VoiceMessage++: Augmented Voice Recordings for Mobile Instant Messaging

GABRIEL HAAS, Ulm University, Germany
JAN GUGENHEIMER, Telecom-Paris/LTCI/IP-Paris, France
ENRICO RUKZIO, Ulm University, Germany

Media (e.g. videos, images, and text) shared on social platforms such as Facebook and WeChat are often visually enriched through digital content (e.g. emojis, stickers, animal faces) increasing joy, personalization, and expressiveness. While voice messages (VMs) are experiencing a high frequent usage, they currently lack any form of digital augmentation. This work is the first to present and explore the concept of augmented VMs. Inspired by visual augmentations we designed and implemented an editor, allowing users to enhance VMs with background sounds, voice changers, and sound stickers. In a first evaluation ($N = 15$) we found that participants used augmentations frequently (2.73 per message on average) and rated augmented VMs to be expressive, personal and more fun than ordinary VMs. In a consecutive step, we analyzed the 45 augmented VMs recorded during the study and identified three distinct message types (*decoration, composition* and *integrated*) that inform about potential usage.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Social content sharing**.

Additional Key Words and Phrases: voice messaging; messenger apps; media sharing

ACM Reference Format:
Gabriel Haas, Jan Gugenheimer, and Enrico Rukzio. 2020. VoiceMessage++: Augmented Voice Recordings for Mobile Instant Messaging. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20), October 5–8, 2020, Oldenburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3379503.3403560

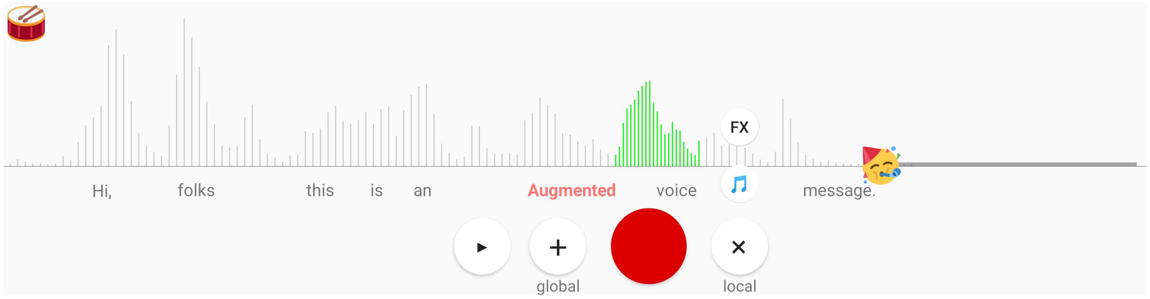

Fig. 1. An illustration of our VoiceMessage++ editor. It shows a recorded message as waveform and text on a shared timeline and the user interface. Added augmentations are background music (march, top left), voice changer ("Augmented") and a sound effect (emoji with party hat). A segment that is played back is temporarily rendered in green.

1 INTRODUCTION

The evolution of mobile network technology has caused mobile messaging to gradually move away from the short message service (SMS). Messenger applications emerged that got rid of the charge per message and the limitation of 160 characters per message. Along with better network coverage and higher transmission rates those rather simple messenger applications have today turned into social platforms, that not only offer the exchange of text messages but also support to exchange other types of media such as images, videos, and voice messages.

Voice messaging was introduced in 2011 by WeChat [36] and since then has established as a widespread means of communication and become part of mobile messaging. Today, voice messaging is available on any major messaging application. VMs combine important characteristics of text messaging and phone calls. By using the voice as a medium, users can easily transmit emotions and express themselves more naturally compared to text messages [11]. Unlike traditional voice mail systems, the mobile push-to-talk implementation embedded in the chat window, is much more immediate and creates a dialogue flow similar to text-based chatting. Due to its asynchronous implementation, it does not require simultaneous availability of two persons, but retains the freedom to respond whenever it fits the receiver.

When analyzing the most used social platforms [22], we found that editing and augmenting capabilities are available for every type of asynchronous exchanged media, except VMs. Texting is supported by the addition of emojis, sometimes also by GIFs. Images offer lots of editing options containing cropping, rotation, overlaying stickers (i.e. small images), overlaying emojis, overlaying text, overlaying drawings, and the addition of static filters. Static filters are "one-click" manipulations, best known for Instagram, which allow to alter the appearance of images within seconds, making the formerly time consuming and technical process of image editing accessible to a wide population. For videos augmentations are mostly the same as for images but rotation and static filters being not always available but often dynamic filters that interact with the content of videos. In 2015, Snapchat first introduced such dynamic filters [20] which add sunglasses or animal features to the faces of people by using computer vision algorithms. Since then, they found widespread use. Such augmentations were shown to increase engagement with the shared content [6] and offer a new layer of personalization and expressiveness [28, 35]. As noted, voice messages are currently lacking any form of augmentation and personalization leaving their full potential untapped. At the current state, they are sent in a push-to-talk manner, often without even offering the ability to further edit or review the recorded content.

This work is the first to present and explore the concept of augmented voice messages. Augmented voice messages are recorded messages that are enriched with additional sounds and altered through filters, similar to what is known from text, images and videos. To evaluate the concept of augmented VMs and explore what type of messages users create, we conducted a user study with 15 participants. The VoiceMessage++ editor we developed, allows to record voice messages, see the transcribed text and apply augmentations to it (see Figure 1). We provide two types of augmentations: a) *filters* as alterations to the recording (e.g. transform a person's voice to sound like a robot) and b) *additions* to the recording (e.g. music, sound effects or complete soundscapes). Both can either be applied and placed *locally* (covering a specified timeframe) or *globally* (covering the whole voice message). We let participants use the system to record and augment voice messages for given situations and according to their own choice. The resulting 45 recorded messages were later analyzed and coded by two authors, unveiling three distinct types of augmented voice messages (*decoration*, *composition* and *integrated*). Overall, we found that participants used on average 2.73 augmentations per message. The focused mainly on *local additions* (i.e., "sound stickers") which have been present in 98% of all recorded VMs. Participants rated augmented VMs to be more expressive and fun than ordinary VMs, and 80% reported that they would like to use such a system.

The main contributions of our work are:

(1) the first introduction of the concept of augmented voice messages as an equivalent to augmented text, images and videos
(2) the design, implementation, and preliminary evaluation of VoiceMessage++, an Android application that allows the recording and augmentation of VMs

A exploratory user study (N = 15) unveiled three types of augmented voice messages (*decoration*, *composition* and *integrated*) that inform about future use of such augmentations, increasing the already high potential of voice messages as a personal and expressive alternative to texting.

2 RELATED WORK

In the following, we give an overview of text augmentations in the form of emojis, image and video sharing as well as audio editing. Furthermore, we provide an analysis of currently available augmentations in popular social platforms.

2.1 Text Augmentation in Instant Messaging

Emojis are an integral part of today's text-based communication. They emerged from Emoticons, which have been combinations of characters that form recognizable symbols such as the famous smiling face ":-)", to graphical representations. The encoding in Unicode led to a wide adoption across the world, making them an "ubiquitous language" [25]. Recently, attempts have been made to make their selection even more diverse, better representing all cultures [23]. By enabling the expression of emotions and non-verbal content, emojis and emoticons have made an important contribution to the success of texting [19].

Tigwell et al. showed variations in people's emoji interpretation [33]. They argue that a faulty interpretation of messages, due to a different understanding of an emoji's meaning, can lead to a communication breakdown resulting in hurt relationships. Despite being not always easy to interpret they are used intensively [14].

Zhou et al. [38] used interviews and observation to examine the use of emojis and graphical stickers in mobile communication. They found that they not only help to express emotions but also self representation and describe the following three distinct strategies of emoji/sticker use. First strategy is to better express emotions and avoid misunderstandings in text, second to use emojis and stickers when text is unable to convey the desired attitude or tone, and third to include pictorial representations when something is impossible to express in text [38].

2.2 Image and Video Sharing

The creation and sharing of images and videos via social platforms is very popular among teens. In fact, within this user group, Instagram and Snapchat today are more often used than Facebook [2]. Hu et al. [18] were one of the first to analyze the users and shared content of the image-centered platform Instagram on a large scale. By coding 1000 crawled photos of 50 pre-selected users, they found eight distinct photo categories (selfies, friends, activities, captioned photos, food, gadgets, fashion, and pets).

Which type of content leads to the most social engagement and how the popular filters provided by Instagram are applied was investigated by Bakhshi et al. [5, 6]. Using algorithmic visual analysis, they found that shared images containing faces are more likely to receive likes and comments. However, the amount of faces, their age and gender had no significant influence on engagement. Subsequently, they explored the concept of image filters that change the look and feel of images such as visually ageing it, making colors more vibrant or adjusting color temperature. The identified motives of users were "to improve aesthetics, add vintage effects, highlight objects, manipulate colors, and making photos appear more fun and unique" [6].

Mc Roberts et al. investigated what content is shared via Snapchat's Story feature, who is the target audience of that content and why users chose that form of content sharing [26]. The "Story" feature, also implemented in other messengers, allows ephemeral sharing of content. The images or short video clips are available to view for 24 hours. They found that this feature is used to capture those moments of users days that seem noteworthy and are contributing to a coherent storyline. While not being explicitly targeted at an audience, the ephemerality of stories offers some protection to try out things that would be preserved forever in other forms of social media.

Voice messaging in particular received sparse scholarly attention despite its high usage (e.g. WeChat reported 6.1 billion voice messages sent per day in 2017 [32]). One of the few papers that include voice messages is the work of Di Cui [13]. He explores the role of mobile instant messaging in the management of intimate relationships using the example of WeChat's multimodal communication. The expressivity and cue-richness of VMs helped to accurately convey emotions and create a stronger bonding for the couples in his study. Combining these characteristics with the positive aspects found in image and video augmentation represents a huge opportunity for voice messages.

2.3 Audio Editing and Processing Speech

To allow the augmentation of voice messages, audio files containing the recorded speech need to be processed and edited. For desktop computers a range of such audio workstations exist, including professional tools such as Adobe Audition [1] or Avid ProTools [4] but also sophisticated free tools such as Audacity [3]. They all allow users to record and edit multi-track audio but the immense amount of features lead to an increasingly complicated user interface that is aimed at professional or ambitious home users. The swift addition of complex filters or specific sound effects at semantically relevant locations is not well supported.

Whittaker and Amento [37] developed a semantic editor that allows the editing of speech recordings using cut and paste actions on a textual representation of recordings instead of the typical acoustic editing directly on the waveform. They showed that semantic editing is more efficient than acoustic editing, even when the speech recognition result was highly inaccurate. This concept was refined and extended for video editing by Berthouzoz et al. [9] to reduce the time-consuming low-level work necessary for placing cuts and creating transitions in video-recorded interviews.

Similar attempts have been made for tools specifically aimed to produce audio stories. The work of Rubin et al. [29] presented an editor focused on the creation of such, aiming to provide higher level workflows. They addressed the problems of navigation in spoken audio, speed up music selection by providing multi-feature search, and convenient music modification to combine recorded speech and background music.

All of the editors presented are still aimed at professionals and, unlike our approach, did not combine the semantic textual representation and the acoustic waveform in a single view.

2.4 Use of Augmentations in Social Platforms

To analyze the current state-of-the-art of media augmentations in social platforms, we took each of the four media types text, image, video, and voice messages into account. We selected the six biggest social platforms in terms of monthly active users. Those are Facebook, YouTube, WhatsApp, Facebook Messenger, WeChat, and Instagram [22]. We combined Facebook and Facebook Messenger as they together provide the same functionality as other platforms (messaging and feed) and excluded YouTube, due to only providing a video feed and not having a dedicated messaging functionality. As a first step we collected all available augmentations for the four media types on each platform. We then identified two dimensions to categorize those augmentations: their scope and the type of augmentation. Augmentations with *global* scope are augmentations that affect the whole medium. That can be the addition of background music to a video or partially transparent overlays to a video such as an added framing or border (e.g. a filmstrip) or a large overlay similar to a photo stand-in or cardboard stand-up which leaves room for only part of the original content. *Local* augmentations affect only small parts of it such as the popular face filters that ages a persons face or adds features such as sunglasses. When considering the type of augmentations there are *additions* that add to the existing content (e.g., overlaying an emoji onto a video) and *alterations* that interact with the content and transform the content (e.g., a video filter that transforms a video into a black-and-white film). All currently available augmentations of Facebook/FB Messenger, WhatsApp, WeChat, and Instagram can be categorized in those two dimensions as shown in Table 1. Notably, voice recordings are absent in the table due to no available augmentations. When combining the available types of augmentation for the four media types and four selected social platforms, the lack of augmentations for voice recording becomes obvious. Table 2 gives an overview which types of augmentation are available in which social platform. While at least three, sometimes all four types of augmentation are provided for other types of media, not any type of augmentation for voice recordings is provided by the social platforms.

	scope	***addition***	***alteration***
text	*global*	strikethrough	bold, italic
	local	emoji, GIF	bold, italic
image	*global*	partially transparent overlay	crop, rotate, filter, dynamic filter
	local	sticker, text, free drawing	face filter (e.g., ageing a persons face)
video	*global*	partially transparent overlay, background music	video filter (e.g., contrast, color)
	local	sticker, text, free drawing	face filter (e.g., ageing a persons face)

Table 1. Media augmentations of social platforms classified into the dimensions of scope (global, local) and type (addition, alteration).

	FB (Messenger)	***WhatsApp***	***WeChat***	***Instagram***
text	global addition global alteration local addition local alteration	global addition global alteration local addition local alteration	local addition	local addition
image	global addition global alteration local addition	global alteration local addition	global alteration local addition	global alteration local addition local alteration
video	global addition global alteration local addition	global alteration local addition	global addition local addition	global alteration local addition local alteration
voice	none	none	none	none

Table 2. Types of augmentation as available in popular social platforms for various types of media.

Furthermore, we inspected VM augmentations outside of the small selection of popular social platforms. There are plugins for messenger applications (e.g., WeChat Voice) and dedicated voice changer apps (e.g., Zoobe) that are able to share recorded messages using voice changers (global alteration). However, within these apps, recordings can only be processed as a whole. Additionally, the audio does not stand on its own but is supported by images or videos. Video centric apps such as Snapchat and TikTok provide the ability to add background music to videos (global addition) and sometimes the use of voice changers in videos (global alteration). Beside being not the main focus, audio augmentations can only be made for a video as a whole. VM++ allows to use global voice filters and background sounds for recordings but also considers the temporal dimension of local sound effects and local voice filters. As a result, we enable more complex forms of personalizing and enriching voice messages and go well beyond existing possibilities.

3 VOICEMESSAGE++

The following section presents the concept of augmented voice messages and the implementation of our editor.

3.1 Concept

We first identified augmentations that can be performed on voice messages. To provide all of the four types of augmentations identified during our analysis, we took inspiration from augmentations available for text, images and videos and looked into techniques used in audio books (also audio drama, radio play or audio theatre). Such recordings can be categorized into three different kind of sounds: language, music, and noise [15]. Audio books use language to convey most of the information: one or multiple speakers communicate the content of the play. Those voices, as the most important part, are supported by music and noises. Musical sounds are the ones that are harmonic and regular. They can be utilized to define the boundaries of an audio presentation (e.g., intro and outro music) and are very effective in setting a mood, similarly to how music is used in movies [12]. So, global sounds such as background music are one type of augmentation (*global addition*). Noises on the other hand cover all non-language and non-musical sounds [15]. They can be further separated into sound effects and background sounds. Sound effects serve the purpose of linking known sounds to actions, e.g., a shattering glass is easily recognized by a listener even without any supporting visuals. Those sound effects are used for *local addition*. Background sounds are a composition from a variety of sound sources, effectively relocating a listener to a designated location, e.g., a meadow close to a river by incorporating the sounds of water, birds, insects and so on. They are a second option for *global addition*. Audio supports the *addition* and layering of various sounds very well. In contrast to images and videos, several audio sources playing simultaneous mix and unify naturally and to a listener appear pretty much indistinguishable from a single recorded soundscape.

Beside those *additions*, filters (*alterations*) are to be provided as well. The most simple filter would be the manipulation of playback speed by changing the sampling rate of the audio file which also leads to an altered perceived pitch, e.g., the popular "chipmunk-voice" when sped up [8]. More advanced filters make use of tempo manipulation, pitch manipulation, bandpasses and echo effects, resulting in filters that convincingly alter a persons voice to sound completely different, e.g., robotic, monster, or alien voices [8]. Such voice filters as *alterations* can be applied either globally (whole message) or locally (specified timeframe of the message such as a single word), serving different purposes and leading to distinct outcomes. Specific examples to all of the four types of VM augmentations are given in Table 3.

3.2 Implementation

Voice message implementations currently do not provide any form of manipulation. In fact, VMs often cannot even be reviewed before sending but only aborted during recording. Conventional audio editors usually compel users to manipulate recordings or music on the waveform level with multiple layers on a joint timeline. This allows for low-level editing but also forces producers to learn a complex interface. Our goal was not to develop a full-blown editor for

	addition	***alteration***
global	background sounds (e.g., waves and seagulls) background music (e.g., cheerful march)	voice filter (e.g., robot voice)
local	sound stickers (e.g., popping bottle, clinking glasses, laughing, ...)	voice filter (e.g., a single word in robot voice)

Table 3. A categorization of augmentations for recorded speech. Global additions (sounds covering the whole duration of the recording) are useful for audio based reproduction of a location or setting the mood via music. Local additions (sounds covering parts the recording) are able to illustrate actions. Global alterations change the whole message to sound different, local alterations change only parts of a recording which can be useful to emphasize a certain word.

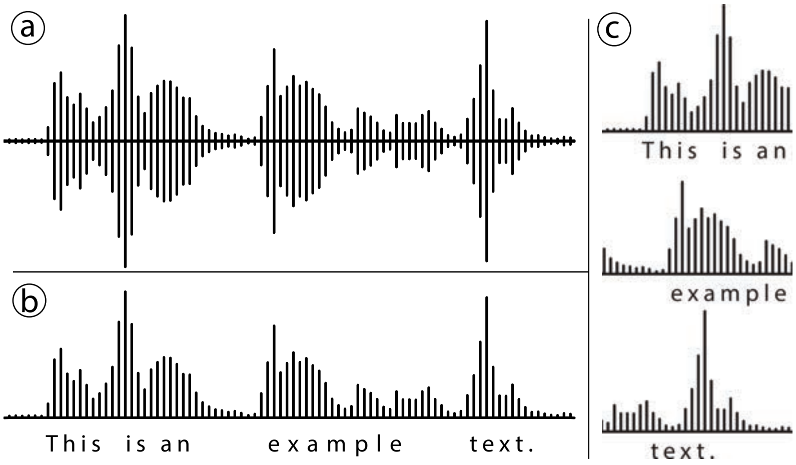

Fig. 2. a) A visualization of a recording as waveform as often used in audio editing tools containing the sentence "This is an example." b) The spoken words within the recording are displayed below the waveform. Words are temporally aligned with the soundwave. c) The same recorded soundwave and temporally aligned words but with two additional line breaks to better fit smartphone displays and allow a useful presentation of longer recordings.

fine-grained audio manipulation but a quick and easy way to create augmented voice messages on a semantic level (the actual spoken words).

We implemented the editor as a mobile application for the Android platform. Instead of using the approach of displaying the recording as a single waveform as prevalent in audio editing (2 a), we decided to combine waveform and spoken words. Therefore, we temporally aligned the spoken words and the acoustic waveform (2 b). Additionally, we break down the recording into multiple lines (see Figure 2 c), each displaying 1500 ms of the recording. This allows to make good use of the portrait orientated displays of smartphones and allows to maintain a high precision for longer recordings compared to a scaling of the timeline. The number of lines (and the resulting lineheight) is based on the length of the initial recording. We render a waveform on top of those lines, visualizing the raw sampled data and therefore the energy of the recorded signal. Google's Cloud Speech-to-Text API [16] is used to convert spoken words to text. Utilizing word timestamps from the recognition result allowed to display the corresponding text below the

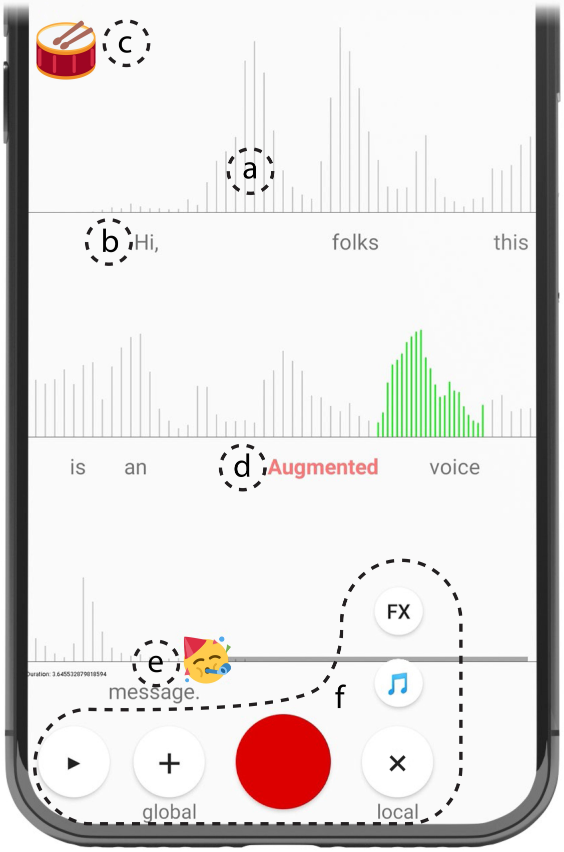

Fig. 3. The main view of our mobile application displaying the spoken message as waveform (a) and text (b) on a shared timeline. The message is broken down to several lines to be consistent with standard text representation and make use of the available display space. Selected background music (*global addition* of a festive march) is represented by the snare-drum icon (c). A *local alteration* on the word "Augmented" is visualized by setting the text in bold and red (d) and a *local addition* (emoji with party hat) (e) concludes the message. On the bottom of the screen are most of the user interface elements located (f) (left to right: play/pause, collapsed global menu, record button and expanded local menu showing buttons to add filters ("FX") or sound stickers ("♫"). The waveform of words is rendered more detailed in green when a word is played back (see "voice").

waveform, forming a link between the spoken words and the energy represented in the waveform. This combination of sound signal and contained words allows for both: easy navigation and exact positioning of elements in the message.

All of the applications user interface is combined in a single main view consisting of the visualization of the recording as waveform and text taking up most of the view, and the controls at the bottom of the view (see Figure 3). The dominant control element of the application is a circular red button used for recording, positioned at the bottom center. When starting the application, it is the only available interface element. To record a new message, this button needs to be pressed during speaking and released after finishing. When a recording is completed, the recording visualization and further control elements are added. Augmentations can be added via the global and local menu located left and right of the recording button (see Figure 3 f). By tapping the global "+" button, the two options "FX" and "BG" become available. "FX" is short for effect and due to it being *global*, it refers to the *alteration* of the whole recording. "BG" is short for

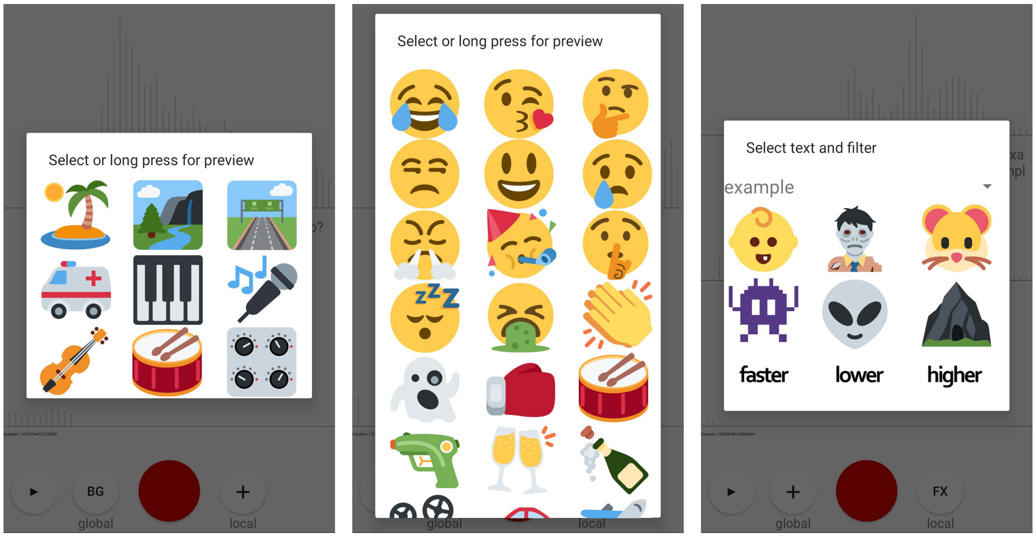

Fig. 4. The three selection dialogs for VM augmentation. Left: The selection dialog for background sounds. It can be accessed by clicking the global "+" button and then selecting "BG". Middle: The selection dialog for sound stickers, opened by clicking the local "+" button and then "♫". Individual sounds are represented by icons. Right: The selection dialog for voice filters. The local dialog has the additional word selection drop-down list at the top that isn't necessary for global voice filters.

background. In the local menu there is also "FX" for *local alteration* and the "♫" as a symbol for *local addition*. Clicking them, opens the corresponding selection dialog (see Figure 4).

As sound is not being glanceable, we provided preview functions for every available augmentation. When opening one of the four selection dialogs (Figure 4), all selectable items can be long pressed to receive a preview playback of the effect or sound in question. The provided augmentations are visually represented in the main view once they have been selected: If one of the nine available background sounds (seaside, forest, traffic, ambulance, slow piano music, engaging hip-hop beat, string orchestra, marching band, and electronic) is selected, it is visualized by an icon in the top left corner (Figure 3 c). A long press removes icon and corresponding background sound. The voice filters (baby, monster, mouse, robot, alien, echo, faster, lower, higher) can either be applied to the whole voice recording (selecting global "+" and "FX") or to individual words (selecting local "+" and "FX"). Global filters are indicated by adding a slightly thicker, red colored bar to the timeline (i.e., the bottom line of the acoustic waveform) of the recording. Local filters are indicated by coloring the corresponding word in red and setting its typeface to bold (Figure 3 d). Each can be removed by a long press on the corresponding interface element.

The available local sound effects (i.e., sound stickers) can be added by clicking the local "+" button and then the "♫" icon. Sound stickers are categorized into *emojis* (representing human-made sounds such as laughing, crying, and snoring), *sounds* (representing object based sounds such as a popping cork, clinking glasses), *quotes* (excerpts from movies or tv), and *songs* (choruses from popular songs). The sound stickers (twelve sounds in each of the four categories *emojis*, *sounds*, *quotes*, and *songs* resulting in a total of 48 sounds) are selected via a dialog, showing a long scrollable grid including all items (4 middle). They are presented as icons that can be positioned on the timeline of the recording (Figure 3 e). The actual icon serves as the beginning position of the sound effect and its duration is visualized via a transparent grey bar. After selection, sound stickers can be tapped to play the corresponding sound effect in isolation or long pressed to remove.

The voice filters are all implemented by processing the recorded data with the Tarsos DSP library [30] which allows real time audio processing in Java and includes digital signal processing algorithms such as WSOLA [34] for tempo manipulation. Background sounds and sound stickers have been collected via freesound.org[1] and YouTube's audio library[2]. Additional sounds from the categories quotes and music were included but will not be published in any form for copyright reasons. The application and individual sounds are available as supplementary material.

4 USER STUDY

To build an understanding of how people are going to use such voice message augmentations, we conducted an exploratory study. We used quantitative analysis on the recordings and logged the use of augmentations. Subsequently we collected qualitative insights regarding strategies and goals of applied augmentations. The study took place inside a separate room of our institution and lasted in between 20 and 30 minutes. Participants received 5 Euro for compensation.

4.1 Participants

We recruited 15 participants (40% female and 60% male), mostly on campus, including students and university staff. The average age was 27.93 years (SD = 5.09). Furthermore, 40% of our participants reported to regularly send VMs (M = 29.33 messages per week, SD = 35.5) and 46.7% reported to regularly receive VMs (M = 23.43 messages per week, SD = 27.64). All of those reported only friends and family members as communication partners for voice messaging.

4.2 Method

Participants received a short introduction to the topic of the study and provided the informed consent. The VoiceMessage++ editor was then explained in detail by doing a walk-through of all available features. To make sure that the concepts and possible manipulations were understood, we used an exemplary scenario. Participants were asked to record a message from an imaginary holiday on the beach and add specific manipulations such as suitable background sounds, a boat horn, and an emphasized word via a local voice filter. Thereafter, three tasks were given to the participants in which they had to imagine a specific situation and send a suitable voice message using our system:

- A private appointment in the evening can not be attended. Record a message to cancel and apologize.
- Today is the birthday of a friend of yours. Send a message to congratulate them.
- Think of a message you would like to send to any person via this application. Start by stating receiver and topic of the message.

For the third task, participants were encouraged to think aloud and tell the instructor if any feature, particular effect or sound that they would like to use was missing. After finishing those tasks the system usability scale (SUS) questionnaire [21] was presented and filled by the participants. Additionally, we asked participants to state how much they agree with the following statements on a scale from 1 (=strongly disagree) to 7 (=strongly agree): *"I would like to use the functions provided by this editor"*, *"I consider voice messages/augmented voice messages to be expressive"*, *"I consider voice messages/augmented voice messages to be personal"*, *"I have fun creating voice messages/augmented voice messages"*, *"I have*

[1] https://freesound.org/
[2] https://www.youtube.com/audiolibrary/

fun sending voice messages/augmented voice messages", *"I have fun receiving voice messages/augmented voice messages"*, *"I find it complex to create voice messages/augmented voice messages"*.

4.3 Results

To explore how and why people used VM augmentation features we used both quantitative and qualitative analysis.

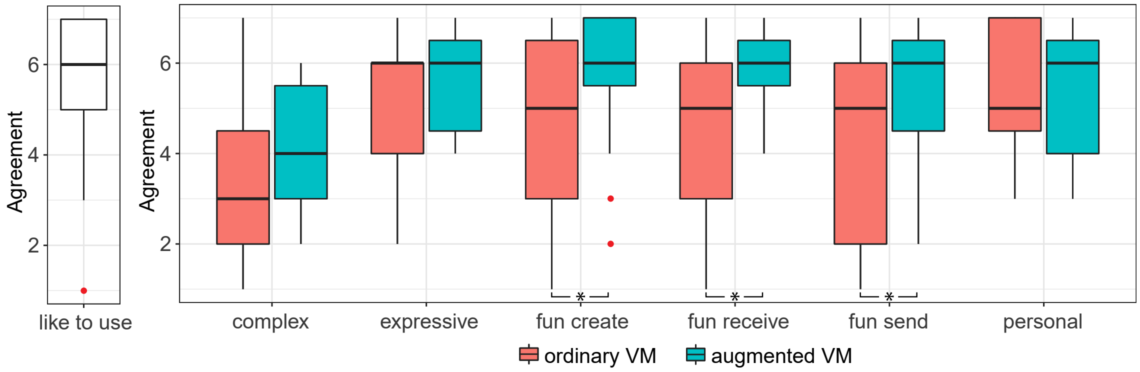

Fig. 5. The resulting boxplots of our single item questions (reported agreement on likert scales ranging from 1 = "Strongly disagree" to 7 = "Strongly agree") as explained in Section *Method*. Pairs marked with '*' were found to be significantly different.

4.3.1 *Quantitative analysis.*

We collected 45 messages that were recorded during the three study tasks. Average message duration was 8.95 seconds (*Mdn* = 7.99 s, *IQR* = 4.41 s) and contained 18.4 words on average (*Mdn* = 17, *IQR* = 9). Participants recorded and edited each message in just under three minutes on average (*Mdn* = 179.10 s, *IQR* = 141.06 s). 46.67% of messages included a global sound, 22.22% included a global voice filter, 24.44% included a local voice filter and 97.79% included local sounds (sound stickers). On average, local voice filters occurred 0.60 times per message and local sounds 2.24 times. In general it can be observed that *addition* was used more frequently than *alteration*. In total, a message contained 2.73 augmentations on average.

To quantify the use of sound stickers as the most used feature, we analyzed the usage frequency of local sounds in the categories *emojis*, *sounds*, *quotes*, and *songs*. Most popular was the emoji category (*n* = 34) followed by sounds (*n* = 32), songs (*n* = 19), and quotes (*n* = 17). We labeled the position of sound stickers within the message for the three cases 'before speech', 'during speech', and 'after speech'. 22.11% of message started with a sound sticker, 71.11% of messages included stickers during the recorded speech, and 84.44% concluded the spoken message with a sound sticker. This is in line with findings regarding emojis in text message, where more than 80% of messages were found to include emojis at the end of a message [31].

After participants used the provided editor, usability was evaluated via the 10-item SUS, with 5 items being framed negatively and 5 being framed positively [10]. A mean score of 81.83 (*Mdn* = 85, *IQR* = 11.25) puts our application in the excellent range [7].

When asked if participants would like to use the features of the provided editor, participants predominantly agreed.To further investigate how augmented VMs differ from ordinary VMs, we asked 6 pairs of questions for augmented VMs and ordinary VMs respectively (12 questions in total). We compared results from each of the 6 paired questions using Wilcoxon's signed rank test. Values below 5% ($p < .05$) are referred to as significant in the following. Participants reported to find ordinary VMs similarly expressive as augmented VMs and reported only slightly higher agreement for augmented VMs being personal. When asked about the complexity of creation, participants found it moderately complex to create both, ordinary VMs and augmented VMs. Participants reported significantly (z = -2.38, p = .017, n = 15, r = 0.62) higher agreement on having fun during the creation of augmented messages (Mdn = 6, IQR = 1.5) when compared to ordinary VMs (Mdn = 5, IQR = 3.5). They also reported they'd have more fun during sending (Mdn = 6, IQR = 2) and receiving (Mdn = 6, IQR = 1) of augmented voice messages when compared to ordinary VMs (sending: Mdn = 5, IQR = 4; receiving: Mdn = 5, IQR = 3). The differences for sending (z = -2.17, p = .030, n = 15, r = .56) and receiving (z = -2.55, p = .011, n = 15, r = .66) were found to be significant. Results are also shown in Figure 5.

4.3.2 Qualitative analysis. To gain further insights, two of the authors carried out a qualitative analysis of the recordings that were created during the user study. We conducted a content analysis as outlined by Lazar et al. [24] and adapted it for augmented voice messages. Two authors first listened to a small part of the recordings individually to get an impression of the created messages and identified codes that describe the intention of individual augmentations (e.g. sets mood, supports spoken message, adds emphasis, adds effect, essential part of message, ...). We then iterated over all 45 augmented VMs in a joint session lasting for approximately 2.5 hours to uncover interesting patterns. Codes were applied for each message augmentation separately (global sounds, local sounds, global voice filter, and local voice filter). Conflicts were resolved via discussion. During this process we discovered three distinct types of messages and again iterated over all messages to label the type of message (*decorations* = 25, *compositions* = 18, and fully *integrated messages* = 2).

Decoration was the most common type of message augmentation. Messages of this type are mostly ordinary voice messages where all of the information is within the spoken words and augmentations are only sparsely used. Sound stickers either support the general theme of the message or a suitable sound was added at the end of the message. To give an example, in a birthday message a popping cork and clinking glasses can be added or a kissing sound can be added at the end of a message which is directed to a romantic partner. The augmentations used within these messages are similar to the typical use of emojis in text messages, e.g. where a smiling emoji concludes the message [31]. Decorated VMs could also be sent without their augmentations but would loose some of their appeal (cf. video figure 1:21).

Composition is the type of augmentation were the informative content is also dominantly within the spoken words but augmentations are so manifold that they completely change the feel of the message. Those messages typically include background sounds, make use of voice filters and add multiple local sounds. For example encouraging, spoken words are underscored with engaging music, the most important words are highlighted by local voice filters. Local sounds are used to make it even more exciting. All of the augmentations are designed to complement and support the content of the message. For this type, when listening to the augmented VM and the bare recording it originated from, they do not appear to be the same (cf. video figure 1:33).

Integrated messages are the type of messages where information is incomplete without their augmentations. They use additional sounds as an integral part of the message such as quotes from movies or the chorus of a song to complement the users spoken words. For example the sender starts with "I want to tell you something:" and the message is finished off by the refrain of a happy birthday song. This type of message requires good knowledge of the available sound stickers as the recording must be planned accordingly. Listening to the original recording without the added augmentations, one cannot make fully sense of it (cf. video figure 1:52).

Beside those three types of messages, we also coded how individual augmentations were used. Background sounds were usually utilized to set and support the general mood of a message. For sending birthday wishes many participants resorted to the energetic festivity of a march while in the "cancel an appointment"-task, a sad piano song was repeatedly selected. Few also used background sounds to fake a soundscape that wasn't their own. A participant used traffic noises as a background and added additional sound stickers such as a starting car and a plane that is flying over to find a credible excuse why the appointment (first task of the study) has to be cancelled.

The third task should encourage creativity. Therefore, it did not predefine the situation and receiver of the message. Participants were instructed to think aloud during this task and especially mention missing features, filters or sound stickers. All stated to send the message to a family member, friend or the partner; only one participant chose a group chat as the receiving end. A popular theme were messages between friends and family without a very specific objective but telling about something they experienced and try to engage in a conversation (n = 9). Those messages are likely helpful to stay in touch rather than just exchanging information. Other topics were messages to make appointments (n = 4), discuss organizational matters (n = 2) and telling a joke (n = 1).

Features that have been mentioned as missed by our participants were the need for volume controls of individual augmentations ($n = 4$), the possibility to add breaks or cut the recording ($n = 4$), the possibility to select multiple words or sentences for local voice filters ($n = 3$), the ability to edit the text directly, e.g. delete words ($n = 1$), and manual controls (tempo, pitch) for voice filters ($n = 1$). Content related remarks where to generally increase the amount of available sounds ($n = 3$) or the addition of specific sounds such as animal sounds. All mentioned features can be integrated into a commercial application.

5 DISCUSSION

By simplifying the procedure of audio editing we were able to provide users a powerful tool to alter their recordings on a mobile device, allowing them to enhance voice messages by adding few augmentations (*decoration*), create hybrid messages consisting of own speech and others words (*integrated*), or even create small works of art in the style of audio books (*composition*).

When only considering the reported editing time of individual voice messages, it may seem to be a long time for the added augmentations. We are confident that the longer time is not attributed to the complexity of the interface - as shown by the excellent SUS rating - but the playful, creative process of audio editing and the exploration of the included sound library. Most of the time was spent browsing and listening to various sound effects while crafting a personal message. We expect that editing duration would significantly drop when such a system is used frequently, users are familiar with the library and be clear about what they want to achieve.

The results of our user study indicate that the concept of voice message augmentations was well received and participants rated augmented VMs to be expressive, personal, and generally more fun compared to ordinary voice messages. This is reflected in the excellent usability rating, the high usage of augmentations and most importantly in the single item questions. Over 80% of the participants reported that they would like to use our system while less than half of them reported to regularly use VMs. Some of this increase is possibly attributed to a novelty effect, but there is much to suggest that the creative process of augmenting voice messages is a highly playful interaction similar to emoji usage and image augmentation that complements current mobile communication. A frequent use of instant messaging is to "keep in touch with friends and family" [27] contrary to the technical interpretation of the term communication, the exchange of information. The augmentation of shared media is often not about transmitting more information but about a digital equivalent of brief, social interactions. These social interactions, like a pat on the shoulder or a hug are an important part of being human and should therefore also be reflected in the tools we use for digital communication.

5.1 Perception of Own Voice

Regarding the implementation of our application, participants were divided in their opinion about voice filters. Some of the participants told us that they do not like their voice altered because the other person should recognize them and know who the message is from. They also mentioned lower comprehensibility as reasons for not liking voice filters. On the other hand, we received user comments explaining that they do not like the sound of their own voice, which is a common finding [17]. Our editor allowed those users to change their voice to be more comfortable with sharing recordings of them. Voices can be changed just a little, to sound more like the person would expect themself to sound or one could change its voice completely, to have to reveal less of one self. In contrast to texting and sharing images (except selfies), it is currently a lot more challenging to stay anonymous when sending a voice recording. Voice filters could potentially offer privacy for VMs. Some aspects of the provided concept would also work completely without recording a voice. In our study a participant recorded silence and added sound stickers to it, creating a audio message without having to speak. Such "audio reactions" can be a nice addition even without the necessity to record speech.

5.2 Limitations

The study presented has some limitations. First, being a lab study it has a lower external validity compared to a field study. Sitting inside a closed room together with the study examiner while recording private messages does not reflect the usual usage behaviour. Second, the selection of voice filters and sounds provided within the editor was limited. In a real-world application those may be more diverse and offer a broader choice but for a prototypical implementation, providing such a selection was impractical. Therefore, we narrowed them down to fit our tasks and still provide a decent selection.

5.3 Future Work

To obtain a holistic evaluation of augmented voice messages and the communicative behaviour of this concept, a consecutive field study is necessary to show that augmented voice messages are a worthwhile expansion to current usage behaviour on social platforms. To tap the full potential of VM augmentation, a sound effect database similar to the popular image database giphy should be established.

6 CONCLUSION

In this work, we introduced and explored the concept of augmented VMs and presented an editor (VoiceMessage++) that allows the augmentation of VMs. We conducted a exploratory study to gain an initial understanding of voice message augmentations. Participants used lots of augmentations and utilized the provided features to create three distinct types of voice message augmentations: *decoration*s, *composition*s and *integrated* messages. Users particularly enjoyed augmenting messages, were able to tell engaging stories and reported to like to use features as provided by our VoiceMessage++ editor.

Augmentation in voice messaging can be a next step in social interaction, further increase the potential of voice messaging and open up new opportunities. The option to add augmentations adds minimal overhead of a single button click when no manipulation is desired but provides lots of possibilities for those occasions where a user wants to create something more special than an ordinary voice message.

ACKNOWLEDGMENTS

We want to thank our participants and the community of freesound.org for their contribution of countless sounds and thus their support in this project. Emojis used are Twemojis[3] by Twitter, Inc. licensed under CC BY 4.0.

REFERENCES

[1] Adobe. 2019. Adobe Audition | Audio recording, editing, and mixing software. https://www.adobe.com/products/audition.html

[2] Anderson, Monica and Jiang, Jingjing. 2018. Teens, Social Media & Technology 2018. https://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[3] Audacity. 2019. Audacity | Free, open source, cross-platform audio software for multi-track recording and editing. https://www.audacityteam.org/

[4] Avid. 2019. Pro Tools - Musiksoftware - Avid. https://www.avid.com/pro-tools

[5] Saeideh Bakhshi, David A. Shamma, and Eric Gilbert. 2014. Faces Engage Us: Photos with Faces Attract More Likes and Comments on Instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’14)*. ACM, New York, NY, USA, 965–974. https://doi.org/10.1145/2556288.2557403

[6] Saeideh Bakhshi, David A. Shamma, Lyndon Kennedy, and Eric Gilbert. 2015. Why We Filter Our Photos and How It Impacts Engagement. In *ICWSM*. AAAI, Palo Alto, 10.

[7] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies* 4, 3 (May 2009), 114–123.

[8] Patrick Bastien. 2003. Voice Specific Signal Processing Tools. In *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Audio Engineering Society, NY, 21. http://www.aes.org/e-lib/browse.cfm?elib=12316

[9] Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2012. Tools for Placing Cuts and Transitions in Interview Video. *ACM Trans. Graph.* 31, 4 (July 2012), 67:1–67:8. https://doi.org/10.1145/2185520.2185563

[10] John Brooke. 1996. *SUS - A quick and dirty usability scale.* Redhatch Consulting Ltd., London.

[11] Barbara L. Chalfonte, Robert S. Fish, and Robert E. Kraut. 1991. Expressive Richness: A Comparison of Speech and Text As Media for Revision. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’91)*. ACM, New York, NY, USA, 21–26. https://doi.org/10.1145/108844.108848

[12] Annabel J. Cohen. 2001. Music as a source of emotion in film. In *Music and emotion: Theory and research*. Oxford University Press, New York, NY, US, 249–272.

[13] Di Cui. 2016. Beyond “connected presence”: Multimedia mobile instant messaging in close relationship management. *Mobile Media & Communication* 4, 1 (Jan. 2016), 19–36. https://doi.org/10.1177/2050157915583925

[14] Instagram Engineering. 2016. Emojineering Part 1: Machine Learning for Emoji Trends. https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad

[15] Gary Ferrington. 1994. Audio Design: Creating Multi-Sensory Images For the Mind. *Journal of Visual Literacy* 14, 1 (Jan. 1994), 61–67. https://doi.org/10.1080/23796529.1994.11674490

[16] Google. 2019. Cloud Speech-to-Text – Spracherkennung | Cloud Speech-to-Text. https://cloud.google.com/speech-to-text/?hl=en

[17] Philip Holzman and Clyde Rousey. 1966. The voice as a percept. *Journal of Personality and Social Psychology* 4, 1 (July 1966), 79–86.

[18] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *Eighth International AAAI Conference on Weblogs and Social Media*. AAAI, Palo Alto, 4. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8118

[19] Albert H. Huang, David C. Yen, and Xiaoni Zhang. 2008. Exploring the potential effects of emoticons. *Information & Management* 45, 7 (Nov. 2008), 466–473. https://doi.org/10.1016/j.im.2008.07.001

[20] Snap Inc. 2015. A Whole New Way to See Yourself(ie). https://www.snap.com/en-US/news/post/a-whole-new-way-to-see-yourselfie

[21] Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. 1996. *Usability Evaluation In Industry*. CRC Press, Boca Raton, FL.

[22] Simon Kemp. 2019. Digital 2019: Global Digital Overview. https://datareportal.com/reports/digital-2019-global-digital-overview

[23] Philippe Kimura-Thollander and Neha Kumar. 2019. Examining the "Global" Language of Emojis: Designing for Cultural Representation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*. ACM Press, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300725

[24] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann, Amsterdam.

[25] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the Ubiquitous Language: An Empirical Analysis of Emoji Usage of Smartphone Users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp ’16)*. ACM, New York, NY, USA, 770–780. https://doi.org/10.1145/2971648.2971724

[3] https://twemoji.twitter.com/

[26] Sarah McRoberts, Haiwei Ma, Andrew Hall, and Svetlana Yarosh. 2017. Share First, Save Later: Performance of Self Through Snapchat Stories. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6902–6911. https://doi.org/10.1145/3025453.3025771

[27] Bonnie A. Nardi, Steve Whittaker, and Erin Bradner. 2000. Interaction and Outeraction: Instant Messaging in Action. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 79–88. https://doi.org/10.1145/358916.358975

[28] Anne Oeldorf-Hirsch and S. Shyam Sundar. 2016. Social and Technological Motivations for Online Photo Sharing. *Journal of Broadcasting and Electronic Media* 60, 4 (Oct. 2016), 624–642. https://doi.org/10.1080/08838151.2016.1234478

[29] Steve Rubin, Floraine Berthouzoz, Gautham J. Mysore, Wilmot Li, and Maneesh Agrawala. 2013. Content-based Tools for Editing Audio Stories. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13)*. ACM, New York, NY, USA, 113–122. https://doi.org/10.1145/2501988.2501993

[30] Joren Six, Olmo Cornelis, and Marc Leman. 2014. TarsosDSP, a real-time audio processing framework in Java. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, Audio Engineering Society, New York, NY, 7.

[31] Channary Tauch and Eiman Kanjo. 2016. The roles of emojis in mobile phone notifications. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*. ACM Press, Heidelberg, Germany, 1560–1565. https://doi.org/10.1145/2968219.2968549

[32] IBG Tencent. 2017. WeChat Data Report. http://blog.wechat.com/2017/11/09/the-2017-wechat-data-report/

[33] Garreth W. Tigwell and David R. Flatla. 2016. Oh That's What You Meant!: Reducing Emoji Misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '16)*. ACM, New York, NY, USA, 859–866. https://doi.org/10.1145/2957265.2961844

[34] W. Verhelst and M. Roelands. 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. IEEE, Piscataway, NJ, 554–557 vol.2. https://doi.org/10.1109/ICASSP.1993.319366

[35] Amy Voida and Elizabeth D. Mynatt. 2005. Six Themes of the Communicative Appropriation of Photographic Images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 171–180. https://doi.org/10.1145/1054972.1054997

[36] WeChat. 2011. WeChat 2.0 for iPhone. https://www.wechatapp.com/cgi-bin/readtemplate?lang=zh_CN&t=page/faq/ios/ios_20

[37] Steve Whittaker and Brian Amento. 2004. Semantic Speech Editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 527–534. https://doi.org/10.1145/985692.985759

[38] Rui Zhou, Jasmine Hentschel, and Neha Kumar. 2017. Goodbye Text, Hello Emoji: Mobile Communication on WeChat in China. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 748–759. https://doi.org/10.1145/3025453.3025800