Prosodic Question Detection: Recognizing Spoken Questions through Intonation

Jonathan Peters
jmp22@sfu.ca
301546055
Simon Fraser Univsersity
Burnaby, British Columbia, Canada

Jimmy Chui
kyc46@sfu.ca
301569797
Simon Fraser Univsersity
Burnaby, British Columbia, Canada

Lucy Zhou
lza168@sfu.ca
301544374
Simon Fraser Univsersity
Burnaby, British Columbia, Canada

ABSTRACT

Spoken language recognition systems have been developed with a heavy emphasis on extracting the lexical features (semantic content) of a sentence while ignoring the prosodic features (intonation). In other words, AI can tell what is being said, but not how it is being said. We propose a system for improving accuracy of Air Traffic Control (ATC) radio transcriptions by detecting questions and non-questions with machine learning. We first extract prosodic features such as pitch and energy and apply regression and Support Vector Machine (SVM) classification methods. This could benefit aviation-related communications by clarifying interrogative communications between the ATC tower and pilots while avoiding the need to learn specific jargon. We also foresee useful applications in generating more expressive transcriptions for investigations or training.

KEYWORDS

Question Detection, Prosody, Affective Computing, Machine Learning, Dataset, Annotation

ACM Reference Format:
Jonathan Peters, Jimmy Chui, and Lucy Zhou. 2025. Prosodic Question Detection: Recognizing Spoken Questions through Intonation. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Currently, speech recognition systems suffer to respond accurately and sympathetically to questions asked in various tones. When asked a rhetorical question, such as "why is it ALWAYS raining?" Amazon's Alexa assistant typically tries to answer factually, where a human would understand the rhetorical tone. This apparent insensitivity to tone is a result of a heavy focus on the development of speech recognition and transcription systems with respect to lexical content. Most speech-recognition datasets are derived from text which is then transformed to synthesized speech [1]. Training of modern deep-learning speech recognition models, such as OpenAI's Whisper model, relies on 680,000 hours of labeled audio data [2]. Whisper relies on supervised training to predict the raw text of transcripts. Many other existing models rely on datasets of machine-generated data, and normalize training data in such a way that removes features of speech such as exclamation or question marks.

These AI models lose much of the context that is needed for immediate understanding of transcriptions by reviewers. In many languages, punctuation is needed to clearly and efficiently communicate ideas through writing [3]. Omitting punctuation or using it incorrectly can completely change the meaning of a sentence. Consider the important distinction between the statement "I am going to turn left" and the question "I am going to turn left?" Clear communication is needed in many fields of work to maintain safety, but perhaps aviation is the most important example of this. In 2024, an estimated 5 billion passengers were safely carried by airlines to destinations around the world [4]. The lives of passengers, pilots, and civilians depend on clear, unambiguous communication.

Whisper has been used to create an annotated dataset of ATC radio recordings with their transcription [5]. However, these transcriptions lack any information about intonation like punctuation, speaker role, or authority, all of which is important for communication. Intonational features are not determined by syntax, and it is difficult to predict intonation from the text of a sentence. For example, different pitch contours can be used on the same phrase to change the meaning completely. If an upwards pitch inflection is used on a sentence like "I can't park there," the listener may understand that the speaker was unaware that they had parked in an illegal spot. However, if the same sentence is said with a downwards inflection, it indicates that they already know this information. However, these rules are not consistent and different levels of authority and confidence may change the tone and the meaning of a question or statement [6].

Table 1: Example transcriptions of ATC audio samples [5]

ID	Transcription
0309	csa three charlie tango runway three one cleared for takeoff...
11616	sky travel one zero one zero confirm ready for departure...
10676	initial one zero zero sierra papa sierra thank you
9401	warsaw good morning lufthansa seven seven nine...
1895	thank you good bye five nine six

Inspired by the need for safer air travel and the fascinating way language works, we seek to develop a system that can accurately identify questions, non-questions, and/or various types of questions. We build upon the ATC dataset [5] by adding manual annotation to approximately 1000 audio samples. We then extract

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

prosodic features including pitch, energy, zero-crossing rate, and Mel-Frequency Cepstral Coefficients (MFCCs) to train a machine learning classifier to predict questions and non-questions. We also aim to better understand if the inconsistent nature of questioning tone allows for accurate prediction from prosodic features alone. Finally, we combine the prosodic and lexical features to determine if tone helps to improve classification accuracy.

2 APPROACH

By limiting our scope to Air Traffic Control (ATC), we faced some unique challenges while developing our system. The first problem we had to solve was data collection. ATC radio is available online in multiple datasets, but many do not have the required length or transcription quality to be used for our purposes. We settled on using an ATC dataset with a Train split of N=11,867 recorded clips with transcriptions from OpenAI's Whisper model. This dataset was created by compiling the ATCO2 Corpus (1-hour test subset) [7] and the UWB-ATCC Corpus [8]. The data was downloaded locally using a custom script that saved audio to the project folder and saved a CSV file containing all audio paths and transcriptions.

To reduce the size of this large dataset to a reasonable size for human annotation, we filtered the 11,867 recorded clips to N=1000 examples. We did this by removing one-or-two-word clips, and then we searched for keywords indicative of questions in aviation communications. See examples of keywords in Table 2. By using a regular expression to perform filtering, we flagged clips as either potential questions or potential statements. Examining our search results, we found a total of 420 clips labeled as suspected questions. To reach our target of 1000 examples, we randomly sampled 580 clips flagged as suspected statements. This gave us a collection of 1000 filtered training data with suspected labels.

Table 2: Question patterns and associated keywords

Regex Pattern	Keyword(s)
`\?`	Question mark
`what|where|...`	WH-words
`is|do|are|can|...`	Yes/No starters
`right|copy|confirm|...`	ATC confirmations
`say again`	Clarification
`request`	Request
`confirm`	Confirmation

We then manually annotated clips using 4 classes. We chose to label clips as either a non-question (nq), a yes-no question (yn), a wh-question (wh), or an implied question (imp) as identified by Hirschberg [6]. If we were unable to determine the type of clip for reasons such as lack of context or poor audio, we labeled it as unknown. Each clip was manually annotated using a custom GUI application to speed up our workflow and reduce typos or incorrect annotations (see Figure 6 in 7). Project members listened to clips and read transcriptions to select the appropriate label. All group members contributed to the annotation. Jimmy has extensive experience flying in VATSIM, an online flight simulator program that includes live interaction with human ATC. See Figure 1 for the distribution of classes.

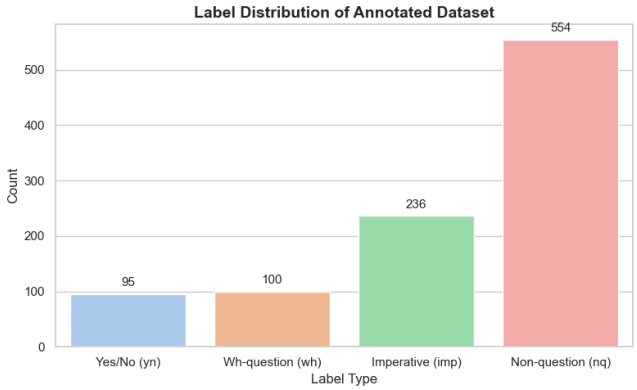

Figure 1: Distribution of Annotated Labels

Project members Lucy and Jonathan did not have extensive air traffic communication phraseology experience, and faced difficulty when annotating clips. Air traffic communication phraseology is globally standardized such that repetitive or non-essential phrases are omitted by pilots and air traffic controllers to streamline communication. This also explains why 50% of questions are implied questions. According to NAV CANADA, the organization responsible for managing Canadian airspace, proper use of phraseology allows fast, effective, and unambiguous communication between air traffic controllers and pilots [9]. Phraseology plays a crucial role in ensuring safety by reducing radio congestion and facilitating the clear and concise exchange of essential information. [9]

For example, in *Table 3* , the phrase "Can you see me on the radar?" and " you seem to have forgotten to turn on your transponder. Can you turn the transponder to..." were omitted because of standardized phraseology. However, the ATC replies with "radar contact" and "thanks for that" as the response to the hidden questions.

After generating the labels required for supervised learning, we then extracted the prosodic features to be used to train our classifier. Before we extracted audio features, we attempted to isolate the prosodic and lexical elements of speech. Chi and de Seyssel demonstrated how applying a low-pass filter to audio can remove frequency bands that contain lexical information in speech [1]. We set our low-pass cutoff frequency to 512Hz, in order to preserve the accurate fundamental frequency in a range of voices while attempting to eliminate many of the vowel and voiced/unvoiced consonants present up to 8000Hz. However, it is impossible to fully separate prosodic and lexical information from natural speech. After listening to audio filtered at 512Hz, we agreed that it was difficult to understand what was being said, but we could still identify tonal aspects such as speaking rate and pitch. We selected 4 prosodic features to be used, including estimated pitch (fundamental frequency, f0), energy/intensity contour, speaking rate (ZCR), and MFCC features. Our pipeline extracted the mean f0 and standard deviation, f0 end slope, mean energy and standard deviation, mean ZCR, and a collection of mean MFCC features. See Figure 2 for the visualized features. We used the Python Librosa library to perform feature extraction and visualization of waveforms and spectrograms [10]. The pYIN algorithm was used to perform f0 estimation. We

Table 3: Conversation between Pilot and ATC
Red: Skipped component of the conversation.
Blue: Reply of the skipped component.

Conversation	Speaker	Message
1	Pilot	Vancouver Terminal, Air Canada 321, two-thousand feet, direct ZN-DER on course. Can you see me on the radar?
	ATC	Air Canada 321, Vancouver Terminal, radar identified, climb flight level 220, continue on course.
2	Pilot	Vancouver Apron, Air Canada 321 at taxiway JA, ready to taxi.
	ATC	Air Canada 321, you seem to have forgotten to turn on your transponder. Can you turn the transponder to SQUAWK MODE CHARLIE?
	Pilot	Wilco, Air Canada 321
	ATC	Air Canada 321, thanks for that, now taxi via JA, J, L, cross runway 13, to holding point L6 for runway 08R.

were specifically interested in the f0 end slope, as it provides useful information about voice inflection, such as the common upward inflection at the end of a question.

Extraction of features for 1000 clips took 3 minutes and 55.9 seconds on a 2021 M1 MacBook Pro. We performed post-processing of the data by removing clips where f0 estimation had failed. This could be due to the poor audio quality of certain clips, muffled speaking from microphone placement, or radio transmission artifacts. After filtering unusable examples, we were left with 984 clips with extracted features, which we saved to a CSV file for easier loading and processing. A standard scaler was applied to these features in preparation for training. 7-fold cross-validation was used to avoid incorrect model evaluation and overfitting. To perform binary classification, we used a logistic regression model and an SVM model from the Python machine learning library sci-kit learn [11]. We tested the same two models on multi-type classification.

3 DATASET

We did not recruit participants to compile a dataset of spoken questions and statements, instead relying on real Air Traffic Controllers in the existing UWB-ATCC and ATCO2 corpus. Although little information exists about these datasets, all clips are in English and the majority of audio use male voices. Some speech is heavily accented. The ATCO2 corpus lists included towers from Sion, Bern, Zurich, Brisbane, Sydney, and others. Our work to perform manual annotation was limited to annotation from project members, removing the need for consent to be gathered from participants. More information and questions addressing risks of harm related to this dataset is available in the Appendix. Our dataset can be visualized by the extracted features using t-SNE dimensionality reduction (see Figure 3).

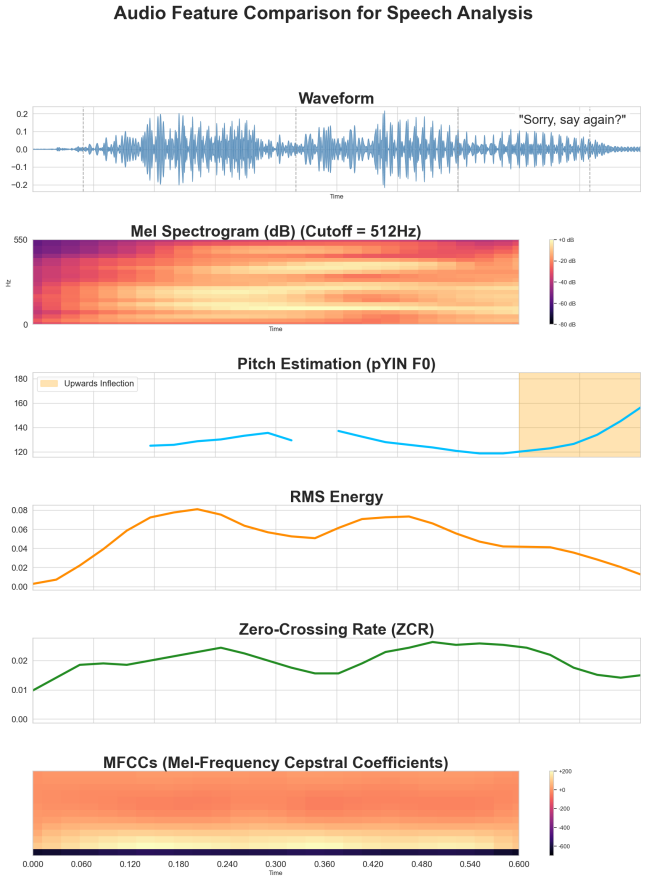

Figure 2: Extracted features after filtering (Low-pass butterworth filter, cutoff = 512Hz)

As mentioned in section 2, our team faced difficulty when annotating the dataset, since extensive ATC phraseology knowledge is required to understand the nature of each clip. Realizing the inconsistent nature of manual annotation, we decided to run a test about inter-rater reliability using Fleiss' Kappa on annotations from three raters across 200 ATC audio samples. The result of the test is $K = 0.4256$, which shows moderate agreement. Therefore, for the remaining 800 clips, we decided to let Jimmy, who has the most experience, finish the annotation.

4 EXPERIMENTS & RESULTS

Our initial model was a binary SVM classifier differentiating between questions and non-questions using only prosodic characteristics of the audio clips. We also trained a SVM classifier that differentiate between the four different types of questions outlined in section 2. We evaluated our models using 7-fold cross validation, whose confusion matrices are shown in Figure 4.

The binary model had an average accuracy of 0.6136 and an average recall of 0.6136, which performs only slightly better than random classification (which would yield an average accuracy of 0.5). The question type classification model had an average accuracy of 0.3571 and average recall of 0.3558.

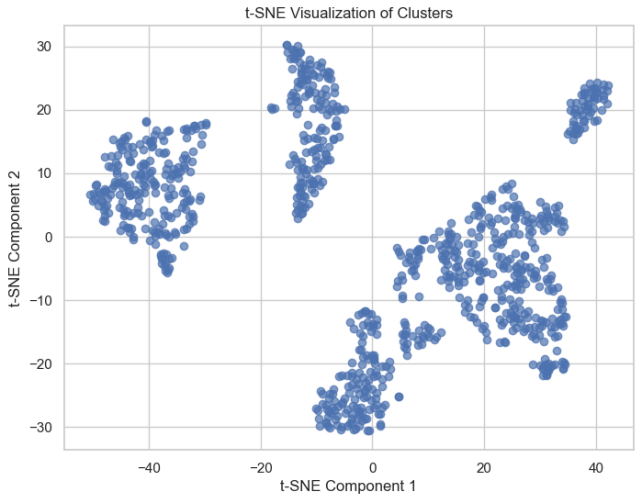

Figure 3: Unlabeled clustering of clips from extracted prosodic features

We then trained a separate binary classifier that differentiates questions from non-questions using both prosodic and lexical characteristics, where textual transcriptions were used together with the extracted audio features to train the model. We evaluated our trained model using 7-fold cross validation, whose confusion matrix is shown in Figure 5.

The multi-modal binary model had an average accuracy of 0.9134 and an average recall of 0.9135, while the multi-modal question type classification model had an average accuracy of 0.8695 and average recall of 0.8365.

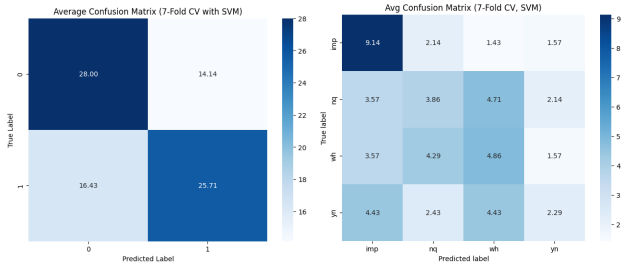

Figure 4: Confusion matrices of models using only prosodic characteristics

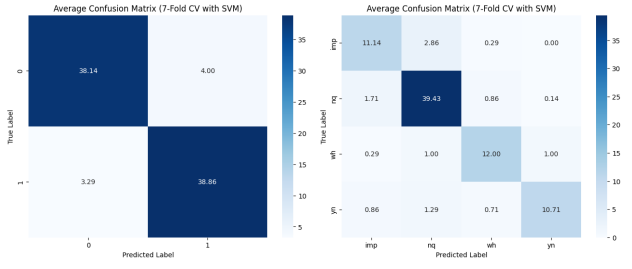

Figure 5: Confusion matrices of models using both textual and prosodic characteristic

5 DISCUSSION

While classifying sentences using prosodic characteristics results in slightly above-average accuracy, such classification still proves to be insufficient as a means of reliably detecting questions in sentences. When it comes to classifying different types of questions, the prosody-only classifier scores even worse, with many misclassifications all across the board (see Figure 4). Given the ambiguity of our sample data, and how even our own annotators had trouble determining whether some of the sentences were questions or answers, it is not a very surprising result. Additional factors, such as varying personal speaking habits or environmental affective influences, could have also contributed to the lackluster classification. For example, in our own dataset, pilots may have felt fatigued or bored sitting in a control tower for hours and repeatedly saying the same sentences, which may have resulted in some operators speaking in a more monotone voice when asking questions. This would then influence the classifier's ability to accurately decipher questions based on tone and pitch.

However, using classifiers trained on both lexical and prosodic data yielded much better results, with an accuracy of 91% in binary classification and 87% in type classification. Given the success of classifiers based on lexical characteristics of sentences [1], it follows that combining two modes of features that can detect questions would result in better classification.

Given the success of our multi-modal classifier, perhaps future work can explore the efficacy of this multi-modal classification of questions in comparison to simply classifying questions using textual data. Moreover, given how our dataset was restricted to only audio from ATC towers, expanding the dataset to include sentences from other various contexts could expand applications of this classification technique to other fields. Although annotation costs restricted us to simple SVM models, should a more holistic and extensive question-based audio dataset be compiled, further exploration using deep learning networks also has the potential to yield results.

6 CONCLUSION

While aspects of audio features can be useful for detecting spoken questions, it alone is not sufficient for accurately identifying questions and is entirely unreliable for classifying different types of questions. Rather, combining prosodic and lexical characteristics yield the best results in both aspects.

REFERENCES

[1] J. Chi, M. de Seyssel, and N. Schluter, "The role of prosody in spoken question answering," 2025.
[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.
[3] F. Suliman, M. Ben-Ahmeida, and S. Mahalla, "Importance of punctuation marks for writing and reading comprehension skills," *Faculty of Arts Journal*, vol. 13, no. 1, pp. 29–53, 2019.
[4] ATAG, "Facts and figures."
[5] J. Tol, "Fine-tuning whisper for air traffic control: 84% improvement in transcription accuracy," *jacktol.net*, Oct. 2024.
[6] J. Hirschberg, "Distinguishing questions by contour speech recognition tasks," in *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989.

[7] J. Zuluaga-Gomez, K. Veselý, I. Szöke, A. Blatt, P. Motlicek, M. Kocour, M. Rigault, K. Choukri, A. Prasad, S. S. Sarfjoo, I. Nigmatulina, C. Cevenini, P. Kolčárek, A. Tart, J. Černocký, and D. Klakow, "Atco2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications," 2022.
[8] L. Šmídl, J. Švec, D. Tihelka, J. Matoušek, J. Romportl, and P. Ircing, "Air traffic control communication (atcc) speech corpora and their use for asr and tts development," *Language Resources and Evaluation*, vol. 53, no. 3, pp. 449–464, 2019.
[9] NAV CANADA, *IFR Phraseology*. NAV CANADA, Apr. 2022. [Online].
[10] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. v. Niekirk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, E. Halvachs, C. Thomé, F. Robert-Stöter, R. Bittner, Z. Wei, A. Weiss, E. Battenberg, K. Choi, R. Yamamoto, C. J. Carr, A. Metsai, S. Sullivan, P. Friesch, A. Krishnakumar, S. Hidaka, S. Kowalik, F. Keller, D. Mazur, A. Chabot-Leclerc, C. Hawthorne, C. Ramaprasad, M. Keum, J. Gomez, W. Monroe, V. A. Morozov, K. Eliasi, nullmightybofo, P. Biberstein, N. D. Sergin, R. Hennequin, R. Naktinis, beantowel, T. Kim, J. P. Åsen, J. Lim, A. Malins, D. Hereñú, S. v. d. Struijk, L. Nickel, J. Wu, Z. Wang, T. Gates, M. Vollrath, A. Sarroff, Xiao-Ming, A. Porter, S. Kranzler, Voodoohop, M. D. Gangi, H. Jinoz, C. Guerrero, A. Mazhar, toddrme2178, Z. Baratz, A. Kostin, X. Zhuang, C. T. Lo, P. Campr, E. Semeniuc, M. Biswal, S. Moura, P. Brossier, H. Lee, W. Pimenta, J. P. Åsen, S. Hyun, I. S, E. Rabinovich, G. Lei, J. Guo, P. S. M. Skelton, M. Pitkin, A. Mishra, S. Chaunin, BenedictSt, S. VanRavenswaay, and D. Südholt, "librosa/librosa: 0.11.0," Mar. 2025.
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

7 APPENDIX

7.1 Dataset: Further Information

7.1.1 Motivation. This dataset was created to develop a system to identify questions using only prosodic features of speech. Specifically, we focus on Air Traffic Control communication to improve clear, concise, and efficient communication. We could not find an existing dataset with ATC audio clips labeled with the type of question.

The authors listed in this report provided manual annotation to Jack Tol's existing ATC Dataset [5]. It was done as part of a research project for CMPT 419 (Affective Computing), a course at Simon Fraser University. No-one received funding for the creation of this dataset.

7.2 Common phase of imply question on air traffic communication

Word	Meaning
SQUAWK IDENT	Can't see you on the radar, click the ident button on your transponder.
SQUAWK MODE CHARLIE	You seems to forget to turn the transponder on, please turn it on.
REQUEST X altidue/heading/speed/direct	May I have X altidue/heading/speed/direct?
CONFIRM?	Verify (clearance, instruction, action, information) given.
HOW DO YOU READ?	Can you hear my transmission clearly?
SAY AGAIN?	Can you repeat all, or specified part of last transmission.

Table 4: Standard Words and Phrases

7.2.1 Composition. The dataset is composed of audio communication between an Air Traffic Control Tower and a Pilot. Some clips contain one speaker, and others contain multiple speakers. A machine-generated transcription accompanies each clip. There are 11,867 instances of audio clips. Our annotated dataset contains a sample of 1,000 instances from the large set. Samples were selected semi-randomly. Geographic coverage of this dataset is unknown, but can be considered global due to the interconnected nature of air travel. Each instance consists of unprocessed audio data, an AI-generated transcription, a question type label, and extracted signal features. Relationships between isolated clips may be identified by the flight number that often is included in recorded audio. We do not have a recommended training/validation split, but we used a ratio of testing = 0.2 for our own classification. All ATC audio is publicly available on encrypted radio frequencies. Identification of individuals is possible but difficult as audio may potentially be traced to specific airports or airline flights. Audio is not timestamped and does not contain names of individuals. No sensitive information is present in this dataset.

7.2.2 Collection. For each instance, audio data is directly observable, the transcription is AI-generated, and question labels were reported by annotation subjects. Features were created via computer algorithms performing signal processing on the audio. The question labels were validated by using multiple subjects, one of whom has simulator experience speaking with ATC. A custom software program was used to collect annotations, which was developed by project member Jonathan Peters. A variety of functional tests were used to verify if the program worked as expected. 420 clips were sampled by a keyword search. 580 clips were sampled uniformly from all remaining instances. Just the researchers were involved in annotation and received no compensation. Annotations were compiled in the time frame of approximately one week. No ethical review processes were conducted as this study is exempt.

7.2.3 Preprocessing/cleaning/labeling. Instances were removed when signal processing failed, for example unable to estimate pitch. Manual labeling of data occurred using a custom GUI application included in the project .zip folder. Raw data is also available in the project folder.

7.2.4 Uses. The dataset has been used to train a unimodal and multimodal classifier to detect questions and non-questions. This dataset could also be used for speaker role detection, i.e., speech coming from a pilot or an Air Traffic Controller. Future users of this dataset should consider the various accents in this dataset to avoid creating a model that could result in bias to one or many accents, and thus groups of people. This dataset should not be used to train pilots and air traffic control due to the lack of context and unverified standards that may be present in the data.

7.3 Group Contributions

7.3.1 Jonathan.

- Set up Github Repository for code and wrote Readme.md
- Created install_dataset, filter_dataset, annotate_dataset, and extract_and_visualize Jupyter Notebook programs
- Created custom GUI annotation app to improve annotation efficiency

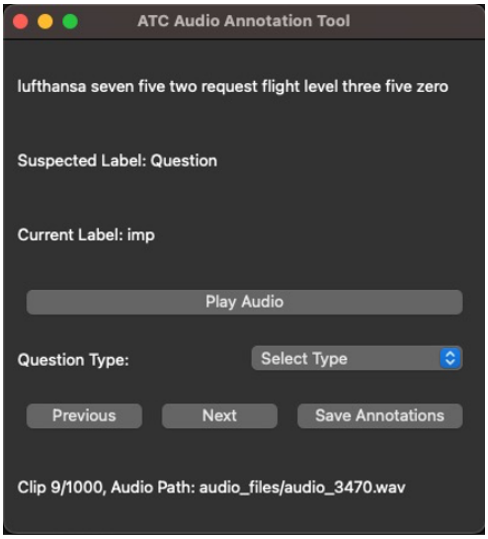

Figure 6: Annotation App Interface

- Worked on research poster by writing and creating figures for Background, Research Goal, Dataset, and Methods sections
- Wrote and created figures for Report sections including Abstract, Introduction, Approach, Dataset, and Appendix

7.3.2 Jimmy.

- Provided insight on air traffic communication phraseology for teammates.
- Annotated 1000 clips that were flag as potential questions, with the mix of non-questions.
- Create the code for model training and result visualization for the poster and the report.
- Coded and calculated the inter-rater agreement scores
- Wrote parts of the report (approach, dataset and result).

7.3.3 Lucy.

- Annotations of 200 audio clips.
- Primarily authored Results section of poster and general editing and formatting of poster.
- Print and manual assembly of presentation poster.
- Authored Experiments & Results, Discussion, and Conclusion of the report.
- General feedback and editing of code and reports.