. However, the emotional naturalness of such speech data is dependent on the ability of actors to imitate. As a result, as compared to the natural database, the emotional components in the acted database are frequently inflated and do not reflect true feelings.

Elicited speech corpus, rather than definitions of emotion state, is where emotions were induced, i.e. performers were asked to improvise conversation in hypothetical settings tailored to elicit certain emotions. This technique is also more operable and can get more natural data than a simulated approach, although it has significant drawbacks. There is no assurance that performers will be able to elicit the desired feeling. Furthermore, not all emotions may be available, and if the speaker is aware that they are being recorded, the emotion conveyed by speakers can be fabricated.

As shown in table 4.3, there are various emotional speech databases based on the three types stated previously. A database of authentic and realistic emotional speech can assist in improving speech emotion detection. However, the majority of current emotional speech databases are acting emotions, and the database size is insufficient. Recognizing the natural speech emotion based on these acted emotional speech datasets is challenging. In addition, languages have an important influence on speech emotion recognition. Currently, the majority of languages offered by such databases are in English. The other major database is in German. This section will provide three types of emotional speech databases.

Elicited Database - English - IEMOCAP According to IEMOCAP [49], four criteria are required in the preparation of a speech corpus: scope (number of speakers, emotional classes, language, etc.), naturalness (acted versus spontaneous), context (in-isolation versus in-context), and descriptors (linguistic and emotional description). They emphasize the need of having appropriate databases with real emotions recorded during conversations rather than monologues. Furthermore, information concerning physiological signals, the aim of speech emotion recognition, and the number of emotion states should be considered to improve the performance of the speech emotion recognition system. The IEMOCAP database was created by recording ten actors in dyadic sessions with markers on their faces, heads, and hands, which provide detailed information about their facial expressions and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts as well as improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration, and neutral state).

Acted Database - English - SAVEE SAVEE database has been recorded as an important prerequisite for the development of an automatic emotion recognition system. The database includes seven emotions, namely anger, disgust, fear, happiness, neutral, sadness, and surprise [150]. Recordings from 4 male (identified as DC, JK, JE, KL) actors in 7 different emotions, 480 British English utterances in total, are contained in the database. Each emotion class contains 15 samples except the neutral class contains 30 samples and a total of 120 samples per class available.

Table 4.3: Current available emotional speech databases [97, 310, 274]

Corpus	Access	Language	Type of database	Size	Emotions
IEMOCAP[49]	Public and free[1]	English	Acted, Elicited	12 hours, 10 actors*5 Sessions*8 Wavs	Nl, Hs, Sd, Ar, Sue, Fr, Dt, Fn, Ed
Berlin Emotional Database(EMO)[48]	Public and free[2]	German	Acted	800 utterances, 10 actors	Ar, Jy, Sd, Fr, Dt, Bm, Nl
Danish Emotional Database(DES)[99]	Public with license fee[3]	Danish	Acted	4 actors, 2 words + 9 sentences+ 2 passages	Ar, Jy,Sd,Sue,Nl
SUSAS[127]	Public with license fee[4]	English	Acted, Natural	16,000 utterances, 32 actors	Four stress levels
FERMUS	Public with license fee[5]	German, English	Automotive environment	2829 utterances,13 actors	Ar, Dt, Jy, Nl, Sd, Sue
OMG Emotion Dataset[28]	Public with License [29]	German, English	Natural	178 Videos, 2725 utterances	Ar, Dt, Fr, Hs, Nl, Sd, Sue
eNTERFACE	Public and free	English	Acted	42 Male, 34 Female	Ar, Dt, Fr, Jy, Sd, Sue
SALAS	Public with license	English	Elicted	20 Subjects	Wide range
Smartkom	Public with license	German	Natural	Total=224, Time=4.5min/p	Nl, Jy, Ar, Sue, Hes, Pg

Abbreviations for emotions: Anger: Ar, Anxiety: Any, Boredom: Bm, Contempt: Ct, Disgust: Dt, Despair: Dr, Excited: Ed, Elation: En, Fear: Fr, Frustration: Fn, Happiness: Hs, Helplessness: Hes, Interest: It, Joy: Jy, Neutral: Nl, Panic: Pc, Pride: Pe, Pondering: Pg, Surprise: Sue, Sadness: Sd, Shame: Se.

Other Abbreviations: H/C: Hot/Cold.

[1] Speech Analysis and Interpretation Laboratory(SAIL), University of Southern California(USC), USA.

[2] Institute for Speech and Communication, Department of Communication Science, the Technical University, Germany.

[3] Department of Electronic Systems, Aalborg University, Denmark.

[4] Linguistic Data Consortium, University of Pennsylvania, USA.

[5] FERMUS research group, Institute for Human-Machine Communication, Technische Universität München, Germany.

Acted Database - German - The Berlin Emotional Speech Database (Emo-DB) Emo-DB consists of about 800 sentences spoken by 10 actors. The German speech samples contained are related to seven emotions: anger, disgust, fear, happiness, boredom, neutral, and sadness [48]. The material was evaluated in an automated listening test and each utterance was judged by 20 listeners with respect to the recognisability and naturalness of the displayed emotion. Because of the high quality of its recording as well as its free availability, it thus has served as the basis for numerous studies and can be used for each gender separately and combined.

4.2.2 Machine Learning Methods for Speech Emotion Recognition

As aforementioned, There are several approaches for recognizing emotions in speech. A wide range of methods, including HMM, K-Nearest Neighbors (k-NN), SVM, neural networks (NN), and many others, are used in speech emotion recognition research. There is yet to be unanimity on which method is best suited for all emotion categorization tasks since every algorithm has both advantages as well as limitations. In this section, it will provide an overview of different methods that are important to emotion categorization. Their limitations are also explained.

k-Nearset Neighbors (k-NN) is a supervised learning technique used in machine learning. k-NN is widely utilized in a range of applications, including economic forecasting, because of its simplicity and robustness. In general, given an unlabeled sample X, the k-NN classifier searches the training data for the K nearest neighbor samples and classifies the sample X with the class label that appears the most frequently in the region of k time [10]. Nonetheless, k-NN has limitations when compared to alternatives. Because it works so slowly, k-NN is called a lazy algorithm. It also takes a long time to compute, which might be a concern with huge datasets. Another key issue in k-NN [34] is determining the value of parameter **k**.

Gaussian Mixture Model (GMM) is a probabilistic model for density estimation that employs a convex mixture of multivariate normal densities [330]. GMMs are deemed ideal for speech emotion for global feature extraction because they are extremely efficient in modeling multi-modal distributions, and their training and testing conditions are considerably less required compared to a typical continuous HMM [90]. In terms of design difficulties, the most critical to be addressed for GMM is the definition of the optimum number of Gaussian components [14].

Support vector machine (SVM) is the state-of-the-art classifier model in speech emotion recognition. An SVM is a supervised machine learning algorithm and is mostly used in classification tasks. The core idea of SVM is to find the best hyperplane which separates the data of one class from those from another class [298]. SVM offers advantages over

GMM and HMM including the global optimality of the training algorithm, and the existence of excellent data-dependent generalization bounds [97]. Nevertheless, their handling of non-separate cases is heuristic. Since there is no systematic way of choosing the kernel functions, and hence, the separation of transformed features is not always guaranteed. In reality, perfect separation of the training data is not recommended in speech emotion recognition, for the purpose of avoiding over-fitting. With respect to review, SVM is widely used in the application field of speech emotion recognition [298, 15, 243]. For example, Hadhami Aouani et al. [15] used standard SVM and proposed DSVM in combination with the use of deep learning methods, in their study. As a result, they achieved an accuracy of 68.25% and 73.01% with 39 MFCC coefficients SVM and 65 MFCC coefficients SVM, respectively, for the speaker-independent classification.

Deep Neural Network (DNN) is a conventional multi-layer perceptron (MLP) with many hidden layers [347], and it is capable of learning high-level representation from raw features and effectively classifying data [125]. It is believed that DNNs can achieve a good performance in machine learning tasks (e.g., speech recognition) with sufficient training data and appropriate training strategies. Thus classification models which are based on DNNs are proposed recently. Kun Han et al. [125] proposed their approach utilizing DNN to estimate the emotional state of an utterance. As for the result of their experiment, the DNN-based approaches outperform the other two, namely, using HMM and SVM respectively with 20% relative accuracy improvement for both unweighted ($0.402 \rightarrow 0.482$) and weighted ($0.451 \rightarrow 0.543$) accuracy. The experimental results indicate that their approach based on DNN significantly promotes the performance of emotion recognition from speech signals. Thus it is promising to use neural networks in emotion recognition tasks.

Sparse Attention Mechanism in neural network for speech emotion recognition[141] can effectively improve the recognition performance. Since the audio signals are time-series data, recurrent neural networks with attention mechanisms, such as Attention Long Short-Term Memory(Attention-LSTM)[348], can offer alternatives. Sparseness measurement can quantify how much energy of a vector is packed into only a few components[140]. Attention mechanism with sparse analysis can provide more accurate predictions because of emotionally salient parts in a sentence[222]. Thus, using sparse Attention-LSTM for speech emotion recognition can effectively improve recognition performance.

4.2.3 User Study - Speech Emotion Recognition in Different Languages

In this user research, SVM, a traditional machine learning approach, and DNN, a deep learning method, are employed to recognize distinct emotions in the SAVEE and Emo-DB databases. The SAVEE database is solely for men, and it was tested on four participants DC, JK, JE, and KL. The audio performed by DC, JK, and JE was recorded at a sample rate of 44.1 kHz. To keep the sampling rate consistent, the sampling rate in the study was set to 44.1 kHz while loading these audio files using LibROSA [3].

A combined model containing speech samples from both the SAVEE and Emo-DB databases is also generated and evaluated using SVM and DNN methods on the English, German, and combined databases. Male and female participants are kept in Emo-DB for experimentation. It operated in a gender-independent situation, which means that the recognition system can combine both male and female individuals. Speech utterances in speech corpora are separated into two groups: 80% for training and 20% for testing.

Six emotional states are identified in this study: Angry, Disgust, Fear, Happy, Neutral, and Sadness. For feature extraction, a collection of spectral characteristics such as Mel-frequency Cepstral Coefficients (MFCC), Zero-crossing rate, and others are used. The experimental findings show that the suggested technique is effective in recognizing emotions, with a recognition rate of about 90% using both English-based corpora and combined corpora, testing on both English and combined corpora. Using both training models, the identification accuracy (about 45%) on the German corpus is far from flawless.

SVM Method

As stated in previous sections, the main principle of SVM is to discover the optimum hype plane that can distinguish data from one class from data from another. SVM transforms the original input feature space into a high-dimensional feature space by applying a kernel function where the data may be linearly divided, allowing the data to be classified [298]. This method makes use of a variety of kernel functions, including linear, nonlinear, radial basis function (RBF), and sigmoid. RBF can be utilized for speech emotion detection not only because it localizes and finishes responses throughout the whole x-axis, but also because it acts well when dealing with multiple features. The kernel function and RBF function are respectively represented by formula 4.1 and formula 4.2.

$$kernel(x, y) = (x, y) \tag{4.1}$$

$$kernel(x, y) = e^{\frac{-||x-y||^2}{(2\sigma^2)}} \tag{4.2}$$

Parameter tuning is another intricate task in SVM. To achieve good results in SVM, parameter tuning was performed in this algorithm. There are basically two parameters in SVM with RBF as kernel functions, namely C and γ. A common way for parameter tuning is to loop over all the parameters defined and run all the combinations of parameters. However, this approach is time demanding and is not advised for huge data sets. GridSearchCV [297] is a straightforward technique to tune parameters at a low cost. The best parameter pairings can be found by simply importing Python's GridSearchCV library, creating a parameter grid (usually with values in [0.01, 0.1, 1, 10, 100]), and providing the algorithms. In this way, the model with the best parameter pairings can be produced.

[3] https://librosa.org/doc/latest/index.html

Table 4.4: The global SER results obtained from English and combined databases using SVM.

Test data Training data	English(SAVEE)	German(Emo-DB)	Combined
English(SAVEE)	0.889	0.470	0.712
Combined(SAVEE+Emo-DB)	0.951	0.517	0.816

Table 4.5: The SER results obtained on trained in each emotion class by the English-based training model test on different speech samples using SVM.

Test data Emotions	English(SAVEE)	German(Emo-DB)	Combined
Angry	0.84	0.39	0.74
Disgust	0.98	0	0.91
Fear	0.97	0.67	0.88
Happiness	0.93	0.67	0.81
Neutral	0.93	0.24	0.871
sadness	1.00	0.80	0.89
Average	0.942	0.462	0.85

The table 4.4 displays the global recognition results acquired from English and the combined data sets using SVM. The accuracy of the model utilizing English training databases is lower than that of the combined training databases. Furthermore, whether SAVEE or combination datasets are used as training data, the accuracy of combined test data is lower than the SAVEE dataset but higher than the Emo-DB dataset, when SVM is utilized.

The table 4.5 shows the SER results acquired on trained in each emotion class by the English-based training model test on various speech samples using SVM. In particular, the training data in this study are solely from the SAVEE database. The highest accuracy can be obtained from the SAVEE test database and the lowest one is from the Emo-DB database. There are two key causes behind these outcomes. On the one hand, the Emo-DB database has fewer users than others. SVM, on the other hand, is not an effective recognition approach for speech emotion recognition.

The table 4.6 displays the results of the combined languages-based training model test on different speech samples using SVM. The training data in this study, in particular, are drawn entirely from the merged database. A similar result was reached as shown in table 4.5. In comparison to table 4.5, the accuracy in table 4.6 is better when SVM is used.

Table 4.6: The SER results obtained on trained in each emotion class by the combined languages-based training model test on different speech samples using SVM.

Emotions \ Test data	English(SAVEE)	German(Emo-DB)	Combined
Angry	0.88	0.50	0.69
Disgust	1.00	0	0.95
Fear	1.00	0.12	0.86
Happiness	1.00	0.40	0.85
Neutral	0.94	0.18	0.86
sadness	1.00	0.67	0.92
Average	0.97	0.312	0.855

Table 4.7: The global recognition rates obtained from English and combined model for the system with combined feature set using DNN.

Training data \ Test data	English(SAVEE)	German(Emo-DB)	Combined
English(SAVEE)	0.989	0.412	0.772
Combined(SAVEE+Emo-DB)	0.923	0.407	0.905

DNN Method

As previously stated, it is also feasible to train a DNN to predict emotional state. DNN model optimization is substantially more challenging than SVM model optimization. Because a DNN is essentially a feed-forward network with numerous hidden layers between its input and output, establishing the number of hidden layers, as well as the number of input and output units, is critical. The number of input units is proportional to the size of the feature vector. The output layer's number is adjusted to the maximum number of emotions N by employing a softmax output layer. Unlike the input and output layers, the number of hidden layers in neural networks is not fixed. According to Jeff Heaton [132] in his book, 'the appropriate number of the hidden layer is generally between the size of the input and number of the output layers,' which serves as an empirically-derived rule-of-thumb for the majority of situations. In this scenario, the number of concealed layers is set to 4. Dropout was used in the technique to overcome the frequent issue in most neural networks, namely over-fitting. The term Dropout refers to dropping out units (both hidden and visible) in a neural network or disregarding neurons during the training phase of a random collection of neurons. In the DNN, a range of Dropout values ranging from 0.0 to 0.9 was employed, and the optimal value of Dropout(0.2) was established by cross-validation.

Table 4.8: The accuracy rates obtained on trained in each emotion class by the English-based training model test on different speech samples using DNN.

Test data Emotions	English(SAVEE)	German(Emo-DB)	Combined
Angry	0.99	0.33	0.70
Disgust	0.98	0	0.75
Fear	0.98	0.20	0.83
Happiness	1.00	0.14	0.77
Neutral	0.96	0.29	0.87
sadness	1.00	0.33	0.81
Average	0.985	0.215	0.788

Table 4.9: The accuracy rates obtained on trained in each emotion class by the combined languages-based training model test on different speech samples using DNN.

Test data Emotions	English(SAVEE)	German(Emo-DB)	Combined
Angry	0.96	0.52	0.95
Disgust	0.91	0.15	0.96
Fear	0.92	0.17	0.85
Happiness	0.90	0.33	0.86
Neutral	0.89	0.30	0.95
sadness	0.84	0.59	0.90
Average	0.903	0.247	0.912

Compare to the table 4.4 4.7, the best recognition rate (98.9%) was obtained with the SAVEE database test on English data utilizing DNN as a classifier model when compared to the SVM approach. The results show no significant difference in performance between SVM and DNN. As a result, determining which classification model is superior in speech emotion identification is difficult. The main reason is that parameter adjustment, such as choosing the number of hidden layers, is far more difficult in a DNN than in an SVM. As a result, the model is not the best model for a DNN. A mix of algorithms then appears to be interesting for further research.

However, the recognition results achieved in the German corpus as test data are far from flawless, with a retention rate of roughly 45%, which is barely half of the retention rate in English. For one thing, the lack of German speech samples leads to an imbalance of English and German data, which further influences the final performance; and for another, a language model related to speech recognition is also a case where syntax and semantics are important, but only speech features are considered in this case.

Table 4.5, table 4.6, table 4.8, and table 4.9 describe the recognition accuracy in each emotion class based on a single English-based and combined training model using SVM and DNN, respectively. The total recognition results are also included in the tables below.

It is clear that the emotion of Disgust was nearly not identified in the German speech corpus, whereas the other five emotions were adequately detected. Because the emotional utterances for creating training models were produced by trained actors, they are false emotions, there may be differences between the acted emotions and genuine emotions. The results for six emotions also show that SVM performs better in terms of recognizing emotions in the German corpus, with total recognition rates of 46.2% and 31.2%.

4.3 User Study: Teachers Emotion Recognition

With the advancement of AI and voice emotion detection, emotion-aware VIs could be employed in the classroom. It is advantageous to classroom teaching to anticipate the teacher's or students' emotions in advance. On the one hand, teachers can modify their instruction according to the student's speech and emotions. Students can also provide comments based on the emotions shown by the teachers throughout their speeches. On the other hand, emotion-aware VIs can improve the teaching experience, if the technology is entrenched in the classroom, especially in remote teaching. This part contains a user study on teachers' speech emotion recognition in classroom teaching.

The teacher's emotional state has a substantial impact on students' willingness to learn and achievement. It can be recognized automatically for feedback or assessment utilizing face expression recognition from video records or physiological monitoring. However, since microphones are commonly utilized in classrooms and lecture halls, the teacher's recorded continuous speech signal provides a readily available source for emotion identification. For such a classroom context, this study proposes an emotion detection system that comprises speaker segmentation and speech emotion identification. In the thesis, it employs a wavelet transform and Kullback-Leibler divergence (WT-KLD) for speaker segmentation, with promising results. Moreover, this work can also achieve promising detection accuracy on recordings from a real classroom using a deep neural network based on the sparse attention method.

4.3.1 Study Background

The mood of the teacher is quite essential in the interaction between teachers and pupils. Frenzel et al. [108, 316, 46] that teachers' emotions influence the quality and success of teaching and learning. Diverse teacher emotions have different effects on pupils [280, 58] and can have a substantial influence on student behavior, particularly motivation in the classroom [122]. Particularly positive emotions from teachers help to increase students' attention and engagement [226]. Recognizing and reflecting on the conveyed emotion may therefore raise instructors' awareness and benefit a successful engagement with their pupils, resulting in a better learning outcome.

Surveys are currently the most established feedback channel, but in addition to the time delay they introduce, they may also suffer from subjectivity and memory effects. In real-time, emotions can currently be detected from facial expressions or certain physiological sensors. Chen et al. [60] obtained facial emotions using a camera array and a Multi-Layer Perceptron network (MLP). Park et al. [244] explored teachers' emotions communicated through facial expression in an immersive virtual teaching simulation. Current methods in facial emotion recognition include hybrid deep-learning approaches, combining for example a convolutional neural network (CNN) and long-short-term memory (LSTM) [169].

Physiological signals such as electroencephalogram (EEG), temperature (T), electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), and respiration (RSP) can also help to recognize and analyze different emotions [296]. For example, Song et al. [304] proposed a novel method using feature extraction and emotion classification from a multichannel EEG based on dynamical graph convolutional neural networks (DGCNN). However, while camera recordings of the face can plausibly be used, EEG signals in a classroom setting seem hardly feasible.

On the other hand, Pentland et al. [254, 253] found that humans actually detect very subtle "secret signals" in other humans' voices. A combination of automatic speech segmentation and speech emotion recognition (SER) based on the analysis of prosodic/acoustic features [212] and pitch or formant analysis [284] for capturing such social speech signals [251] is therefore proposed. For detecting the teacher's emotions in class from their live speech signals, it needs robust methods for speaker segmentation [309] and speech emotion recognition [97, 222]. In this study, the WT-KLD method was proposed to segment teachers' voices, and the LSTM networks with an attention mechanism were utilized to distinguish teachers' emotions.

4.3.2 Implementing Emotion Recognition from Continuous Speech

The implemented system contains two main components: speaker segmentation and speech emotion detection. The teacher's voice is isolated from quiet or background noise in the first stage using the WT-KLD approach, which is detailed below. The emotion is then recognized by deep learning methods using MFCC, ZCR, roll-off frequency, spectral centroid, and spectral bandwidth. Figure 4.3 depicts a schematic diagram of the processing processes in this system.

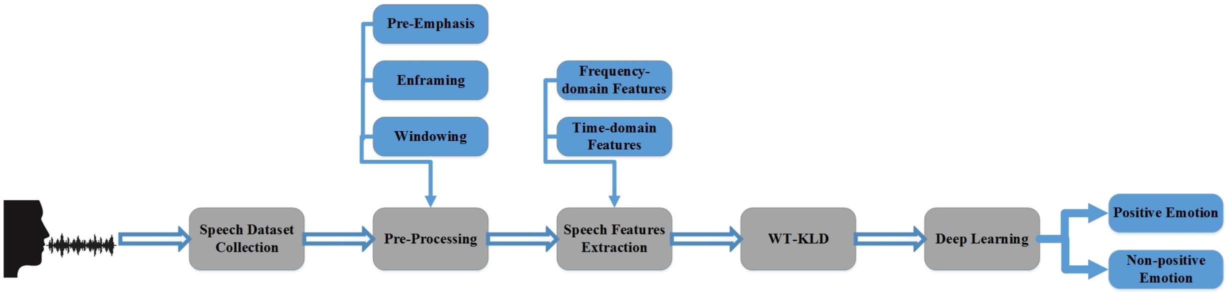


Figure 4.3: Global processing steps in the implemented system

Speaker Segmentation Speaker segmentation in this system serves the purpose of separating the teacher's voice from other noise or silence and thereby identifying the segments of data, in which emotion will be recognized. Since this system did not intend to train the system on a specific voice, it used an unsupervised method, as demonstrated in figure 4.4. The procedure is divided into two stages: a Wavelet Transform and detection based on Kullback-Leibler-Divergence (KLD).

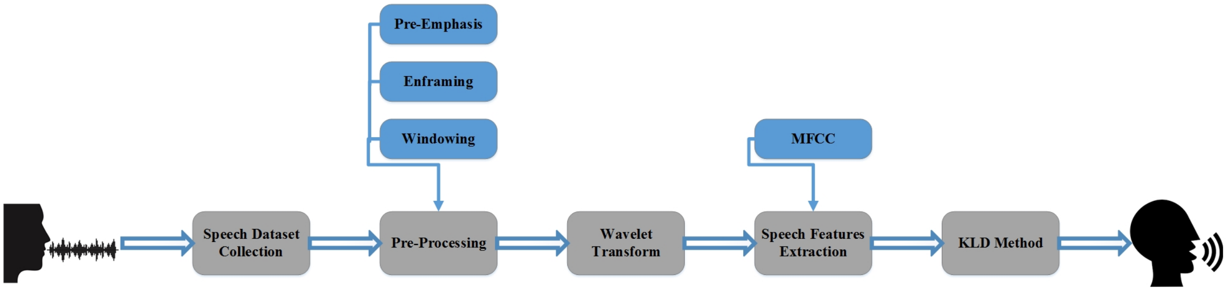


Figure 4.4: Processing steps for Speaker Segmentation in detail

It used certain pre-processing procedures and a wavelet transform to reduce noise. Pre-emphasis, enframing, and windowing are all part of pre-processing. It uses a filter to highlight higher frequencies in the pre-emphasis to balance the influence of the glottal source and energy radiated from the lips. Furthermore, as speech signals are non-stationary, their properties vary in relatively small time intervals. Enframing and windowing assist in maintaining the signal in a more stable form. It picked a frame size of 25ms and a frame step of 10ms for this system. It then employed Daubechies (D4) wavelets [265, 338] and its inverse transform to successfully remove noise in the audio data. The Kullback-Leibler

Divergence (KL-Divergence) can be utilized to reliably segment the teacher's speech from the entire lecture voice data.

KL-divergence [342] is then applied to determine the change points between speakers and silence: The MFCC is allowed to detect these change points since the symmetric Kullback-Leibler distance between two adjacent frames of MFCCs detects whether this frame is a change point between speech and non-speech. When all of the change frames have been identified, the K-means algorithm is used to categorize them and thereby segment the teacher's voice data.

Speech Emotion Recognition In the next stage, it then analyzes the speech segments to identify emotions. Fortunately, it is only concerned with positive emotions in this target domain, so it can simply utilize a binary classification for positive or negative parts. It employs time-domain features such as STE, ZCR, pitch, and formant, as well as frequency-domain features such as MFCC(see figure 4.5).

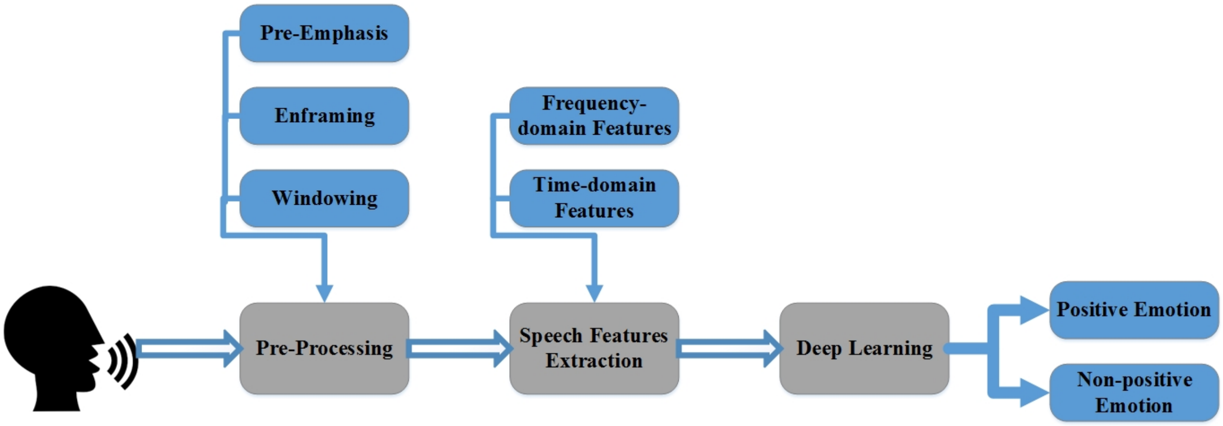


Figure 4.5: Processing steps used for speech emotion recognition

The Support Vector Machine (SVM) method by Vanik and Cortes [329] is often used in SER systems. Alternatively, Gaussian Mixture Models (GMM) describe the distribution of speech feature parameters by a weighted linear combination of Gaussian probability density functions, and can also be used to detect different emotions. Deep neural networks, particularly feed-forward neural networks with more than one hidden layer, have also been used to identify different emotions more accurately [322]. Moreover, Bi-LSTM based on a sparse attention mechanism can also be utilized to recognize different emotions. It can overcome the shortage of DNNs and achieve better performance.

Sparse Attention LSTM In general, DNN for speech emotion recognition based on different speech features can effectively achieve accurate recognition. However, a silent frame or low emotion expression frame in speech may present a negative impact on speech emotion recognition. Thus, Bi-LSTM based on sparse attention mechanism(see figure 4.6), capable of reducing silent frame and low emotion expression frame[222] can provide efficient emotion prediction.

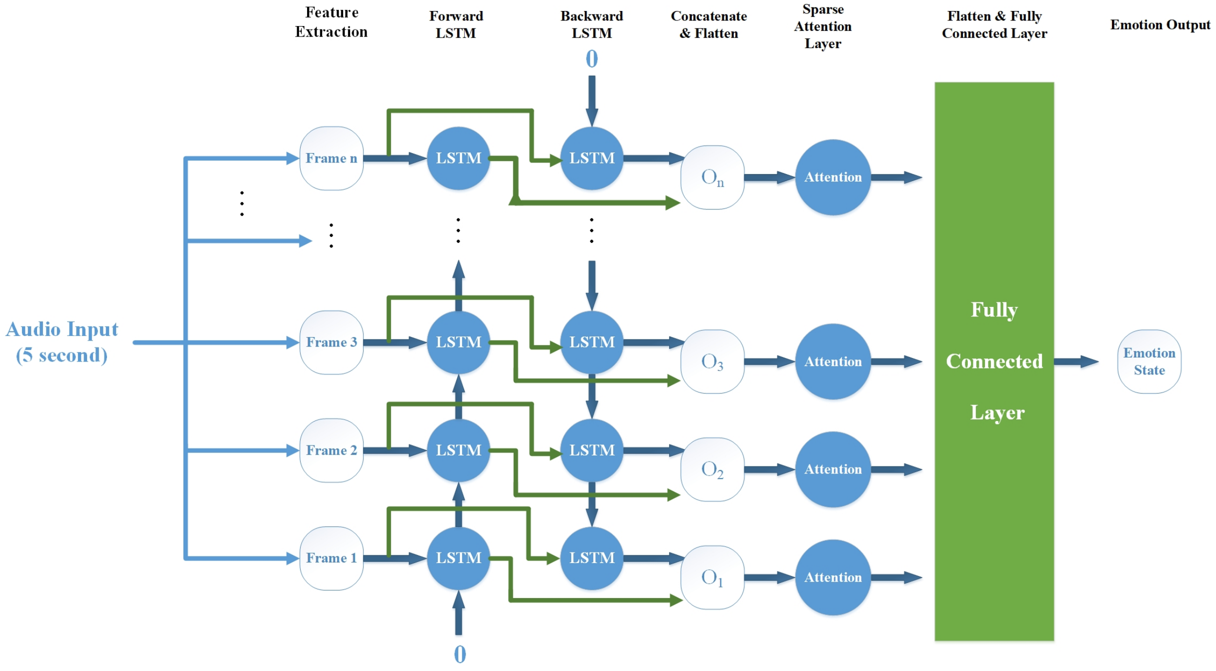


Figure 4.6: The structure of Sparse Attention-Bi-LSTM

As illustrated in figure 4.6, the 5-second sentence will be regarded as input voice data. Features extraction and analysis will be conducted after input of voice data and preprocessing. Bi-LSTM can be effectively handled the timing of voice data and it can also learn much information from LSTM processing. The output from bi-LSTM after concatenating and flattening can represent the different frames of emotion information and this information may exist in low emotion expression frames. Combing sparse based on frame information and attention mechanism can, more efficiently, eliminate the silent frames and low emotion frames. The fully connected layer is based on feed-forward neural networks and it can achieve the final emotion state.

4.3.3 Study Results

Data Collection For this experiment, it collected realistic data from 4 teachers (2 female), each in 2 (English-speaking) classroom lectures of about 45 minutes (a total of 6 hours). A video camera filmed their faces and also recorded their voice. During the lecture, students occasionally engaged in discussions with the teacher. In the audio data, the teacher's voice is most prominent, but noise such as whispering and the sound of the recording computer is also contained. it then used the Elan[4] annotation software to manually label the teachers' voice segments and their facially expressed emotion as a ground truth.

Results and Analysis For measuring the accuracy of this system, it split the audio data into training data and testing data at a ratio of 4:1. it then used 5-fold cross-validation to evaluate the final performance. Accuracy is also described using the F1-score[5].

Table 4.10: Results for the speaker segmentation: A-D refers to the 4 teachers, while 01 and 02 represent the different lectures.

Name	*Time*	*Accuracy*	*F1-score*
A01	45min	0.7642	0.8664
A02	45min	0.6968	0.8213
B01	45min	0.8782	0.9352
B02	45min	0.8801	0.9362
C01	45min	0.9598	0.9795
C02	45min	0.7906	0.8831
D01	45min	0.9565	0.9778
D02	45min	0.9389	0.9685

Speaker Segmentation The average accuracy of the speaker segmentation was about 80% and the average F1 score was above 0.85. Table 4.10 shows the results for each recording, which exhibit a certain variation between teachers. However, in general, this confirms that MFCC was a suitable speech characteristic for detecting the teachers' voice and that KL-Divergence based on MFCC could effectively recognize and segment the speech signal, using D4 Daubechies wavelets for noise reduction.

In the audio data of recording C01, the accuracy (0.9598) and the F1-score (0.9795) are the highest. Although this recording contained the students' voices quite prominently towards the end of the class (see figure 4.7 left), it still produced this good result. The accuracy (0.6968) and F1-score (0.8213) for recording A02 are the lowest. Figure 4.7 right shows this recording and an inspection by ear revealed that it contained parts, in which the students and the teacher's voices mixed and made noise elimination and segmentation virtually impossible.

Speech Emotion Recognition As shown in tables 4.11, 4.12,4.13 and 4.14, different algorithms in combination with different features achieved different accuracies. The comparison shows that in general, the DNN and sparse attention-Bi-LSTM method performed

[4] https://archive.mpi.nl/tla/elan

[5] https://deepai.org/machine-learning-glossary-and-terms/f-score

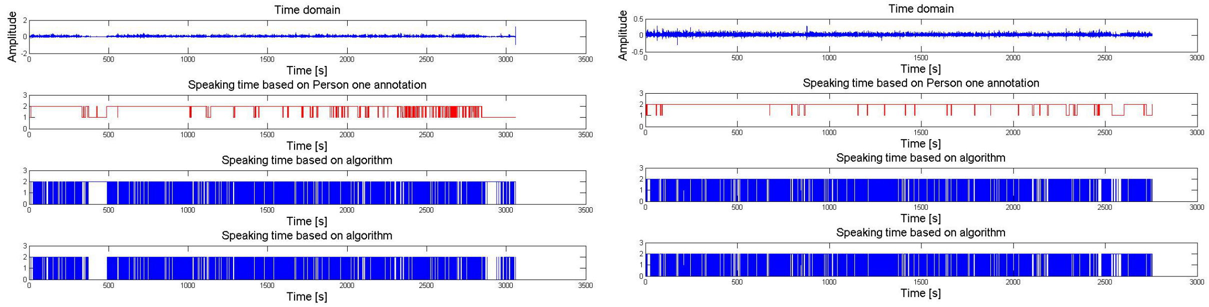


Figure 4.7: Result of the speaker segmentation in recording C01 (left) and A02 (right)

better than the traditional machine learning methods SVM and GMM. Table 4.11 shows that SVM can provide at least comparable results in the best case, but at a high computational cost due to the use of kernel functions and a longer time for training.

Table 4.11: Experimental results for the SVM classifier with different combinations of MFCC, ZCR, and Short term energy.

Group	*Feature Combination*	*Accuracy*	*F1-score*
G1	ZCR+Energy	0.75	0.53
G2	MFCC	0.83	0.69
G3	MFCC + ZCR + Energy	0.87	0.76

Table 4.12: Experimental results for the GMM classifier with different combinations of MFCC, ZCR, and Short term energy.

Group	*Feature Combination*	*Accuracy*	*F1-score*
G1	ZCR+Energy	0.80	0.60
G2	MFCC	0.76	0.52
G3	MFCC + ZCR + Energy	0.78	0.62

The recognition rate of GMM in table 4.12 ranges from 0.52 to 0.80 and the accuracy is the lowest among the three methods. Since GMM is an unsupervised learning algorithm based on probability estimation, the existing emotion labels in the training data were not used during training. Moreover, this method is based on the assumption that every feature space of the training data is independent and identically distributed, which is not true for the first 13 MFCC extracted. The accuracy for DNN (see table 4.13) was highest, ranging from 0.85 to 0.88. From these three tables, it can also see that the prosodic features ZCR and energy perform slightly worse than the spectral feature MFCC. The best accuracy is achieved using the DNN and combining prosodic and spectral features. Moreover, the accuracy of GMM in table 4.12 shows the highest increase of more than

Table 4.13: Experimental results for the DNN classifier with different combinations of MFCC, ZCR, and Short term energy.

Group	*Feature Combination*	*Accuracy*	*F1-score*
G1	ZCR+Energy	0.85	0.59
G2	MFCC	0.88	0.74
G3	MFCC + ZCR + Energy	0.88	0.75

Table 4.14: Experimental results for the Sparse Attention-Bi-LSTM classifier with different combinations of MFCC, ZCR, and Short term energy.

Group	*Feature Combination*	*Accuracy*	*F1-score*
G1	ZCR+Energy	0.88	0.83
G2	MFCC	0.90	0.87
G3	MFCC + ZCR + Energy	0.95	0.92

21% and the F1 Score of GMM also increases from 0.41 to 0.54. Bi-LSTM-based sparse attention mechanism for speech emotion recognition can gain a higher accuracy than the other three methods, ranging from 0.88 to 0.95. This result also demonstrates that sparse attention-Bi-LSTM can effectively solve the shortage of DNN and traditional methods and a better performance system can be obtained.

4.3.4 User Study Discussion

In this study, it mainly focused on speaker segmentation and speech emotion recognition in classroom teaching. Based on the WT-KLD method, it can achieve the teachers' speech in their lecture and the teacher's voice can be used for their emotional analysis. In speech emotion recognition, it can overcome the shortage of traditional methods and achieve promising results. In GMM and SVM, it cannot achieve high accuracy because the traditional machine learning method cannot learn much more emotional information from the training database. Although the DNN can achieve the high-accuracy, it can not overcome the silent frame and low emotion expression frame in the input audio data. The Bi-LSTM network with a sparse attention mechanism can effectively overcome these issues and obtain promising results in speech emotion recognition.

In this system, it exists certain limitations. Not too many teachers' voice data were collected in this study and the study only focused on positive and non-positive emotions. Moreover, it may exist incorrect labeling in emotion annotation although the annotation is based on manual labeling. On the other hand, inaccurate speaker segmentation may influence the final speech emotion recognition. Therefore, it is quite important to improve speaker segmentation accuracy and collect audio data for teaching.

For now, all of these studies were done based on collecting voice data of teaching data. In the future, it intends to iterate on the speech processing chain and eventually provide a fully automatic real-time system based on the proposed structure, which can, for example, effectively assess the teachers' real-time emotions, but also provide interaction among the teachers, students, and system. it expects that such a system will be generally applicable in a wide variety of contexts when measuring emotional states.

4.4 Summary

The chapter was primarily concerned with speech emotion recognition, which encompassed speech signal pre-processing, speech feature extraction, speaker segmentation, and speech emotion recognition. More specifically, pre-processing and speech feature extraction is the first and most critical steps in speech emotion detection, and appropriate features can help enhance the accuracy of speaker segmentation and speech emotion recognition.

Due to the limitations of supervised methods and the difficulty of obtaining big datasets in speaker segmentation, an unsupervised technique is employed to recognize the speaker's voice. Traditional unsupervised approaches, such as HMM method, can segment a speaker's voice based on the difference in speech energy. It provides an unsupervised speaker identification approach based on Wavelet Transform and Kullback-Leibler-Divergence in this user research. The wavelet transform can be used to minimize voice noise and the Kullback-Leibler Divergence (KL-Divergence) can be used to consistently separate the speech of a speaker from the rest of the voice data. Promising results are achieved in the interview and classroom teaching scenarios.

Emotional speech databases and recognition systems are introduced in Speech Emotion Recognition. The three types of emotional speech datasets are natural databases, acted databases, and elicited databases. Acted datasets make up the vast bulk of emotive speech databases. Traditional machine learning methods like SVM and deep learning methods like DNN may also be used to identify speech emotions. User research of SER in several languages using SVM and DNN is given. A promising result can be obtained in this user study.

Another used case for speaker segmentation and voice emotion identification is teacher emotion recognition. A similar WT-KLD-based unsupervised approach was suggested, and this method can be utilized to recognize teachers' voices in long-term classroom instruction. In addition, a recurrent neural network (LSTM) with an attention network was developed to identify the instructors' voice emotions, and good results were obtained in the end.

Chapter 5

Emotion-Aware VIs Exploration

As aforementioned, emotional speech has the potential to enhance the users' experience within VIs, as well as speech signals that can create a bridge between users and VIs. Although there are many techniques for VIs, not too many emotion-aware techniques can be employed in practical VIs applications. With the advancement of artificial intelligence (AI) and deep learning techniques, it is now possible to apply deep learning algorithms to understand users' speech emotions and then respond appropriately. However, due to the limitations of the acted emotional speech database, recognizing the speaker's actual emotions is still challenging. It is particularly difficult to determine whether users have the ability to simulate real-life emotions. Furthermore, it is unclear if users will be able to track future emotion-aware VIs. Once emotion-aware VIs are capable of perceiving users' emotions and interacting with their responses according to their actual feelings, dealing with those emotions is the next challenge. In light of these concerns, it is feasible to explore potential emotion-aware VI applications using the research survey. In this section, it mainly explores the potential emotion-aware VIs application based on the above-mentioned issues. Furthermore, a study questionnaire is presented to investigate users' views toward emotion- and personality-aware voice assistants and the results demonstrate that three basic user types (*Enthusiasts*, *Pragmatists*, and *Sceptics*) exist in all cultures.

5.1 Emotion-Aware Voice Assistants

Emotion-Aware Voice Assistant refers to a new type of emotion-aware VIs system that provides emotionally intelligent personal assistant capabilities. More specifically, the new VIs are capable of understanding and responding to users' emotions utilizing speech emotion detection and speech emotion synthesis technologies. Emotion-aware techniques, as previously indicated in chapter 1.4, can facilitate interaction experience between users and VIs. On the one hand, emotions play a critically significant part in everyday interpersonal interactions [170]. On the other hand, various applications based on emotional artificial technologies, such as emotion-aware autonomous systems [335] and emotion-aware voicebots [221], can be applied in VIs, which can also improve users' experience while engaging with VIs. There is a great deal of information presented in this section on emotion-aware voice assistants and the goal they are intended to achieve.

5.1.1 The Future Emotion-Aware Voice Assistants

As previously stated, emotion-aware voice assistants can comprehend and react to users' emotions. Furthermore, this voice-bot can recognize people's personalities while simultaneously having its own personality. In other words, future emotion-aware voice assistants will be capable of communicating with users like humans, grasping their personalities and emotions, and then providing suitable responses. Moreover, depending on the circumstances, the gadgets may actively interact with humans. For instance, when people return to their homes, the voice assistants can provide proactive greetings such as "Hello, welcome to your home! How were you today?". Then these voice bots can identify your emotions or sentiments using speech emotion recognition technology and provide appropriate replies based on your present emotion and answers. In this process, speech recognition and synthesis, speech emotion recognition and synthesis, and speech personality recognition and synthesis are employed in VIs and these techniques offer the possibility of developing future emotion-aware voice assistants.

Currently, speech recognition and synthesis have already been implemented in voice assistants. Users can use voice commands to ask their assistants questions, operate home automation devices and media playback, and handle other basic chores like email, to-do lists, and calendars [139]. With the progress of AI, speech emotion recognition, and synthesis, NLP, voice-bots can detect and respond to users' spoken emotions. Due to the limitations of emotional speech databases, not too many emotion-aware techniques can be utilized in voice assistants. Yet, it is feasible to detect and interpret users' emotions via NLP techniques [188]. Furthermore, individualized voice assistants will play a major part in future voice assistants, therefore speech personality recognition and synthesis will be accessible in future VIs. On the one hand, distinct personalities require different speaking strategies in human communication. Voice assistants with varied personalities, on the other hand, might be beneficial when dealing with people in various scenarios. Additionally, future voice assistants will be capable of knowing users' mental health status utilizing speech signal processing [33].

5.1.2 The Limitations of Current Voice Assistants

As stated earlier in chapter 5.1.1, emotion-aware voice assistants can recognize and respond to users' emotional speech. However, due to several limitations, existing emotion-aware voice assistants are unable to identify users' real-time emotions. For one reason, the majority of existing emotional speech datasets are acting emotions, making it impossible to determine people's true feelings. For another, current speech emotion recognition methods [8] are not robust enough to attain high accuracy with existing data. Because of these constraints, contemporary VIs struggle to distinguish between genuine and disguised emotions. Furthermore, it is difficult to demonstrate that users can emulate real-world speech emotions. Based on this issue, the thesis investigates whether users can trick emotion-aware voice assistants.

How to produce emotional speech and reply to users with appropriate emotion is another challenge in current voice assistant development. Although deep learning has the potential to produce emotive voices, determining the synthesized emotional speech remains problematic. On one side, normal individuals find it difficult to categorize emotional speech since various people may have different speech emotion labels, particularly in the absence of context or real scenarios. On the other side, due to algorithm limitations, it is difficult to produce emotional speech that is similar to human emotional speech. Even if these concerns have been resolved, another challenge is determining how to respond to people's emotions. The thesis proposed user research to investigate how to cope with users' emotions, particularly negative emotions.

5.2 User Study: Ability to Mimic Real Emotions

Emotions can be important in the interactions with future VIs. As mentioned in chapter 1.4, although the effectiveness of existing speech emotion identification systems is increasing, it is not easy to recognize users' natural vocal emotions using deep learning algorithms. Based on this issue, the user research evaluated whether the emotion-aware voice-bot can communicate with users organically and if the voice-bot can properly interpret users' spoken emotions. Participants engaged with an emotion identification system built inside a website it constructed, delivering voice samples for five fundamental emotions: neutral, happy, sad, angry, and fearful. The study employed an open-source emotion-in-voice detector to offer feedback on whether users could effectively play out the desired emotions. This study discovered that it is difficult for users to mimic all five emotions, although it is unclear if this is due to people's inability to act emotions or to the detector's performance. However, the study can assist in the collection of labeled data for subsequent analysis and the training of neural networks for emotion recognition in voice signals. This part covers a use case regarding whether users can trick emotion-aware VIs, and the main work in this part consists of emo-voice website building and data analysis of speech emotions gathered during the experiments. The majority of the results, including figures and tables, are based on referenced paper [208].

5.2.1 Study Background

Voice-based user interfaces and conversational agents (which shall refer to as VoiceBots) such as Alexa, Siri, and Google Home have become increasingly popular in recent years. VoiceBots frequently employ artificial intelligence (AI) in the implementation of relatively complicated HCI systems, such as a voice-controlled robot [138] or a mental health VoiceBot[277]. They have evolved into an effective means of communication between people and robots. However, building stronger speech interfaces for a more genuine dialogue with the VoiceBot will eventually necessitate some kind of emotional awareness, as this is an important part of human-to-human communication.

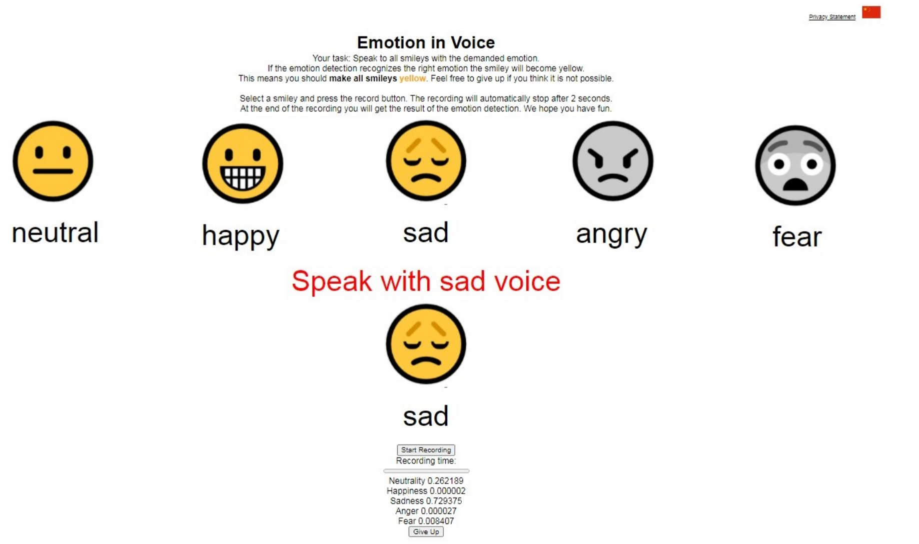
Emotions can be detected in two ways: traditionally, voice-bots can identify users' emotions using standard speech recognition (SR) and natural language understanding (NLU) algorithms [30]. This method detects emotion by emphasizing what is stated. A more modern technique called Speech Emotion Recognition (SER) [289] involves analyzing the users' voice and recognizing emphhow things are uttered. This allows the voice-bot to detect users' emotions based on their voice signals. The basic processing steps for SER are the extraction of appropriate speech characteristics followed by emotion recognition utilizing classic machine learning (ML) approaches such as Gaussian Mixture Models (GMM), Support Vector Machines (SVM), or Artificial Neural Networks (ANN) [97]. With the recent developments in AI, detecting speech emotion using deep learning architectures [104] has become a feasible alternative.

However, a major challenge for all these ML methods is to obtain accurately labeled data for different speech emotions and to provide ground truth for learning. Currently, there are two types of speech-emotion databases - acted-emotion datasets and induced-emotion datasets. For an acted-emotion dataset, researchers asked actors to perform different speech emotions, as in the SAVEE dataset [129, 130]. The alternative approach is to elicit authentic speech emotions. Research in psychology typically induces emotions by showing pictures or videos which can be used to arouse the intended emotions. In the IEMOCAP dataset [49], actors performed selected emotional scripts (acting emotions) and also improvised hypothetical scenarios designed to elicit specific types of true (i.e. non-acted) emotions.

Considering these challenges and the general shortage of training data, it became curious to find out whether regular users (i.e. non-actors) are able to mimic five basic emotions (neutral, happy, sad, anger, fear) and whether they manage to trick emotion recognition into detecting the intended emotion. A web page was set up as shown in Figure 5.1 and recruited a small number of participants to record their voices in five basic emotions. As shown in Figure 5.1, the emoticons in the upper row can be clicked to select the emotion to enter. In the beginning, all emoticons were gray. After successfully mimicking an emotion, the corresponding emoticon becomes yellow. The web page provided feedback on whether the emotion was successfully recognized. It can count the number of trials until success in each emotion and found that participants were not able to successfully mimic all five basic emotions. However, it was not clear whether this was a failure of the emotion detector it had used or the users' actual inability. As a side effect, this experiment also provided a small labeled data set for acted emotions from regular users, which it was hoping to use for training future SER prototypes.

5.2.2 Study Design

In this study, it set up the web page in figure 5.1 to conduct the study. The page shows five clickable emoticons in a row, plus the currently selected emotion in the center below, as well as the recognition result. Participants can select an emotion from the five given basic emotions and then are asked to say something with the selected emotion. In this study,



Emotion in Voice

Your task: Speak to all smileys with the demanded emotion.
If the emotion detection recognizes the right emotion the smiley will become yellow.
This means you should **make all smileys yellow**. Feel free to give up if you think it is not possible.

Select a smiley and press the record button. The recording will automatically stop after 2 seconds.
At the end of the recording you will get the result of the emotion detection. We hope you have fun.

Speak with sad voice

Recording time:

```
Neutrality 0.262189
Happiness 0.000002
Sadness 0.729375
Anger 0.000027
Fear 0.008407
```

Figure 5.1: The EmoVoice web page for collecting voice samples [208]

it mainly explores whether it is possible for participants to act on all basic emotions and whether it is possible to trick the emotion-aware VoiceBot. The goal of the study was to find out whether participants would be able to act all basic emotions and in consequence, could trick an emotion-aware VoiceBot.

Apparatus

In this study, it utilized the SER package OpenVokaturi[1] version 3.4, a one-layer neural network, to detect participants' emotions in the voice samples recorded on this website. The voice data was collected by the built-in microphone of their own computers and uploaded to the university web server. The sampling rate of all recorded voice signals was set as 48 kHz. Matlab 2017a was utilized for data analysis afterward.

Participants

This study recruited 26 participants (13 males) from the experimenters' personal networks to join the experiment. Most of them had experience in using computer recording. They joined the study after the web page link was sent by email, but this study did not have any means to connect the collected recordings to the email invitations, as they were free to join whenever and from wherever they wanted. On this web page, participants can choose any language they like and speak anything they want.

Study Conduction

Participants were informed in the privacy statement that the study was completely anonymous and participants did not even need to provide demographic data. The study proce-

[1] https://vokaturi.com/

dure was approved by the local ethics review board of LMU Munich. The only instruction for the participants was to select an emotion and then speak with the selected emotion, so they were free to use any language they liked and to say anything they wanted. Participants can read the introduction and guidance about this study when they open the link they received. After acquainting relevant details, they could choose one of five basic emotions they like. There was no particular order and the central emoji would change to what they had chosen in the top row. When they clicked the "start recording" button, 2 seconds of voice data was recorded and uploaded to the university web server. Participants could try as often as they wanted to mimic each emotion. Upon success, the corresponding emoji in the top row would turn from grey to yellow. This meant that they had successfully acted on the selected emotion. They could then choose another emotion and continue. If they found certain emotions too difficult to imitate, they could click a "give up" button and end the user study. When they could make all five emojis become yellow, it shows that participants can be able to successfully mimicked all emotions in the user study.

5.2.3 Study Results

This study only captured the number of trials each participant spent on each emotion. As a derived measure, the system calculated the success rate as the inverse of the number of trials until success, or as 0 in case they gave up. For both, it calculated descriptive statistics, i.e., mean and standard deviation. Statistics methods, like average, and standard deviation can be used to analyze whether participants are able to mimic all basic emotions and which emotions are difficult for them to act. A Wilcoxon Signed Rank Test (WSRT) [340] was then used to determine whether there were any significant differences between the different emotions, regarding the success rate and the number of trials. The study's expectation was that all emotions would be equally well recognized. However, it found a significant difference in the number of trials between "neutral" and "happy" (p=0.0284) and between "happy" and "fear" (p=0.0479). Table 5.1 shows the differences between "neutral" and all other emotions regarding success rates. From the study, it only found a significant difference in the number of trials between "neutral" and "happy" (p=0.0284) and between "happy" and "fear" (p=0.0479). Evaluating the success rate, however, showed significance between "neutral" and all other emotions. Table 5.1 shows the p-values for "neutral" against all other emotions regarding success rates.

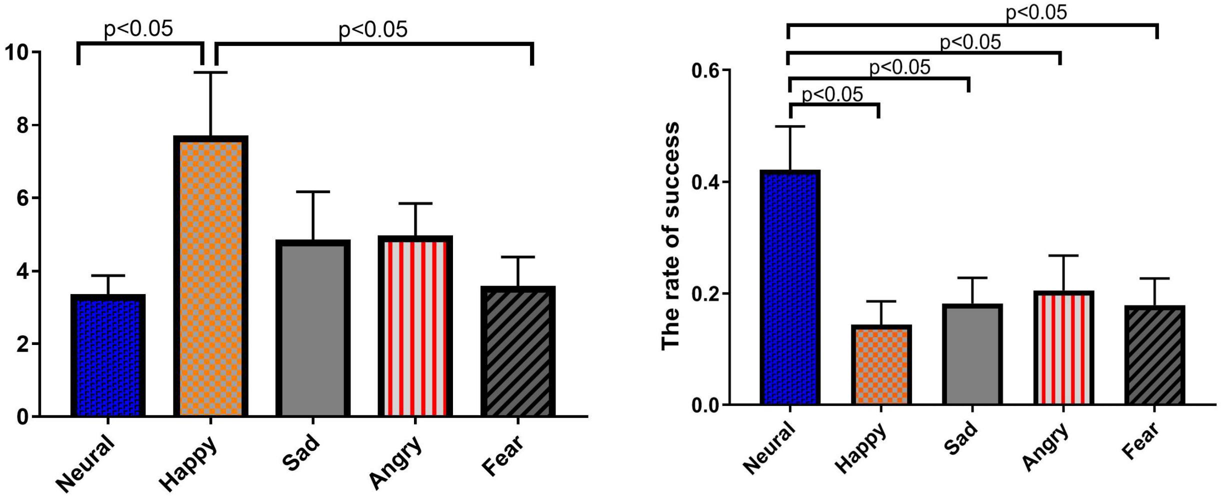
Table 5.1: Result of a Wilcoxon signed rank test on the success rate between "neutral" and all other emotions [208]

	Neutral	*Happy*	*Sad*	*Angry*	*Fear*
Neutral	-	0.0029	0.0109	0.0354	0.0107

Figure 5.2 shows the number of trials and the success rates for each emotion. If participants did not try a specific emotion, no success rate was calculated. it can see that not

all participants were able to imitate all five emotions equally well. It seemed easy for them to mimic the "neutral" emotion but difficult to act the "happy" emotion. In addition, the study found that only a few participants chose the "give up" button when they could not imitate a certain emotion. Most participants could not act on all five basic emotions, even if they tried many times.

In the WSRT hypothesis test, the p-value of the participants' attempt times between neutral and happy is 0.0126. It shows that the data from neutral and happy emotions are meaningful and independent. Moreover, the p-value of their trial times between happiness and fear is also below 0.05. It demonstrates that these data from happy and fearful emotions are also meaningful and independent. In addition, as it showed in Table 5.1, the study found that the p-value between the neutral emotion and other emotions is mostly below 0.05. It means that they are also meaningful and independent in the participants' attempt success rate data.



(a) Number of trials until success [208]

(b) Success rates [208]

Figure 5.2: Number of trials and success rates of all 26 participants for all basic emotions they tried to mimic.

In the following figures, the data shown above draws a mixed image: While most participants succeeded well in mimicking a "neutral" emotion. From these histograms and tables, it could illustrate that the collection data is meaningful and can be used to analyze participants' mimic-emotion ability. Based on Figure 5.2a and Figure 5.2b, it can state that participants can be able to imitate "neutral" emotion easily and mimic "Happy" hardly. Moreover, it can show that it is impossible for users to act five basic emotions from the study's results.

5.2.4 Study Discussion

In this study, not most participants succeeded to mimic all five basic emotions. However, it can not simply claim that the SER system OpenVokaturi it used is not good enough for emotion detection. Instead, the main reason is that it may have failed because the participants were unable to successfully act out each emotion, i.e. to properly fake it. The result of the study hence suggests that it may in general be difficult to “cheat” the emotion detector, at least OpenVokaturi, with acted emotions. This may be a strong hint for the detection of true emotion. On the other hand, the detection results vary very much which is not in accordance with the hypothesis that the detector detects the true emotion. The users’ emotions should not change too much over the short time of the study. It is possible that the study itself affected the user’s emotional state. Success in entering the demanded emotion could have made participants happy, while failure could have made them angry. It was unable to verify such effects as the data basis was not big enough. However, even with a larger data basis, such effects can only be seen if the emotion detection reports accurately true emotions, which is not guaranteed.

An interesting question for future research is whether it is possible to build an emotion-in-voice detector, which detects acted emotion and not the true emotion. After collecting a bigger corpus of data with the existing system, it will become possible to train a neural network with it. If the users will be more successful in mimicking emotions with the new detector, it can achieve the goal and can claim to have built a detector for acted emotion. However, there are more general reasons to be skeptical. A 40-millisecond voice sample is probably not enough for humans to judge the emotion in this voice sample. Humans normally need at least a few words to judge the emotion in a voice. Maybe it will take even other methods or longer samples to detect acted emotion.

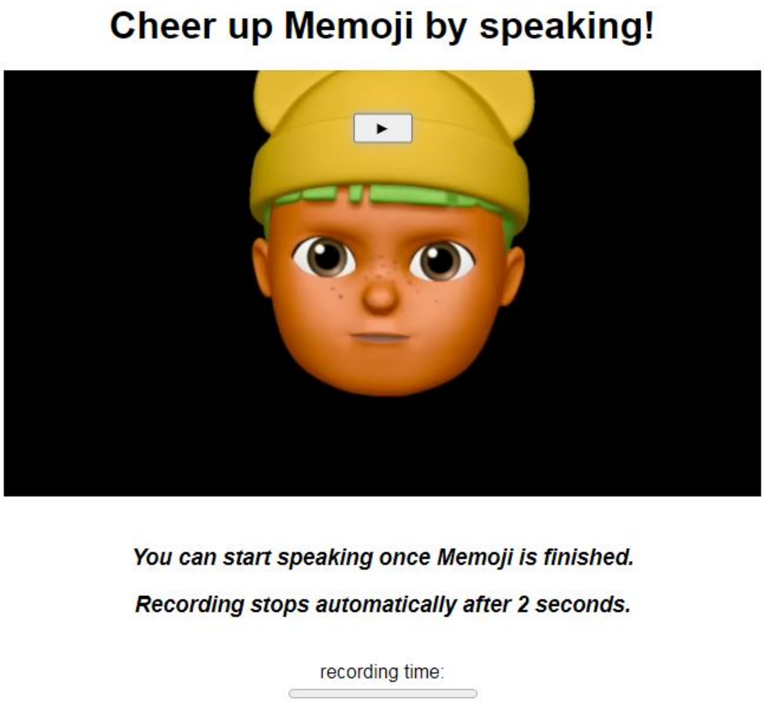
5.3 User Study: How to deal with User’s Emotion

It is critical for virtual assistants to recognize and respond to users’ emotions. With the progress of VAs, it is now feasible to detect the emotions of users. How to respond to customers’ emotions has become one of the most difficult difficulties for modern VAs. As speech technology advances, voice assistants (VAs) are becoming increasingly significant in people’s daily lives. They can serve as a conduit for communication between users and computers. In general, VAs can be found in both mobile and stationary devices, and users can interact with them simply by speaking or using voice commands. VAs can identify users’ emotions by applying semantic analysis or speech emotion recognition. Responding to users’ emotions, especially negative emotions, becomes more challenging. Human emotion reaction approaches can be used to change the roles of users and VIs. The major focus of this work is on how to respond to emotional inputs. This study generated three avatar emojis (angry, sad, and afraid) that may depict distinct emotions using animation and sound. 52 people were requested to engage in this user study, and they used emotional vocal input to try to turn the emojis into a desired emotional state (pleasant sentiments). According to the findings, users usually used neutral emotion to respond to these three unpleasant feelings, and there is a gender difference in emotional reactions. However, since the study's experiments largely used male voices as emotional stimuli, gender differences in responsiveness to negative emotions are unclear. This section presents a user study on how to deal with a user's emotions. The main work in this part consists of the Memoji website development and data analysis in the user study. Most of the results, such as figures, are based on the citation paper [209].

5.3.1 Study Background

Voice Assistants (VAs), which are embedded in the smartphones or smart home gadgets, are becoming increasingly popular in the current daily lives. VAs may be found in both mobile and stationary devices, and users can simply interact with them using voice or speech commands. Users can communicate with VAs by voice or speech, and VAs can identify users' spoken orders after they speak them. Then, like humans, VAs will deliver their replies based on their comprehension, but they will not say anything thereafter. In general, contemporary VAs are unable to recognize and communicate their own emotions when speaking with people. With advancements in speech signal processing and emotion detection, it is now feasible to recognize users' emotions from voice input [8, 103]. However, dealing with consumers' emotions becomes extremely difficult. Some studies argued that self-reported and contemporaneous expression can assist computers in efficiently sensing, recognizing, and responding to human emotional communication [259, 260, 256]. Psychologists performed emotional transformation research in 310 clinical and 130 sub-clinical cases [246] using Pascual-Leone and Greenberg's sequential model of emotional processing or its associated measure [247]. Nevertheless, it still exists certain issues to deal with users emotions in VAs. Concretely, it is still incredibly difficult for VAs to respond to users' emotions without any prior knowledge, even while VAs can understand users' emotions. It means VAs need to take certain strategies or follow certain rules to deal with users' emotions.

This study presented a role-swapping technique to investigate these concerns. In other words, in this study, the roles of VAs and humans are reversed. Participants will take part in the study and develop their own strategies for dealing with unpleasant emotions in VAs. The goal of this study is to turn negative emotions in VAs into good or neutral feelings. Participants can speak with negative-emotion VAs using a range of emotive voices. The website in Figure 5.3 was built on a university server, and participants can join this user study by clicking on the study web link. The memoji in the figure 5.3 will emerge at random, and the participants' goal is to turn negative memoji into positive memoji. They have five attempts to use any emotional voice. Participants can enter their emotional voice onto that website, and their voice will be instantly recorded. After five attempts, the participants' voices will be analyzed for mood and speech qualities. According to the final results, the majority of participants will pick the neutral feeling to react to all of the unpleasant emotions. However, there is a variation between the genders in how they react to certain unpleasant emotions.



Cheer up Memoji by speaking!

You can start speaking once Memoji is finished.

Recording stops automatically after 2 seconds.

recording time:

Figure 5.3: The emoji user study web page for collecting voice samples [209]

5.3.2 Study Design and Conduction

The goal of this user study is to discover user strategies of users who are confronted with an avatar's negative emotions. Their mission was to transform the avatar's feeling into a pleasant emotion, such as happiness. The data obtained comprises of voice recordings and user feedback.

Recruiting Participants Participants were recruited using the author's personal network and university's user email list. 52 people signed up for the study after receiving an email with a link to the study's website. There was no attempt to link the issued email invitation to the recordings of the attendees. Participation was available at any time and from any location.

Data Collection and Privacy The website's privacy policy assured participants that the research was fully anonymous. Furthermore, the statement advised them of the information gathered. The data comprises of voice recordings from participants, questionnaire responses, and demographic information. The latter simply contains the participants' age and gender information. Participants were also informed that their voice samples will be evaluated using SER software based on a neural network.

User Instructions Aside from the microphone test, the landing page briefed the participant about the research in general. They were instructed to check their microphone and grant access to it if necessary. More information regarding the future research came if the subject completed the microphone test. It notified the participant that they would be speaking with a Memoji, that the Memoji was depressed, and that they should attempt to cheer it up. They were also informed that a questionnaire would follow each video sequence, and that this combination of a video sequence followed by questions would recur three times. The button to start the research was activated beside the supplementary information text, and the participant may advance to the first video sequence. When the participant pressed the Start Study button, the movie began to play automatically if the browser settings permitted it. If the participant's browser settings did not enable the video to start automatically, they were provided a play button that they could hit to start the movie manually. After a video had finished playing, the participant might respond to Memoji orally at any moment. After 2 seconds, the recording began and ended automatically. When the recording was finished, the next video was immediately loaded and begun. The participant was taken immediately to the questions page after the last video of one Memoji series. The same questions were asked for each series, with the answers tailored to the Memoji's present mood. Only in the first questionnaire were participants asked for demographic information. After responding to the questions, the participant might go on to the next Memoji video sequence by hitting the “Cheer up next Memoji” button. After seeing all three sequences of Memoji films and answering the accompanying questions, the research came to an end. The participant presented an informative text alerting them that the research was coming to an end and that they could exit their browser.

5.3.3 Study Results

The study was conducted over three weeks and involved 52 participants. Six individuals did not finish the trial, leaving 46 valid data records for analysis. The research data evaluation, including stimulus validation and reaction evaluation to the various Memoji moods utilizing standard and new SER methodologies is presented in this part. Furthermore, user methods and views are examined based on the questionnaire responses of the participants.

Demographic Data

In terms of demographics, the participants had an average age of 30.48, with the youngest person being 11 years old and the oldest participant being 69 years old. Instead than choosing an age range, individuals were asked to enter their precise age. Figure 4.1 depicts the participants' ages as a histogram. Participants were also asked to select their gender from three options: female, male, or other. There were 22 female and 24 male participants since no one chose the option “other”. As a result, the gender distribution is nearly equal between male and female. The average age of all female participants (30.41) is about matched with the average age of all male participants (30.54).

Validity of Stimulus

Participants were asked to answer questions after each film episode. The first four questions were designed to determine whether the stimulus was effective. The objective and expectation was for participants to view the Memoji as being in one of the negative emotions (sad, angry, or scared) at the start and cheerful at the conclusion. To summarize, participants received the stimulus as intended and expected by the author. The findings are examined in depth in the sections that follow. These questions were answered by all 46 participants. Regarding the beginning, they selected one of five possibilities on a likert scale ranging from not sad/angry/frightened at all (option 1) to extremely sad/angry/frightened (option 5). For the last questions, participants picked one of five Likert-scale alternatives ranging from not pleased at all (option 1) to very happy (option 5). The following inquiries were made:

- **RQ1:** How sad/angry/frightened did Memoji's voice sound in the beginning?
- **RQ2:** How sad/angry/frightened did Memoji's face look in the beginning?
- **RQ3:** How different age groups influence the preferences of VAs?
- **RQ3:** How happy did Memoji's voice sound at the end?
- **RQ4:** How happy did Memoji's face look at the end?

In overall, the sound of the voice seemed to have convinced the participants of Memoji's mood more than the appearance on its face. The findings of the questions concerning the beginning reveal that it was able to create a stimulus that was viewed as intended by the author by the majority of the participants. According to the mean values, the most compelling stimulus is sad, the second most convincing is afraid, and the third most convincing is furious.

In terms of the final stimulus, it appears that the face appearance has the same or higher mean values as the voice sound. This might imply that Memoji's expression was slightly more believable than its voice tone at the end. For queries regarding the beginning, it is the opposite way around. For them, it seemed that the sound of Memoji's voice was more convincing than the expression on its face. Except for the pleasant voice stimulus at the end of the furious Memoji, the majority of participants understood the stimulus as the author intended. Based on the results, it can be concluded that the sad Memoji best accomplished the author's purpose of providing a sad stimulus at the start and a joyful stimulus at the conclusion. Aside from the tragic Memoji, it can be seen that the stimulation at the start was more successful than the stimulus at the end. According to the mean values, the order from most compelling to least convincing stimulus stays unchanged; it is first sad, then terrified, and finally furious. One possible reason for these results is that the Memoji films were recorded by the author himself. The findings could have been different if a professional actor had been requested to film the videos and act out the appropriate emotions. Furthermore, it is vital to note that the author fabricated the emotions throughout the recordings. He only pretended to be upset, furious, or scared. Better stimulus values may have been obtained if the author had felt the appropriate emotion throughout the recording. Furthermore, the author's subjective perception is that the shift from sad to joyful is best experienced in the sad Memoji films. The ends of the terrified and enraged Memoji are not regarded as credible. However, the overall results are consistent with the author's aims.

The participants' assessment of the stimulus revealed that the avatar's mood was regarded as intended. Figure 5.4 depicts the average emotion recorded by Vokaturi. The large standard deviations show that individuals' emotions are widely dispersed. Vokaturi reports five values for five fundamental emotions, with four numbers being modest and just one being high. When it exclude values below 15% as noise, it get Figure 5.5, which reveals that the majority of answers were supplied with neutral emotion.

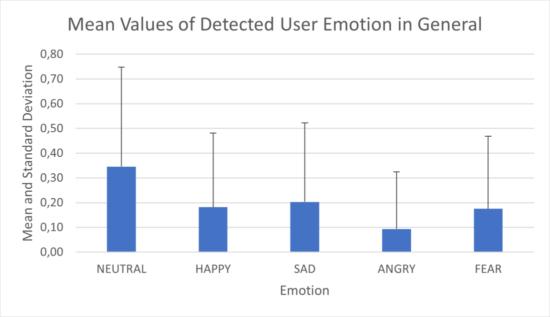


Figure 5.4: Emotion in the participants' voice (analyzed with Vokaturi) averaged over the avatar's mood [209]

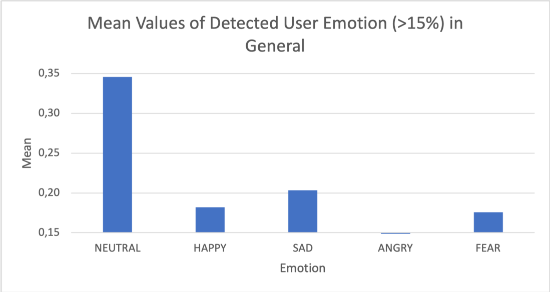


Figure 5.5: Vokaturi reports at least small values for all emotions, which it considers as noise. The figure above takes only values above 0.15 into account [209]

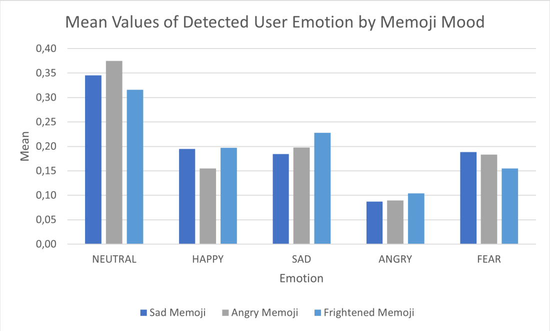


Figure 5.6: Emotion in the participant's voice (analyzed with Vokaturi) for the three avatar's moods [209]

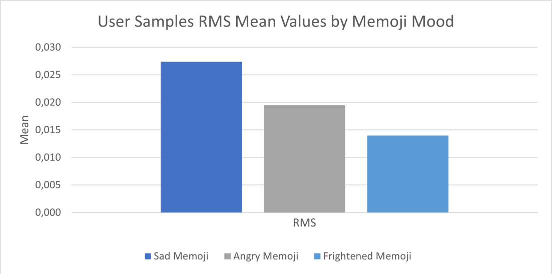


Figure 5.7: Mean RMS values for voice samples for the different avatar moods [209]

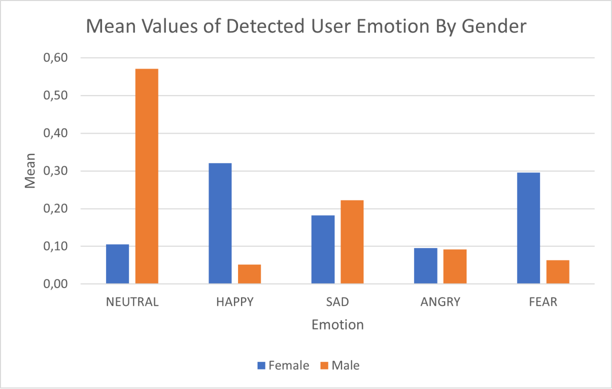


Figure 5.8: Emotion in the participants' voice in different genders (analyzed with Vokaturi) [209]

Figure 5.6 depicts the average emotion of users for the three negative avatar moods. The emotional response to the mood of the various avatars is similar. However, it discovered a difference in RMS (root mean square) (see 5.7). Figure 5.8 depicts the users' emotion by gender averaged over the three avatar moods. It demonstrates a clear gender difference. We did not evaluate the importance of this finding because it is unclear how accurate Vokaturi's findings are. The documentation for Vokaturi indicates that "the accuracy on the five built-in emotions is 76.1%" [2].

5.3.4 Study Discussion

In this study, three different negative Memojis were established on the internet and served as emotional stimuli. Participants were invited to cheer up the Memoji and their voices were recorded during the conversation. The initial assumption is that variances in user reactions reflect differences in the three avatar emotions. However, the majority of participants picked neutrality as their reaction to negative Mimojis. It demonstrates that the reactions' emotions were independent of the stimuli's emotions. On the one hand, people may suppress their true feelings and just react to Momojis with neutral emotions. On the other hand, Open-Vokaturi has certain limitations, and the existing SER system cannot identify users' emotions in real-world settings. Furthermore, the negative Mimojis cannot elicit participants' inner emotions, and participants believe that neutral emotions can restore their happiness.

Another assumption is that there is no gender difference in emotional reactions. Speech emotions are unaffected by language or culture [249], and there was no reason to suspect a gender difference. On the contrary, a gender difference in genders was achieved. Although it is improbable, this foundation could have occurred by accident. The other possibility is that the speech emotion detector has a gender bias. Male sample voices outnumber

[2] https://developers.vokaturi.com/q-a

female sample voices in Vokaturi's training database. In fact, Vokaturi's training databases[3] includes the Berlin Database of Emotional Speech (Emo-DB[4] [47]), which has five female and five male speakers, and SAVEE[5], which contains voice samples of four males. This raises the question of whether gender-balanced training databases in SER are required. According to Vogt et al. [331], gender differentiation increases the accuracy of SER.

A further explanation could be that there is a gender difference in emotional reactions. As a result, a female voice assistant should react differently to emotions than a male voice assistant. In the other words, male users may require a different approach than female users to cheer them up. Although there are clues that this could be the case [293], this setting in voice assistants would not be acceptable for users since it indicates that emotion-aware voice assistants also require gender awareness, which would result in gender differences in technology. The general question is whether it exists a difference in conversations between males and females, males and males, or females and females. If this is correct, the question is whether this gender difference should be incorporated into future voice assistants.

In a UNESCO report[6] and in the media[7], gender issues are a critical concern for voice assistants. Moreover, a genderless voice is being developed to eliminate gender bias in AI[8]. A Follow-up study could include both male and female stimuli. Alternatively, the stimulus could also come from a gender-neutral voice or that of a comic character. Which choice to select is determined by whether humanity prefers gendered voice assistants or non-gender voice assistants. These issues will need to be addressed in future research.

5.4 Summary

Speech signals can help users engage with VIs, and emotion-aware techniques can improve the user experience in the VIs. However, there are several challenges in the VIs application, such as detecting imitation emotions and dealing with user emotions. On the one hand, most emotional speech databases are performed emotion databases, making it difficult for the SER system to discern the user's true feelings. On the other side, selecting the appropriate emotions to reply to users might be difficult. Due to these concerns, this chapter investigates prospective VIs applications and creates a study questionnaire for future emotion-aware VIs.

How to trick future emotion-aware VIs is the first challenge. Although the effectiveness of existing speech emotion detection systems is improving, it is difficult to recognize users' natural voice emotions using deep learning algorithms. A user study was designed to evaluate if the emotion-aware VIs can communicate with users organically and whether VIs can properly interpret users' spoken emotions. A small number of participants (26)

[3] https://developers.vokaturi.com/algorithms/annotated-databases

[4] http://www.expressive-speech.net/emodb/

[5] http://kahlan.eps.surrey.ac.uk/savee/

[6] https://unesdoc.unesco.org/ark:/48223/pf0000367416

[7] https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html

[8] https://www.genderlessvoice.com/

were asked to mimic five basic emotions and an open-source emotion-in-voice detector was used to provide feedback on whether their acted emotions were recognized as intended. In this user study, it found that it was difficult for participants to mimic all five emotions and that certain emotions were easier to mimic than others. However, it remains unclear whether this is due to the fact that emotion was only acted or due to the insufficiency of the detection software. As an intended side effect, a small corpus of labeled data for acted emotion in the speech was collected, which can extend and eventually use as training data for future emotion detection.

The second challenge is dealing with the user's emotions, particularly negative emotions. There is a growing body of research in HCI on detecting users' emotions. Once it is possible to detect users' emotions reliably, the next question is how an emotion-aware interface should react to the detected emotion. The first step is finding out how humans deal with the negative emotions of an avatar. The hope behind this approach was to identify human strategies, which it is possible to mimic in an emotion-aware voice assistant. A user study in which participants were confronted with an angry, sad, or frightened avatar was presented. The task in this user study was to make the avatar happy by talking to it. The participants' voice signal was recorded and the voice data was analyzed. The results show that users predominantly reacted with neutral emotion. However, there are also gender differences, which opens a range of questions.

Chapter 6

User Preference for Emotion-Aware Voice Assistants

Voice Assistants (VAs) are becoming increasingly common in modern life, however, it is difficult for existing VAs to converse with users in a more natural manner. With the advancement of speech and AI technology, it is currently feasible for VAs to perceive users' emotions and personalities, which may tremendously enhance the user experience. Ongoing research in speech emotion recognition indicates that voice assistants can be emotion-and personality-aware in the near future, just as natural language processing has improved dialogues with chatbots. As AI and SSP improve, more speech techniques can be implemented in the VAs. The most difficult challenge, though, will be creating future voice assistants that are acceptable to the majority of people. In other words, user attitudes toward these novel speech technologies are vital in the development of future VAs. Despite the fact that most researchers concentrated on state-of-the-art speech techniques, the societal implications, ethical boundaries, and security and privacy concerns regarding future VAs remained unexplored. Moreover, the attitudes of users toward these new speech techniques are unclear. Based on these issues, this chapter focuses on the research of users' attitudes toward and preferences for emotional-aware VAs across different cultures. The study in this chapter, in particular, used a survey to investigate the aforementioned difficulties. This questionnaire covers demographic information, questions on technology, and social issues in various contexts. The study's goal is to acquire a better understanding of future VA users' requirements and preferences. It investigates users' preferences and attitudes regarding emotion- and personality-aware virtual assistants (VAs) among German, Chinese, and Egyptian users. Furthermore, the study investigated the user's acceptance of specific speech styles as well as various ethical, security, and privacy concerns. The main work in this chapter comprises the construction of a questionnaire, data analysis in the survey, and the establishment of user types using the clustering method. Almost all of the results are based on the citation paper [209], including figures and tables.

6.1 Research Questions

Emotion-and personality-aware VAs can enhance a user's mental health [173], enable safer car travel [133], and increase trust in the device [35]. To construct these highly technological VAs, particularly on human concerns in VA design, it is essential to comprehend the users' requirements, preferences, acceptability, etc. The majority of the participants in this study are from various nations such as Germany, China, and Egypt, among others. Accordingly, it is possible to investigate if the findings in the survey results are influenced by different cultures. Furthermore, it is essential to investigate whether there is a gender difference in the answers to the various questions in the questionnaire. Based on these concerns, the following research questions [209] were presented and they should be addressed in this chapter.

- **RQ1:** What is the general user attitude toward existing speech techniques if they can be integrated into voice assistants?
- **RQ2:** How does gender difference influence the preferences regarding VAs, such as certain speech techniques?
- **RQ3:** How do cultural variations impact preferences for VAs, particularly in terms of ethics, privacy, and security?
- **RQ4:** Is it possible to identify culture-independent user groups? These user types can be considered while designing future VAs.

6.2 Study Background

VAs offer a natural and intuitive type of interaction that is modeled after human-to-human conversations. In certain cases, VAs outperform traditional interfaces since they do not require physical space for interaction and just utilize voice. They are also hygienic during the present pandemic circumstances since there is nothing to touch, and VAs are believed to be simple to use due to the usage of spoken language. The previous study has demonstrated that users' perceptions of a voice are quite significant [233]. Users tend to interact with computers as though they were chatting with a human [189, 233, 292]. This prompted the development of speech technologies or products such as Google Duplex [192], in which VAs can have a human-sounding voice rather than a synthetic one. If computer voices become practically indistinguishable from human voices, a new set of challenges arises, such as calibrating trust and expectation.

Some researchers are already attempting to consider psychological and social aspects such as affect and emotion, trust, credibility, and the relationship environment [336, 333]. Furthermore, emotion- and personality-awareness is essential for realistic human communication; hence, there is a significant amount of research in the HCI field to achieve these abilities, especially for voice interfaces [333, 332, 41]. However, there are various unanswered problems on the subject, including user acceptance of such technology, specifics on how the voice assistant should respond to the user's emotions, and potential ethical constraints. Furthermore, some of these questions will elicit different replies based on other social criteria such as gender, age group, and culture. This section expands on the context of the study based on various issues and challenges, such as existing speech technologies and varied user groups.

6.2.1 Voice AI and Voice Assistants

Voice AI and Voice Assistants Since voice assistants were first invented more than four decades ago, and with the basic technical challenges solved, they have become part of daily life for many families [266, 139], making human-to-VAs communication more realistic. It is possible to develop hyper-personalized, intelligent agents by combining the capabilities of artificial intelligence with conversational agents, as proposed by Zhou et al. [358]. With the advancement of Voice AI, numerous different techniques can now be applied to various modern voice assistants, which can currently comprehend human speech and reply via synthetic voices [139]. The monitoring and evaluation of human mental health using speech signal analysis [43, 320] can be implemented in future VAs. Moreover, some studies developed dialect voice conversion techniques based on phonetic posteriorgrams [112, 9] or deep learning algorithms [355]. These approaches have the potential to provide dialect voice assistants in the future. Furthermore, advances in emotion recognition and synthesis from speech signals [162, 289, 25, 279], could enable future voice assistants to recognize users' emotions and converse with them in an appropriate emotional speech. AffectAura [213] is one of the first emotion-aware assistants, automatically collecting emotional signals over time to assist users in reflecting on their emotional well-being. Future VAs could also perceive users' personalities and generate their own personality based on the current technique of personality detection [264, 204] and synthesis [18]. Using a questionnaire, this chapter investigated user acceptance and preference for potential speech technologies that could be employed in future VAs based on various voice AI technologies.

The relationship with Voice Assistants Clark et al. [66] discovered a substantial distinction in user expectations between interacting with a purely functional assistant and building a connection with a separate entity, with the functional viewpoint clearly dominating these days. It is a step forward in this study toward developing the new form of dialogue proposed there: constructing VAs in such a way that they do not aim to replace human equivalents, but rather optimize a functional dialogue toward more trust and less friction. Emotion- and personality-aware VAs, for example, have the potential to improve Human-Computer Interaction (HCI), as demonstrated by McDuff et al. [113]. Emotion-aware voice assistants can assist users to perceive their emotions and become more self-aware, which can result in improved well-being and mood [148, 312]. McRorie et al. [216], for example, explored agents constructed based on psychological concepts of personality, while Völkel et al. built a more current model [334]. According to opintions [41, 358], VAs with particular personality features are viewed as more trustworthy, and in driving conditions, they can boost safety by lowering stress [41, 133]. It has been demonstrated by

Braun et al. [40] that users gain more from using VAs when their personalities match, but current VAs use the same personality for all users. Klein et al. [167] conducted a survey on the user acceptance of such VAs for a German user group, however, the questionnaire was mostly concerned with privacy problems. This study intends to improve their work by investigating users' perceptions regarding the emotional and personality aspects of VAs across three cultures.

6.2.2 Studying Diverse User Groups

When attempting to enhance the user experience for varied user demographics [68], it has been recognized in HCI that a one-size-fits-all strategy is sometimes insufficient. Understanding the expectations and preferences of various user groups will lead to more specialized interface designs that will better meet each group. Users can be distinguished not just by basic demographic parameters such as age, gender, or education, but also by their cultural background. This section describes how previous research handles culture (or does not) and then covers broad techniques for user types.

Cultural Characteristics VAs are employed in a variety of areas throughout the world. However, most of them were designed for Western users [200]. In HCI studies, Western participants are over-represented, accounting for 73% of total research findings. Particularly, participants from African countries and other regions of the world are particularly underrepresented. HCI progress in China has been somewhat slow in recent decades, as Ma et al. [194] demonstrated. Furthermore, culture as an HCI concern is far from being prevalent. Culture studies began to take place in 1971 in psychology [95, 219], but it wasn't until 1997 that intercultural HCI studies began to emerge [101], followed by a variety of theoretical frameworks [77, 308, 301, 134] and comparative cultural studies [106, 102]. Hofstede et al. [135] devised a well-known framework for intercultural comparison, which was utilized to analyze consumer behavior across cultures. Rather than comparing two civilizations, Kamppuri et al. [158] suggest that human factors research should study how culture and technology interact. Hence, this study conducts its research in Africa, Asia, and Europe, concentrating on similarities rather than cultural differences and the relation between culture and culture-independent user groups.

User Groups Aside from culture, users can be grouped according to a variety of features. For example, users can be classified into six profiles based on their sharing and privacy attitudes on Online Social Networks (OSNs), with design implications for each user type [339]. Moreover, Behrooz et al. [239] created an interactive framework (IUGA) that leverages group discovery primitives to explore the user space by constructing labeled groups of people that are related in kind. User groups are constructed by combining obtained user data, which includes basic demographic information (such as gender, age, and native language) and user interests. To establish a user group, many users need to have the same values for some of the attributes [239]. The user modeling community has devised a variety of ways for characterizing such user groups. It chose K-medoids for dividing and grouping the participants in this study since it was seeking first knowledge of prospective different user groups across cultures and did not have a specific operationalization in mind. Future research might improve on this strategy by applying hierarchical methods to establish a more thorough ontology of user groups at different levels of abstraction [109].

6.3 Study Design

The questionnaire was originally established in English and then translated into German, Chinese, and Arabic. LimeSurvey[1] was utilized to design the survey. Its translation was validated in terms of language and cultural appropriateness by native speakers. These people are also acquainted with the cultural backgrounds of Egypt, China, and Germany, as it includes those who grew up in each of these nations. This study employed non-biased phrasing and gave participants more space to express themselves through open-ended questions. The original English questionnaire is provided in the table below(See Table 6.2).

The collected data was anonymous, and all participants gave informed consent to their anonymous responses being stored and used. In the study, participants who wanted to be compensated by shopping vouchers had to complete a separate form. After compensations had been issued, this contact information was removed from a separate questionnaire. This section mainly discusses the questionnaire design, which includes three dimensions (technology, social, and context).

6.3.1 Determining Cultural Identity

It is difficult to associate a person with a culture [136]. In particular large nations, such as China, there are a number of different ethnic groups with different cultural backgrounds [2], which makes nationality an unreliable criterion. Furthermore, in some cases, participants may even have parents from two different cultures, or they may live in a third culture. Therefore, the study tried to analyze participants' culture by requesting their native language, as this was considered to be a significant predictor of their upbringing and thus cultural socialization.

6.3.2 Questionnaire Structure

The questionnaire for this study was designed utilizing the notion of a question area for the subject of voice assistant emotions and personalities. Dimensions and subcategories are also used to partition the question area. The dimensions and categories were generated through collaborative brainstorming sessions among various LMU Media Informatics researchers. In addition, some information was acquired from current media conversations concerning

[1] https://www.limesurvey.org/

[2] https://en.wikipedia.org/wiki/List_of_ethnic_groups_in_China, last accessed December 19, 2022.

VAs. The actual questions were created through various refining phases to cover the whole question area in a balanced way while limiting the total questionnaire to a reasonable size. This procedure yielded three dimensions: a technological dimension, a social dimension, and a contextual or situational dimension. The questionnaire was structured accordingly, and the final questionnaire had 39 items.

Demographics Survey questions on demographics include questions on age, gender, mother language, self-assessment of computer proficiency, and voice assistants previously used by the participant. Participants were asked to fill out these questions at the beginning, and their answers can help with the subsequent analysis of questionnaire data. This section of the demographic data focuses primarily on the gender and mother language of the participants.

Technology Dimension In the technology dimension, the type of the questions was divided into seven areas in the study: speech technology in general, speech recognition, speech synthesis, speech emotion detection, speech emotion synthesis, speech personality detection, and speech personality synthesis. These seven categories were eventually utilized to form the top level of the questionnaire.

Social Dimension Acceptability, interaction details, privacy and security, relationship to the voice assistant, and ethics/morals are the categories of questions from the social dimension. Technology acceptance relates to whether users are enthusiastic about it or skeptical about these speech technologies. Considerations concerning how to implement VAs, such as which voice to use and if the device should be programmable or self-adjusting, are included in the interaction details. Concerns about who has access to speech data, as well as the abilities of a voice assistant, fall under the topic of privacy and security (e.g., for money transfers). The relation category includes whether the voice assistant should be regarded as a servant or a friend, whether it should respect corporate or family hierarchy, and whether it should be capable of conversing in local dialects. In the ethical and moral realm, children's use of voice assistants such as Alexa, which includes a kid mode [3], can be a concern. As an example, the questionnaire asked about ethical constraints, including using the voices of the deceased. This subject garnered media attention shortly following this survey[4]. Additionally, a questionnaire that asks about sexual activity (flirting) and religious practice (praying) might seem inappropriate in some cultural contexts.

Context Dimension Context dimensions, in general, represent the various contexts, such as public, semi-public, private, and semi-private environments. Contextual preferences varied according to whether users preferred a private voice assistant in their homes like

[3] https://www.amazon.com/Echo-Dot-4th-Gen-Kids/dp/B084J4QQK1, last accessed December 19, 2022.

[4] https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice, last accessed December 19, 2022.

Table 6.1: Distribution of questions over categories projected on two dimensions (social and technological, merging context) of the question space [207]

Technology	Acceptance	Relationship	Interaction	Ethics	Privacy	Questions
General	1	5	1	1	1	9
Voice recognition	1	1	1	1	3	7
Voice synthesis	-	-	5	3	-	8
Emotion detection	1	-	1	1	1	4
Emotion synthesis	1	-	2	2	1	6
Personality detection	1	-	-	-	-	1
Personality synthesis	-	1	2	-	1	4
Questions	5	7	12	8	7	39

Alexa or a public voice interface in an elevator. Some circumstances, however, cannot be classified as either private or public. Voice assistants become semi-private when they are used with a mobile phone while using public transportation, or when they are used in private houses with visitors. Similarly, a public voice assistant becomes semi-public if it is restricted to a few users. Finally, there are four categories under context dimension. There are many questions that can be included in the questionnaire based on each category of context dimension.

Questionnaire Summary The technical dimension was composed of seven categories, the social dimension was composed of five, and the context dimension was composed of four. In case all combinations of factors had been considered, it would have resulted in 140 ($7 \times 5 \times 4$) questions at least. This questionnaire, however, would have been too long to locate volunteers who were prepared to answer all questions. In contrast, the questionnaire in this section chose to highlight acceptance and concerns expressed by the community, such as the VAS' default gender voice[5], which is addressed in the media. Further, participants are asked to answer questions about interactions and ethics in this survey.

As depicted in Table 6.1, there are 9 questions in general speech technologies and only one question in speech personality detection. Furthermore, this table contains 12 questions on interaction details and 5 questions on acceptance. The final questionnaire comprised 46 questions, seven regarding demographic information and 39 about user preferences. All the final questions are showed in Table 6.2. This questionnaire was completed by participants in about half an hour on average.

[5] https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html, last accessed December 19, 2022

Table 6.2: A table of all questions in the English version of the questionnaire [207]

	questions on demographic data
A1.	What is your native language?
A2.	How old are you? Please state your age in years.
A3.	Please select your gender.
A4.	What is your professional field or field of study?
A5.	How would you rate your computer skills?
A6.	Which voice assistants have you used before?
A7.	How often do you use voice assistants?
	questions on general technology
B1.	Voice assistants are becoming more and more common. What is your attitude towards voice assistants?
B2.	Do you prefer a standardized voice assistant with the same voice for everyone (like Alexa or Siri) or a voice assistant that has a unique voice exclusively for you?
B3.	Do you want to have conversations with your voice assistant that are not task-related e.g. chitchat?
B4.	Voice assistants can be used by anyone who is able to verbally communicate regardless of age. Do you think there should be an age restriction on the usage of voice assistants by minors?
B5.	Do you want to have a child-mode, so children can play and learn with voice assistants
B6.	Do you want to use a voice assistant that supports your mental health?
B7.	How do you want your voice assistant to deal with your mental health?
B8.	Do you think in some situations (e.g. during the COVID-19 pandemic) voice assistants could substitute conversations with human beings?
B9.	Your voice assistant has to know about your preferences to become personalized. How do you want your voice assistant to learn about your preferences?
	questions on voice recognition
C1.	Do you want to be identified by your voice assistant?
C2.	Do you feel comfortable using voice assistants, knowing they record every interaction with you and send the collected data to the server of a company?
C3.	You use your voice assistant together with other members of your household. How should the voice assistant prioritize multiple or contradicting inputs coming from different people at the same time?
C4.	Do you think it is ethically justifiable to allow voice assistants to respond according to hierarchical structures instead of treating everybody equally? This could lead to e.g. needing your confirmation for your child's commands or prioritizing the commands of the voice assistant's owner.
C5.	A personalized voice assistant that fits your preferences and needs must collect at least some data. What type of data may your voice assistant collect and store, so you still feel comfortable?
C6.	When you use your voice assistant and other people are talking in the same room, their unintentional voice input is also recorded and stored by your voice assistant. Do you think these people should be informed that their voice may be recorded?
C7.	You ask your voice assistant to look up vegan recipes. Your voice assistant replies: "You mentioned last month that you do not like carrots, I, therefore, looked up recipes without carrots for you. Here are my suggestions". The voice assistant adapted its suggestions to the information you gave in the past. Do you want your voice assistant to refer to the information you gave in the past?
	questions on voice synthesis
D1.	What age do you generally prefer for the voice of a voice assistant?

D2.	What gender do you generally prefer for the voice of a voice assistant?
D3.	Which voice do you prefer for a voice assistant designed for children?
D4.	Do you want your personal voice assistant to be able to speak and understand dialects?
D5.	It is possible to synthesize any voice, even the voice of existing people. Which voice do you choose for your voice assistant?
D6.	If there was the possibility to recreate the voice from a person that has passed away, do you think it would be acceptable to use their voice?
D7.	Most voice assistants come with a female voice by default. A news organization stated: “Women have been made into servants once again. Except this time, they’re digital.” Do you agree or disagree?
D8.	Do you believe voice assistants should respond differently depending on the users age, e.g. by using child friendly language?
	questions on emotion detection
E1.	A voice assistant that is able to detect emotions responds more considerately and naturally. Do you want to use a voice assistant that is able to detect your emotions?
E2.	What kind of insight do you want on the emotions your voice assistant detects?
E3.	Do you believe a voice assistant could manipulate you (e.g. for commercial purposes) when being able to detect and interpret your emotions?
E4.	Do you want your voice assistant to store the emotions it detects, so it can track your emotional well-being?
	questions on emotion synthesis
F1.	There are linguistic parameters (e.g. words, phrases) and paralinguistic parameters (e.g. pitch, intensity) in a voice that can be interpreted as emotions. How do you want your personal voice assistant to express emotions?
F2.	Imagine you had a bad day. Your voice assistant detects sadness. How do you want your voice assistant to react to how you feel?
F3.	You use your voice assistant to get a news update. Do you want the device to adapt to the content, e.g. in tone or expressed emotions?
F4.	The emotions of your voice assistant could influence your mood. Which emotions should your voice assistant be allowed to express?
F5.	Do you believe a voice interface should be able to show affection to you?
F6.	Do you want your voice assistant to change which emotions it shows when other people (e.g. guests, roommates, partner, etc.) are present?
	questions on personality detection
G1.	Do you want to use a voice assistant that is able to detect and adapt to your personality?
	questions on personality synthesis
H1.	In what kind of relationship do you want your personal assistant to interact with you?
H2.	You share a voice assistant with other members of your household. Do you prefer to have a device with one personality per household or one that shows different personalities depending on the person using it?
H3.	Active voice assistants initiate conversations, make suggestions, and state their opinion. Passive voice assistants do not initiate conversations but rather wait for your commands and they do not make suggestions without being asked. Do you want your personal voice assistant to be more active or passive?
H4.	You use your mobile voice assistant at home in your room. You leave the house to take the bus. In the bus you decide to use your voice assistant again. How do you want the personality of your voice assistant to adapt to the new surrounding?

6.4 Study Results

In order to collect data, the questionnaire was made available online and distributed through the experimenters' private networks. The study supplemented recruiting with snowball sampling, in which initial participants were requested to circulate the link in order to generate a more varied sample. Participants had the option of receiving credits for their studies in return for completing the questionnaire or winning one of ten shopping vouchers. Additionally, participants in the survey were self-selected, which meant they had the option to stop and not complete the questionnaire. To counteract this possible bias, questionnaires with any empty response fields were totally eliminated. The survey was completed by 576 people, however only 364 completed responses were used for this study.

6.4.1 Demographics

According to the mother language results, there are three different types of cultures. There were 149 German participants, 76 women and 73 men ranging in age from 17 to 64 years ($M = 26.2, SD = 8.30$), and 125 of them had prior VA experience. As for the 102 Chinese participants, 56 were female and 46 were male, ranging in age from 18 to 56 years ($M = 28.8, SD = 5.27$), and 94 had previous VA experience. Additionally, 39 of the 80 Egyptians identified as female, while 41 identified as male, ranging from 20 to 67 years of age ($M = 30.6, SD = 7.73$), and 70 having prior experience with VAs. Among the participants in this survey, twelve selected English as their mother tongue while twenty-one selected another language. Gender was roughly equal across all participants, as illustrated in Table 6.3.

Among the demographic data, there were 171 men (47.9%) and 186 women (52.1%). In terms of mother tongues, 76 (22.3%) identified Arabic as their native tongue, 102 (28.0%) identified Chinese, 11 (3.3%), 148 (40.9%) identified German, and 20 (5.5%) identified other languages. Furthermore, it was reported that three quarters of the participants were proficient in computers, and the majority of them had used voice assistants before. The statistics results are shown in Table 6.3.

Table 6.3: Age (mean and standard deviation) and gender (number and percentage) distribution by mother tongue for 364 participants [207]

Mother	Age (Mean, SD)	Male	Female	Total
German	$M = 26.2, SD = 8.30$	73 (49.0%)	76 (51.0%)	149
Chinese	$M = 28.8, SD = 5.27$	46 (45.1%)	56 (54.9%)	102
Egyptian	$M = 30.6, SD = 7.73$	41 (51.3%)	39 (48.8%)	80
other	$M = 26.6, SD = 6.97$	18 (54.6%)	15 (45.5%)	33
Overall	$M = 27.9, SD = 7.51$	178 (48.9%)	186 (51.1%)	364

General Attitude

What are the users' attitudes towards voice assistants? The following figure depicts the overall attitude toward voice assistants (see Figure 6.1). From the figure 6.1, most participants (58.5%) are positive about voice assistants, while a minority (19.2%) are negative.

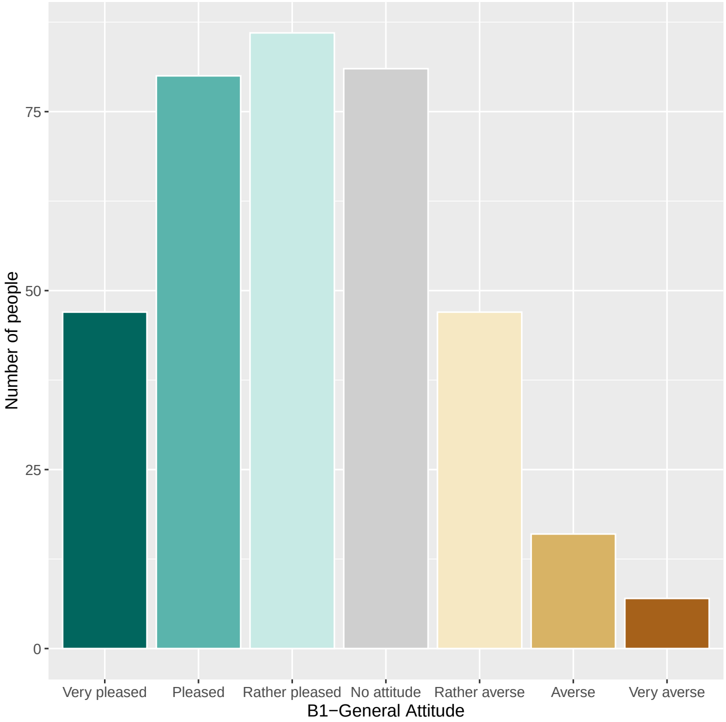


Figure 6.1: General attitude (B1) towards voice assistants for all participants [207]

The chart in Figure 6.2 shows the results of all questions that used Likert-scale-based. Except for two questions, all of them are answered on a 5-point scale from 'Yes' to 'No'. H3 in the questionnaire varies from 'active' to 'passive', whereas B2 goes from 'standardized' to 'tailored'. Bars move to either side of the chart depending on the answers of the question (the left side indicates positive attitude, the right side indicates negative attitude), but all bars move both ways. From the Figure 6.2, the answer of **RQ1** demonstrates that users' attitudes towards virtual assistants are not common, and VAs design have to based on these varying opinions.

Figure 6.3 illustrates the correlations between each pair of Likert-scale questions. It is clear from the matrix that many question pairings are correlated substantially. From the participants' perspective, there are interrelated groups of questions. Answers to most questions are not homogeneous, but span a wide range.

According to this, it may be possible to find certain groups of users answering certain questions homogeneously, which implies that it exists different types of users. This will be covered in further depth in the following part. Figure 6.4 depicts the user attitudes regarding emotion (E1) and personality detection (G1). Based on these two figures on the

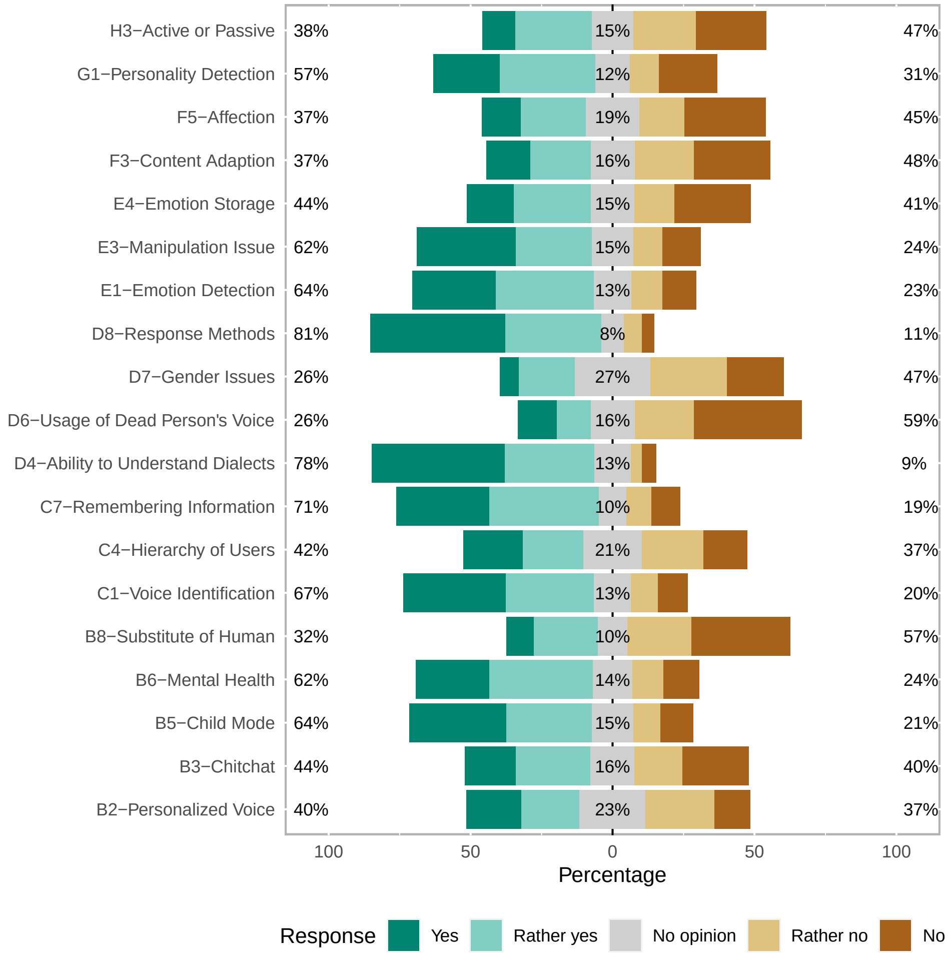


Figure 6.2: Overview of values for all Likert-based questions. The numbers on the gray field in the middle show the percentage of participants with no opinion. The numbers on the right and left side show the percentage of positive (yes and rather yes) and negative (no and rather no) answers respectively. The overview shows that the questions were answered in both directions [207]

diagonal, emotion (E1) is clearly associated with personality detection (G1). In terms of number, the majority of people replied ’rather yes’ to both questions. In both questions, Germans answered ’no’ in the majority.

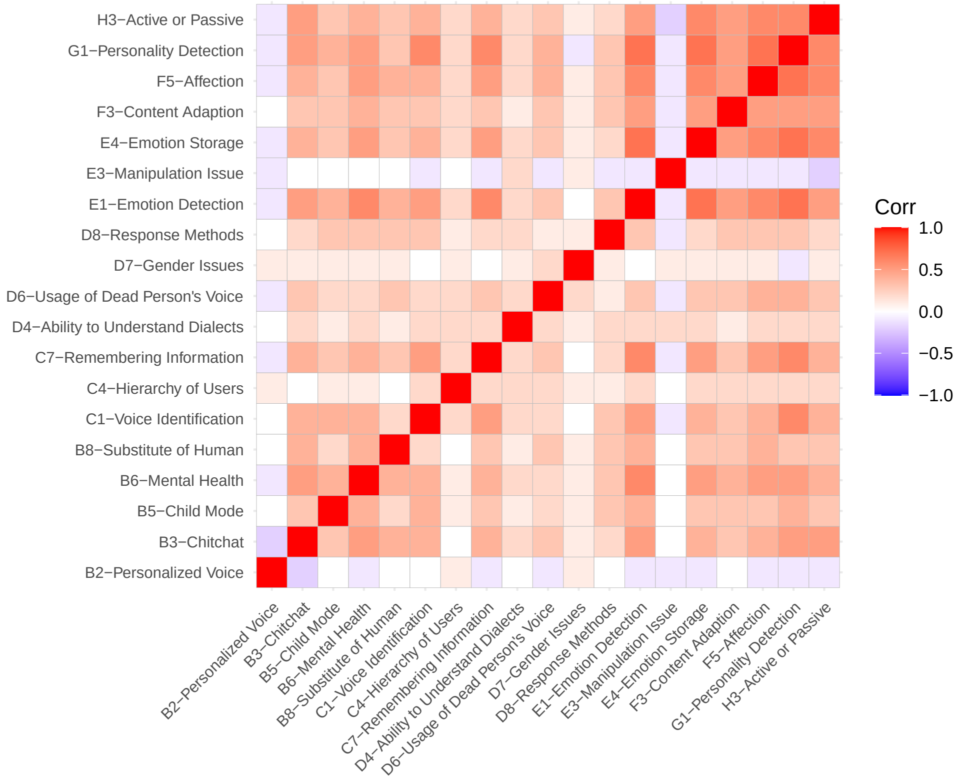


Figure 6.3: Correlation matrix for all Likert-based questions. The overview (Figure 6.2) shows that the questions were answered in great variation. The correlated pairs in the matrix tell that a participant who answered one question in a certain way also answers other question in the same way. From this it can conclude that there might be distinct types of users [207]

6.4.2 Gender Influence

In this study, all Likert-based questions were assessed using the Wilcoxon signed-rank test and the result summary was presented in Table 6.5. As shown in Table 6.5, P-values under 5% significance were found in only five of the 21 questions. It should be noted, however, that when seeking for a 5%-significance level, one of twenty tests may declare significance by chance. Consequently, the gender comparison was tested using a Bonferroni correction and a significance value of 5%/21 =0.0024 for the 21 items. The difference between the two

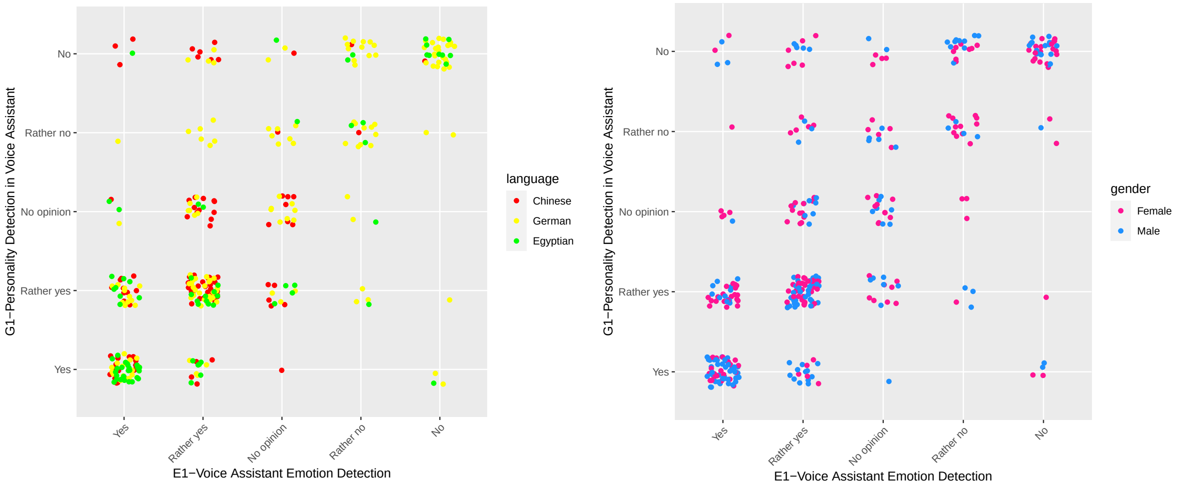


Figure 6.4: Attitude towards emotion and personality detection across cultures (left) and gender (right). The dots are distributed randomly around their integer coordinates to not cover each other. Every dot represents a participant. The data is correlated as most dots lie on the diagonal. In the language plot some dependencies on culture are suggested, while there is a more even distribution for gender [207]

questions (D7, D8) was statistically significant. As a result, it claims that the difference between genders is minor.

Gender problem (D7), which asks about attitudes toward female voice as default, is one of the biggest differences between the genders. Compared to males in the answer, females are more doubtful of this. Given that the questions are gender-specific, it is not surprising that answers vary by gender. A gender-specific distribution of responses is shown on the left side of Figure 6.5.

In addition, there is a large gender difference when it comes to the question of whether a voice assistant should respond differently to a child (D8). The gender distribution of the responses is shown on the right side of Figure 6.5. As can be seen in Figure 6.5, the responses are divided by gender. Females do not consider children to be adults, according to their opinions. However, this result should be interpreted with caution since there was a comparable question (B5) asking if there should be a kid mode, and there was no significant gender difference.

Differences by Culture

Instead of demonstrating vast differences between cultures, the aim of this study was to better understand diverse user demographics. Significant differences in answers across cultures are not necessarily explained by culture. It is not necessary for the culture to explain the difference in answers, even if the answers differ significantly between cultures. A different distribution of user groups or different wealth levels may also explain the differences

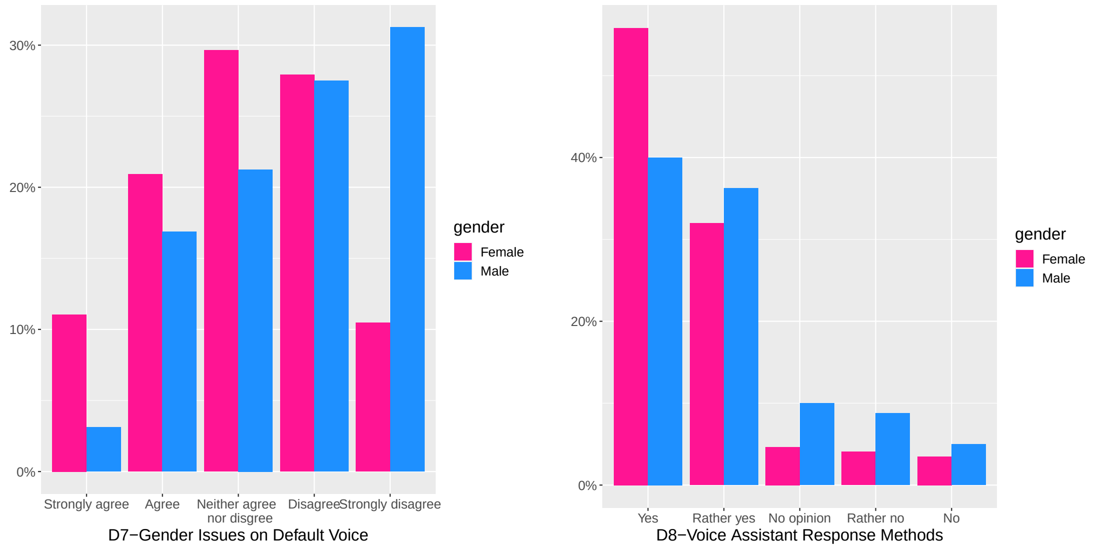


Figure 6.5: Opinion on default female voice (D7) (left) and whether children should get a different response (D8) (right) by gender. These are the only two questions where it found significant differences by gender [207]

between these cultures. It is possible that these cultures differ for other reasons, including different distributions of user groups or different wealth levels.

It is also important to be cautious when making cultural claims based on the number of respondents who started but didn't complete the survey. Several participants reported being bored by the lengthy questioning. The survey may have also been abandoned by some participants because some questions provoked them. In this scenario, several viewpoints would have been lost, resulting in distorted results. This study excluded all results from participants who reported English or another language as their mother tongue due to their small numbers. A table containing the means for the Likert-scale-based questions regarding cultures is available at table 6.4. The Wilcoxon-signed-rank test results are presented in Table 6.5. Due to the fact that 21 items were compared in three ways, the significance threshold was set at 5%/63 using a Bonferrroni adjustment.

Table 6.5 compares general attitudes (B1) across cultures. The attitude toward voice assistants varies from culture to culture. This question is displayed across cultures in Figure 6.6. Among Chinese, most have a positive view of the technology, most Egyptians have a balanced view, and few Germans have a negative view.

The right side of Figure 6.6 shows responses to the question whether women are degraded by a female voice. In all cultures, the majority opposes the statement, but the Chinese agree much more than the Egyptians, and the Germans fall somewhere in the middle.

Table 6.4: Mean values of Likert scale questions by language and gender and for all [207]

Likert Scale Questions	**German**	**Arabic**	**Chinese**	**Male**	**Female**	**All**
B2. Personalized Voice	2.82	2.82	2.98	3.01	2.81	2.90
B3. Chitchat	3.44	2.88	2.52	2.97	3.06	3.01
B5. Child mode	2.63	2.08	2.22	2.43	2.26	2.34
B6. Mental health	2.81	2.24	2.18	2.51	2.45	2.48
B8. Substitute of human	3.94	3.56	2.83	3.61	3.40	3.51
C1. Voice identification	2.58	1.76	2.18	2.20	2.34	2.27
C4. Hierarchy of users	3.07	2.46	2.91	2.76	3.02	2.89
C6. Informing others	1.97	1.68	2.24	1.54	1.58	1.98
C7. Remembering information	2.50	1.93	2.04	2.08	2.42	2.25
D4. Ability to understand dialects	1.89	1.54	2.14	1.87	1.91	1.89
D6. Usage of dead person's voice	3.86	3.90	2.97	3.50	3.65	3.57
D7. Gender issues	3.40	3.60	3.08	3.62	3.06	3.34
D8. Response to children	1.58	2.11	1.68	2.06	1.67	1.86
E1. Emotion detection	1.90	2.87	2.15	2.30	2.52	2.41
E3. Manipulation Issue	2.07	2.16	3.05	2.38	2.43	2.41
E4. Emotion Storage	3.50	2.73	2.79	2.91	3.24	3.08
F3. Content adaption	4.02	2.43	2.70	3.14	3.31	3.23
F5. Affection	3.74	2.86	2.68	3.13	3.32	3.23
G1. Personality detection	3.15	2.23	2.37	2.52	2.89	2.71
H3. Active or passive	3.95	2.76	2.47	3.09	3.33	3.21

Table 6.5: p-values of Wilcoxon signed-rank test for Likert-scale-based questions comparing gender and the three cultures. Values below 0.001 are given as **<.001**. Values at or below 0.05 are shown in **blue**. Values below 0.0024 (0.05/21) for gender and 0.0009 (0.05/63) for culture show significant differences after a Bonferroni correction and are colored in **green** [207]

Likert Scale Questions	**Gender**	**Chinese -German**	**Chinese -Egyptian**	**Egyptian -German**
B1. General Attitude	.157	**<.001**	**.005**	**.039**
B2. Personalized Voice	.145	.366	.472	.997
B3. Chitchat	.529	**<.001**	.216	**.006**
B5. Child mode	.296	.056	**.007**	**<.001**
B6. Mental health	.674	**<.001**	.442	**<.001**
B8. Substitute of human	.102	**<.001**	**.002**	.169
C1. Voice identification	.160	.089	**<.001**	**<.001**
C4. Hierarchy of users	.081	.303	**.016**	**.002**
C6. Informing others	.500	**<.001**	**<.001**	.233
C7. Remembering information	**.018**	.107	**.008**	**<.001**
D4. Ability to understand dialects	.699	.049	**<.001**	**<.001**
D6. Usage of dead person's voice	.367	**<.001**	**<.001**	.342
D7. Gender issues	**<.001**	**.044**	**.005**	.181
D8. Response to children	**<.001**	**<.001**	.928	**<.001**
E1. Emotion detection	.106	**<.001**	.750	**<.001**
E3. Manipulation Issue	.679	**<.001**	**<.001**	.980
E4. Emotion Storage	**.033**	**<.001**	.361	**<.001**
F3. Content adaption	.281	**<.001**	.062	**<.001**
F5. Affection	.201	**<.001**	.641	**<.001**
G1. Personality detection	**.003**	**<.001**	.073	**<.001**
H3. Active or passive	.091	**<.001**	.236	**<.001**

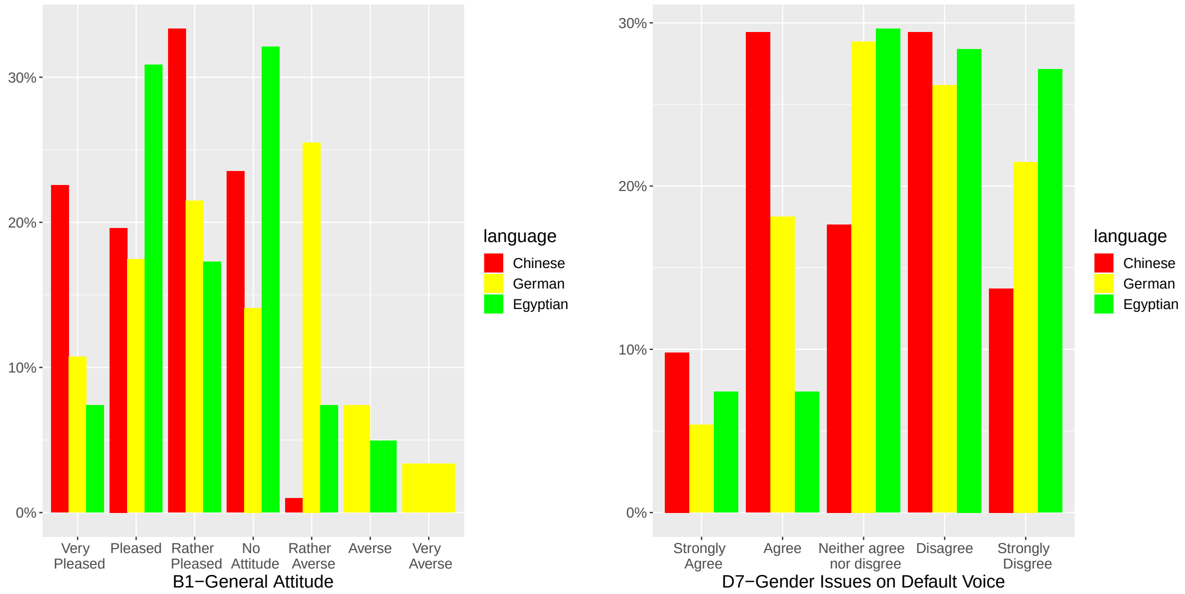


Figure 6.6: Left: General attitude towards voice assistants across cultures. The Chinese have a rather positive attitude, the Egyptians' attitude is balanced, and Germans are more skeptical. Right: Answers for the gender issue across cultures. Egyptians see less problems in a female default voice [207]

6.4.3 User Types: *Enthusiasts*, *Pragmatists*, and *Sceptics*

Results for all participants are correlated, as shown in section 6.4.1 and Figure 6.3. This, together with the fact that there are participants on both sides of the scale in Figure 6.2, strongly suggests that different types of users exist. On the basis of the findings in this study, cluster detection was performed. In order to explore clusters, all Likert-scale questions were used. In particular, there are a lot of them, including B1, B3, B5, B6, B8, C1, C4, C6, C7, D4, D6, D7, D8, E1, E3, E4, F3, F5, G1. Therefore, 19-dimensional space was explored to find clusters. There is a uniform range of numeric values for each of these Likert-scale questions.

Clustering is defined as the grouping of data points, according to Jain et al .[151], and the two primary approaches that are regarded acceptable for a clustering analysis of Likert-scale data are K-means and agglomerative hierarchical clustering [220, 61, 198]. It was the same approach that was used to identify user groups in [109]. In general, a principle component analysis (PCA) [85] is performed before a cluster discovery search is conducted to minimize the number of dimensions in the algorithms. A version of K-means known as K-medoids clustering [153] has been proposed as being more robust against noise and outliers, and Shamsuddin et al. [295] have demonstrated this to be true. Consequently, K-medoids were utilized using PCA and the PAM (Partitioning Around Medoids) method [159]. Both K-means and K-medoids have a critical parameter, K, which can be selected using the elbow technique, silhouette coefficient algorithm, or gap statistics algorithm [349]. The

K-medoids cluster identification in this part was based on both the elbow method (shown on the left side of Figure 6.7) and gap methodology (shown on the right side of Figure 6.7) which recommended $k = 3$. The final results are shown on the left side of Figure 6.8.

After analyzing the clusters, the next step was to determine their meanings. A closer examination revealed that the three groups represent people who are enthusiastic, neutral, or suspicious about emotion- and personality-aware voice assistants. To test this assumption, the study plotted each group's general attitude (B1). This result is shown on the right hand side of Figure 6.8 and appears to validate the hypothesis. These user groups were subsequently referred to as *Enthusiasts*, *Pragmatists*, and *Sceptics*, respectively. As a final consideration, culture and gender were taken into account when assessing cluster membership. Figure 6.9 displays the results. The distribution of user types among cultures varies despite the fact that they are distributed equally by gender. There appears to be a large majority of *Enthusiasts* or *Pragmatists* Chinese, with only a small number of *Sceptics*. Conversely, the Germans have the highest percentage of *Sceptics*. Approximately half of Egyptians are positive, while the other half are *Pragmatists* or *Sceptics*.

The Dunn Index was calculated to determine if automatically discovered clusters partitioned data better than clusters based on language or gender. The Dunn Index [241, 59] determines the distance between data within each cluster by comparing the shortest distance between data from different clusters. A higher Dunn Index generally corresponds to a smaller cluster diameter and a wider cluster spacing. There is a Dunn Index of 0.116 in gender clusters. According to the Dunn Index *Dculture*, cultural groups are ranked at 0.091. The Dunn Index is higher for users of type *Dusertype* than for users of type *Ausertype* and *Busertype*. Instead of separating users by their gender or culture, it is more descriptive to categorize them by types of users.

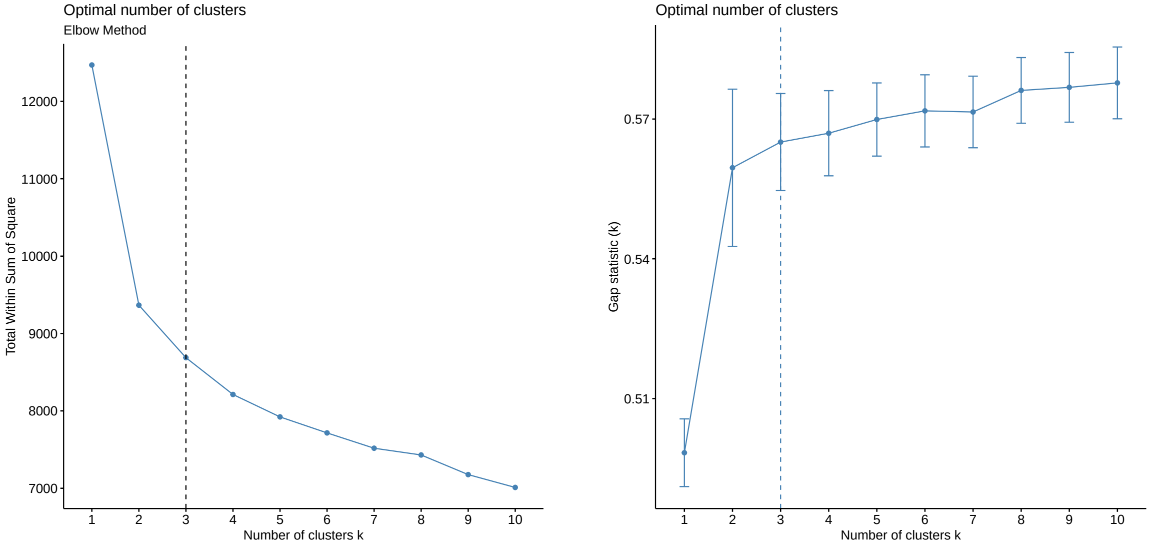


Figure 6.7: k-value estimation with the elbow method (left) and gap statistics (right). Both methods suggest 3 for the number of clusters [207]

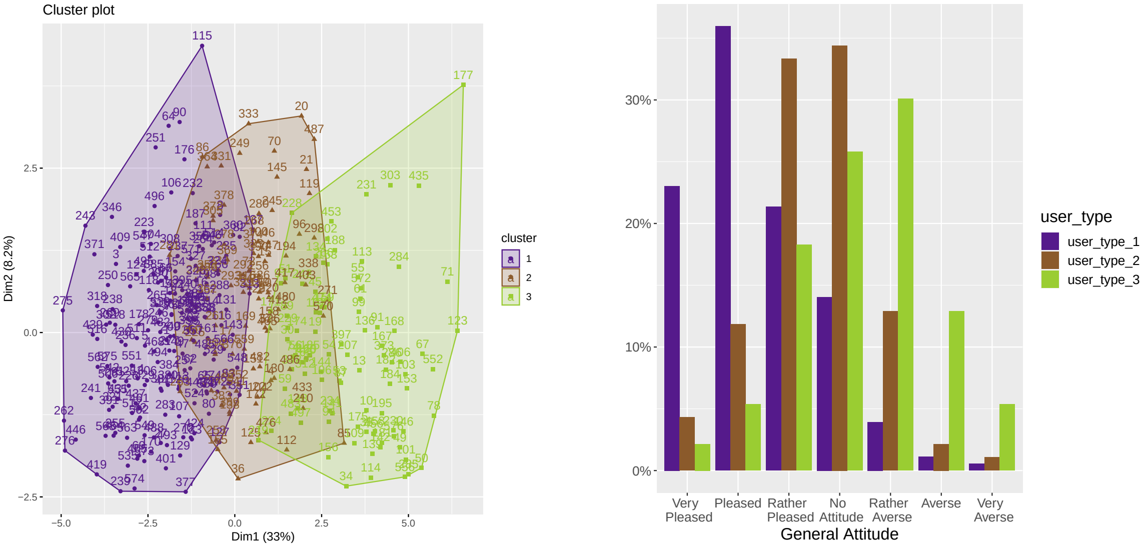


Figure 6.8: Left: Visualization of the three clusters detected by K-medoids cluster detection via PCA. Right: Distribution of general attitude (B1) by membership in a cluster. This bar chart allows to associate cluster member to attitude and gives the clusters a meaning. The three clusters represent positive, neutral and skeptical users [207]

6.4.4 Selected Results

This section discusses the most valuable insight from the study results based on their relevance in current VA research and their timely relevance.

Dialects In the survey (D4), most participants expressed a preference for voice assistants that speak several languages. Essentially, a dialect question is about how it feels to be understood and to be able to communicate in “my language”. In spite of the fact that there are no differences based on gender, there are differences based on culture (see Figure 6.10). Since Arabic is spoken in many countries, each with its own dialect, Egyptians seemed to be the most interested in dialects. Amazon’s Alexa platform supports five distinct dialects of English and three distinct dialects of Spanish. The demand for dialects will lead to future voice assistants speaking additional dialects.

Voice of People who Passed Away The purpose of this study was to establish ethical boundaries. Participants were asked about the possibility of synthesizing existing voices (if recordings exist), including those of deceased individuals (D6). A person may be speaking to you through a member of their family or a renowned individual, such as a famous actor, entertainer, or state official. These people cannot be asked for consent anymore. Whether it is ethical to resurrect people in this manner is also debated. In spite of this, it did not explain or provide examples of such features. Figure 6.11 shows the final result. Participants are overwhelmingly opposed to the concept. According to the results by

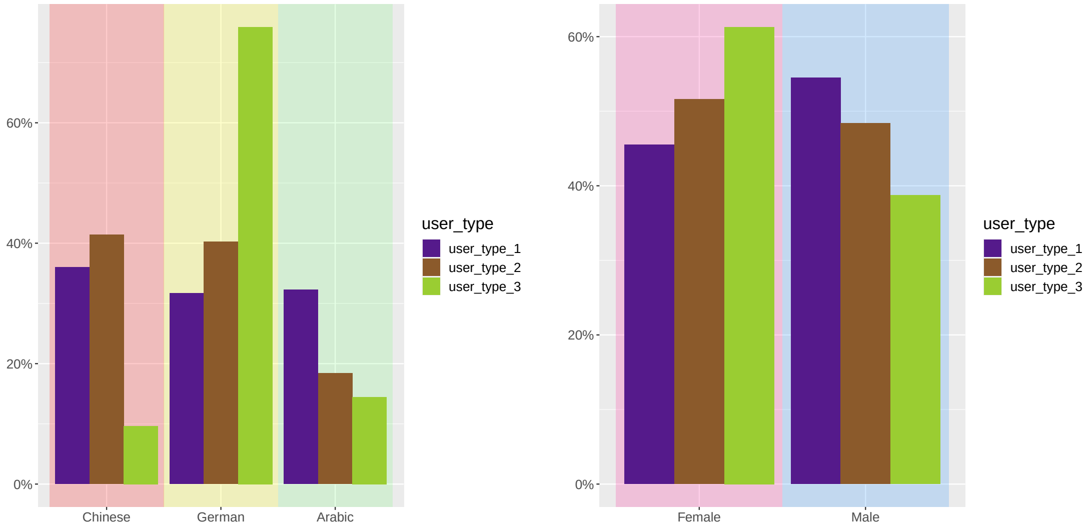


Figure 6.9: Cluster membership by culture (left) and gender (right). The violet bars represent the *Enthusiasts*, the brown bars the *Pragmatists*, and the light green bars the *Sceptics*. Every culture has all types of users, however the distribution varies. For gender, the user types are almost equally distributed [207]

culture, Egyptians are the most concerned.

Taking Care of Mental Health The importance of taking care of mental health is outlined in the following paragraph. A delicate subject like mental health deserves special attention. It is important to keep health issues, especially psychological ones, private in order to maintain good relationships. Both voice assistants and doctors formed similar levels of trust, and it was interesting to observe whether one was more trustworthy than the other. This diagram shows the results for question B6 in Figure 6.12. It is common for people to want their mental health taken care of by a voice assistant. Moreover, this implies that an artificial intelligence's diagnosis can be very trusted.

Private and Public Use In this paragraph, it explored whether a smartphone voice assistant should change its personality when on public transportation rather than in a private setting (H4). The possible answers were:

- Less emotional, but more serious personality
- Funnier personality, to show people around me, that I am a fun person
- More sensitive to my emotions, so the device will not annoy or stress me in addition to the environmental stress
- More neutral personality like Siri or Alexa to keep my preference private

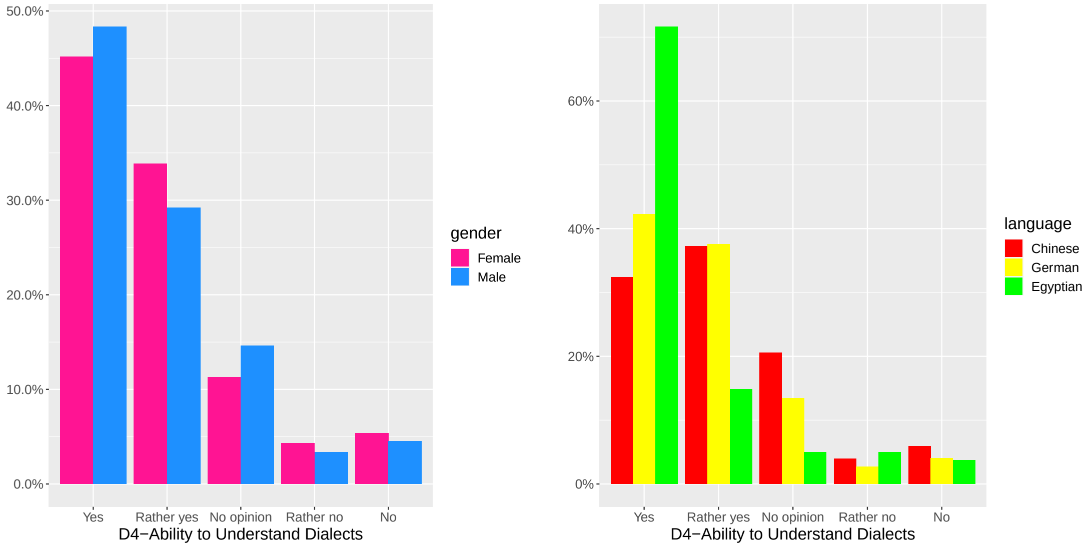


Figure 6.10: Results for question D4 on whether a voice assistant should be able to speak dialects. The left side shows the results by gender, the right side by culture [207]

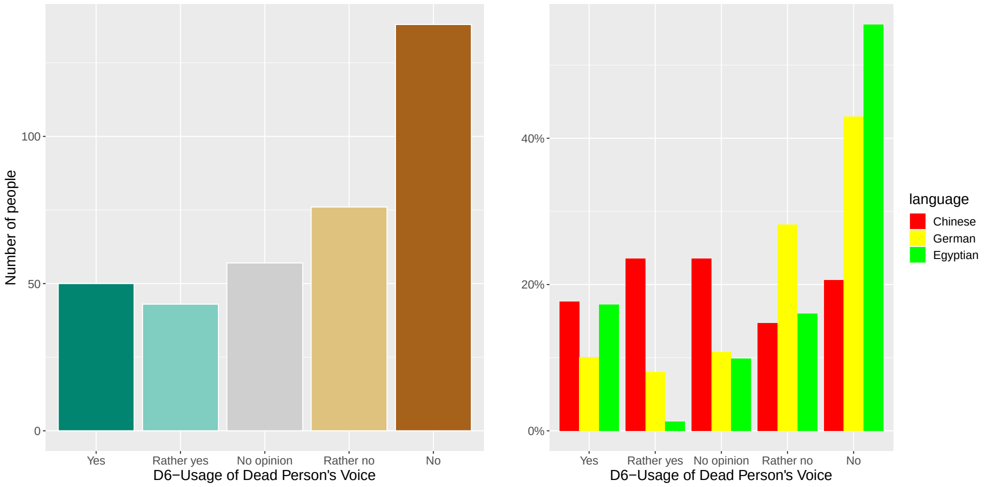


Figure 6.11: Results for question D6 on the usage of voices from people who passed away. The left side shows the results for all, the right side by culture [207]

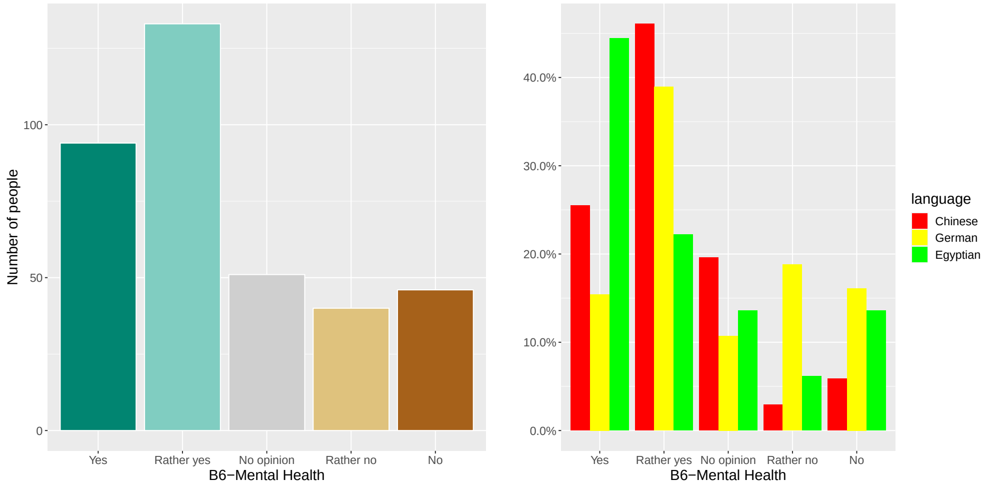


Figure 6.12: Results for question B6 on whether a voice assistant should take care of the user's mental health. The left side shows the results for all, the right side by culture [207]

- No change in personality
- Other

Figure 6.13 shows the results regarding VAs personality adaptation. It is the desire of the majority of people for their public personalities to be neutral. The second largest group is made up of people who are unconcerned with their privacy and do not wish to alter their personality in any way. Personalized voice assistants aren't the only devices that reveal user information through personal communication devices. Using a mobile device's personalized ringtone, the mobile device already knows about the problem.

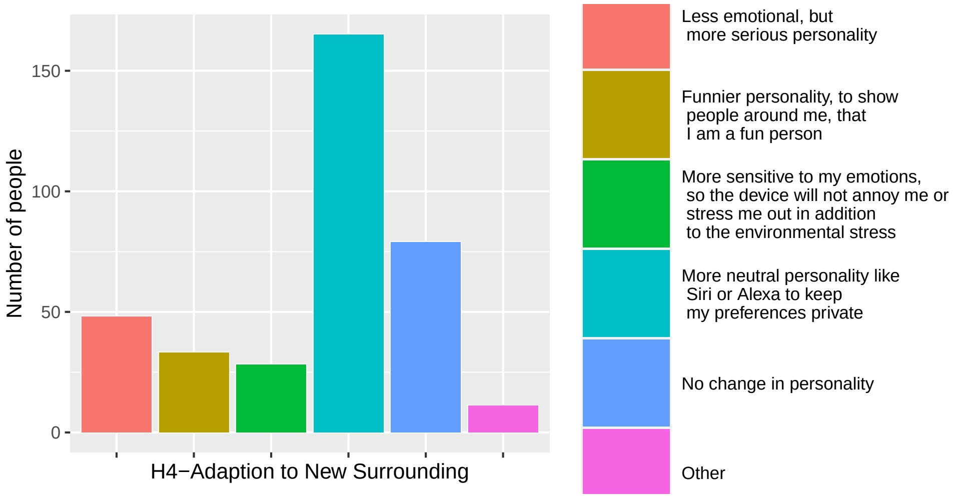


Figure 6.13: Results for question H4. Most people want to have a neutral voice assistant personality in public [207]

Obeying Hierarchies On the Likert-scale ranging from ‘yes’ to ‘no’, responses to the question of whether the voice assistant should obey hierarchies are evenly divided. There seems to be no consensus on this issue. Despite male participants’ tendency to obey hierarchies, there are no apparent differences between genders (see Figure 6.14 left side). Answers by culture are not evenly distributed (see Figure 6.14 right side). Egyptians reach a peak when they vote ‘yes’, Chinese when they vote ‘no opinion’, and Germans when they vote ‘very no’.

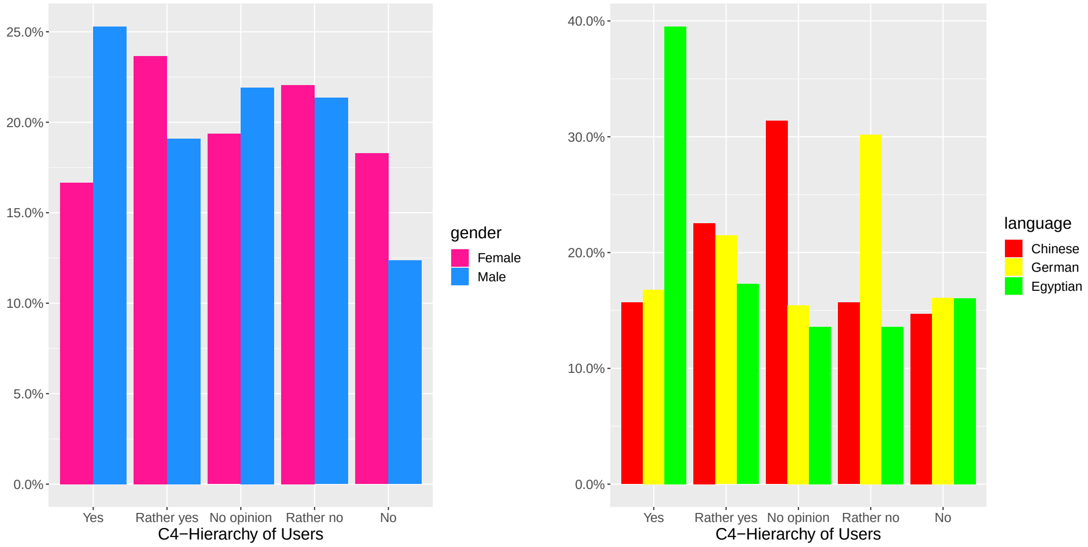


Figure 6.14: Results for question C4 on whether a voice assistant should obey hierarchies. The left side shows the results by gender, the right side by culture [207]

It is important to consider the user hierarchy for voice assistants. Imagine that a child tells a voice assistant to raise the temperature in a smart house controlled by a voice assistant. A suggestion is made by the father to lower the heating. Children are able to override the father’s commands with existing voice assistants since they always execute the last instruction. At this point, it may be beneficial for the voice assistant to be aware of the hierarchy. If the mother enters, it may raise the question of whether she is on the same level as the father, on a lower level, or on a higher level. Their dilemma is the same as their grandparent’s. Although they may be hidden or unsaid, most families have hierarchy levels (possibly complicated). Using a hierarchy-aware VAs, hierarchies can be made explicit, making future VAs much more intelligent.

6.5 Study Discussion and Analysis

This study’s findings provide answers to the beginning questions, such as users’ attitudes, gender differences, and cultural influences. In terms of users’ attitudes toward future emotion- and personality-aware VAs (**RQ1**), it is not difficult to find that most participants are willing to accept these VAs based on Figure 6.1 and Figure 6.4. Except for two obvious gender-related criteria, there does not appear to be a significant difference between genders for the **RQ2**. Therefore, there is no point in developing gender-specific Voice Assistants. This study will also attempt to identify whether users' preferences for VAs are influenced by cultural differences (**RQ3**). As a result of the results, there appear to be some intercultural differences, but given the uncertainties surrounding these judgments, these differences appear to be minor. Contrary to these differences, there was no clear design principle that could allow for customizing voice assistants accordingly. Finally, cross-cultural attitudes toward emotion-and personality-aware VAs can be divided into three user types. This implies that emotion-aware Voice Assistants can be universally designed and used across cultures. Language and culture do not seem to influence emotions or attitudes towards voice assistants using these features. The cluster analysis revealed that attitudes can be more accurately characterized than culture, language, or gender by categorizing users as *Enthusiasts*, *Pragmatists*, and *Sceptics* (**RQ 4**). In what ways can this information be utilized to improve VAs? In general, these three clusters differ primarily in how much they agree that their Voice Assistants are emotional. Consequently, voice assistants will almost certainly need to be programmable in terms of their emotional behavior and emotion detection. A scaling system would allow the user to select the level of emotionality of the assistant, and instructions such as "Alexa, fewer emotions please" would be possible.

A simple user profile, however, does not indicate a default level of emotionality (e.g., language, culture, gender). To improve emotion awareness in future VAs, a low emotionality setting just above neutral should be used as the starting point, followed by periodic questions asking the user how they feel. As an alternative, it is possible to configure directly via preferences or the startup dialog, but it is less likely to be used, since it does not provide users with any immediate benefit. According to its ultimate solution, the VA would be able to self-adjust to an optimal level of emotionality based on the perception of its users. There will, however, be a substantial amount of work to be done in the future.

Additionally, this study left many unanswered questions, and perhaps suggested new ones, such as how to handle emotions detected by VAs. In order to reverse or counteract them, what intensity should they aim for? Based on Völkel et al. [332] recent work, new research has begun investigating the desired characteristics of VAs. However, for scalability across markets, a multicultural perspective is required. Moreover, it is unclear how indirect data from conversations can be used to determine a match's personality. Several recent studies have investigated the difference between artificial and actual emotions by [208] and [156]. In some cases, it is unclear which ones VAs should respond to (Chapter 5.3).

Additionally, this study has several limitations. Despite the fact that there are many participants from three different cultures, it is still difficult to establish a rather general basis that can be helpful to future VAs. The study should recruit more participants from other countries, including the United States and American countries. Furthermore, the majority of participants come from university networks, where academics are overrepresented. Additionally, there were also some participants who felt provoked by some questions, and some averse opinions were not taken into account. As a result, positive attitudes may be more prevalent. Moreover, public opinions can change over time, so a questionnaire conducted over several years may produce different results. It is, however, a temporary evaluation of user attitudes toward emotion-aware voice assistants because the field of voice assistants and emotion recognition is currently developing very rapidly.

6.6 Summary

Despite the increasing use of virtual assistants in daily life, designing future emotion-aware VAs is a challenge. This issue can be resolved by considering the attitudes of users toward future emotion- and personality-aware voice assistants. Taking these factors into account, this chapter presented a study about users' attitudes toward future VAs.

In addition to the research question about user attitudes, this chapter also proposed the other two questions - whether culture influences VA preferences, and whether gender influences VA preferences. In order to find the answers, three dimensions of technology, society, and context were considered in the design of the survey. A total of 46 questions were included in this questionnaire, 7 relating to demographic information and 39 pertaining to user preferences. Furthermore, 364 participants from Germany, China, and Egypt were invited to participate in this study to explore variations and similarities in attitudes. The results of this study show that there is not too much gender difference and there are cultural differences. Moreover, clusters are evident in the Likert-scale questions, since most of the Likert-scale questions correlate with one another. According to cluster analysis, there are three basic types of users (positive, neutral, and skeptical) across cultures. Future VIs applications can be designed based on this finding.

Chapter 7

Conclusion and Outlook

This thesis exploits the several issues in VIs, especially emotion-aware techniques in VIs. As it well known, it is feasible for VIs to detect users' emotional states using speech emotion recognition and generate the emotional voice using speech emotion synthesis. However, due to the limits of existing emotional speech datasets, speech emotion recognition techniques cannot reliably identify users' emotional states, especially in the wild. It is also difficult to distinguish genuine or disguised emotions. Based on these issues, the thesis investigated the teachers' speech emotion in the classroom teaching using neural network with attention mechanism. Moreover, it presented a user study which explores whether users can trick emotion-aware voice assistant, as well as how the VIs react to users' emotions. In addition, the thesis also investigates the users' attitudes and preferences towards the future emotional-aware VAs across cultures using a questionnaire survey.

In the final chapter, it summarizes the research contribution and answers the research questions that were addressed throughout the thesis. Furthermore, it presents a perspective towards future research directions in this chapter.

7.1 Discussion and Conclusion

Due to the shortage of existing VIs, the thesis explores several aspects of emotion-aware VIs. Speech signal analysis can be employed in VIs because it can provide an interactive bridge between users and VIs. Speech emotions are also important in VIs, and emotion-aware technologies that can be implemented in VIs can improve the user experience. Furthermore, based on the existing limitations of VIs, this thesis attempts to investigate prospective VIs applications using two user studies. More specifically, it conducts brief user research to see if people can trick an emotion-aware VI. Once future emotion-aware systems can detect and synthesize users' emotions, devising a strategy for coping with users' emotions is another potential application. This section focuses mostly on the thesis summary and user study discussion based on the aforementioned.

7.1.1 Speech Feature Analysis (TQR1)

As previously stated in section 3, speech features analysis techniques can be implemented in future VIs to enhance the users' experience in VIs. On the one hand, speech features analysis can be employed in voice recognition, speech synthesis, speech augmentation, as well as other speech technologies, and these speech technologies play a significant part in VIs. On the other hand, voice features analysis could be utilized to study the user's emotional states, mental health, social behavior, etc., which is critical for future VIs. Based on their significance, the thesis presented user research on gauging restoration effects using speech features analysis. Compared to other measurements such as attention scales or response tests, speech features analysis is both less obtrusive and more accessible.

In the user study as presented in Chapter 3, short-time energy and zero-crossing rate in the time domain and MFCC in the frequency domain are correlated with the attentional capacity measured by traditional ratings, and thus speech features analysis can be potentially utilized to detect and evaluate the restorative effect in the same way as traditional measurements, such as attention scales or response test. However, due to the constraints of the number of participants and the experimental circumstances, it is difficult for the participants to be properly exhausted or relaxed, and much more data is necessary to verify the experimental results. Furthermore, due to the impact of noise and unpleasant surroundings, participants are unable to convey their true feelings while being observed by experimenters. A chatbot based on speech signal analysis could improve the assessment process even further by eliminating the human experimenter and generating better audio from participants who feel under less surveillance. After iterating on the speech features analysis chain, it is possible to provide a fully automatic assessment system based on acoustic characteristics in the future, which can effectively assess the restorative effects of VREs in automated driving while also providing attention measurements in other study setups.

7.1.2 Emotional Speech in VIs (TQR 2)

Speech emotion recognition has become essential for future emotion-aware VIs. With the advancement of voice emotion detection and synthesis, emotion-aware approaches can be able to create an interactive bridge between users and VIs while also increasing the user's interaction experience. Generally, speech emotion recognition consists of speech signal preprocessing, speech feature extraction, speaker recognition, and speech emotion recognition. The first and most significant step that can help in feature extraction is preprocessing. It was involved in speech pre-emphasis, normalization of vocal tract length, framing, windowing, etc. Speaker recognition can help to get the speaker's voice, which is crucial for detecting the speaker's emotions. Considering the limits of supervised approaches and the difficulties of getting large databases in speaker recognition, unsupervised methods provide an alternative for recognizing the speaker's voice. This thesis proposes an unsupervised speaker recognition method based on WT-KLD. The wavelet transform can be utilized to minimize noise, and the KL-Divergence can be employed to separate the voices of the speakers from the rest of the audio data. Promising results are achieved in the interview and classroom teaching scenarios. However, not all of the results are highly accurate. On the one hand, wavelet transform is not the ideal approach for reducing noise in speech, even if it is currently the best choice. On the other hand, various speech features that are associated with distinct speech and can be employed in speaker detection may exist. Finding high-performance and efficient unsupervised speaker recognition algorithms is essential for future research.

Once the speaker's voice has been obtained, recognizing their emotion becomes another challenge. In this part, it proposes a novel speech emotion recognition system based on a Bi-LSTM network with an attention mechanism in classroom teaching context. This method can overcome the limitations of previous traditional methods and produce promising results. In general, in speech emotion recognition, the previous speech frame emotion might correlate with the subsequent one, and the recurrent neural network may successfully tackle this issue. Furthermore, each speech frame may contain a variety of speech emotions, and the attention mechanism can assist in determining which speech frames contain the primary emotion in speech emotion recognition. Despite the fact that this study collected 45 minutes of classroom instruction data, there isn't a lot of collected teacher voice data. Furthermore, even if the annotation is based on hand labeling, there is a possibility of inaccurate labeling in emotion annotation. Incorrect speaker segmentation may have an impact on emotion detection in the final speech. After iterating on the speech processing chain, it is possible to eventually provide a fully automatic real-time system based on the proposed structure, which can, for example, effectively assess the teachers' real-time emotions while also providing interaction between the teachers, students, and system.

7.1.3 Emotion-Aware VIs (TQR 3 and 4)

Although there are not many emotion-aware VI applications available right now, emotion-aware techniques are still a hot research topic. With the progress of artificial intelligence, voice emotion recognition, and synthesis, it is possible to incorporate these technologies into the present VIS. The new VIs that use these techniques can recognize users' emotions and respond to them with suitable emotions. However, because of the limitations of the present acted emotional speech database, recognizing people's true emotions is the first challenge. It is especially difficult to determine whether users can imitate real-life emotions. In the other words, it's unknown if users can be able to fool future emotion-aware VIs. Once emotion-aware VIs are capable of recognizing users' emotions and engaging with their reactions based on those emotions, the next hurdle is dealing with those emotions. The thesis develops two user studies based on these two difficulties. One is the capacity to simulate actual emotions, and the other is coping with the emotions of users.

Emotional speaking is rather crucial in naturally interacting between users and VIS. Users' spoken emotions can be correctly detected by VIs, and VIs can respond to users with suitable emotions. However, it is questionable if existing emotion recognition systems correctly detect such performed emotions, or rather the speaker's actual feeling. In the thesis, user research was designed to see whether users can trick emotion-aware VIs or whether users cannot simulate true emotions. A small group of participants (26) were instructed to imitate five fundamental emotions, and an open-source emotion-in-voice detector (OpenVokaturi) was utilized to determine if their performed emotion was recognized as intended. According to the findings of this study, it was difficult for participants to mimic all five emotions, and certain emotions were easier to mimic than others. Participants, on the other hand, may find it difficult to "cheat" the emotion-aware VIs. On the one hand, because most of the emotional databases in this system are performed emotion databases, OpenVokaturi may not be a competent spoken emotion detector. On the other hand, the experiments last around 2 seconds, making it difficult for participants to communicate their true feelings within this time. Furthermore, the study itself may have an impact on the participants' emotional state.

Dealing with the user's emotions, particularly negative emotions, is another challenge. As previously stated, if users' emotions can be consistently recognized, the following question is how an emotion-aware interface should react to the discovered emotion. The thesis offered a strategy for alternating the roles of users and VIs based on human emotional reaction strategies. Three avatar emojis (angry, sad, and scared) were created to communicate certain emotions with animation and sound. The user research included 52 participants who attempted to transform the emojis into a desired emotional state (positive feelings) using emotional voice input. The results suggest that users mostly employed neutral emotion to react to these three unpleasant emotions, and there is a gender difference in emotional reaction. However, because study tests primarily employed male voices as emotional stimuli, variations in reaction to negative emotions by gender remain unknown.

7.1.4 Users' Attitudes Towards the Future Emotion-Aware Voice Assistants (TQR 5)

As AI progresses, more speech techniques are being included in the VIS, such as emotion- and personality-aware technologies. The issues in VIs, such as the societal consequences, ethical boundaries, and general user attitudes remain unexplored. The thesis presented here investigates users' attitudes toward and preferences for emotionally aware VAs in three distinct cultures. This survey employed an online questionnaire with 364 participants to investigate differences and similarities in attitudes in Germany, China, and Egypt. According to the findings, the majority of participants have a favorable impression of emotion- and personality-aware voice assistants. Furthermore, there are just a few statistically significant variances for gender, and the differences are primarily cultural. On the one hand, most participants, regardless of gender, are eager to embrace new technology. However, there are disparities among cultures due to differences in cognition, knowledge, privacy, security, ethics, and so on. On the other hand, the participants are from the university's network, and the majority of them are over-represented. Another major disadvantage of surveys is self-report bias.

Furthermore, utilizing a cluster analysis, the thesis demonstrates the three primary user types (positive, neutral, and skeptical) across all cultures. The three clusters differ mostly on their level of agreement with their Voice Assistants having emotional characteristics.

Furthermore, the three clusters created in the thesis can be used to construct future VIs.

7.1.5 Conclusion

Emotion-aware VIs of the future are ones that can recognize users' emotions and respond emotionally to them. Speech features analysis and speech emotion detection are critically important in this emotion-aware VIs. On the one hand, speech features analysis can be employed in future automated vehicles, such as monitoring restorative effects in-car. Speech emotion recognition, on the other hand, can be utilized to identify user's speech emotion, such as teachers' emotion detection based on LSTM network with attention mechanism, and this technology can be implemented in VIs to detect users' emotions. Furthermore, identifying imitated and actual emotions, as well as dealing with users' emotions in emotion-aware VIs, can improve the user interaction experience in VIs.

The results of Chapter 6 reveal that most participants welcome emotion- and personality-aware VAs in general. It is for this reason that SER and some speech techniques are valuable. As aforementioned, attention restorative effects can be measured using speech features analysis. These techniques can also be embedded into future VIs to detect users' restoration effects in other contexts. Furthermore, speech emotion recognition is quite important for future VIs, and LSTM with attention mechanisms can help improve recognition accuracy. In the case of mimicking emotion-aware VIs, not all participants will be able to successfully mimic certain emotions, such as "Happy" or "Angry". Future emotion-aware VIs can benefit from this foundation. Additionally, despite the fact that no acceptable approach exists for emotionally responding to users, gender differences and speech features analysis may provide a clue for future studies.

7.2 Outlook

As AI and voice technology progress, more and more speech functionalities are being integrated in emotion-aware VIs. However, there are certain technical issues that need to be addressed currently. On the one hand, it is critical to detect the true emotions based on the present acted emotions. With the advancement of deep learning algorithms and speech emotion recognition, it is now feasible to detect speech emotions with high accuracy. However, recognizing users' emotions is difficult, especially in a real-world environment. It is difficult to improve real emotion detection using acted emotional speech databases since most emotional speech databases are imitated emotions. On the other side, although voice emotion synthesis techniques exist, it still requires an effective approach to respond to users' feelings in emotion-aware VIs. As previously stated, future emotion-aware VIs can be able to recognize users' emotions and reply to them with appropriate emotional voices.

As these problems are overcome, future emotion-aware VIs can be able to perceive users' feelings and respond emotionally to them. Emotion-aware VIs, in particular, can first record and identify users' voices, and then recognize users' voice emotions using speech emotion recognition and natural language understanding algorithms. Following that, VIs will interpret the users' voice and reply appropriately depending on dialog management. Moreover, VIs will reply to users with appropriate emotions. In addition to that, future VIs may be included in the driverless car, and users' mental and restorative states can be recognized. Users can have a fantastic time engaging with VIs in this system. Furthermore, this VIs can be implemented in classroom education, which can assist in the interaction between students and their teachers when their emotions are appropriately recognized. Ultimately, future emotion-aware VIs will not only facilitate effective communication with users, but also provide constructive feedback after monitoring their mental states, social behaviors, speech emotions, etc.

List of Figures

List of Tables

Bibliography

[1] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.

[2] I. Abdic, L. Fridman, D. McDuff, E. Marchi, B. Reimer, and B. Schuller. Driver frustration detection from audio and video in the wild. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1354–1360. Springer, 2016.

[3] C. Abras, D. Maloney-Krichmar, and J. Preece. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction.*, 37(4):445–456, 2004.

[4] A. G. Adami. Automatic speech recognition: From the beginning to the portuguese language. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 1–73. Springer, 2010.

[5] P. K. Adelmann and R. B. Zajonc. Facial efference and the experience of emotion. *Annual review of psychology*, 40(1):249–280, 1989.

[6] A. Agarwal and A. Meyer. Beyond usability: evaluating emotional response as an integral part of the user experience. In *Extended Abstracts of the 2009 CHI Conference on Human Factors in Computing Systems*, pages 2919–2930. ACM, 2009.

[7] C. Ai, H. Zhao, R. Ma, and X. Dong. Pipeline damage and leak detection based on sound spectrum lpcc and hmm. In *6th International Conference on Intelligent Systems Design and Applications*, volume 1, pages 829–833. IEEE, 2006.

[8] M. B. Akçay and K. Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[9] T. Akiyama, S. Takamichi, and H. Saruwatari. Prosody-aware subword embedding considering japanese intonation systems and its application to dnn-based multi-dialect speech synthesis. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 659–664. IEEE, 2018.

[10] M. Ali, A. H. Mosa, F. A. Machot, and K. Kyamakya. Emotion recognition involving physiological and speech signals: A comprehensive review. *Recent advances in nonlinear dynamics and synchronization*, pages 287–302, 2018.

[11] F. Allison, M. Carter, M. Gibbs, and W. Smith. Design patterns for voice interaction in games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 5–17. ACM, 2018.

[12] S. Amiriparian and B. Schuller. AI hears your health: Computer audition for health monitoring. In *International Conference on ICT for Health, Accessibility and Wellbeing*, volume 1538, pages 227—-233. Springer Nature, 2022.

[13] G. An, S. I. Levitan, J. Hirschberg, and R. Levitan. Deep personality recognition for deception detection. In *Interspeech*, pages 421–425. ISCA, 2018.

[14] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[15] H. Aouani and Y. Ben Ayed. Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder. In *4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–5. IEEE, 2018.

[16] A. C. Arevian, D. Bone, N. Malandrakis, V. R. Martinez, K. B. Wells, D. J. Miklowitz, and S. Narayanan. Clinical state tracking in serious mental illness through computational analysis of speech. *PLoS One*, 15(1):1–17, 2020.

[17] M. Athineos and D. P. Ellis. Frequency-domain linear prediction for temporal features. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266. IEEE, 2003.

[18] M. P. Aylett, A. Vinciarelli, and M. Wester. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing*, 11(2):361–372, 2017.

[19] J. A. Bachorowski. Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57, 1999.

[20] A. A. Badr and A. K. Abdul-Hassan. A review on voice-based interface for human-robot interaction. *Iraqi Journal for Electrical And Electronic Engineering*, 16(2):91–102, 2020.

[21] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.

[22] L. R. Bahl, P. F. Brown, P. V. de Souza, and M. Picheny. Acoustic markov models used in the tangora speech recognition system. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP' 88)*, pages 497–498. IEEE Computer Society, 1988.

[23] K. Bahreini, R. Nadolski, and W. Westera. Towards real-time speech emotion recognition for affective e-learning. *Education and information technologies*, 21(5):1367–1386, 2016.

[24] Z. Bai and X. Zhang. Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99, 2021.

[25] A. Baird, S. Amiriparian, and B. Schuller. Can deep generative audio be emotional? towards an approach for personalised emotional audio generation. In *IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[26] P. Banerjee, B. Chakraborty, and J. Banerjee. Procedure for cepstral analysis in tracing unique voice segments. In *2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 351–356. IEEE, 2015.

[27] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[28] P. Barros, N. Churamani, E. Lakomkin, H. Sequeira, A. Sutherland, and S. Wermter. The OMG-Emotion Behavior Dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1408–1414. IEEE, 2018.

[29] P. Barros, E. Lakomkin, H. Sequeira, A. Sutherland, and S. Wermter. Access to the omg emotion dataset. https://www2.informatik.uni-hamburg.de/wtm/OMG-EmotionChallenge, 2018.

[30] R. Batish. *Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger*. Packt Publishing Ltd, 2018.

[31] R. Beale and C. Peter. The role of affect and emotion in HCI. In *Affect and Emotion in Human-Computer Interaction*, pages 1–11. Springer, 2008.

[32] D. Beirl, Y. Rogers, and N. Yuill. Using voice assistant skills in family life. In *Computer-Supported Collaborative Learning Conference (CSCL)*, pages 96–103. International Society of the Learning Sciences, Inc., 2019.

[33] C. Bérubé, T. Schachner, R. Keller, E. Fleisch, F. v Wangenheim, F. Barata, and T. Kowatsch. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: systematic literature review. *Journal of Medical Internet Research*, 23(3):e25933, 2021.

[34] N. Bhatia et al. Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*, 2010.

[35] T. Bickmore and J. Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 396–403. ACM, 2001.

[36] W. Biesmans, N. Das, T. Francart, and A. Bertrand. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):402–412, 2016.

[37] D. Bone, C. C. Lee, T. Chaspari, J. Gibson, and S. Narayanan. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine*, 34(5):196–195, 2017.

[38] D. Bonet, G. Cámbara, F. López, P. Gómez, C. Segura, and J. Luque. Speech enhancement for wake-up-word detection in voice assistants. *arXiv preprint arXiv:2101.12732*, 2021.

[39] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[40] M. Braun, A. Mainz, R. Chadowitz, B. Pfleging, and F. Alt. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11. ACM, 2019.

[41] M. Braun, A. Mainz, R. Chadowitz, B. Pfleging, and F. Alt. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–11. ACM, 2019.

[42] S. Brave and C. Nass. Emotion in human-computer interaction. *Human-Computer Interaction Fundamentals*, pages 53–68, 2009.

[43] S. Brederoo, F. Nadema, F. Goedhart, A. Voppel, J. De Boer, J. Wouts, S. Koops, and I. Sommer. Implementation of automatic speech analysis for early detection of psychiatric symptoms: What do patients want? *Journal of psychiatric research*, 142:299–301, 2021.

[44] R. N. Brewer, L. Findlater, J. Kaye, W. Lasecki, C. Munteanu, and A. Weber. Accessible voice interfaces. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 441–446. ACM, 2018.

[45] M. Buchanan. Behavioural science: Secret signals. *Nature News*, 457(7229):528–530, 2009.

[46] I. Burić and A. C. Frenzel. Teacher emotional labour, instructional strategies, and students' academic engagement: a multilevel analysis. *Teachers and Teaching*, pages 1–18, 2020.

[47] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *9th European Conference on Speech Communication and Technology*, volume 5, pages 1517–1520. ISCA, 09 2005.

[48] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520. ISCA, 2005.

[49] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[50] M. Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[51] J. T. Cacioppo and W. L. Gardner. Emotion. *Annual review of psychology*, 50(1):191–214, 1999.

[52] E. Cambria, A. Livingstone, and A. Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.

[53] A. Caranica, H. Cucu, C. Burileanu, F. Portet, and M. Vacher. Speech recognition results for voice-controlled assistive applications. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE, 2017.

[54] C. Carreiras, A. Lourenço, H. Aidos, H. P. da Silva, and A. L. Fred. Unsupervised analysis of morphological ecg features for attention detection. In *Computational Intelligence*, pages 437–453. Springer, 2016.

[55] G. Castellano, L. Kessous, and G. Caridakis. Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, pages 92–103. Springer, 2008.

[56] K.-h. Chang. *Speech Analysis Methodologies towards Unobtrusive Mental Health Monitoring.* PhD thesis, UC Berkeley, 2012.

[57] K.-h. Chang, D. Fisher, and J. Canny. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense*, 2011.

[58] M. Chang and J. Taxer. Teacher emotion regulation strategies in response to classroom misbehavior. *Teachers and Teaching*, 27(5):353–369, 2021.

[59] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. NbClust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61(1):1–36, 2014.

[60] S. Chen, J. Dai, and Y. Yan. Classroom teaching feedback system based on emotion detection. In *9th International Conference on Education and Social Science (ICESS)*, pages 940–946. The International Society for Engineers and Researchers (ISER), 2019.

[61] L. K. Cheng, A. Selamat, M. H. M. Zabil, M. H. Selamat, R. A. Alias, F. Puteh, F. Mohamed, and O. Krejcar. Comparing the accuracy of hierarchical agglomerative and k-means clustering on mobile augmented reality usability metrics. In *2019 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 34–40. IEEE, 2019.

[62] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, pages 1–18, 2021.

[63] B. Cimprich, M. Visovatti, and D. L. Ronis. The attentional function index—a self-report cognitive measure. *Psycho-oncology*, 20(2):194–202, 2011.

[64] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini. The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6):365–388, 2020.

[65] L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, J. Edwards, and B. R. Cowan. The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4):349–371, 2019.

[66] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, V. Wade, and B. R. Cowan. *What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents*, page 1–12. ACM, 2019.

[67] P. Clarkson and P. J. Moreno. On the use of support vector machines for phonetic classification. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP' 99)*, pages 585–588. IEEE, 1999.

[68] A. Cooper, R. Reimann, D. Cronin, and C. Noessel. *About face: the essentials of interaction design*. John Wiley & Sons, 2014.

[69] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[70] R. Craggs and M. Wood. A two dimensional annotation scheme for emotion in dialogue. In *Proceedings of AAAI spring symposium: exploring attitude and affect in text*, pages 1–6. AAAI Press, 2004.

[71] N. Cummins, A. Baird, and B. W. Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

[72] J. R. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802, 2007.

[73] F. Cutugno and R. Savy. Correlation between segmental reduction and prosodic features in spontaneous speech: the role of tempo. In *Proceedings of the 12th International Congress of Phonetic Sciences (ICPhS)*, pages 471–474. International Phonetic Association, 1999.

[74] J. Dai, V. Vijayarajan, X. Peng, L. Tan, and J. Jiang. Speech recognition using sparse discrete wavelet decomposition feature extraction. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0812–0816. IEEE, 2018.

[75] A. R. Damasio. Review: Toward a neurobiology of emotion and feeling: Operational concepts and hypotheses. *The Neuroscientist*, 1(1):19–25, 1995.

[76] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey. Fundamentals, present and future perspectives of speech enhancement. *International Journal of Speech Technology*, 24(4):883–901, 2021.

[77] D. L. Day. Shared values and shared interfaces: The role of culture in the globalisation of human-computer systems. *Interacting with Computers*, 9(3):269–274, 1998.

[78] J. R. Deller Jr. Discrete-time processing of speech signals. In *Discrete-time processing of speech signals*, pages 908–908. IEEE Press, 1993.

[79] J. Deng and F. Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, pages 1–1, 2021.

[80] G. Deshpande and B. Schuller. An overview on audio, signal, speech, & language processing for covid-19. *arXiv preprint arXiv:2005.08579*, 2020.

[81] H. Detjen, S. Faltaous, S. Geisler, and S. Schneegass. User-defined voice and mid-air gesture commands for maneuver-based interventions in automated vehicles. In *Proceedings of Mensch und Computer 2019*, pages 341–348. ACM, 2019.

[82] N. Dey. *Intelligent speech signal processing.* Elsevier, 2019.

[83] L. S. Dhupati, S. Kar, A. Rajaguru, and A. Routray. A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings. In *2010 IEEE International Conference on Automation Science and Engineering*, pages 917–921. IEEE, 2010.

[84] R. F. Dickerson, E. I. Gorlin, and J. A. Stankovic. Empath: a continuous remote emotional health monitoring system for depressive illness. In *Proceedings of the 2nd Conference on Wireless Health*, pages 1–10. ACM, 2011.

[85] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning*, ICML '04, page 29. ACM, 2004.

[86] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara. Speech emotion recognition 'in the wild' using an autoencoder. In *Interspeech*, pages 526–530. ISCA, 2020.

[87] S. D'Mello and R. A. Calvo. Beyond the basic emotions: what should affective computing compute? In *Extended Abstracts of the 2013 CHI Conference on Human Factors in Computing Systems*, pages 2287–2294. ACM, 2013.

[88] L. Docio-Fernandez and C. Garcia-Mateo. *Speech Production*, pages 1290–1295. Springer US, 2009.

[89] D. Dojchinovski, A. Ilievski, and M. Gusev. Interactive home healthcare system with integrated voice assistant. In *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 284–288. IEEE, 2019.

[90] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

[91] S. P. Dubagunta, B. Vlasenko, and M. M.-D. Doss. Learning voice source related information for depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE, 2019.

[92] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[93] M. Egger, M. Ley, and S. Hanke. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science*, 343:35–55, 2019.

[94] P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992.

[95] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[96] P. E. Ekman and R. J. Davidson. *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[97] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[98] S. Emerich and E. Lupu. Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks. In *Interspeech*, pages 1168–1172. ISCA, 2011.

[99] I. S. Engberg and A. V. Hansen. Documentation of the danish emotional speech database (des). *Internal AAU report, Center for Person Kommunikation, Denmark*, page 22, 1996.

[100] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman. Front-end speech enhancement for commercial speaker verification systems. *Speech Communication*, 99:101–113, 2018.

[101] V. Evers and D. Day. The role of culture in interface acceptance. In *IFIP TC13 Interantional Conference on Human-Computer Interaction (INTERACT'97)*, pages 260–267. Springer, 1997.

[102] S. B. Eysenck and A. M. Abdel-Khalek. A cross-cultural study of personality: Egyptian and english children. *International Journal of Psychology*, 24(1-5):1–11, 1989.

[103] M. S. Fahad, A. Ranjan, J. Yadav, and A. Deepak. A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, 110:102951, 2021.

[104] H. M. Fayek, M. Lech, and L. Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[105] M. Feidakis. A review of emotion-aware systems for e-learning in virtual environments. *Formative assessment, learning data analytics and gamification*, pages 217–242, 2016.

[106] O. Frandsen-Thorlacius, K. Hornbæk, M. Hertzum, and T. Clemmensen. Non-universal usability?: A survey of how usability is understood by chinese and danish users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 41–50. ACM, 2009.

[107] M. Franěk, D. Šefara, J. Petružálek, J. Cabal, and K. Myška. Differences in eye movements while viewing images with various levels of restorativeness. *Journal of Environmental Psychology*, 57:10–16, 2018.

[108] A. C. Frenzel, B. Becker-Kurz, R. Pekrun, T. Goetz, and O. Lüdtke. Emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology*, 110(5):628, 2018.

[109] E. Frias-Martinez, S. Y. Chen, and X. Liu. Survey of data mining approaches to user modeling for adaptive hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(6):734–749, 2006.

[110] S. Furui. *Digital speech processing, synthesis, and recognition.* CRC Press, 2018.

[111] G. Gainotti. Neuropsychological theories of emotion. *The neuropsychology of emotion*, pages 214–236, 2000.

[112] Z. Gan, R. Wang, Y. Yu, and X. Zhao. Voice conversion from Tibetan Amdo dialect to Tibetan U-tsang dialect based on StarGAN-VC2. In *2020 International Conference on Big Data Economy and Information Management (BDEIM)*, pages 184–187. IEEE, 2020.

[113] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan. Towards understanding emotional intelligence for behavior change chatbots. In *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 8–14. IEEE, 2019.

[114] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

[115] L. H. Gilpin, D. M. Olson, and T. Alrashed. Perception of speaker personality traits using speech signals. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM, 2018.

[116] L. N. Girouard-Hallam, H. M. Streble, and J. H. Danovitch. Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, 3(5):1118–1131, 2021.

[117] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[118] H. P. Greeley, E. Friets, J. P. Wilson, S. Raghavan, J. Picone, and J. Berg. Detecting fatigue from voice using speech recognition. In *2006 IEEE International Symposium on signal processing and information technology*, pages 567–571. IEEE, 2006.

[119] J. J. Gross. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26, 2015.

[120] J. Guo. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, 31(1):113–126, 2022.

[121] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal processing magazine*, 36(6):111–124, 2019.

[122] G. Hagenauer, T. Hascher, and S. E. Volet. Teacher emotions in the classroom: associations with students' engagement, classroom discipline and the interpersonal teacher-student relationship. *European Journal of Psychology of Education*, 30(4):385–403, 2015.

[123] M. G. Hall, A. V. Oppenheim, and A. S. Willsky. Time-varying parametric modeling of speech. *Signal Processing*, 5(3):267–285, 1983.

[124] O. K. Hamid. Frame blocking and windowing speech signal. *Journal of Information, Communication, and Intelligence Systems (JICIS)*, 4(5):87–94, 2018.

[125] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, pages 223–227. ISCA, 2014.

[126] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan. Ordinal learning for emotion recognition in customer service calls. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6494–6498. IEEE, 2020.

[127] J. H. L. Hansen and S. E. Bou-Ghazale. Getting started with susas: a speech under simulated and actual stress database. In *5th European Conference on Speech Communication and Technology*, pages 1743–1746. ISCA, 1997.

[128] S. Haq and P. J. Jackson. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global, 2011.

[129] S. Haq, P. J. Jackson, and J. Edge. Audio-visual feature selection and reduction for emotion classification. In *Proc. Auditory-Visual Speech Processing*, pages 185–190. ISCA, 2008.

[130] S. Haq, P. J. Jackson, and J. Edge. Speaker-dependent audio-visual emotion recognition. In *AVSP 2009–International Conference on Audio-Visual Speech Processing*, volume 2009, pages 53–58. ISCA, 2009.

[131] A. O. Hatch, S. S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Interspeech*, pages 1471–1474. ISCA, 2006.

[132] J. Heaton. *Introduction to neural networks with Java.* Heaton Research, Inc., 2008.

[133] J. Hernandez, D. McDuff, X. Benavides, J. Amores, P. Maes, and R. Picard. Autoemotive: Bringing empathy to the driving experience to manage stress. In *Proceedings of the 2014 Companion Publication on Designing Interactive Systems*, DIS Companion '14, page 53–56. ACM, 2014.

[134] M. Hertzum, T. Clemmensen, K. Hornbæk, J. Kumar, Q. Shi, and P. Yammiyavar. Usability constructs: a cross-cultural study of how users and developers experience their use of information systems. In *International Conference on Usability and Internationalization*, pages 317–326. Springer, 2007.

[135] G. Hofstede. The globe debate: Back to relevance. *Journal of International Business Studies*, 41(8):1339–1346, 2010.

[136] A. Holliday. Complexity in cultural identity. *Language and Intercultural Communication*, 10(2):165–177, 2010.

[137] M. A. Hossan, S. Memon, and M. A. Gregory. A novel approach for mfcc feature extraction. In *4th International Conference on Signal Processing and Communication Systems*, pages 1–5. IEEE, 2010.

[138] B. House, J. Malkin, and J. Bilmes. The voicebot: a voice-controlled robot arm. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 183–192. ACM, 2009.

[139] M. B. Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[140] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9):1457–1469, 2004.

[141] P. W. Hsiao and C. P. Chen. Effective attention mechanism in dynamic models for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2526–2530. IEEE, 2018.

[142] C. C. Hsu and L. W. Ku. Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues. In *Proceedings of the 6th international workshop on natural language processing for social media*, pages 27–31. IEEE, 2018.

[143] Z. Huang, M. Dong, Q. Mao, and Y. Zhan. Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804. ACM, 2014.

[144] Z. Huang, J. Epps, and D. Joachim. Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6549–6553. IEEE, 2020.

[145] IBM. Pioneering speech recognition. Website, 2011. https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/.

[146] Y. A. Ibrahim, J. C. Odiketa, and T. S. Ibiyemi. Preprocessing technique in automatic speech recognition for human computer interaction: an overview. *Annals of the University of Craiova, Mathematics and Computer Science Series*, 15(1):186–191, 2017.

[147] E. Inwood and M. Ferrari. Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: A systematic review. *Applied Psychology: Health and Well-Being*, 10(2):215–235, 2018.

[148] E. Isaacs, A. Konrad, A. Walendowski, T. Lennig, V. Hollis, and S. Whittaker. *Echoes from the Past: How Technology Mediated Reflection Improves Well-Being*, page 1071–1080. ACM, 2013.

[149] C. E. Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280, 2007.

[150] P. Jackson and S. ul haq. Surrey audio-visual expressed emotion (SAVEE) database, 2011.

[151] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[152] M. Jalil, F. A. Butt, and A. Malik. Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In *International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, pages 208–212. IEEE, 2013.

[153] X. Jin and J. Han. *K-Medoids Clustering*, pages 564–565. Springer US, 2010.

[154] V. S. Johnston. *Why we feel: The science of human emotions.* Perseus Publishing, 1999.

[155] B. H. Juang and L. R. Rabiner. Automatic speech recognition–a brief history of the technology development. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf, 2004.

[156] P. Juslin, P. Laukka, and T. Bänziger. The mirror to our soul? Comparisons of spontaneous and posed vocal expression of emotion. *Journal of Nonverbal Behavior*, 42:1– 40, 2018.

[157] C. Kaiser and R. Schallner. The impact of emotional voice assistants on consumers' shopping attitude and behavior. *Wirtschaftsinformatik 2022 Proceedings*, 2022.

[158] M. Kamppuri, R. Bednarik, and M. Tukiainen. The expanding focus of HCI: case culture. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 405–408, 2006.

[159] L. Kaufman and P. J. Rousseeuw. *Partitioning Around Medoids (Program PAM)*, pages 68–125. John Wiley & Sons, Inc., 2008.

[160] A. Keerio, B. K. Mitra, P. Birch, R. Young, and C. Chatwin. On preprocessing of speech signals. *International Journal of Signal Processing*, 5(3):216–222, 2009.

[161] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub. A review on speech emotion recognition: Case of pedagogical interaction in classroom. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–7. IEEE, 2017.

[162] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[163] H. Kim and J.-S. Park. Automatic language identification using speech rhythm features for multi-lingual speech recognition. *Applied Sciences*, 10(7):2225, 2020.

[164] J. Kim and E. André. Emotion recognition using physiological and speech signal in short-term observation. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 53–64. Springer, 2006.

[165] S. Kim, N. Kwon, H. O'Connell, N. Fisk, S. Ferguson, and M. Bartlett. "How are you?" Estimation of anxiety, sleep quality, and mood using computational voice analysis. In *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5369–5373. IEEE, 2020.

[166] N. Kimura, M. Kono, and J. Rekimoto. SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11. ACM, 2019.

[167] A. M. Klein, A. Hinderks, M. Rauschenberger, and J. Thomaschewski. Exploring voice assistant risks and potential with technology-based users. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020)*, pages 147–154. Science and Technology Publications, 2020.

[168] P. R. Kleinginna and A. M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379, 1981.

[169] B. Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018.

[170] E. A. Konijn and H. C. Van Vugt. Emotions in mediated interpersonal communication: Toward modeling emotion in virtual humans. In *Mediated interpersonal communication*, pages 114–144. Routledge, 2008.

[171] S. G. Koolagudi, Y. Murthy, and S. P. Bhaskar. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, 21(1):167–183, 2018.

[172] O. Korn, L. Stamm, and G. Moeckl. Designing authentic emotions for non-human characters: A study evaluating virtual affective behavior. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 477–487. ACM, 2017.

[173] T. Kostoulas, I. Mporas, O. Kocsis, T. Ganchev, N. Katsaounos, J. J. Santamaria, S. Jimenez-Murcia, F. Fernandez-Aranda, and N. Fakotakis. Affective speech interface in serious games for supporting therapy of mental disorders. *Expert Systems with Applications*, 39(12):11072–11079, 2012.

[174] Z. Kövecses. *Emotion concepts*. Springer Science & Business Media, 2012.

[175] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*, 84:65–75, 2012.

[176] J. Krajewski, D. Sommer, T. Schnupp, T. Laufenberg, C. Heinze, and M. Golz. Applying nonlinear dynamics features for speech-based fatigue detection. In *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*, pages 1–5. ACM, 2010.

[177] J. Krajewski, R. Wieland, and A. Batliner. An acoustic framework for detecting fatigue in speech based human-computer-interaction. In *International Conference on Computers for Handicapped Persons*, pages 54–61. Springer, 2008.

[178] L. Lamy. Beyond emotion: Love as an encounter of myth and drive. *Emotion Review*, 8(2):97–107, 2016.

[179] B. Lance and S. Marsella. Glances, glares, and glowering: how should a virtual human express emotion through gaze? *Autonomous Agents and Multi-Agent Systems*, 20(1):50–69, 2010.

[180] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects*, pages 1–8. Bordeaux, 2007.

[181] O. Lartillot, P. Toiviainen, and T. Eerola. A matlab toolbox for music information retrieval. In *Data analysis, machine learning and applications*, pages 261–268. Springer, 2008.

[182] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller. Augmenting generative adversarial networks for speech emotion recognition. *arXiv preprint arXiv:2005.08447*, 2020.

[183] S. Latif, J. Qadir, A. Qayyum, M. Usama, and S. Younis. Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356, 2020.

[184] H. Lausberg and H. Sloetjes. Coding gestural behavior with the neuroges-elan system. *Behavior research methods*, 41(3):841–849, 2009.

[185] J. E. LeDoux. Emotional processing, but not emotions, can occur unconsciously. *The nature of emotion: Fundamental questions*, 1994:291–292, 1994.

[186] A. Lee, K. Oura, and K. Tokuda. Mmdagent—a fully open-source toolkit for voice interaction systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8382–8385. IEEE, 2013.

[187] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. Emotion recognition based on phoneme classes. In *8th International Conference on Spoken Language Processing*. ICSLP, 2004.

[188] D. Lee, K.-J. Oh, and H.-J. Choi. The chatbot feels you-a counseling service using emotional response generation. In *2017 IEEE international conference on big data and smart computing (BigComp)*, pages 437–440. IEEE, 2017.

[189] J.-E. R. Lee and C. I. Nass. Trust in computers: The computers-are-social-actors (casa) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives*, pages 1–15. IGI Global, 2010.

[190] K. F. Lee, H. W. Hon, M. Y. Hwang, S. Mahajan, and R. Reddy. The sphinx speech recognition system. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP' 89)*, pages 445–448. IEEE, 1989.

[191] P. Lei, M. Chen, and J. Wang. Speech enhancement for in-vehicle voice control systems using wavelet analysis and blind source separation. *IET Intelligent Transport Systems*, 13(4):693–702, 2019.

[192] Y. Leviathan and Y. Matias. Google duplex: An AI system for accomplishing real-world tasks over the phone, 2018.

[193] J. Li, Y. Ma, P. Li, and A. Butz. A journey through nature: Exploring virtual restorative environments as a means to relax in confined spaces. In *Creativity and Cognition*, pages 1–9. ACM, 2021.

[194] J. Li, Y. Ma, and C. Ou. Cultivation and incentivization of HCI research and community in china: Taxonomy and social endorsements. In *CHI '19 Workshop: HCI in China*. ACM, 2019.

[195] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 312–317. IEEE, 2013.

[196] Z. Lian, B. Liu, and J. Tao. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021.

[197] J. Liang, S. Chen, J. Zhao, Q. Jin, H. Liu, and L. Lu. Cross-culture multimodal emotion recognition with adversarial learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4000–4004. IEEE, 2019.

[198] K. C. Lim, A. Selamat, M. H. Mohamed Zabil, Y. Yusoff, M. H. Selamat, R. A. Alias, F. Puteh, F. Mohamed, and O. Krejcar. A comparative usability study using hierarchical agglomerative and k-means clustering on mobile augmented reality interaction data. In *Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques*, pages 258–271. IOS Press, 2019.

[199] N. Lim. Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative Medicine Research*, 5(2):105–109, 2016.

[200] S. Linxen, C. Sturm, F. Brühlmann, V. Cassau, K. Opwis, and K. Reinecke. How weird is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, 2021.

[201] R. Lippmann, E. Martin, and D. Paul. Multi-style training for robust isolated-word speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'87)*, volume 12, pages 705–708. IEEE, 1987.

[202] S. Liu, A. Mallol-Ragolta, and B. W. Schuller. Covid-19 detection with a novel multi-type deep fusion method using breathing and coughing information. In *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1840–1843. IEEE, 2021.

[203] Z. Liu, B. Hu, L. Yan, T. Wang, F. Liu, X. Li, and H. Kang. Detection of depression in speech. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 743–747. IEEE, 2015.

[204] Z. Liu, A. Rehman, M. Wu, W. Cao, and M. Hao. Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features. *IEEE Transactions on Multimedia*, 2020.

[205] H. Lövheim. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses*, 78(2):341–348, 2012.

[206] B. T. Lowerre. *The harpy speech recognition system.* Carnegie Mellon University, 1976.

[207] Y. Ma, Y. Abdelrahman, B. Petz, H. Drewes, F. Alt, H. Hussmann, and A. Butz. Enthusiasts, pragmatists, and skeptics: Investigating users' attitudes towards emotion- and personality-aware voice assistants across cultures. In *Proceedings of Mensch und Computer 2022*, pages 308–322. ACM, 2022.

[208] Y. Ma, H. Drewes, and A. Butz. Fake moods: Can users trick an emotion-aware voicebot? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–4. ACM, 2021.

[209] Y. Ma, H. Drewes, and A. Butz. How should voice assistants deal with users' emotions? *arXiv preprint arXiv:2204.02212*, 2022.

[210] Y. Ma, J. Li, H. Drewes, and A. Butz. You sound relaxed now - Measuring restorative effects from speech signals. In *IFIP Conference on Human-Computer Interaction*, pages 585–594. Springer, 2021.

[211] K. Mahajan, D. R. Large, G. Burnett, and N. R. Velaga. Exploring the effectiveness of a digital voice assistant to maintain driver alertness in partially automated vehicles. *Traffic injury prevention*, 22(5):378–383, 2021.

[212] S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *9th European Conference on Speech Communication and Technology*, pages 1–4, 2005.

[213] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski. Affectaura: An intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 849–858. ACM, 2012.

[214] D. McEvoy and B. R. Cowan. The importance of emotional design to create engaging digital HCI learning experiences. *On the Horizon*, 9(5):1–6, 2016.

[215] K. McRae and J. J. Gross. Emotion regulation. *Emotion*, 20(1):1, 2020.

[216] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin, and C. Pelachaud. Evaluation of four designed virtual agent personalities. *IEEE Transactions on Affective Computing*, 3(3):311–322, 2012.

[217] A. Mehrabian. *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies.* Oelgeschlager, Gunn & Hain, 1980.

[218] H. L. Meiselman. *Emotion measurement*. Woodhead publishing, 2016.

[219] B. Mesquita and N. H. Frijda. Cultural variations in emotions: a review. *Psychological bulletin*, 112(2):179, 1992.

[220] C. Michalopoulou and M. Symeonaki. Improving likert scale raw scores interpretability with k-means clustering. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 135(1):101–109, 2017.

[221] W. Mieleszczenko-Kowszewicz, K. Warpechowski, K. Zieliński, R. Nielek, and A. Wierzbicki. Tell me how you feel: Designing emotion-aware voicebots to ease pandemic anxiety in aging citizens. *arXiv preprint arXiv:2207.10828*, 2022.

[222] S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[223] A. Mohamed, D. Yu, and L. Deng. Investigation of full-sequence training of deep belief networks for speech recognition. In *Interspeech*, pages 2846–2849. ISCA, 2010.

[224] A.-r. Mohamed, G. Dahl, and G. Hinton. Deep belief networks for phone recognition. *Nips workshop on deep learning for speech recognition and related applications*, 1(9):39, 2009.

[225] G. Mohammadi and A. Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284, 2012.

[226] S. Moskowitz and J.-M. Dewaele. Is teacher happiness contagious? A study of the link between perceptions of language teacher happiness and student attitudes. *Innovation in Language Learning and Teaching*, 15(2):117–130, 2021.

[227] S. Murali, F. Rincon, and D. Atienza. A wearable device for physical and emotional health monitoring. In *2015 Computing in Cardiology Conference (CinC)*, pages 121–124. IEEE, 2015.

[228] C. M. Myers, A. Furqan, and J. Zhu. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–9. ACM, 2019.

[229] D. G. Myers. Theories of emotion. *Psychology: Seventh Edition, New York, NY: Worth Publishers*, 500, 2004.

[230] A. M. Narayanan and A. Bertrand. Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection. *IEEE Transactions on Biomedical Engineering*, 67(1):234–244, 2019.

[231] S. Narayanan and P. G. Georgiou. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203–1233, 2013.

[232] F. Nasirian, M. Ahmadian, and O. K. D. Lee. AI-based voice assistant systems: evaluating from the interaction and trust perspectives. In *23rd Americas Conference on Information Systems, AMCIS 2017*, pages 1–10. AIS, 2017.

[233] C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. ACM, 1994.

[234] M. Neumann and N. g. Thang Vu. Cross-lingual and multilingual speech emotion recognition on english and french. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773. IEEE, 2018.

[235] S. Nooteboom. The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673, 1997.

[236] K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50, 1987.

[237] H. Ohly, M. P. White, B. W. Wheeler, A. Bethel, O. C. Ukoumunne, V. Nikolaou, and R. Garside. Attention restoration theory: A systematic review of the attention restoration potential of exposure to natural environments. *Journal of Toxicology and Environmental Health, Part B*, 19(7):305–343, 2016.

[238] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30):13354–13359, 2010.

[239] B. Omidvar-Tehrani, S. Amer-Yahia, and A. Termier. Interactive user group analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM ’15, page 403–412. ACM, 2015.

[240] A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions.* Cambridge University Press, 1990.

[241] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3):487–501, 2004.

[242] V. Pammer-Schindler, E. Harpstead, B. Xie, B. DiSalvo, A. Kharrufa, P. Slovak, A. Ogan, J. J. Williams, and M. J. Lee. Learning and education in HCI: A reflection on the SIG at CHI 2019. *Interactions*, 27(5):6–7, 2020.

[243] Y. Pan, P. Shen, and L. Shen. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108, 2012.

[244] S. Park and J. Ryu. Exploring preservice teachers' emotional experiences in an immersive virtual teaching simulation through facial expression recognition. *International Journal of Human–Computer Interaction*, 35(6):521–533, 2019.

[245] T. Park and C. Lim. Design principles for improving emotional affordances in an online learning environment. *Asia Pacific Education Review*, 20(1):53–67, 2019.

[246] A. Pascual-Leone. How clients "change emotion with emotion": A programme of research on emotional processing. *Psychotherapy Research*, 28(2):165–182, 2018.

[247] A. Pascual-Leone and L. S. Greenberg. Emotional processing in experiential therapy: Why" the only way out is through.". *Journal of Consulting and Clinical Psychology*, 75(6):875, 2007.

[248] R. Pekrun, T. Goetz, A. C. Frenzel, P. Barchfeld, and R. P. Perry. Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ). *Contemporary educational psychology*, 36(1):36–48, 2011.

[249] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, 37(4):417–435, 2009.

[250] A. Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, volume 5, pages 1–5. ACM, 2004.

[251] A. Pentland. Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine*, 24(4):108–111, 2007.

[252] A. Pentland. *Honest signals: how they shape our world.* MIT press, 2010.

[253] A. Pentland. To signal is human: Real-time data mining unmasks the power of imitation, kith and charisma in our face-to-face social networks. *American scientist*, 98(3):204–211, 2010.

[254] A. Pentland and A. Esposito. Secret signals. *Nature*, 457(2009):528–530, 2009.

[255] C. Peter and R. Beale. *Affect and emotion in human-computer interaction: From theory to applications.* Springer Science & Business Media, 2008.

[256] R. Picard. Computers that recognize and respond to user emotion. In *International Conference on User Modeling*, pages 1–26. Springer, 2003.

[257] R. W. Picard. Affective computing for HCI. In *Proceedings of HCI International on Human-Computer Interaction: Ergonomics and User Interfaces*, pages 829–833. ACM, 1999.

[258] R. W. Picard. *Affective Computing.* MIT press, 2000.

[259] R. W. Picard. Toward computers that recognize and respond to user emotion. *IBM systems journal*, 39(3.4):705–719, 2000.

[260] R. W. Picard and J. Klein. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers*, 14(2):141–169, 2002.

[261] R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[262] P. Podder, T. Z. Khan, M. H. Khan, and M. M. Rahman. Comparative performance analysis of hamming, hanning and blackman window. *International Journal of Computer Applications*, 96(18):1–7, 2014.

[263] F. B. Pokorny, B. Schuller, P. B. Marschik, R. Brueckner, P. Nyström, N. Cummins, S. Bölte, C. Einspieler, and T. Falck-Ytter. Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach. In *Interspeech*, pages 309–313. ISCA, 2017.

[264] T. Polzehl, S. Möller, and F. Metze. Automatically assessing personality from speech. In *IEEE 4th International Conference on Semantic Computing*, pages 134–140. IEEE, 2010.

[265] D. Popov, A. Gapochkin, and A. Nekrasov. An algorithm of daubechies wavelet transform in the final field when processing speech signals. *Electronics*, 7(7):120, 2018.

[266] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–12. ACM, 2018.

[267] A. Poushneh. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, 58:102283, 2021.

[268] K. M. Prabhu. *Window functions and their applications in signal processing.* Taylor & Francis, 2014.

[269] S. PS and G. Mahalakshmi. Emotion models: a review. *International Journal of Control Theory and Applications*, 10(8):651–657, 2017.

[270] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019.

[271] K. Qian, M. Schmitt, C. Janott, Z. Zhang, C. Heiser, W. Hohenhorst, M. Herzog, W. Hemmert, and B. Schuller. A bag of wavelet features for snore sound classification. *Annals of Biomedical Engineering*, 47(4):1000–1011, 2019.

[272] L. R. Rabiner and R. W. Schafer. *Introduction to digital speech processing*. Now Foundations and Trends, 2007.

[273] F. Rakotomalala, H. N. Randriatsarafara, A. R. Hajalalaina, and N. M. V. Ravonimanantsoa. Voice user interface: Literature review, challenges and future directions. *The System Theory, Control and Computing Journal*, 1(2):65–89, 2021.

[274] S. Ramakrishnan. Recognition of emotion from speech: A review. *Speech Enhancement, Modeling and recognition–algorithms and Applications*, 7:121–137, 2012.

[275] S. Ramakrishnan and I. M. El Emary. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52(3):1467–1478, 2013.

[276] A. Reddy, A. B. Kocaballi, I. Nicenboim, M. L. J. Søndergaard, M. L. Lupetti, C. Key, C. Speed, D. Lockton, E. Giaccardi, F. Grommé, H. Robbins, N. Primlani, P. Yurman, S. Sumartojo, T. Phan, V. Bedö, and Y. Strengers. Making everyday things talk: Speculative conversations into the future of voice interfaces at home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, 2021.

[277] S. Revathy, N. R., and R. K. J. Health care counselling via voicebot using multinomial naive bayes algorithm. In *5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1063–1067. IEEE, 2020.

[278] J. A. Rincon, A. Costa, P. Novais, V. Julian, and C. Carrascosa. A new emotional robot assistant that facilitates human interaction and persuasion. *Knowledge and Information Systems*, 60(1):363–383, 2019.

[279] G. Rizos, A. Baird, M. Elliott, and B. Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3502–3506. IEEE, 2020.

[280] D. Rodrigo-Ruiz. Effect of teachers' emotions on their students: some evidence. *Journal of Education & Social Policy*, 3(4):73–79, 2016.

[281] S. Rosen. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 336(1278):367–373, 1992.

[282] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[283] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[284] B. V. Sathe-Pathak and A. R. Panat. Extraction of pitch and formants and its analysis to identify 3 different emotional states of a person. *International Journal of Computer Science Issues (IJCSI)*, 9(4):296, 2012.

[285] A. Sayem. Speech analysis for alphabets in bangla language: automatic speech recognition. *International Journal of Engineering Research*, 3(2):88–93, 2014.

[286] K. R. Scherer. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000.

[287] M. Schrepp. User experience questionnaire handbook. https://www.ueq-online.org/Material/Handbook.pdf, 2015.

[288] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03).*, volume 2, pages 1–4. IEEE, 2003.

[289] B. W. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

[290] B. W. Schuller, H. Coppock, and A. Gaskell. Detecting covid-19 from breathing and coughing sounds using deep neural networks. *arXiv preprint arXiv:2012.14553*, 2020.

[291] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–43, 2021.

[292] K. Seaborn and J. Urakami. Measuring voice ux quantitatively: A rapid review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–8. ACM, 2021.

[293] S. Seo. When female (male) robot is talking to me: Effect of service robots' gender and anthropomorphism on customer satisfaction. *International Journal of Hospitality Management*, 102:103166, 2022.

[294] K. Shama, A. Krishna, and N. U. Cholayya. Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *Journal on Advances in Signal Processing (EURASIP)*, 2007:1–9, 2006.

[295] N. R. Shamsuddin and N. I. Mahat. Comparison between k-means and k-medoids for mixed variables clustering. In *Proceedings of the 3rd International Conference on Computing, Mathematics and Statistics*, pages 303–308. Springer, 2019.

[296] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.

[297] Y. Shuai, Y. Zheng, and H. Huang. Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV. In *IEEE 9th international conference on software engineering and service science (ICSESS)*, pages 449–453. IEEE, 2018.

[298] M. Sinith, E. Aswathi, T. Deepa, C. Shameema, and S. Rajan. Emotion recognition from audio signals using Support Vector Machine. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 139–144. IEEE, 2015.

[299] G. M. Slavich, S. Taylor, and R. W. Picard. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22(4):408–413, 2019.

[300] P. Slovak, A. N. Antle, N. Theofanopoulou, C. D. Roquet, J. J. Gross, and K. Isbister. Designing for emotion regulation interventions: an agenda for HCI theory and research. *arXiv preprint arXiv:2204.00118*, 2022.

[301] A. Smith and F. Yetim. *Global human–computer systems: cultural determinants of usability.* Oxford University Press, 2004.

[302] V. Soman and A. Madan. Social signaling: Predicting the outcome of job interviews from vocal tone and prosody. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–4. IEEE, 2010.

[303] M. Song, E. Parada-Cabaleiro, S. Liu, M. Milling, A. Baird, Z. Yang, and B. W. Schuller. Supervised contrastive learning for game-play frustration detection from speech. In *International Conference on Human-Computer Interaction*, pages 617–629. Springer, 2021.

[304] T. Song, W. Zheng, P. Song, and Z. Cui. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[305] R. O. Stanley and G. D. Burrows. Varieties and functions of human emotion. *Emotions at work: Theory, research and applications in management*, pages 3–19, 2001.

[306] S. S. Stevens and J. Volkmann. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, 53(3):329–353, 1940.

[307] B. H. Su, C. M. Chang, Y. S. Lin, and C. C. Lee. Improving speech emotion recognition using graph attentive bi-directional gated recurrent unit network. In *Interspeech*, pages 506–510. ISCA, 2020.

[308] H. Sun. Exploring cultural usability. In *Proceedings. IEEE International Professional Communication Conference*, pages 319–330. IEEE, 2002.

[309] Y. X. Sun, Y. Ma, K. B. Shi, J. P. Hu, Y. Y. Zhao, and Y. P. Zhang. Unsupervised speaker segmentation framework based on sparse correlation feature. In *2017 Chinese Automation Congress (CAC)*, pages 3058–3063. IEEE, 2017.

[310] M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[311] M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[312] L. Taber and S. Whittaker. Personality depends on the medium: Differences in self-perception on snapchat, facebook and offline. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13. ACM, 2018.

[313] E. S.-H. Tan. Entertainment is emotion: The functional architecture of the entertainment experience. *Media psychology*, 11(1):28–51, 2008.

[314] L. Tarantino, P. N. Garner, and A. Lazaridis. Self-attention for speech emotion recognition. In *Interspeech*, pages 2578–2582. ISCA, 2019.

[315] V. Tassiello, J. S. Tillotson, and A. S. Rome. “Alexa, order me a pizza!”: The mediating role of psychological power in the consumer–voice assistant interaction. *Psychology and Marketing*, 38(7):1069–1080, 2021.

[316] J. L. Taxer and A. C. Frenzel. Brief research report: The message behind teacher emotions. *The Journal of Experimental Education*, 88(4):595–604, 2020.

[317] L. Tian, S. Oviatt, M. Muszynski, B. C. Chamberlain, J. Healey, and A. Sano. Emotion-aware Human–Robot Interaction and Social Robots. *Applied Affective Computing*, 2022.

[318] M. Tkalcic, A. Kosir, and J. Tasic. Affective recommender systems: the role of emotions in recommender systems. In *Proceedings of the 5th ACM conference on Recommender systems (RecSys’ 11)*, pages 9–13. ACM, 2011.

[319] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila. Stress detection through speech analysis. In *14th International Conference on Signal Processing and Multimedia Applications*, pages 394–398. SPIE publications, 2018.

[320] P. Tonn, Y. Degani, S. Hershko, A. Klein, L. Seule, and N. Schulze. Development of a digital content-free speech analysis tool for the measurement of mental health and follow-up for mental disorders: Protocol for a case-control study. *JMIR Research Protocols*, 9(5):e13852, 2020.

[321] P. Totterdell and D. Holman. Emotion regulation in customer service roles: testing a model of emotional labor. *Journal of occupational health psychology*, 8(1):55, 2003.

[322] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[323] S. Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang. Emotional speech synthesis with rich and granularized control. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7254–7258. IEEE, 2020.

[324] A.-J. van Kesteren, R. op den Akker, M. Poel, and A. Nijholt. Simulation of emotions of agents in virtual environments using neural networks. *IEEE Transactions on Magnetics*, pages 137–147, 2000.

[325] R. Vergin and D. O'Shaughnessy. Pre-emphasis and speech recognition. In *Proceedings of 1995 Canadian Conference on Electrical and Computer Engineering*, volume 2, pages 1062–1065. IEEE, 1995.

[326] S. Vhaduri, T. Van Kessel, B. Ko, D. Wood, S. Wang, and T. Brunschwiler. Nocturnal cough and snore detection in noisy environments using smartphone-microphones. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7. IEEE, 2019.

[327] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[328] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1061–1070. ACM, 2008.

[329] N. Vlassis and A. Likas. A kurtosis-based dynamic approach to gaussian mixture modeling. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 29(4):393–399, 1999.

[330] N. Vlassis and A. Likas. A greedy em algorithm for gaussian mixture learning. *Neural processing letters*, 15(1):77–87, 2002.

[331] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiaion. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1123–1126. European Language Resources Association (ELRA), 2006.

[332] S. T. Völkel, D. Buschek, M. Eiband, B. R. Cowan, and H. Hussmann. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, 2021.

[333] S. T. Völkel, P. Kempf, and H. Hussmann. Personalised chats with voice assistants: The user perspective. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20. ACM, 2020.

[334] S. T. Völkel, R. Schödel, D. Buschek, C. Stachl, V. Winterhalter, M. Bühner, and H. Hussmann. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–14. ACM, 2020.

[335] H. Vögel, C. Süß, T. Hubregtsen, E. André, B. Schuller, J. Härri, J. Conradt, A. Adi, A. Zadorojniy, J. Terken, J. Beskow, A. Morrison, K. Eng, F. Eyben, S. Al Moubayed, S. Müller, N. Cummins, V. Ghaderi, R. Chadowitz, R. Troncy, B. Huet, M. Önen, and A. Ksentini. Emotion-awareness for intelligent vehicle assistants: A research agenda. In *IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, pages 11–15. IEEE, 2018.

[336] S. T. Völkel, R. Schödel, and H. Hussmann. Designing for Personality in Autonomous Vehicles: Considering Individual's Trust Attitude and Interaction Behavior. In *Workshop "Interacting with Autonomous Vehicles: Learning from other Domains" at CHI 2018.* ACM, 2018.

[337] G. Wadley, V. Kostakos, P. Koval, W. Smith, S. Webber, A. Cox, J. J. Gross, K. Höök, R. Mandryk, and P. Slovák. The future of emotion in human-computer interaction. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM, 2022.

[338] B. Wieland, K. Urban, and S. Funken. *Speech signal noise reduction with wavelets.* PhD thesis, Ulm University, 2009.

[339] P. J. Wisniewski, B. P. Knijnenburg, and H. R. Lipford. Making privacy personal: Profiling social network users to inform privacy education and nudging. *International Journal of Human-Computer Studies*, 98:95–108, 2017.

[340] R. Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.

[341] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai. End-to-end emotional speech synthesis using style tokens and semi-supervised training. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA–ASC)*, pages 623–627. IEEE, 2019.

[342] Y. Yang, H. Zhang, W. Tu, H. Ai, L. Cai, R. Hu, and F. Xiang. Kullback–Leibler divergence frequency warping scale for acoustic scene classification using convolutional neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 840–844. IEEE, 2019.

[343] S. Yeh, Y. Lin, and C. Lee. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689. IEEE, 2019.

[344] S. H. Yella, A. Stolcke, and M. Slaney. Artificial neural network features for speaker diarization. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 402–406. IEEE, 2014.

[345] S. Yoon, S. Byun, and K. Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE, 2018.

[346] K. S. Young, C. F. Sandman, and M. G. Craske. Positive and negative emotion regulation in adolescence: links to anxiety and depression. *Brain sciences*, 9(4):76, 2019.

[347] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide. Feature learning in deep neural networks-studies on speech recognition tasks. *arXiv preprint arXiv:1301.3605*, 2013.

[348] Y. Yu and Y. J. Kim. Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database. *Electronics*, 9(5):713, 2020.

[349] C. Yuan and H. Yang. Research on k-value selection method of k-means clustering algorithm. *Multidisciplinary Scientific Journal*, 2(2):226–235, 2019.

[350] S. Zepf, T. Stracke, A. Schmitt, F. van de Camp, and J. Beyerer. Towards real-time detection and mitigation of driver frustration using SVM. In *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 202–209. IEEE, 2019.

[351] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):1–25, 2018.

[352] T. Zhang and J. Wu. Speech emotion recognition with i-vector feature and RNN model. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pages 524–528. IEEE, 2015.

[353] X. Zhang, J. Bai, and W. Liang. The speech recognition system based on bark wavelet MFCC. In *8th International Conference on Signal Processing*, volume 1. IEEE, 2006.

[354] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller. Snore-GANs: Improving automatic snore sound classification with synthesized data. *IEEE Journal of Biomedical and Health Informatics*, 24(1):300–310, 2019.

[355] G. Zhao, S. Ding, and R. Gutierrez-Osuna. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In *Interspeech*, pages 2843–2847. ISCA, 2019.

[356] J. Zhao, X. Mao, and L. Chen. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, 47:312–323, 2019.

[357] N. Zheng, T. Lee, and P.-C. Ching. Integration of complementary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, 14(3):181–184, 2007.

[358] M. X. Zhou, G. Mark, J. Li, and H. Yang. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(2-3):1–36, 2019.