SPEECH-TO-TEXT CONVERSION FOR EFFECTIVE AUDIO TRANSCRIPTION IN MASS MEDIA INDUSTRY OPERATIONS

[1]Isiaka, O.S; [2]Ibraheem, A.F; [3]Bolaji-Adetoro, D.F; [4]Saka, T.O
[1,3&4]Department of Computer Science, Institute of Information and Communication Technology, Kwara State Polytechnic, Ilorin; [2]Department of Mass Communication, Institute of Information and Communication Technology, Kwara State Polytechnic, Ilorin.
[1]isiakaosalman2@gmail.com, +2348035886758; [2]fibraheem660@gmail.com, +2348068107279; [3]bolajiadetorofunsho@gmail.com, +2348038270619; [4]sakataju70@yahoo.com, +2348164073100

ABSTRACT

Audio transcription is critical in the day-to-day operations of a business, reporting or organization. To avoid disputes, almost all media industries keep a record of daily activities such as interviews, meetings, and presentations. One of the best ways to keep such records is to keep a written record of every word said during the activities. However, because most activities are rushed, it is challenging for media officers, secretaries or attendees to listen and take written accounts of what is said at the same time. To cope with the challenges, it is easier to record the audio and then convert it to texts. This is where Speech-to-Text (STT) transcription comes in handy.STT transcription enables the media to quickly convert spoken words, numbers, or acronyms to text. STT, for short, transcription has a wide range of applications, the most common of which are media documentation or reports, legal proceeding logs, education, and entertainment.

Keywords: Audio transcription, verbatim, speech-to-text conversion, media industry operations, intelligent verbatim, Hidden Markov Model

I. INTRODUCTION

Speech-to-Text (STT) and Automatic Speech Recognition (ASR) are crucial technologies in fields like dictation software, voice-activated assistants, call center automation, and transcription services. STT analyzes audio signals and uses machine-learning algorithms to recognize words and convert them into texts. ASR, on the other hand, analyzes syntax, semantics, intonation, and pauses. Both technologies use machine-learning algorithms trained on large datasets, allowing them to recognize speech patterns and make accurate predictions (Hinton et al., 2012). They offer benefits such as improved accessibility for hearing impaired individuals, increased efficiency in call center operations, and enhanced user experiences for voice-activated assistants (Iosifov, Iosifova & Sokolov, 2020).

Mass media play a vital role in society by ensuring that accurate information is disseminated without misinterpretation or misunderstanding. Transcription is a process that converts a spoken language or recorded audio into a written or digital text, often used in interviews, academic research, conversations, or event recordings. It breaks down barriers and makes information accessible to all, including those with hearing impairments or language barriers (Dahl, Yu, Deng & Acero, 2012). This process promotes inclusion and diversity, ensuring everyone can benefit from the wealth of knowledge and entertainment available today.

Transcription is essential in today's communication-driven world, as it helps to avoid misunderstandings, clarify key points, and capture every detail with precision. It is essential in fields like legal, medical, and mass communication for accurate record-keeping and reporting. It ensures the preservation of truth and a clear understanding of the world, enabling faster reading and review of audio content (Frank,Catherine, Kaitlyn & Daniel, 2016; Chen, Wang, Chen & Wang, 2020). Transcription saves time when taking notes during meetings, lectures, or interviews, allowing individuals to focus on active listening and engaging in discussions. It assists in having an accurate written record.

Transcription is effective in the media industry in various ways including audio files, video files, and written materials. Human transcription involves real people deciphering audio files and converting them into texts, while automated transcription uses speech recognition to convert audio to text quickly. Human transcription is more accurate due to its ability to decipher heavy accents and industry jargons, and is more effective with tough audio, including background noise and multiple speakers (Kim, Seltzer, Li & Zhao, 2018). Automated transcription requires editing with an editing tool, but high-quality audio files with one or two speakers, with no accents or complicated jargons, and minimal background noise will produce more accurate transcripts.

II. REVIEW OF RELATED TEXT

According to Frank, Catherine, Kaitlyn & Daniel (2016), cognitive modelers use verbal protocol analysis for data testing, and new tools allow automatic transcription of audio. This research compares the time it takes to transcribe a movie using Google's subtitle tool, corrected from the tool, and by hand. It recommends the use of Google subtitles as a starting point for spoken transcription, with guidelines for when and how to utilize them. It also recommends headphones, and automated spelling correction in text editors. Weiner et al. (2016) present two feature extraction methods for dementia detection: manual and fully automatic. The manual pipeline employs manual transcriptions, whereas the fully automatic pipeline uses transcriptions generated by automatic speech recognition (ASR). According to the study, manual transcription is unnecessary for diagnosing dementia on the ILSE corpus because ASR technologies have access to eight thousand hours of untranscribed interviews, allowing automatic detection of dementia using acoustic and linguistic variables.

For analysis, Stuckey (2014) converts spoken text data from interviews into written form. This entails de-identifying individuals and transcribing the data, which has an impact on the correctness and reliability of the analyzed data. The first step in the analysis is to examine the initial interview to assess participants' replies and drive refinement. Transcripts are ready for coding, and it is critical to recognize the responsibility of dealing with qualitative research material both during and after the interview. Wang, Yang, Li, Sadhu & Hermansky (2019) present two ways for automatically identifying mistakes in manual voice transcription: utilising a biased language model to discriminate between hypotheses and transcription, and exploiting mismatches between acoustic classifiers and forced alignment and posteriors. On a genuine Mandarin dataset, both methods demonstrated equivalent accuracy in detecting transcription mistakes. Combining these ideas might be interesting in the future.

III. SPEECH-TO-TEXT (STT) ALGORITHMS

Speech-to-Text (STT) is a computer science field involving Linguistics, Mathematics, and Statistics. It involves speech input, feature extraction, vectors, decoder, and word output, using acoustic, pronunciation, and language models (Olena et al., 2021). STT technology is evaluated based on word error rate (WER) and speed, with factors like pronunciation, accent, pitch, volume, and background noise affecting WER. The goal is to achieve human parity, with Lippmann's research estimating a four percent WER (Rista & Kadriu, 2020). Algorithms and computation techniques, such as speech recognition, text comprehension, and machine learning, improve STT and transcription accuracy.

a) **Natural language processing (NLP):** Natural Language Processing (NLP) is a subfield of artificial intelligence that helps machines to understand human language for repetitive tasks like translation, summarization, ticket classification, and spell checking. It combines computational linguistics, machine learning, and deep learning models. Computational linguistics uses language translators, STT synthesizers, and speech recognition software while machine learning trains computers to understand human language features like sarcasm, metaphors, and grammar. Deep learning uses neural networks to teach computers to learn and think like people.

b) **Hidden Markov Models (HMM):** A probabilistic model known as Hidden Markov Model (HMM) is used for natural language processing tasks like tagging, named entity recognition, and machine translation. HMMs provide observations that reflect time series of observations by using the probability distribution of each state at each time step. The model's state is reliant on its prior state.

c) **N-grams:** An N-gram is a group of n consecutive written elements used in text analytics for sentiment analysis, classification, and production. N-gram grammars, an N-th order Markov language model, estimate the probability of an event based on the occurrence of N-1 other symbols. They support open vocabulary applications and come in various forms, including unigrams, bigrams, trigrams, and higher order N-grams.

d) **Neural networks:** Neural networks are networks of artificial neurons that receive input, change state and generate output. These interconnected groups use mathematical models for information processing based on connectionist communication. Neutral networks can be simple models, but are often connected to specific learning rules.

e) **Speaker Diarization (SD):** Speaker diarization involves segmenting and co-indexing audio recordings to identify distinct speakers, enabling speaker-attributed STT transcription and speech recognition.

IV. TRANSCRIPTION IN MEDIA

Transcription is a crucial aspect of the media industry, converting audio and video contents to texts for analysis and editing. As the media industry expands and competition in the industry increases, transcription plays a significant role in achieving success in broadcast media. While some programmes have scripts, others are pre-recorded without scripts, making it easier to analyze (Hsiao, Can, Ng, Travadi & Ghoshal, 2020). Transcription provides detailed text for analysis, making it easier to comply with broadcasting regulations. This is the main role of transcription in broadcast media (Romanovskyi et al., 2020).Transcription in media may be used for editing, verbatim quotation, intelligent verbatim, and phonetic materials, with each having its advantages and disadvantages.

a) **Edited transcription:** Edited transcription formalizes and edits a complete script for readability, conciseness, and clarity. It addresses grammatical mistakes, slangs, and incomplete sentences, corrects spelling and punctuation, and enhances the spoken words' formal sound.

b) **Verbatim transcription:** Verbatim transcription is the written form of spoken language extracted from video and audio files, capturing every sound and indicating pauses. It is crucial for accurate translations in legal-related recordings. Non-verbatim transcription removes background noises, pauses, and throat clearing, cleaning up incomplete sentences.

c) **Intelligent verbatim transcription:** Intelligent verbatim transcription removes fillers, repetitions, and non-standard words to create a concise, readable transcript while maintaining participants' voice and intended meaning. It avoids irrelevant sentences, off-topic remarks, and irrelevant pauses, coughing, and noises, ensuring clear communication without interference with the speaker's voice.

d) **Phonetic transcription:** Phonetic transcription involves using phonetic symbols to represent spoken words in written form, using the International Phonetic Alphabet (IPA). This process is useful for preserving dialects and pronunciation changes over time, such as in period movies or dialect differences.

V. MASS MEDIA INDUSTRIES METHODS OF TRANSCRIPTION

Transcription in mass media industries especially those ones focusing on investigative reporting, which requires evidence, demands great skills to be able to deliver high quality transcriptions. This personal expertise requires a lot of dedication in terms of time and cash (Yu, Zeiler, & Kolossa, 2022). These constraints make manual transcription unsuitable for large-scale projects requiring a fast turn-around. Transcription can take place in the following ways:

a) **Audio Transcription:** Professional transcription is a skill developed over the years of practice. Essential rules include listening to the entire recording before writing to understand spoken content, especially in unfamiliar accents. Understanding context and avoiding misunderstandings over homophones is crucial. Editing transcription for mistakes and grammar enhances efficiency. Touch-typing techniques maximize speed, accuracy, and comfort. Hand cramps can hinder accuracy, but modern software uses foot pedals to control audio. This helps to acknowledge jargons and abbreviations, especially in medical sectors, and double-check for proper terminology, paragraph meaning and accuracy.

b) **Video Transcription:** Spoken content transcription is a lengthy but rewarding process using four key methods, each with its advantages and disadvantages, depending on the specific situation for which it is being adopted.

   i. Transcription apps for mobile phones: Mobile phones provide an easily portable tool for capturing people’s speech on the go. In addition to most smartphones’ built-in speech-to-text application, there are a variety of transcription apps available for download from the various app stores.

ii. Free online video transcription: Search-engine queries reveal various free transcription tools online, but quality can vary. It is important to proofread captions for errors. Automatic YouTube captioning offers up to eighty percent accuracy, depending on video quality, but not all languages are supported.

iii. Transcription software for desktop computers: Mac and PC users can access online transcription tools and download desktop software, enabling remote access without internet connection.

iv. Captioning services: Professional captioning services and localization providers offer superior results, security, and confidentiality compared to free software solutions.

c) **Group Conversations Transcription:** Transcribing multiple people's conversations in a transcript can be challenging, especially with frequent interruptions. To ensure clarity, transcribe each person on a separate line, indicate simultaneous speech with the same time stamp and add an 'Inaudible' tag if commotion makes it impossible to hear.

d) **Automated Transcription**: Software-based transcription reduces time and costs but has limitations in transcribing regional accents. Results depend on AI technology and machine learning capabilities, and even the best machines are not 100% accurate.

VI. AUDIO TRANSCRIPTION GENERATION PROCESS

There are several alternatives available when making an audio transcript. One option is to use automated transcription, in which audio is automatically turned into text using speech recognition software. Although the method is faster and less expensive, it has some limitations. The AI software may find it difficult to distinguish between speakers in a recording with a large number of participants (Zhang et al., 2020). It may also mispronounce the names of places and speakers struggle with accents. Based on the limitations mentioned above, automatic transcribing does not provide accurate results; hence, the transcript should be reviewed before keeping or distributing it. Another option is manual or human-generated transcription. A human transcriber listens to the audio and converts it to text. Since this transcription option is often significantly more accurate, the transcript may not need to be edited (Tanaka, 2019). The only downside of manual transcription is that it takes more time and costs more.

a) Automated Generation: STT software listens to audio files and produces an editable, verbatim transcript on a particular device. The software does this through voice recognition. A computer programme utilizes linguistic algorithms to separate audio signals from spoken words and translate those signals into text using Unicode characters. STT conversion uses a multi-step, complex machine learning algorithm. When sounds are made by speaking, several vibrations are generated. These vibrations are detected by STT technology, which then transforms them into a digital language using an analog-to-digital converter. The analog-to-digital converter receives an audio file as a source and measures and filters the waves to extract the necessary sounds. The sounds are matched to phonemes after being separated into hundredths or thousandths of seconds. A phoneme is a unit of sound that is utilized in every language to distinguish one word from another. For instance, the English language comprises about 40 phonemes. The phonemes are then tested using a mathematical model, using a network of well-known sentences, words, and phrases as comparison points. Depending on the most likely interpretation of the audio, the text is subsequently shown as text or as a computer-based demand.

b) **Manual Generation:** Transcription of audio file is required for effective operations in mass media industries, research, marketing, planning, management, and teaching. It is possible to hire a transcriptionist because the process is straightforward. Learning to make a transcript is not challenging, but it requires a lot of effort. The following steps can be taken to improve transcription quality and accuracy.

i. *Prepare Transcription Materials*: The transcription of audio or video sources is a full-time job that requires the use of specific tools, supplies, and equipment. A master list of elements that must be present in order for correct and timely transcripts to be produced is kept on hand by experienced transcriptionists. Investing in high-quality items like comfortable chairs, transcription software, noise-cancelling headphones, and foot pedals will pay off in the long run.

ii. *Understand the Transcription Requirements*: Before moving on to the next stage, test the menu navigation and functionality of the transcription programme. Transcribing different materials will take less time and effort as a result. Depending on the project, decide if full verbatim or clean verbatim transcription is needed. Use thorough notes to capture the core of an interview so that readers may understand the information without reading the complete transcript.

iii. *Listen to the Audio Recording*: It is essential to learn how to write a transcript in a quiet, enclosed area with few outside distractions. Transcriptionists typically finish their transcription late at night when everyone must have fallen asleep. Prepare your materials and headphones, pay attention while listening, adjust the volume and playback speed as necessary as possible, make a note of any unfamiliar words, and ensure that they make sense in the context of the sentence. Make use of dynamic range compressor technology to reduce background noise.

iv. *Create an Initial Draft*: For accurate transcription, change the playback speed to match the speed of the keyboard. Avoid spelling and grammar mistakes, but make sure drafts are copied and edited several times. Use shorthand and abbreviations to accurately transcribe every word from the recording. Just keep in mind the golden rule of transcription: don't add or omit, and don't fix grammar mistakes.

v. *Double Check the Draft to Ensure Accuracy*: Writing better transcripts requires rereading drafts and looking for spelling mistakes. Instead of editing for grammar or word choice, use a dictionary or a spell checker. The transcription must be as precise as the original recording.

vi. *Format the Transcription*: Format the transcription by adding title, page numbers, and distinct paragraphs after the second pass. Avoid excessive text in one paragraph. Use "[sic]" for grammatical mistakes, "(inaudible)" for unclear words, and "*" or "***" for missing word(s). Add timestamps for each speaker. Use ellipses or "pause" for pauses to maintain consistency. Consider nonverbal communication, such as clapping or laughter, and avoid interpreting nonverbal cues.

vii. *Finalize the Transcription*: For the last time, listen to the original audio recording while reading the updated transcript to evaluate the process correctness, dependability and quality.

VII. INTEGRATION OF AUTOMATED AND MANUAL TRANSCRIPTIONS

Both manual and automated transcriptions have advantages for media sectors, legal, corporations, and people. Manual transcriptions are more dependable and precise because trained humans understand the intricacies of audio media tone and pitch. Automated transcription is inexpensive and rapid, making it ideal for people and enterprises who prefer to conduct transcriptions themselves (Wang, Wang & Lv, 2019). It boosts efficiency and gives real-time transcriptions, while adding translation enables materials to be done in different languages. Both methods have advantages for businesses striving to increase their productivity and access to written materials.

Manual transcription method is inefficient in terms of time and resources. A trained human transcriptionist can take up to ten hours to transcribe a single hour of audio recordings due to changes in circumstances. The process can be fast for simple topics, but if the subject matter is robust or complex, the transcriber may require a lengthy period of time to get the job done. Language context is a challenge for automated transcription methods, resulting in inaccuracies in transcribed contents. They have difficulty distinguishing between speakers, mixing up names, and spelling non-English terms (McDermott, 2018). Furthermore, they lack built-in context sensitivity, making them difficult for quick talks and context sensitivity in voice technology solutions.

Automatic transcribing software, which is currently available, simplifies the choice between manual and automatic transcriptions by addressing individual demands, constraints, and nature of work. The decision is based on the specific demands of the company or organization.In many cases, combining these two methods will produce the best and most detailed results (Iosifova, Iosifov, Rolik & Sokolov, 2020). A human transcriptionist, for example, may begin the work with automated transcription software, followed by manual transcription in which the expert uses his skills and knowledge to correct any mistake that may occur from automated transcription in order to improve the overall quality of work.

VIII. AUDIO TRANSCRIPTION STYLES

The two main styles used in audio transcription are full verbatim and clean verbatim.

i. **Full verbatim**: Full verbatim transcripts provide context through non-speech sounds, making them valuable in group interviews, focus groups, legal transcribing and meetings. It can be used to directly quote sources of information, conduct focus groups or interview from research studies, prepare legal papers, and give formal remarks. Full verbatim is ideal for transcripts needed for evidence purposes. Verbal signals are very handy for transcribing interviews with personalities or possible suspects (Chan, Jaitly, Le & Vinyals, 2016). In full verbatim, the transcriptionist transcribes every word in the audio, including hesitations like “um” and “ah” and false starts. This transcription style also contains pauses, laughter, and slang language such as “gonna.”

ii. **Clean verbatim**:Clean verbatim transcription is the type that modifies audio or video files to make them easier to read as well as comprehend. Filler words, non-speech sounds, and throat clearing are removed to make each sentence less stressful to read. Other transcription service providers, such as edited transcription and intelligent verbatim, also adhere to clean verbatim styles (Chan, Jaitly, Le & Vinyals, 2016). Many businesses value clean verbatim because it efficiently conveys the crucial information without the use of extraneous or filler words. It also removes grammatical errors and colloquialisms to make the content relatively simple to understand. This method can be used in academic, medical and focus groups, without leaving out workshops, depending on the situation.

IX. CONCLUSION

Audio transcription is used in a variety of sectors to create content and collect data for analysis, and it aids businesses, media professionals, and researchers by transforming audio recordings into text format. In the mass media, especially broadcasting, STT transcription technologies are utilized to expand content offerings and reach a larger audience. This technique saves time and allows editing teams to concentrate on their strengths. In digital asset management, artificial intelligence and machine learning are also used to increase organizational efficiency, reduce search time, and save cost.To remain relevant and competitive, businesses and the media industries are embracing digital change. Transcription is being revolutionized by new technologies like artificial intelligence, which is utilized for a range of purposes such as generating written records for media, functioning as legal proceedings, providing accessible resources to students, and recording speeches or presentations. Correct transcription is essential for effective communication in the media, legal, and higher education sectors in order to provide accurate information to people, professionals, and students in general. The results of these findings suggest that combining manual and automatic transcription yields high quality results for the speech-to-text process.

X. REFERENCES

Chan, W., Jaitly, N., Le, Q. &Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Proceedings in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4960-4964. https://doi.org/10.1109/ICASSP.2016.7472621

Chen, Y., Wang, W., Chen, I., & Wang, C. (2020). Data techniques for online end-to-end speech recognition. *ArXiv, abs/2001.09221,* pp. 1–5.

Dahl, G., Yu, D., Deng, l. &Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition;audio, speech, and language processing, IEEE Transactions, 20(1), 30-42. https://doi.org/10.1109/TASL.2011.2134090

Frank E.R., Catherine B., Kaitlyn E. & Daniel G. (2016). An update on automatic transcription vs. manual transcription. *In proceedings of International Conference on Cognitive Modeling (ICCM)*.

Hinton, G., Deng, l. Yu, D., Dahl, G. & Mohamed, A. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine, IEEE, 29*(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597

Hsiao, R., Can, D., Ng, T., Travadi, R., &Ghoshal, A. (2020). Online automatic speech recognition with listen, attend and spell model. *IEEE Signal Processing Letters, 27*, 1889-1893. https://doi.org/10.1109/LSP.2020.3031480

Iosifov, I., Iosifova, O. &Sokolov, V. (2020). Sentence segmentation from unformatted text using language modeling and sequence labeling approaches. 335-337. https://doi.org/10.1109/PICST51311.2020.9468084

Iosifova, O., Iosifov, I., Rolik, O. &Sokolov, V. (2020). Techniques comparison for natural language processing. *In proceedings of International Workshop on Modern Machine Learning Technologies and Data Science, 2631*(1), 57–67. https://doi.org/10.5281/zenodo.3895814

Kim, S., Seltzer, M. L., Li, J., & Zhao, R. (2018). Improved training for online end-to-end speech recognition systems. 2913–2917. https://doi.org/10.21437/interspeech.2018-2517

McDermott, E. (2018). A deep generative acoustic model for compositional automatic speech recognition. In *Proceedings of Neural Information Processing Systems (NeurIPS) Workshop: Interpretability and Robustness in Audio, Speech, and Language*, 1–17.

Olena L., Levgen L., Volodymyr S., Oleh R. & Igor S. (2021). Analysis of automatic speech recognition methods. *Cybersecurity Providing in Information and Telecommunication Systems*, 252-257

Rista, A. &Kadriu, A. (2020). Automatic speech recognition: a comprehensive survey. *SEEU Review*. 15(2), 86-112. https://doi.org/10.2478/seeur-2020-0019

Romanovskyi, O., Iosifov, I., Iosifova, O., Sokolov, V., Kipchuk, F. &Sukaylo, I. (2020). Automated pipeline for training dataset creation from unlabeled audios for automatic speech recognition. Data Engineering and Communications Technologies, 83, 25–36. https://doi.org/10.1007/978-3-030-80472-5_3

Stuckey L.H. (2014). The first step in data analysis: transcribing and managing qualitative research data. *Journal of Social Health and Diabetes*, *2*(1), 6-8.

Tanaka, T., Masumura, R., Moriya, T., Oba, T., &Aono, Y. (2019). A joint end-to-end and DNN-HMM hybrid automatic speech recognition system with transferring sharable knowledge. *Interspeech*, 2210–2214. https://doi.org/10.21437/interspeech.2019-2263

Wang D., Wang X.&Lv S. (2019). An overview of end-to-end automatic speech recognition.*Symmetry*, 11(8) 1–26. https://doi.org/10.3390/sym11081018

Wang, X., Yang, J., Li, R., Sadhu, S., &Hermansky, H. (2019). Exploring Methods for the Automatic Detection of Errors in Manual Transcription. *Interspeech*.

Weiner J., Frankenberg C., Telaar D., Wendelstein B., Schroder J., & Schultz T. (2016). Towards automatic transcription of ILSE – an interdisciplinary longitudinal study of adult development and aging. *In Proceedings of International Conference on Language Resources and Evaluation (LREC'16)*.

Yu W, Zeiler S &Kolossa D. (2022). Reliability-based large-vocabulary audio-visual speech recognition. *Sensors (Basel)*. 22(15), 5501-5512. https://doi.org/10.3390/s22155501

Zhang, S., Gao, Z., Luo, H., Lei, M., Gao, J., Yan, Z. &Xie, L. (2020). Streaming chunk-aware multihead attention for online end-to-end speech recognition. 2142-2146. https://doi.org/10.21437/Interspeech.2020-1972