Journal of Artificial Intelligence Practice (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2024.070310
ISSN 2371-8412 Vol. 7 Num. 3

# *Problems in the Optimization Work of Speech-Text Auto-Recognition and Relevant Possible Solutions*

**Xiwen Qin**

*University of Shanghai for Science and Technology, Shanghai, 200093, China*

***Keywords:*** ASR; audio annotation; speech-to-text; psycholinguistics; Artificial Neural Network

***Abstract:*** This thesis explores the problems occurred in the annotation work of language audios and possible solutions after analysis and judgement. After Part One Introduction of the industry and Part Two clarification of research methods, Part Three delves into various actual issues encountered in the ASR optimization work and their influence. It utilizes and analyzes real-world investigation data to pinpoint these issues and their impact on the effectiveness of ASR. Part Four examines the solutions of the possible problems proposed, one of which is Cohen's Kappa metrics being successfully applied in an experiment. Part Five is the study of the real application of the methods. This section first explores the generation and optimization problems from a psycho-linguistic perspective before finding out various methods and plans that could enhance the accuracy and efficiency of the annotation process. The goal of this thesis is to provide readers with a comprehensive understanding of both the current situation and further direction of audio annotation. By analyzing current challenges and exploring potential advancements, this thesis is dedicated to provide readers with a thorough understanding of the current state of audio annotation and its future trajectory. It contributes valuable insights that can pave the way for more robust and efficient audio annotation practices, ultimately leading to improved performance in ASR systems.

## 1. Introduction

### 1.1. Brief clarification about the (1) Data Annotation Industry and (2) Audio Annotation

The **data annotation** industry focuses on labeling and structuring raw data to make it usable for machine learning algorithms. We can see the AI models, such as Google Gemini, and ChatGPT. They are all products trained by annotators through various kinds of data. We have text annotation, image annotation, and audio annotation. [1] The relevant images shown on the models when a specific order is given are the results classified by annotators and sent to the models for machine learning. The audios generated are the outcomes of labeling, interception, and text checking.

**Audio annotation** is a specific type of data annotation that deals with labeling audio files to improve the algorithm of ASR (Automated Speech Recognition). This can involve various tasks like: *Speech transcription, Speaker identification, Audio quality evaluation, Sentiment analysis.* [2]

### 1.2. Current market situation of Data Annotation

The data annotation industry is experiencing a period of significant growth, driven by several factors:

**Surging Demand for AI**: The increasing adoption of Artificial Intelligence (AI) across various sectors like healthcare, automotive, and retail is fueling the need for high-quality training data. This data requires human annotation to be effective for AI models.

**Growth of Big Data**: The exponential growth of data generation necessitates efficient methods for organizing and labeling it. Data annotation plays a crucial role in making this vast data usable for machine learning applications.

**Advancement in Technology**: The development of new data annotation tools and platforms is streamlining the process, making it faster and more cost-effective. This is further driving market expansion.

**High Growth Rate**: Market research suggests the global data annotation market size was valued at around $0.8 billion in 2022 and is projected to reach $3.6 billion by 2027, with a compound annual growth rate (CAGR) exceeding 33%.

**Increased Demand Across Industries**: The demand for data annotation services is rising across various industries, including:

*Autonomous Vehicles*: Training self-driving cars requires annotating vast amounts of video data for object recognition and scene understanding.

*Healthcare*: Annotating medical images like X-rays and CT scans helps develop AI for disease diagnosis and treatment recommendations.

*Customer Service Chatbots*: Annotating customer service interactions allows chatbots to understand user intent and provide better support.

*Media and Entertainment*: Speech-to-text conversion and content moderation for platforms like Youtube rely heavily on audio annotation.

## 2. Research Methods

### 2.1. Research Design

The type of research design employed here is the mix of qualitative analysis and quantitative analysis.

Qualitative analysis is used for the exploration of the problems using survey questionnaires to observe the difficulties happened on annotators. It focuses on non-numerical data to understand meanings, experiences, and phenomena. Therefore qualitative analysis method helps to understand the problems and identify the solutions of audio annotation.

Quantitative analysis is used further for the experiment about disagreement between annotators on the same inconsistency problem. After the data was collected, applies Cohen's Kappa metrics. This research relies on measurable data to objectively assess situations and minimize subjectivity.

### 2.2. Participants and Data Collection Methods

The participants for the study is purposive selected. The participants are all annotators involving or once involved in speech annotation projects. A questionnaire survey is made for data collection when it's concerned about the research of ASR optimization problems. Contents are difficulties encountered when working for ASR projects and relevant advice. There are several choices in part of the questionnaire which one could choose once. Data is collected based on result of form filling.

When it comes to method testing, the data is collected from real-world experiment. Details are

shown in the sixth point of Part 4.

### 2.3. Data Analysis Methods

The former data analysis is based on the understanding and observation according to the feedback data demonstrated in the survey results given from the participated candidates. The latter is based on the application of formula on the experimental data in order to observe the resulting effects by comparing between the actual value and preset interpretation standards.

## 3. Problems Concluded and Possible Influences

The problems are explored from the results of the collected data concluded in Figure 1.

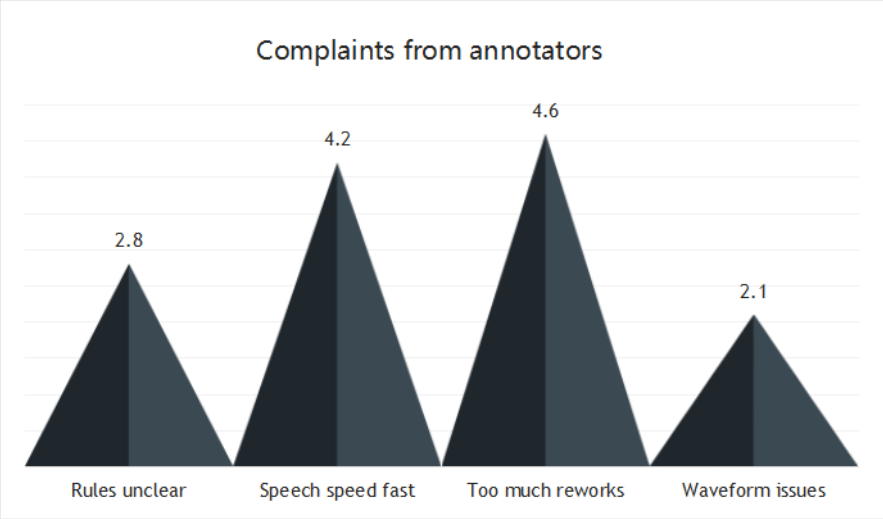


Figure 1: Complaints from Annotators

### 3.1. Inaccuracy between Speech and Text

*Accents and Dialects*: Speakers with strong accents or regional dialects can be challenging for annotators to transcribe accurately. The written text might not perfectly capture the nuances of spoken language.

*Background Noise*: Audio recordings with background noise (e.g., traffic, music) can make it difficult to discern speech clearly, leading to errors in transcription.

*Overlapping Speech*: Conversations with multiple speakers talking simultaneously can be very difficult to annotate accurately, as it's challenging to distinguish individual voices.

Besides, some sounds were not heard by the writer due to their high speed and low volume. Some of them can be heard, but due to the writer's insufficient listening and language skills. Some modal particles are mixed into words, resulting in mishearing, word errors, more words and less words, and typos.

In phonetics we have coarticulation, which means two sounds share with each other to arrive at an effect of mixing, for example lamb, bomb, and map. It is similar to the concept of phonology called assimilation, which is a key area forcing annotators to mishear and misinterpret the audios, causing inconsistency between speech and text. Some of the annotators may use phonetic symbols in replacement of the words that are not activated in their system. **Assimilation** in phonology refers to a sound change process where a sound becomes more similar to a neighboring sound within a word or across words in connected speech. This essentially means one sound is influenced by

another, making pronunciation easier and smoother. Assimilation can be divided into two kinds. (1) Progressive Assimilation: The first sound influences the following sound. (e.g., "pin cushion" - the /n/ at the end of "pin" becomes more like the /k/ sound at the beginning of "cushion") When the words are spoken quickly and in a blurred sound environment, it's easy for annotators, especially those non-native speakers of the target language, to mark as "unclear" segment. (2) Regressive Assimilation: The second sound influences the first sound. (e.g., "hot chocolate" - the /t/ at the end of "hot" becomes more like the /ʃ/ sound at the beginning of "chocolate") Also there's a type of assimilation called **coalescent** assimilation which means that a speaker pronounces abnormally according to his dialect or language characteristics, an example is "Do you smoke?" being pronounced as /dʒu sməʊk/. This example can be easily perceived and transcribed as "Jew smoke", instead of the correct version, causing misunderstanding and confusion of annotators.

### 3.2. Tagging Errors

*Incomplete or Missing Tags*: Annotators might accidentally miss certain sounds or events within the audio, resulting in incomplete annotations.

*Wrong Tagging*: Some speaker roles are difficult to distinguish. Sometimes one person's voice is of both high and low frequency during the whole speech process, while another person's voice is very similar when speaking to him, making it difficult to distinguish. Male's voices are sometimes very identical.

### 3.3. Interjection Deviation

*Filler Words and Disfluencies*: Filler words like "um," "uh," or stutters can be challenging to handle during annotation. Some annotators might choose to include them, while others might omit them, leading to inconsistencies.

*Non-verbal Cues*: Laughter, coughs, sighs, and other non-verbal cues present within the audio don't always translate well into text annotations. These nuances require additional annotation methods beyond simple transcription.

### 3.4. Audio Interception Problems

Some waveform isn't shown with full detail, and the maximization of the waveform sight is limited on most of the annotation platforms. The annotation tool's amplification limit for audio ripples is insufficient, resulting in interceptions such as noise, cut-off sounds, and less interception.

The challenges above can significantly impact the quality of training data for AI models. Inaccurate annotations can lead to poorly performing speech recognition systems or AI applications that misinterpret audio data. However, one important aspect that cannot be ignored is the simplicity of annotation standards. Because of the high complexity of some annotation tasks, the manager may choose to set complex and incomprehensible standards for annotators, and this can be especially hard when junior annotators first get to know some annotation tools. Managers can choose to conclude and simplify the standards, at the same time meet the client's demand.

### 3.5. Possible Influential Outcomes

(1) The machine learning algorithm may not be improved. Sometimes the acceptance party has very low accuracy requirements for recognizing text modifications, which may cause some recognition problems in the actual speech-to-text output.

(2) Inaccurate speaker recognition of the output text will lead to confusion in understanding.

(3) Problems such as missing words can cause incomplete or ambiguous semantics, therefore affecting understanding.

## 4. Possible Solutions

### 4.1. Recruiting Outstanding Talents

*Language Proficiency*: Instead of just B2, consider a tiered system based on specific needs. For complex tasks, aim for C1 or even native fluency.

*Dialect Expertise*: Specify the required dialects beyond "native speaker level." *Skills-based Testing*: Design targeted tests that assess not just language skills but also audio annotation abilities like transcription accuracy, sound identification, and adherence to guidelines.

### 4.2. Investing More in Training

At present, the training given to freelancers usually consists of meeting videos, and some document specifications, and they are asked to practice it. Many problems will arise during this process, which is a waste of time and unnecessary quality inspection costs.

*Interactive Training*: It's beneficial to move beyond passive learning with video lectures, and develop interactive modules, quizzes, and simulations that actively engage freelancers.

*Mentorship Programs*: You can pair experienced annotators with new hires for personalized guidance and feedback.

*Performance-based Incentives*: You can offer bonuses or rewards tied to task completion speed and quality metrics to motivate continuous improvement.

*Remote Collaboration Tools*: You can facilitate online workshops and Q&A sessions for geographically dispersed freelancers. [3]

### 4.3. Boosting Management Strategies

It's suggested to develop a comprehensive annotation style guide with clear examples and decision trees for ambiguous situations. Also, you can invest in a dedicated QA team with strong audio annotation expertise and analytical skills for consistent review and feedback. Furthermore, you can implement periodic assessments for both annotators and quality inspectors to identify areas for improvement and ensure adherence to standards.

### 4.4. Building Powerful Annotation Tools

*Advanced Audio Players*: Features like variable playback speed, audio looping, and spectrogram visualization can be integrated for detailed analysis of sound waveforms.

*Collaborative Annotation Tools*: Real-time collaboration between annotators on the same audio clip for efficient problem-solving can be enabled.

*Machine Learning Assistance*: AI-powered tools for automatic transcription, speaker diarization, and error detection can be utilized to streamline the annotation workflow.

### 4.5. Selecting a Playing Device

There is no need for headphones or external amplifiers, therefore, the most suitable speakers which won't emphasize high frequencies or ignore low frequencies, and can filter noise, can be chosen. You can invest in professional studio monitors with a flat frequency response to ensure an accurate representation of the audio spectrum, and implement tools for noise cancellation or

background noise suppression to enhance audio clarity during annotation.

### 4.6. Solving with Complex Inconsistencies

*Escalation Process:* It's suggested to establish a clear escalation process for situations where multiple annotators disagree and a quality inspector cannot provide a definitive resolution.

Inter-Annotator Agreement (IAA) Metrics: It's advised to use established metrics like Cohen's Kappa to quantify agreement between annotators and pinpoint areas requiring further investigation or consensus building. Cohen's Kappa statistic measures how often two raters agree with each other after accounting for the likelihood they'd agree by chance, and it's experimented in reality and proved as a valuable tool in this situation where annotators agree and disagree on identifying speakers. [4]

When it happens that there are two speakers that an annotator cannot differentiate in an audio clip, there are possibilities that some annotators choose Speaker 1, and others choose Speaker 2.

An experiment is made aiming to prove the efficiency of Cohen's Kappa metrics application, and find out the reasons and solutions dealing with disagreement. Two annotators were invited to annotate the same audio clip in English which contains 64 ambiguous segments that are difficult to identify the speaker roles. Their annotation results of the ambiguous audio segments are the main data source for experiment. Results are shown in Table 1:

Table 1: Annotation Results of Ambiguous Audio Segments

| | A1- S1 | A1- S2 | |
|---|---|---|---|
| A2- S1 | 15 | 1 | 16 |
| A2- S2 | 38 | 10 | 48 |
| | 53 | 11 | 64 |
| | | | |

A1= Annotator 1, A2= Annotator 2

S1= Speaker 1, S2= Speaker 2

The formula of Cohen's Kappa is in Figure 2:

$$\frac{Po - Pe}{1 - Pe}$$

Figure 2: Cohen's K formula

The calculation process of Cohen's Kappa application is as follows:

$$Po= (TP+TN)/N=(15+10)/64\approx 0.39$$

$$Pe=P_1+P_2=(TP+FN)*(TP+FP)/N^2+(TN+FN)*(TN+FP)/N^2=(15+38)*(15+1)/64^2+(10+38)*(10+1)/64^2\approx 0.21+0.13=0.34$$

Cohen's K= (0.39-0.34)/(1-0.34)≈0.08 (**slight agreement)**.

From the result of the experiment, it can be concluded that the two annotators don't agree with each other much when doing the ambiguous audio segments. **Reasons** are found as follows:

(1) Audio Quality Issues:

Low quality audio: Background noise, static, or muffled voices can make it difficult to distinguish between speakers, especially if their voices have similar characteristics.

Poor recording conditions: Improper microphone placement or inadequate recording equipment can distort voices and hinder accurate identification.

(2) Speaker Characteristics:

Similar vocal qualities: If speakers have similar accents, pitches, or speaking styles, it can be challenging to differentiate them, even for trained annotators.

Background conversations: Overlapping speech or conversations in the background can create confusion and lead to misidentification of the primary speaker.

(3) Annotator Factors:

Inconsistent training: If annotators haven't received consistent training on speaker identification guidelines, they might use different criteria for assigning speaker labels.

Fatigue or inattention: Annotating audio for extended periods can lead to fatigue and decreased focus, potentially resulting in errors and inconsistencies.

(4) Task Complexity:

Lengthy or complex audio clips: Annotating long or intricate audio files with frequent speaker changes can be overwhelming and prone to mistakes.

Unclear instructions: Ambiguous or incomplete labeling guidelines can leave room for interpretation and disagreement among annotators.

**Solutions** created according to the reasons concluded from the experiment results are as follows:

- Ensure high-quality audio recordings.
- Provide clear and comprehensive training materials for speaker identification.
- Regularly review and update annotation guidelines.
- Consider using multiple annotators for challenging audio clips.
- Implement quality control measures to identify and address inconsistencies.
- Use better amplifiers.

Therefore, tracking Kappa scores over time can help assess the effectiveness of training interventions or improvements in audio quality. Kappa can be used as part of a quality control process to flag audio clips with high levels of disagreement for further review or adjudication by a senior reviewer. By identifying areas of low agreement through the convenient Kappa analysis, one can find exact problems in time, analyse the reasons behind, and tailor retraining programs to address specific speaker identification challenges faced by annotators.

## 5. The Psycho-linguistic View and Real Application of Solutions

Before discussing how to apply the solutions to real work, it’s necessary to first take a look at the challenges during the ASR generation and optimization process in a neurolinguistic perspective. [5]

### 5.1. Limitations of Attention-based ANNs in Speech-to-Text Recognition Compared to Human Attention

The speech recognition procedure uses lots of different models, one of the most largely-used is the Artificial Neural Network (ANN) model. ANNs are interconnected group of nodes, inspired by a simplification of neurons in a brain. They have been an applicable computational tool because of their effectiveness in natural language processing. [6] It can be applied to check processing, speech-to-text transcription, oil exploration data analysis, weather prediction, facial recognition and so on. The connections are the most important aspect of the model. The effectiveness of connections can be modified by learning rules although the network’s structure is fixed. [7]

While both humans and Attention-based Artificial Neural Networks (ANNs) utilize a form of "attention" mechanism for speech-to-text tasks, there are significant limitations in ANNs compared to human auditory attention.

(1) Human Auditory Attention:

Humans can dynamically shift their attention based on context, speaker cues, and real-time understanding of the conversation. They can focus on specific sounds (e.g., speaker change) while still processing the overall message. Humans leverage their existing knowledge and understanding of the world to fill in gaps, disambiguate homophones, and interpret the speaker's intent. Humans can integrate situational context to enhance speech comprehension and inform their attention.

(2) Limitations and consequences of Attention-based ANNs:

The attention mechanism in ANNs is typically pre-defined during training. The model focuses on specific features based on the training data, and it lacks the flexibility to adapt its attention dynamically like humans. ANNs primarily rely on the immediate audio input for processing. They struggle to incorporate broader context or prior knowledge into their attention, which can lead to errors in ambiguous situations. ANNs trained on biased datasets can inherit those biases in their attention patterns. ANN is unable to fully consider context can lead to mistakes when differentiating homophones that rely on surrounding words for meaning. ANNs struggle to filter out background noise as effectively as humans, leading to errors in transcription. ANN lack of broader context awareness can make it difficult for ANNs to understand subtleties like sarcasm or speaker intent. [8]

## 5.2. Human Auditory Issues

Building on the previous analysis, here's a deeper dive into common annotator problems in speech annotation projects, along with some additional issues, explored through a neurolinguistic lens.

(1) Speech-Text Inaccuracy:

Stuttering, restarts, and self-corrections pose challenges for accurate transcription. The brain might anticipate a correction and miss the initial disfluency. Furthermore, each of the brain hemispheres can be divided into 4 lobes: frontal, temporal, parietal and occupital lobe. The auditory cortex is the part of the temporal lobe that processes auditory information in humans and many other vertebrates. When multiple speakers talk simultaneously, the auditory cortex struggles to segregate individual voices, leading to transcription errors.

(2) Speaker Role Tagging Errors:

Emotional states can alter vocal characteristics, making speaker identification more difficult. The amygdala, which processes emotions, might heighten the emotional content and distract from speaker cues. Distinguishing between speakers with similar voices relies on subtle acoustic cues processed in the auditory cortex. Age, accent, and even emotional state can influence these cues, making speaker identification challenging. The brain uses prosodic or suprasegmental features like pitch and intonation to identify speaker changes. However, subtle variations or fast turn-taking can lead to errors in assigning speaker roles.

(3) Ignorance of Filler Words

The brain prioritizes understanding the core meaning of the speech. Filler words, lacking semantic content, might be unconsciously filtered out during annotation. Like reading, the brain focuses mainly on the notional words, instead of grammatical words. Hearing is the same, though the filler words are perceived by ears, the Wernick’s area usually ignores and focuses on content words for more convenient understanding. Broca’s and Wernick’s area are crucial for speech production and perception. Fatigue or stress can affect their function, leading to errors in processing of speech during annotation tasks.

(4) Neurolinguistic consideration: Mirror Neuron System (MNS)

A neuron fires when the organism observes the same action performed by another. The neuron “mirrors” others’ behavior just like its acting itself. They are situated in supplementary motor area

and medial temporal cortex. Empathy and compassion can force an hyperactive MNS. When an annotator's Mirror Neuron System is over-activated because of sympathy, empathy or common interest after comprehension of audio clips, it could lead to annotators getting "caught up" in the speaker's emotions and losing focus on the task.

### 5.3. How to Apply the Given Solutions in Reality?

(1) Do a pre-training test

Whether it is the native language or a foreign language, the candidate must take a language test, preferably a listening test. Those who have reached B2 or above can participate in the project.

The test can evaluate the candidate's syntax and semantics level, and train their patience and endurance since audio annotation is a repetitive task that might seem monotonous for annotators. Test question types can be set to find errors, fill the gaps etc. If the project involves a specific domain (e.g., medical), include domain-specific vocabulary and terminology in the listening test.

(2) Long-term training

The first training should last at least two days. A new platform can be created to allow the independent development of a simulation system, conduct standardized teaching and simulation drills, identify whether employees' answers are correct, evaluate and optimize answers, and point out errors. You can incorporate peer review sessions where trainees can learn from each other's work and improve their annotation skills. Also, paid training gives employees motivation. After training, a test can also be set.

An example of a comprehensive training **plan** for the new-coming annotators goes as follows:

**Speech Annotation Training Plan (3 Days)**

Target Audience: Individuals who have passed the language proficiency assessment.

Training Goal: Equip trainees with the necessary skills and knowledge to excel as speech annotators.

Training Structure:

*Day 1: Introduction & Core Concepts (5 hours)*

Welcome & Introductions (1 hour): Introduce trainers, trainees, and the overall training program.

Project Overview (1.5 hours): Explain the purpose of the speech annotation project, target audience, and the role of annotators.

Speech Annotation Fundamentals (2.5 hours): Cover core concepts like annotation tools, file formats, different types of speech annotation (e.g., transcription, speaker tagging).

*Day 2: Hands-on Practice & Simulation Drills (5.5 hours)*

Standardized Teaching & Drills (4 hours): Deliver standardized instruction on specific annotation tasks using the simulation system. Trainees will complete practice drills with varying difficulty levels.

Automated Feedback & Error Analysis (1.5 hours): Utilize automated scoring to provide trainees with immediate feedback on their performance. Analyze common errors and discuss strategies for improvement.

*Day 3: Advanced Techniques & Best Practices (4.5 hours)*

Advanced Annotation Techniques (2 hours): Introduce advanced features of the annotation tools and techniques for handling complex audio (e.g., overlapping speech, background noise).

Quality Control & Error Prevention (2.5 hours): Discuss common pitfalls in speech annotation and strategies for ensuring high-quality work.

(3) Annotation Tool Construction

The procedure can include providing standard playback controls (play, pause, rewind, fast-forward) along with adjustable playback speed for granular control, allowing users to type

annotations directly in sync with the audio, with timestamps automatically generated, allowing users to set markers at specific points in the audio file for quick referencing and segmentation, integrating search functionality to locate specific words or phrases within the transcribed text, providing options for easy navigation through lengthy audio files, enabling real-time or asynchronous collaboration for multiple annotators working on the same project. Moreover, for the waveform representation, a scrollable waveform representation of the audio file can be displayed. For better visualization, a spectrogram can be applied on the tool, take the tool "praat" as an example.

Praat is a free software designed for speech analysis, with a powerful built-in annotation tool. Analyzers can easily work on the audios with the demonstration of spectrogram, intensity, and pitch perceived automatically, as shown by Figure 3.

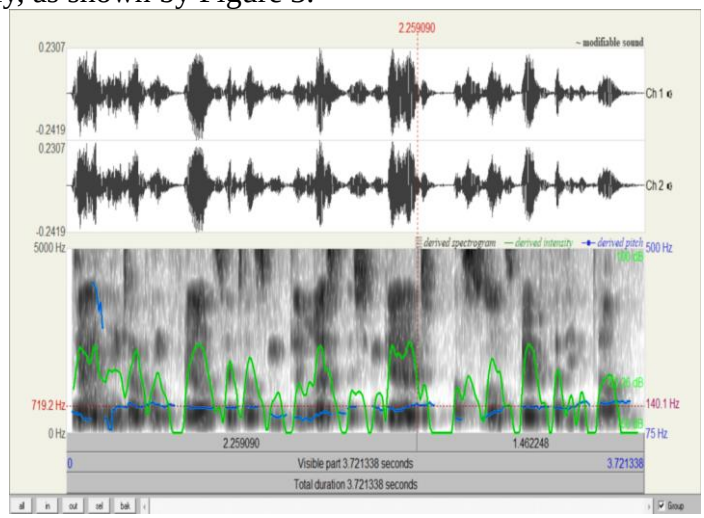


Figure 3: General Demonstration of Praat

The audio analyzed on the picture is recorded in male's voice, so the pitch ranges from about 105hz to 140hz. If the annotation projects are majors in those kind of audios, a low-frequency audio equipment can be applied for annotators focusing on every sound, as some low voice segments are hard to identify and transcribe. Also, the timbre of male's voice isn't as varied as female's voice, that's why annotators find hard to differentiate the speaker roles.

The trend of intensity tells annotators which part should be noticed. In this picture, the intensity is unstable because of the environment and equipment of recording. Low intensity parts combined with low frequencies can be ignored. Moreover, if the trend goes very high meaning high intensity, annotators can be aware of health protection.

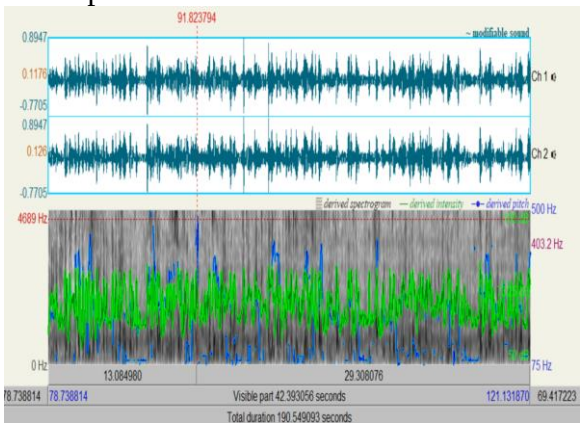


Figure 4: Concrete Demonstration of Audio Clips

The audio in Figure 4 is recorded in female's voice. The intensity is stabler than that of the first picture because this audio is recorded by high-quality devices with low noise and strong sound insulation measures. However, we can see the pitch varies to a great extent because there's a background music. When talking about female's voice, the pitch shown on the picture varies from 190hz to 400hz. Female voice is easier for annotators to identify speakers as the timbre of sound is various. Women speak with varying tones and full sentiments in daily life, while men speak with a flat tone. The former is more noticed by annotators when doing transcription jobs.

Praat enjoys a more detailed view on seconds with a preciseness of six decimal places after the number of second, compared with the common annotation tools for example Figure 5.

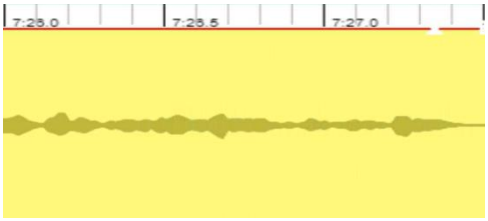


Figure 5: Time Decimals in Another Tool

The maxim of visualization is 0.000045 seconds, which suits detailed annotation, as in Figure 6.

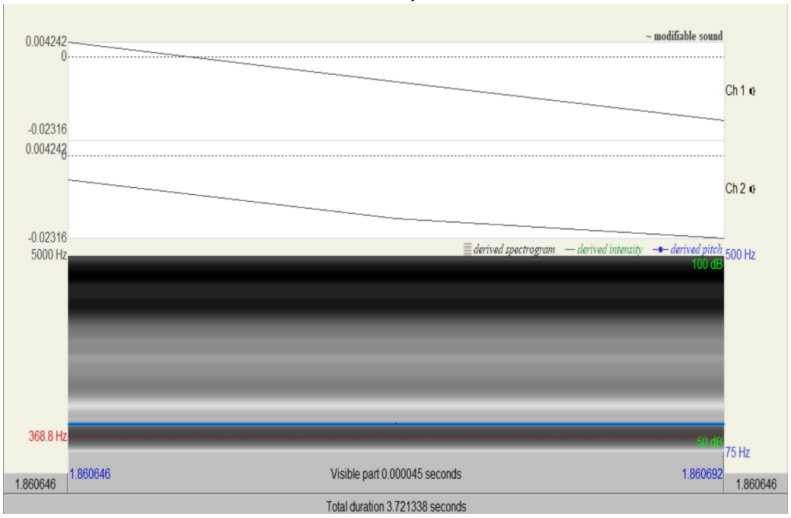


Figure 6: Praat's Maxim of Visualization

(4) Utilization of Acoustic Equipment

If there is strong noise, it will cause damage to the annotator's psychological states and ears. Low-frequency headphones may jar the ears if worn for a long time. High-frequency headphones may block low-frequency sounds. External amplifiers on the computer usually ignore low frequencies. Look for amplifiers or acoustic equipment with a flat frequency response that accurately reproduces all audible frequencies (20 Hz - 20 kHz) without undue emphasis on highs or lows, to avoid altering the audio signal and potentially distorting speech. Consider options with adjustable gain controls to fine-tune the volume without sacrificing audio quality.

(5) Cognitive Function Optimization

Schedule regular breaks to combat fatigue and ensure a comfortable workspace to optimize cognitive function. Mental and auditory fatigue can significantly impact concentration and accuracy in speech annotation. High attention in annotation tasks are essential. Scheduling short breaks every 2 hours allows annotators to rest their eyes, ears, and minds, returning to their tasks ready to focus.

(6) Dealing with inconsistencies using Cohen's Kappa

When inconsistencies occurred, managers can arrange two or more annotators to do the same

task, gather results, and apply Cohen’s Kappa metrics to find out the correct way of annotation as proved by an experiment before. After that, a senior reviewer can establish a process for adjudication to determine the most accurate annotation in case of disagreement. Managers can refine the project's style guide to clarify any ambiguities or address areas where interpretations differ.

## 6. Conclusion

This thesis sheds light on the current shortcomings of audio annotation in language processing, specifically for Automatic Speech Recognition (ASR) systems. By analyzing real-world data, it identifies key challenges hindering ASR optimization. It then explores potential solutions, including the successful application of Cohen's Kappa metrics for measuring annotation agreement. Furthermore, the thesis delves into psycho-linguistic perspectives and tools like Praat to enhance annotation accuracy and efficiency. Ultimately, this research aims to equip readers with an understanding of the current state and future directions of audio annotation. The thesis paves the way for more robust and efficient annotation practices, leading to significant improvements in ASR performance.

## References


*[1] Text annotation. Papers with Code. (n.d.). https://paperswithcode.com/task/text-annotation*

*[2] What is audio annotation, what are the applications and benefits. clickworker.com. (2023, January 16). https://www.clickworker.com/ai-glossary/audio-annotation/*

*[3] Four key metrics for ensuring data annotation accuracy | telus international. (n.d.). https://www.telusinternational.com/insights/ai-data/article/data-annotation-metrics*

*[4] AI, S. (2021, December 15). Inter-annotator agreement: An introduction to cohen’s kappa statistic. Medium. https://surge-ai.medium.com/inter-annotator-agreement-an-introduction-to-cohens-kappa-statistic-dcc15ffa5ac4*

*[5] Ahlsén, E. (2006). Introduction to neurolinguistics. John Benjamins.*

*[6] Sussex Publishers. (n.d.). How the brain’s mirror neurons affect empathy. Psychology Today. https://www.psychologytoday. com/intl/blog/emotional-freedom/202206/how-the-brains-mirror-neurons-affect-empathy*

*[7] Author links open overlay panelEdmondo Trentin a, a, b, AbstractIn spite of the advances accomplished throughout the last decades, Bridle, J. S., Chen, W. Y., Chung, Y. J., Elman, J. L., Franco, H., Jang, C. S., Bell, A. J., Bengio, Y., Bourlard, H., Cerf, P. L., Chang, P. C., Cosi, P., Cybenko, G., Davis, S. B., Mori, R. D., … Hertz, J. (2001a, February 27). A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing. https://www.sciencedirect. com/science/article/abs/pii/S0925231200003088*

*[8] Zacarias-Morales, N., Pancardo, P., Hernández-Nolasco, J. A., & Garcia-Constantino, M. (2021, January 28). Attention-inspired artificial neural networks for Speech Processing: A Systematic Review. MDPI. https://www.mdpi.com/2073-8994/13/2/214*