Smart Transcription

Understanding Telephone Calls at a Glance

Cornelius Glackin
neil.glackin@intelligentvoice.com
Intelligent Voice Ltd
London, UK

Nazim Dugan
Intelligent Voice Ltd
London, UK
nazim.dugan@intelligentvoice.com

Daragh Cahalane
Intelligent Voice Ltd
London, UK
gary.cahalane@intelligentvoice.com

Nigel Cannings
Intelligent Voice Ltd
London, UK
nigel.cannings@intelligentvoice.com

Julie Wall
University of East London
London, UK
j.wall@uel.ac.uk

ABSTRACT

The Intelligent Voice SmartTranscript is an interactive HTML5 document that contains the audio, a speech transcription and the key topics from an audio recording. It is designed to enable a quick and efficient review of audio communications by encapsulating the recording with the speech transcript and topics within a single HTML5 file. This paper outlines the rationale for the design of the SmartTranscript user experience. The paper discusses the difficulties of audio review, how there is large potential for misinterpretation associated with reviewing transcripts in isolation, and how additional diarization and topic tagging components augment the audio review process.

CCS CONCEPTS

• **User interface design**; • **Speech recognition**; • **Speech and audio search**; • **Topic modelling**;

KEYWORDS

UI/UX Design, Audio Review, Automatic Speech Recognition, Diarization, Topic Tagging

ACM Reference Format:
Cornelius Glackin, Nazim Dugan, Daragh Cahalane, Nigel Cannings, and Julie Wall. . Smart Transcription: Understanding Telephone Calls at a Glance. In *Proceedings of* . ACM, New York, NY, USA, Article , 4 pages. https://doi.org/

1 INTRODUCTION

A normal audio or video file is a serial access medium whereby in order to access certain audio (including speech) contained within it, it is necessary to listen to or to watch the file at its original speed (or close to it) until that data is found. Thus, for someone tasked with listening to an audio file or watching a video file to search for certain words, phrases or topics (e.g. a paralegal in a discovery process), such a task is time consuming and fatiguing. In contrast, a paper transcript can be relatively quickly speed-read by a human at rates in excess of 700 words per minute, i.e. in a fraction of the time and the effort. Hence, having a good quality transcript is a quicker way to review large quantities of speech audio.

Human transcription of audio is typically more accurate than automatic speech recognition (ASR), but it is generally much slower. It can take a reasonably skilled human transcriber 8 hours to transcribe each single hour of original audio. However, even if a human could perform the task at real-time speeds, it would still take 1,000 hours to transcribe 1,000 hours of speech audio. For many real-world use cases, such as the aforementioned paralegal discovery task, this is simply not feasible. What is required is not just to have accurate ASR, but to have it running at many times real-time speeds. This is what many commercial speech recognition applications are required to do. The author's commercial ASR offering operates at approximately 200 times real-time speed on a single GPU (based on NVIDIA Quadro Volta GV100 32GB GPU). This means that the aforementioned paralegal's transcription task of 1,000 hours of speech audio, can be handled in 5 hours, and with additional GPU resources it can be accomplished linearly faster with data parallelization.

This research stemmed from a desire to improve the plain text transcripts that are produced by an ASR engine. This paper describes the technological enhancements that have been developed to facilitate the task of understanding large quantities of speech audio. This involves the consideration of the user experience (UX), i.e. incorporating the important functionalities of speaker separation, topic-tagging, searchable content, and ease of interaction with the transcript. The resulting SmartTranscript [Glackin et al. 2019] is searchable, portable, intuitively interactive, clearly separates the speakers and highlights topics of interest within the audio. All of which, enhance the UX.

2 SPEECH AND NLP COMPONENTS

The SmartTranscript is the combination of an audio recording, a transcript generated using ASR, and a list of topics that are automatically identified using Natural Language Processing (NLP).

2.1 Voice Activity Detection

It should be understood that traditional ASR should only be applied to actual speech. In an audio recording there are often long silences and background noises that should not be processed by the speech recognizer. For this reason, Voice Activity Detection (VAD) is a necessary pre-processing step for identifying segments of audio for processing by the ASR engine. VAD typically works

by thresholding acoustic energy. This task is also an opportunity for GPU-based ASR inferencing to segment the audio into smaller chunks for batch processing and optimally loading GPU memory, i.e. for faster processing of the audio through the ASR engine.

2.2 Automatic Speech Recognition

Briefly, ASR, also known as Speech-to-Text is simply the process of converting speech audio (labelled by VAD) to text. There are various ways in which this can be achieved and there is a long history of the evolution of ASR development. This began with template matching approaches, leading to Hidden Markov Models, Gaussian Mixture Models, and then to end-to-end Deep Learning approaches [Huang et al. 2014]. More recently, Convolutional Neural Network-based approaches [Glackin et al. 2018; Tóth 2014] are the state of the art, in recognition of the transferability of automated feature extraction in image recognition to automated feature extraction in the spectro-temporal domain. Recent claims of human parity are now making it possible to use automated transcripts as a cost-effective way of not only keeping a record of voice-based interaction within institutions (often for legal and compliance purposes), but they are also an important investigative tool in forensic investigations e.g. insider trading and fraud investigations [Xiong et al. 2016].

2.3 Diarization

A grammatically unstructured audio conversation, without an indication of who is speaking when, is difficult to interpret, and for practicality purposes it is important to be able to identify who said what during a conversation. Often institutions merge channel-separated audio to mono and compress the audio format to mp3 to save on space. This makes the task of separating speakers very problematic and this task (called diarization), is very much a research problem in its own right. Diarization for mono telephone recordings is typically a template matching problem rather than recognition-based. Most commercial approaches are based on voice biometrics, requiring the identification of speakers by their acoustic signatures (voice prints). Extraction of the signature needs to be reliably enrolled in a database and then segments of the audio are compared to the speakers enrolled in the database. One of the main issues with diarization is the presence of back-channeling. Back-channeling is the technical term given to very short speaker changes. It usually involves feedback from a secondary speaker to the primary speaker. which defines the listener's comprehension and interest in what is being said. Obviously in normal telephone conversation flow, the roles of primary and back-channel speakers alternate. In a typical conversation, this behavior results in segments that are too small to reliably template-match to the enrolled speaker database. However, all of this can be avoided with a properly configured telephone call recording system, which enables ASR to be processed independently in two separate channels. The transcribed speech segments are time-stamped and the two transcripts are time-aligned and merged.

2.4 Topic Tagging

The task of monitoring the context of the spoken word in telephone conversations is particularly problematic for automated systems because telephone conversations are more acoustically challenging due to encoding, down-sampling and the compression inherent in telephone communication. Furthermore, conversational speech is more chopped and broken compared to the spoken word in, for example, presentations, dictations and face to face conversations. Even with accurate transcription and diarization, the ability of speakers to understand transcripts without the accompanying audio requires a significant cognitive load to understand the context of the interaction. The ultimate goal would be the ability to summarize conversations in the same way that humans relay the details of conversations to one another. Whilst NLP-based summarization of language research is progressing, reliable summarization of conversations is currently not available.

Even with ASR inferencing at several hundreds of times faster than real-time, the sheer amount of transcription that is generated on a daily basis for some businesses can be unfeasible to review properly. This can lead to large stores of transcripts that are not properly exploited for the rich source of information they contain. NLP has an important role to play in identifying important topics, named entities, in an automated reliable way. An ASR system with a decent language model should be able to reliably transcribe named entities, proper nouns, and words with larger sequences of phonetic symbols more accurately than smaller words. By decent language model, we mean one that has been trained on an extremely large corpus of real-world speech transcription, so that the conditional probabilities of the sequences of words are accurately calculated.

The tagging process employed in the SmartTranscript not only picks out long words but our proprietary algorithm also automatically identifies phrases. The real challenge is the optimization of the search for phrases, based on the huge number of phrases in the phrase database. This task is achievable in real-time with CPU-based parallel programming. The resulting 'tagged' phrases and keywords contained in the telephone transcript are then presented in the SmartTranscript. The most complex words are listed first as measured by total character count, regardless of the number of words in the topic. This data-driven approach to topic tagging often finds unexpected topics of interest [Hodson 2016]. This can be of particular importance when monitoring telephone conversations for unlawful or adverse practices, where parties of the telephone conversation may use coded words or covert behavior. Even without willful obfuscation by the parties to a conversation, unanticipated or unexpected matters may be discussed that by definition could not form part of a keyword search approach, as they represent blind spots and unknown unknowns.

3 THE SMART TRANSCRIPT

Figure 1 presents the SmartTranscript. The current design has been iteratively developed based on user feedback with the aim of making the SmartTranscript intuitive to use.

The SmartTranscript is an HTML5 document with audio and interaction functionality embedded into it. The choice of HTML5 as the basis for the transcript was made in an effort to make the SmartTranscript portable. In this way, it does not require any external software other than a standard HTML5 compliant browser, the reader can download the transcript [Glackin et al. 2019] using the reference and verify these portability claims.

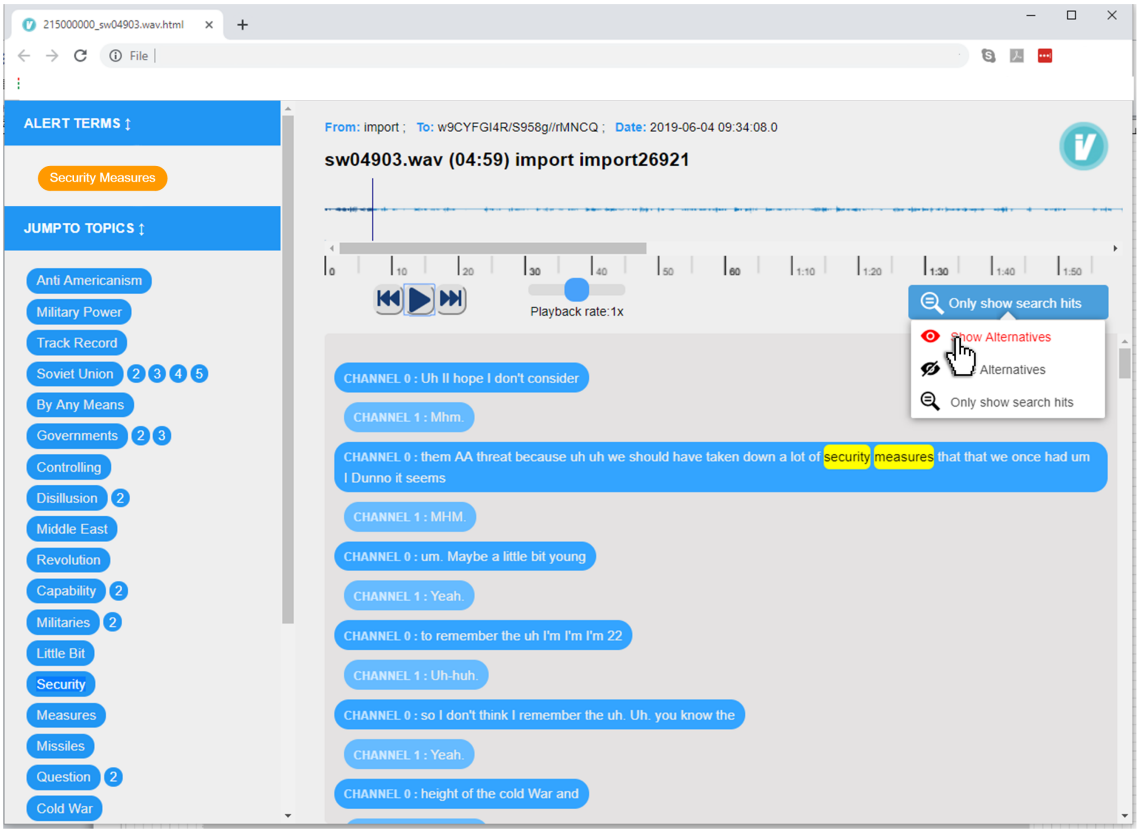

Figure 1: The SmartTranscript is an HTML5 document that contains an interactive transcript (main pane), list of key topics (left pane) and the original audio (right pane top)

The SmartTranscript is organized into three main panels, the panels contain the key topics (left) that have been automatically identified from the transcript, the audio waveform (top right) and the speaker separated transcript below (bottom right). Every part of the transcript is interactive, with the result of navigating to any part of the transcript using the mouse. The audio controls enable simultaneous playing of the audio whilst the words being uttered are highlighted in yellow in the transcript. Clicking on any of the key listed topics (JUMPTO TOPICS in Fig. 1) highlights the transcript five seconds before the topic is mentioned and plays the audio from that particular point. Similarly, the audio waveform can be navigated, as can any word in the transcript itself. In this way, the design promotes freedom of navigation to the user. An additional important design decision is the ability to search and this is a benefit of the SmartTranscripts HTML5 medium, in that it can be searched like any other web page, typically by using Ctrl+F.

Portability has also been considered in regards to the integration aspects of the SmartTranscript. Another significant usecase involves the ability to integrate the SmartTanscript easily into a review platform. This functionality would enable you to search 1,000s of SmartTranscripts simultaneously. You will notice that Figure 1 also shows the Alert Term: 'Security Measures', this is coming from an eDiscovery platform as an example blanket search term of interest, that say several thousand transcripts can be searched for.

The latest version of the SmartTranscript also provides alternative words to those included in the transcript. For example if a word is transcribed with less than certain probability, this means that the ASR engine is acknowledging that the word transcribed might be incorrect, and provides possible alternatives which are now surfaced in the SmartTranscript, see Fig. 1. Alternative words within the SmartTranscript help disambiguate the following cases:

- Acoustically indeterminate - Similarity: "My baby's breathing heavily..." versus "My baby's bleeding heavily..."
- Semantically indeterminate - Spoken words: "...she'll get the clause out..." versus "...she'll get the claws out..."
- Acoustically Indeterminate - Pronunciation: "On" versus "Off" as per checklist response for the Deicers on the Cockpit Voice Recorder for Air Florida Flight 90 which crashed on January 13, 1982 in Washington D.C. [Channel4 1996].

The encapsulation of the audio within the SmartTranscript also helps with disambiguation:

- Semantically Indeterminate - Emphasis: "He'd *kill* us if he got the chance..." versus "He'd kill *us* if he got the chance...". The crux of the entire plot for an Academy Award-nominated movie about covert audio collection [Coppola 1974], the emphasis placed by the speaker on two successive words in a collected snippet of conversation.

- Contested Intelligibility: one person hears, or claims to hear, intelligible words. Another person does not, or claims not to.

4 DISCUSSION

There is a well-known design principle that is summarized by the phrase "what is beautiful is usable" [Tractinsky et al. 2000]. Interestingly, the theory indicates that aesthetics seem only to affect the lasting impression of the product, rather than the actual usability. Nevertheless, whilst the usability of the product is the main priority, significant effort has been made to make the SmartTranscript attractive. To that end, the SmartTranscript has undergone numerous iterations to improve its aesthetics.

Perfect information does not necessarily lead to desired decision making according to the theory regarding the role of information accuracy and its effects on planned behavior [Ajzen et al. 2011]. As an example, presenting a population of people with information about alcohol consumption and its effects on human physiology does not make any difference to drinking habits. Similarly, we have found that in the context of ASR accuracy, that perfect transcription alone falls short of facilitating efficient review of audio transcripts. It is in fact the additional design elements in the SmartTranscript that have more of an impact in terms of usability and information management. Many of these features address the tenets of management information systems identified in [Warren 2014]:

- To combat information overload, as the volume of information increases.
- To ease context switching, in particular, for users who face frequent interrupts in their work.
- To be supported in information integration, across a variety of applications.

To address information overload, we point to the use of topics as being a way to summarize and compress the information contained in the transcript. There is no quicker way to quickly get a sense of what a document contains than looking at the key topics in the left pane.

With regards to context switching, the karaoke highlighting associated with the play/pause facility acknowledges that people are often interrupted in audio review tasks. This highlighting of where playback is paused aids the ability of the user to rapidly refamiliarize themselves with the point in the conversation from where they left off. If refamiliarization with the recording as a whole is first required, the list of topics enable this in a matter of seconds.

The SmartTranscript does not rely on proprietary software that needs to be installed on edge devices before it can be used. It has instead been designed to be used in any browser and as such contains standard browser functionality such as the ability to search for text within the transcript. Additionally, the ease by which the SmartTranscript has already been integrated into e-Discovery Monitoring and Review platforms has been demonstrated. Warren et al. [Warren 2014] provide guidelines for achieving management information system best practice by advocating unified approaches to managing files such as text, email, and urls. This ethos has been adhered to by combining audio and transcripts in one file.

5 CONCLUSIONS

This paper presented a summary of the efforts and technological enhancements that we have made in order to address the current lack of usability of telephone speech transcripts. Audio review of large volumes of telephone calls is an arduous process that is characterized with human reviewers prone to sharp decreases in effectiveness. Recorded speech often features background noise, reverberation and a mixture of near and far-field microphone capture, which often ensures that the speech recognition accuracy is not perfect. The diarization task of who is speaking when in a recording is also imperfect due to back channeling changes of speaker that offer inadequate sample lengths to accurately identify speaker change. Overlapping speech is also an issue for accuracy purposes but also from a representational standpoint because the transcribed word is inherently turn-based.

Near perfect transcription and diarization help to interpret what was said in a recording, but it should be realised that they are interpretations, missing elements of non-verbal communication, gestures, and importantly emotion. In the presented user experience, it is precisely the encapsulation of the conversation topics and diarized transcript together with the original audio or video recording within a single html file that makes the Intelligent Voice SmartTranscript an invaluable too we believe is the only way to optimally perform the task of audio review.

REFERENCES

Icek Ajzen, Nicholas Joyce, Sana Sheikh, and Nicole Gilbert Cote. 2011. Knowledge and the prediction of behavior: The role of information accuracy in the theory of planned behavior. *Basic and applied social psychology* 33, 2 (2011), 101–117.

Channel4. 1996. Survival in the Sky, Episode 2, Deadly Weather. https://www.youtube.com/watch?v=Ww-NHTsIuMY.

Francis Ford Coppola. 1974. The Conversation. American Zoetrope and Paramount Pictures.

Cornelius Glackin, Nazim Dugan, Daragh Cahalane, Nigel Cannings, and Julie Wall. 2019. Intelligent Voice Smart Transcript Example. http://tinyurl.com/y4ntdjz7.

Cornelius Glackin, Julie Wall, Gérard Chollet, Nazim Dugan, and Nigel Cannings. 2018. TIMIT and NTIMIT Phone Recognition Using Convolutional Neural Networks. In *International Conference on Pattern Recognition Applications and Methods*. Springer, 89–100.

Hal Hodson. 2016. *Prisoners' code word caught by software that eavesdrops on calls*. NewScientist. https://www.newscientist.com/article/mg23030762-200-prisoners-code-word-caught-by-software-that-eavesdrops-on-calls/.

Xuedong Huang, James Baker, and Raj Reddy. 2014. A historical perspective of speech recognition. *Commun. ACM* 57, 1 (2014), 94–103.

László Tóth. 2014. Convolutional deep maxout networks for phone recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*.

N Tractinsky, AS Katz, and D Ikar. 2000. What is beautiful is usable. *Interacting with Computers* 13 (2000), l27–145.

Paul Warren. 2014. Personal information management: The case for an evolutionary approach. *Interacting with Computers* 26, 3 (2014), 208–237.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256* (2016).