Technical Disclosure Commons

Defensive Publications Series

July 2021

On-Device and System-Wide Audio Live Captioning with Language Translation

Brian Kemler

Danielle Cohen

Nadav Bar

Yoni Tsafir

Benny Schlesinger

See next page for additional authors

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

Kemler, Brian; Cohen, Danielle; Bar, Nadav; Tsafir, Yoni; Schlesinger, Benny; Belozovsky, Anna; Ramanovich, Michelle Tadmor; Xia, Bingying; Tickner, Simon; Barbello, Brandon Charles; Upstill, Trystan; Fitoussi, Hen; Laurent, Eric; and Wilson, Jonathan D., "On-Device and System-Wide Audio Live Captioning with Language Translation", Technical Disclosure Commons, (July 19, 2021)
https://www.tdcommons.org/dpubs_series/4458

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Inventor(s)
Brian Kemler, Danielle Cohen, Nadav Bar, Yoni Tsafir, Benny Schlesinger, Anna Belozovsky, Michelle Tadmor Ramanovich, Bingying Xia, Simon Tickner, Brandon Charles Barbello, Trystan Upstill, Hen Fitoussi, Eric Laurent, and Jonathan D. Wilson

This article is available at Technical Disclosure Commons: https://www.tdcommons.org/dpubs_series/4458

On-Device and System-Wide Audio Live Captioning with Language Translation

Abstract:

This publication describes techniques and apparatuses that enable an electronic device (e.g., a smartphone) to provide on-device (e.g., offline), system-level (e.g., operating system), live captioning with language translation in a language that a user can choose (select). Therefore, the smartphone enables the user to read live captioning in the language of their choice without relying on an internet connection, cellular data, or any wired and/or wireless communication with a remote server. Also, the smartphone enables the user to read live captioning in the language of their choice on any medium with audio content supported by the smartphone.

Keywords:

Automatic speech recognition, ASR, multilingual, language translation, live caption, recurrent neural network, RNN, RNN transducer, RNN-T, pulse-code modulation, PCM, machine learning, ML, artificial intelligence, AI, machine translation

Background:

Electronic devices (e.g., smartphones) may display media with audio content (e.g., news, movies) with closed captioning, often provided in a single language. To include closed captioning for pre-recorded media with audio content, the content provider (e.g., a news channel) may manually generate captions (e.g., text and/or metadata) and embed the captions into the media. The content provider, however, may not be able to provide closed captioning in live (real-time) feeds, for example, when conducting a live interview, covering a breaking and/or a developing

story, and so forth. Due to the financial cost associated with the generation of the captions, individuals, small organizations, and content providers with limited financial resources may not be able to provide captions for their audio content. As a result, media with audio content frequently do not include closed captioning in any language. Consequently, hearing-impaired people cannot access the information included in the audio content. Furthermore, people with good hearing also appreciate closed captioning of media with audio content because they may want to consume information by reading the closed captioning instead of, or in addition to, listening to the audio content.

In addition, even though our world has become increasingly interconnected, people often fail to understand media with audio content provided in a language they do not understand. For example, assume Jane, a native English speaker, enjoys songs of the Belgian musician *Stromae*. Jane, however, speaks little French. As such, Jane fails to understand what *Stromae* may be singing, saying in a social media post, stating in an interview, and so forth.

Jane may choose to use a translation service to understand a French speaker using closed captioning in English. Existing translation services, however, often use application software that require an adequate audio input signal, and which may not operate for low or turned-off volume levels of the smartphone. In addition, the translation services may not work when a user is offline.

Description:

This publication describes techniques and apparatuses that enable an electronic device (e.g., a smartphone) to provide on-device (e.g., offline), system-level (e.g., operating system(OS)), live captioning with language translation in a language that a user can choose (select). Therefore, the smartphone enables the user to read live captioning in the language of their choice without

relying on an internet connection, cellular data, or any wired and/or wireless communication with a remote server. Also, the smartphone enables the user to read live captioning in the language of their choice on any medium with audio content supported by the smartphone. Figure 1 illustrates an example technique of an OS of the smartphone supporting live captioning with language translation.

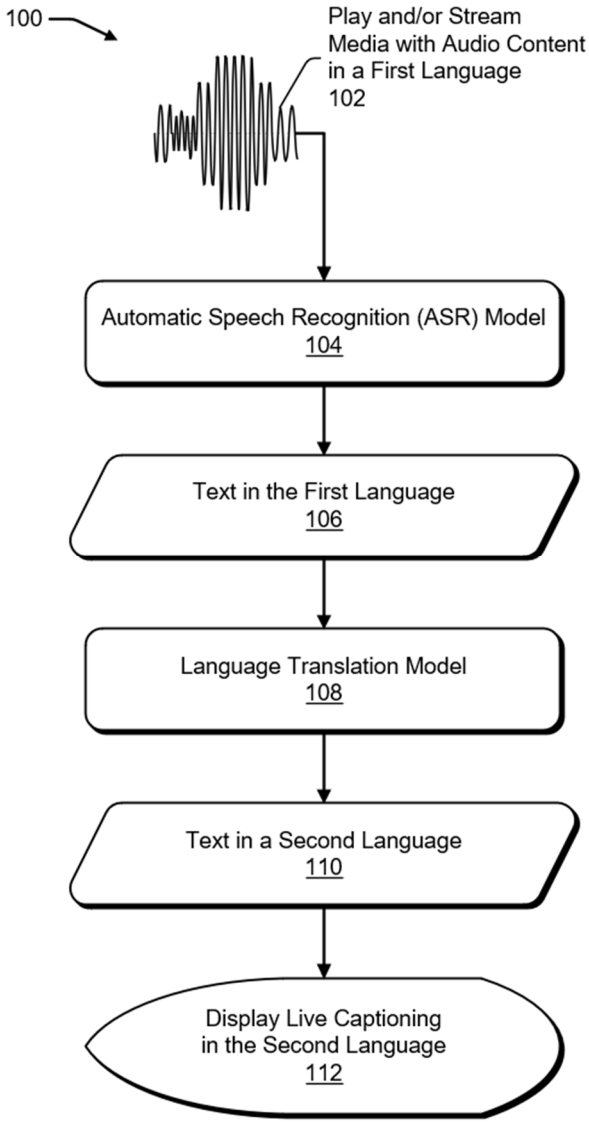

Figure 1

Figure 1 depicts a block diagram illustrating a technique 100 implemented on a smartphone with an OS that can automatically generate and display on a display screen of the smartphone live

captioning with language translation without relying on a content provider to support closed captioning.

At stage 102, the smartphone plays and/or streams media with audio content in a first language. First-party and/or third-party applications can create, generate, transmit, and/or support the media with audio content in the first language. In aspects, the media with audio content in the first language can be audio-only or audio and visual, including telephony, podcasts, radio, television news, non-silent movies, video conferencing, social media that include speech, songs, operas, readings, virtual assistants, navigation applications, and any media with speech and/or lyrics. The OS of the smartphone may enable the first-party and third-party applications to utilize at least one audio application programming interface (API) to access a raw audio byte stream(s) of a system audio output of the smartphone without disrupting playback.

Nevertheless, respective developers of the first-party and third-party applications may control whether the OS or any applications supported by the OS can access the raw audio byte stream of the system audio output of the smartphone and/or the audio content provided and/or supported by their respective first-party and third-party applications. Further, the user may also control which of the first-party and third-party applications can access the raw audio byte stream of the system audio output of the smartphone. In one aspect, the user may choose to forgo the generating and displaying on the display screen of the smartphone live captioning with language translation while they use, for example, the telephony. Suppose the developers of the first-party and third-party applications and the user give the OS permission to do so. In that case, the OS may generate the raw audio byte stream by utilizing a pulse-code modulation (PCM) of a sampled analog signal to digitally represent the media with audio content in the first language. Accessing the raw audio byte stream of the system audio output enables the OS to generate live captioning

with language translation regardless of an output volume level of the smartphone. As such, the user can read live captioning with language translation while playing and/or streaming the media with audio content in the first language in a high-level volume, medium-level volume, low-level volume, or muting the smartphone, as is further described below.

Assume the user utilizes the smartphone to consume the media with the audio content in the first language (e.g., French). Regardless of the volume of the smartphone, at stage 104, the OS feeds (transfers, pipelines) the raw audio byte stream into an automatic speech recognition (ASR) model. The ASR model differentiates speech and/or lyrics from other audible (e.g., music) and/or visual inputs (e.g., video) of the media with audio content in the first language. The ASR model may utilize at least one machine-learned (ML) model. The ML model(s) may be a recurrent neural network (RNN) transducer (RNN-T) model for speech recognition, a text-based RNN model for unspoken punctuation, a convolution neural network (CNN) model for sound events classification (e.g., music, background noise, sound effects), or a combination thereof. Subsequently, at stage 106, the ASR model outputs and/or generates a text in the first language. The text of the first language may also include punctuation, symbols (e.g., €, $, %, @), tags (e.g., applause, music), pictograms (e.g., Emojis), and so forth.

In one aspect, using the ASR model, the OS processes the raw audio byte stream and converts, outputs, and/or generates speech and/or lyrics into text with an associated confidence level. The confidence level may vary, in part, due to a type of speech (e.g., commonly used words, official language, dialect, accent, jargon, slang), a content provider (e.g., low or high audio quality), speech pattern (e.g., mumbling, disfluency, fast talking), and so forth. A developer of the OS may initially train the ASR model on a server before deploying the ASR model to the

smartphone. The initial model training increases a quality of the ASR model while still protecting a user's privacy because the user has yet to use their smartphone.

In addition, the smartphone includes necessary resources (e.g., processors, memory) to further train the ASR model as the user utilizes their smartphone and may do so offline to protect the user's privacy. For example, assume Jane instead uses the smartphone of Figure 1 to communicate with a French-speaking friend, Gabriel, using a first-party or a third-party video conferencing application software. Gabriel may use a distinct French vocabulary, speak in a particular French dialect, have a specific speech pattern, and so forth. Jane's smartphone (e.g., on-device) can further train the ASR model to increase the confidence level in converting Gabriel's spoken French words into text. Inputs to the ASR model are Gabriel's audible spoken French words, whereas outputs of the ASR model are Gabriel's spoken French words converted into text. Consequently, the confidence level in converting Gabriel's spoken French words into text may increase after several communication sessions between Jane and Gabriel. Similarly, if Gabriel also uses the smartphone of Figure 1 during these communications with Jane, Gabriel's smartphone can also train the ASR model to increase a confidence level in converting Jane's spoken English words into text. In the latter case, inputs to the ASR model are Jane's audible spoken English words, whereas outputs of the ASR model are Jane's spoken English words converted into text.

After converting the audio content in the first language to text in the first language, at stage 108, the OS uses a language translation model to translate the text in the first language (e.g., French) to text in a second language (e.g., English). The language translation model may be a statistical machine translation (SMT), a rule-based machine translation (RBMT), a hybrid machine translation (HMT), a neural machine translation (NMT), or a combination thereof.

Subsequently, at stage 110, the language translation model outputs and/or generates a text in the second language (e.g., English). The text in the second language may also include punctuation, symbols, tags, pictograms, and so forth that are suitable for the second language. For example, depending on the first (e.g., French) and the second (e.g., English) language, the language translation model may also convert quotation marks, wherein <<*bonjour*>> may be translated to "hello" or "good morning." As another example, some languages (e.g., Spanish) may place an inverted question (¿) mark at a beginning of a sentence in addition to an ordinary question (?) mark at an end of the sentence.

Finally, at stage 112, the smartphone displays live captioning in the second language. Depending on the second language, the smartphone may display text, characters, punctuation, symbols, tags, pictograms, and so forth, left-to-right (e.g., English, French), right-to-left (e.g., Arabic, Hebrew), or top-to-bottom (e.g., Mandarin Chinese, Japanese). Figures 2A and 2B illustrate the user triggering live captioning with language translation.

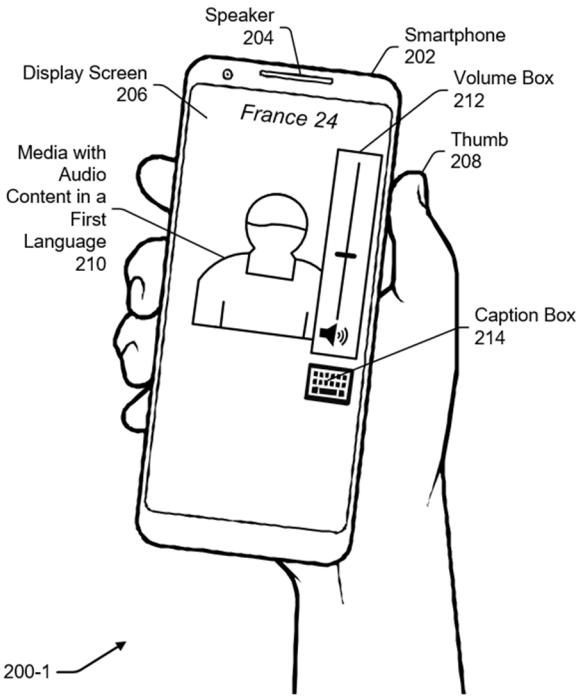

Figure 2A

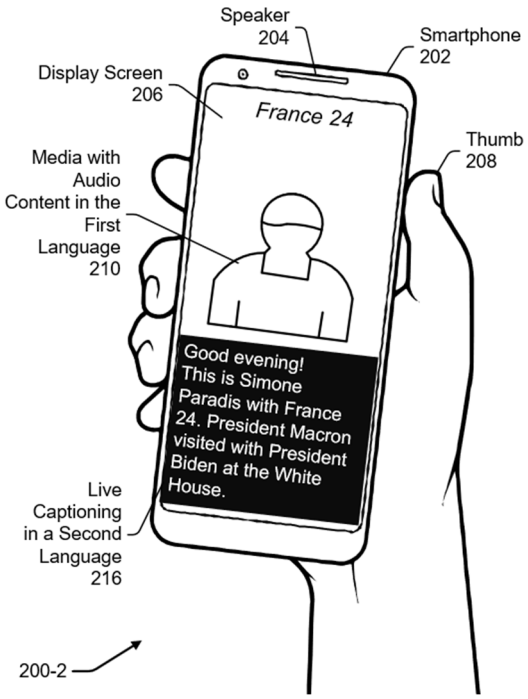

Figure 2B

Figure 2A and Figure 2B illustrate example environments 200-1 and 200-2, respectively, and this disclosure describes the example environments in the context of Figure 1. For example, assume an English-speaking user utilizes their smartphone 202 with a speaker 204 and a display screen 206 to stream media with audio content in a first language 210 (illustrated *France 24*), and they struggle to understand a French-speaking news anchor. To understand the anchor of *France 24*, the user may trigger live captioning with language translation in a second language 216 of Figure 2B (e.g., English).

To trigger live captioning with language translation, the user may press their thumb 208 on a volume control (e.g., a side button) of the smartphone 202. After pressing the side button, the display screen 206 displays a volume box 212 and a caption box 214, as is illustrated in Figure 2A. The user may utilize the side button to increase a volume of, decrease the volume of, or mute the smartphone 202. The user may also tap the caption box 214 to display live captioning with language translation in the second language 216. Thus, regardless of the volume level of the smartphone 202, the user can consume the media with audio content in the first language 210 by reading live captioning in the second language 216.

In one aspect, at a system level (e.g., OS), the user can set a default language translation, for example, English. In such a case, regardless of the first language (e.g., French, German, Mandarin Chinese), when the user taps the caption box 214, the display screen, by default, displays live captioning in English. However, in another aspect, the user may opt out of setting a default language translation at the system level. Instead, the user can select a language translation on a case-by-case basis. In the latter case, when the user taps the caption box 214, the display screen 206 may display a list of second languages and prompts the user to select the second language to be used in live captioning.

Additionally, or alternatively, the smartphone 202 may enable the user to trigger live captioning with language translation using a voice-activation and/or a voice-assistant feature. For example, the user may trigger live captioning with language translation by saying phrases, like "Assistant, display live captioning translated in the second language," "Assistant, mute the smartphone and display live captioning translated in the second language," "Assistant, display live captioning in the first language (e.g., French) alongside live captioning in the second language (e.g., English)," or other well-defined phrases.

Figure 2B illustrates the live captioning in the second language 216 with small white text inside a black text box. The user, however, may customize a size, font, and color of the text and a color of the text box depending on their needs, vision limitations (e.g., color blindness, poor vision), and/or preferences. In addition, the user may also configure the smartphone 202 to play aloud the live captioning in the second language while muting the media with audio content in the first language. Therefore, the techniques and apparatuses described herein enable visually impaired people, hard of hearing people, and unimpaired people to consume media with audio content in the first language with live captioning (and/or audible playing) in the second language.

References:

[1] Bolella, Danny. "On-Device Video Subtitle Generation in SwiftUI." Heartbeat, March 19, 2020. https://heartbeat.fritz.ai/on-device-video-subtitle-generation-on-ios-with-swiftui-and-ml-kit-6ed28dd2734c.

[2] Tadmor-Ramanovich, Michelle, and Nadav Bar. "On-Device Captioning with Live Caption." Google AI Blog, October 29, 2019. https://ai.googleblog.com/2019/10/on-device-captioning-with-live-caption.html.

[3] Patent Publication: CN106156012A. A kind of method for generating captions and device. Priority Date: June 28, 2016.

[4] Patent Publication: US20120010869A1. Visualizing automatic speech recognition and machine. Priority Date: July 12, 2010.

[5] Patent Publication: US20190244606A1. Closed captioning through language detection. Priority Date: May 26, 2017.

[6] Patent Publication: KR20210028041A. Electronic device and Method for controlling the electronic device thereof. Priority Date: September 3, 2019.

[7] Wang, Xiong, Zhuoyuan Yao, Xian Shi, and Lei Xie. "Cascade RNN-Transducer: Syllable Based Streaming On-Device Mandarin Speech Recognition with a Syllable-To-Character Converter." 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 15-21, doi: 10.1109/SLT48900.2021.9383506.