NLP701/805 final
Whispering in the OR: End-to-End Speech-to-Intent for Surgical Command Understanding

Ayah Al-Naji
ayah.al-naji@mbzuai.ac.ae

Abstract

We developed an end-to-end speech command understanding system to enable hands-free communication between robotic assistants and surgeons. Our system combines an enhanced BERT classifier for intent classification with OpenAI's Whisper model for automatic speech recognition (ASR). To replicate realistic operating room conditions, we developed a domain-specific dataset of 120 spoken surgical commands: 60 were manually recorded (30 clean and normal-speed, 30 fast and quiet), and 60 were enhanced using Audacity with additional noise, reverb, and combined Reverb-Bass-Treble effects. Clear speech was accurately transcribed by Whisper (WER = 0.28, 63.3%), but it performed poorly when distorted (WER = 1.17, 3.4%). Our BERT model classified intents with 15.8% accuracy using these transcriptions, as opposed to nearly 31.7% when trained on reference text. These findings demonstrate that downstream comprehension is significantly impacted by ASR quality. Our results underline the significance of noise-robust and domain-adapted training for clinical deployment and show that transformer-based models are feasible for speech-driven surgical assistance. The source code[1] and the dataset[2] are publicly available.

1 Introduction

Voice-based interaction in surgical robotics is a practical way to improve accuracy, efficiency, and safety in the operating room. Conventional control methods, such as hand-operated instruments or foot pedals, require physical contact and divert surgeons' attention away from the patient. Natural speech communication reduces cognitive load, enhances workflow efficiency, and enables hands-free control between surgeons and robotic systems (Deepa, 2022).

Automatic speech recognition (ASR) has advanced significantly thanks to deep learning models such as Deep Speech (Hannun et al., 2014), Deep Speech 2 (Amodei et al., 2016), and Listen, Attend, and Spell (Chan et al., 2016). Transformer-based architectures (Vaswani et al., 2017) further improved performance, leading to modern systems like Whisper (Radford et al., 2023), which have strong multilingual and noise-resistant capabilities. But most ASR models are trained on generic datasets such as TIMIT (Garofolo et al., 1993), and they perform inconsistently in medical settings with high levels of background noise, specialized terminology, and accent variation (Alharbi et al., 2021).

Due to its highly sensitive and vital nature, it is very tough to create ASR systems for medical or surgery-related applications. Decision-making or care could directly be affected through transcription or intention misunderstanding errors. **This study intends to cover two major problems associated with this field:**

- **Limited medical speech data:** Recording actual surgeries is not easy or shareable because of privacy and ethical reasons. As such, actual training and testing become less realistic because of simulated data being used instead.
- **High reliability demands:** due to the delicacy of the medical industry, even minor transcription errors can have dire repercussions. As a result, it is crucial to confirm that Whisper functions dependably in the high-stakes acoustic environments found in operating rooms.

We combine a refined BERT model (Devlin et al., 2019) for intent classification of spoken surgical commands with Whisper for speech transcription

[1] https://github.com/ay-78787/NLP
[2] https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent

in order to address these gaps. We assess the robustness of Whisper and its impact on downstream understanding using a domain-specific dataset of 120 commands, which we manually recorded, comprising 60 commands and 60 enhanced with realistic distortions using Audacity.

Contributions:

- Create a Whisper–BERT pipeline to comprehend medical commands.
- Create and assess a domain-specific dataset with 120 commands.
- We perform a comprehensive robustness analysis of Whisper (ASR) and investigate how transcription errors affect downstream intent classification.
- We introduce baseline models for both ASR and intent classification, enabling a clear comparison against our fine-tuned BERT classifier.

In the following section, we review related work on automatic speech recognition and medical speech processing, highlighting how our approach differs from existing systems in both methodology and evaluation scope.

2 Related Work

Automatic Speech Recognition (ASR) has advanced significantly, largely due to the widespread adoption of deep learning methodologies and end-to-end modeling strategies. Initial ASR systems, including Deep Speech and Deep Speech 2 (Hannun et al., 2014; Amodei et al., 2016), utilized neural representations, which removed the requirement for manually crafted features; this enabled a direct correlation between audio input and textual output. Furthermore, while transformer-based architectures (Vaswani et al., 2017; Watanabe et al., 2018) incorporated self-attention mechanisms, which significantly enhanced both scalability and multilingual capabilities, attention-based models, including Listen, Attend, and Spell (LAS) (Chan et al., 2016), progressively refined the alignment between spoken phonemes and their corresponding written representations. These advancements have led to the creation of large models like Whisper (Radford et al., 2023). Whisper was trained on hundreds of thousands of hours of speech data, which was not ideally supervised. This training helped it perform well across different languages and in various acoustic environments.

Many studies have focused on improving the performance of automatic speech recognition (ASR) systems in the presence of background noise and when speech characteristics undergo changes. Data augmentation methodologies, as evidenced by prior studies (Ko et al., 2015; Park et al., 2019; Balam et al., 2020), coupled with multi-condition training, have demonstrated efficacy in mitigating overfitting and augmenting the capacity to generalize across diverse acoustic environments. However, most current evaluations utilize open-domain benchmarks, such as LibriSpeech or Common Voice. These benchmarks do not fully capture the acoustic and linguistic complexities found in clinical settings. Medical automatic speech recognition (ASR) systems must handle complex vocabulary, different speaker accents, and background noise from medical equipment. These challenges are often overlooked in standard ASR research (Alharbi et al., 2021; Malik et al., 2021). Although Whisper demonstrates strong generalization across various areas, its performance in medical and surgical settings has not been thoroughly investigated.

Conversely, the application of natural language processing (NLP) within the healthcare sector has predominantly centered on text-centric applications, encompassing information retrieval, diagnostic assistance, and documentation processes. Although models such as BioBERT (Lee et al., 2019) and ClinicalGPT (Wang et al., 2023) have undergone adaptations for the comprehension of biomedical texts, investigations conducted by Deepa (Deepa, 2022) and Sasanelli et al. (Sasanelli et al., 2023) have explored the utility of NLP tools in the context of clinical decision-making and orthopaedic surgical procedures. Only limited work has examined real-time spoken command interpretation for medical robotics. Conventional ASR was used by Zinchenko et al. (Zinchenko et al., 2017) to implement a voice-controlled robotic system, but it was not robust to noise and semantic interpretation. In order to close the gap between speech recognition and intent understanding in surgical assistance, our work builds on these efforts by assessing Whisper's dependability under simulated operating-room conditions and combining it with a refined BERT classifier (Devlin et al., 2019).

3 Method

We develop a complete system that turns spoken surgical commands into predicted intent categories.

As shown in Figure 1, the pipeline has two main parts. The first part is the Whisper ASR model, which transcribes all audio inputs, including our manually recorded commands and the augmented versions made with Audacity. The second part is a fine-tuned BERT classifier that uses the transcriptions to guess the intended action. We also include baseline models for both ASR and intent classification to give context and make the evaluation of the pipeline more thorough.

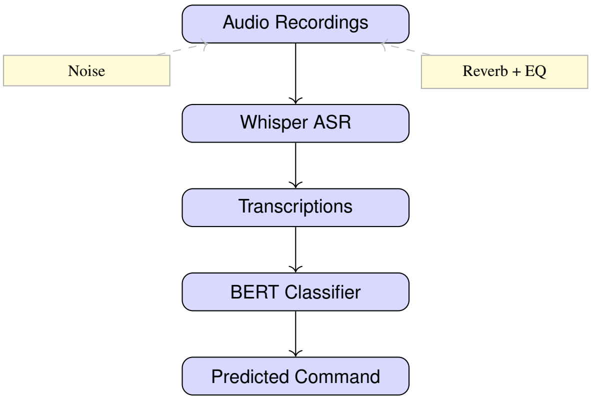


Figure 1: Audio pipeline: recordings are augmented with noise/reverb effects, transcribed by Whisper, and classified by BERT into medical commands.

3.1 Whisper ASR Transcription

We use OpenAI's Whisper model (Radford et al., 2023) to transcribe all audio recordings into text. We used two Whisper variants: **Whisper-small** as our primary ASR model and **Whisper-tiny** as a baseline. Whisper-small provides stronger noise robustness and higher transcription accuracy, while Whisper-tiny offers a lightweight comparison point with substantially fewer parameters.

All audio files (`.m4a` or `.wav`) are loaded from the data set directories and passed to Whisper using the standard transcription call `model.transcribe(audio_path)`. For each file, we extract the predicted text (`result["text"]`) and store it together with the filename and corresponding intent label in a CSV file. This CSV serves as the basis for both the evaluation of ASR (Section 3.2) and the classification of the downstream intent using BERT (Section 3.3).

We evaluated both Whisper models under five acoustic conditions: *(i) normal clean speech*, *(ii) fast-and-quiet speech*, *(iii) noise-augmented speech*, *(iv) reverb–bass–treble–augmented speech*, and *(v) large Kaggle medical speech dataset*. This setup lets us see how different kinds of acoustic distortion affect the quality of transcription and how model capacity (small vs. tiny) affects robustness.

3.2 ASR Robustness Evaluation

To evaluate Whisper's transcription quality, we test the model across five acoustic conditions: *(i) normal clean speech*, *(ii) fast-and-quiet speech*, *(iii) noise-augmented speech*, *(iv) reverb–bass–treble–augmented speech*, and *(v) the large Kaggle medical speech dataset*. We manually created a ground-truth dictionary that linked each audio file name to its corresponding standard reference transcription for our custom data set.

We compute the Word Error Rate (WER) using the `jiwer` library, applying the standard `wer(reference, hypothesis)` function to each Whisper output. We also report exact-sentence accuracy, treating a transcription as correct when `WER == 0`. For every data set condition, we calculate the mean WER and the exact-match rate to summarize Whisper's performance.

This evaluation has two main goals.

- **Measuring robustness:** It measures how Whisper's transcription accuracy drops when there are different types of acoustic distortion, like fast speech, background noise, and echo.

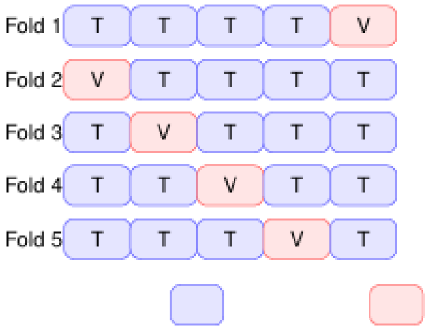
- **Establishing an upper bound for downstream performance:**It sets an upper limit on how accurately intent classification can be performed because transcription errors are directly passed into the BERT classifier.

3.3 Intent Classification with BERT

For downstream intent understanding, we fine-tune a transformer-based classifier built on `bert-base-uncased` (Devlin et al., 2019). Each input to the model is a Whisper-small transcription, and the output is one of three intent labels:

- Request Instrument.
- Adjust Device.
- Request Information.

We prepare the data by using scikit-learn's `LabelEncoder` to change categorical labels into numerical IDs and using the Whisper-small transcription as the input for the classifier. We use `BertTokenizer` to break the text into tokens,



Fold 1	T	T	T	T	V
Fold 2	V	T	T	T	T
Fold 3	T	V	T	T	T
Fold 4	T	T	V	T	T
Fold 5	T	T	T	V	T

Figure 2: Illustration of 5-fold cross-validation. Each fold uses four subsets for training (T) and one for validation (V). Final performance is averaged across all folds.

with truncation, padding, and a maximum sequence length of 64 tokens. The encoded examples are put into a PyTorch dataset and used to improve a `BertForSequenceClassification` model with three output labels.

We used the HuggingFace `Trainer` API to fine-tune the model with standard hyperparameters: three training epochs, a batch size of 8, and a learning rate of 2×10^{-5}. We use 5-fold cross-validation to test the model and ensure it generalizes fairly.

Note: We apply intent classification baselines and cross-validation exclusively to our surgical command dataset, while using the Kaggle dataset for cross-domain ASR evaluation without intent classification baselines.

3.3.1 Training Procedure: 5-Fold Cross-Validation

Because our custom dataset contains only 120 samples, we adopt a 5-fold cross-validation strategy to obtain stable and unbiased estimates of BERT's performance. The dataset is partitioned into five equally sized folds. In each round, four folds are used for training while the remaining fold serves as the validation set. This process cycles until every fold has served once as the validation split.

Figure 2 illustrates the cross-validation setup. By exposing the model to multiple train–validation splits, we reduce the risk of overfitting and obtain a more reliable estimate of downstream performance.

3.3.2 Keyword-Based Baseline

We use a rule-based intent classifier called `baseline_predict`, which utilizes keyword lists created manually from our dataset. "Scalpel," "forceps," and "needle holder" are examples of terms that map to "Request Instrument." "Increase," "lower," "brightness," and "turn off" are examples of terms that map to "Adjust Device." "Show," "pressure," "what is," and "reading" are examples of terms that map to "Request Information." This baseline provides the BERT classifier with a clear and straightforward point of reference.

3.3.3 Majority-Class Baseline

We make a simple baseline for each cross-validation split that always predicts the most common label in that fold's training set. This sets a minimum level that any trained classifier should be able to reach.

3.3.4 Training on Augmented Data

To investigate the impact of acoustic augmentation on downstream classification, we integrate transcriptions from three datasets: the clean/fast–quiet dataset, the noise-augmented dataset, and the Reverb–Bass–Treble dataset. We combine all the examples into a single 120-sample corpus after standardizing the column names. This corpus is saved as `audio_transcriptions_all.csv`. Next, we repeat the fine-tuning and cross-validation process, this time using the Whisper-small transcriptions from the combined dataset.

3.3.5 Kaggle Classification Setup

To assess the classifier's capacity to generalize beyond our regulated surgical domain, we evaluate BERT on the *Medical Speech, Transcription, and Intent* dataset from Kaggle. We use the dataset's clean human-written transcriptions as input text and report absolute performance metrics to assess cross-domain generalization. Note that we do not apply baseline comparisons to the Kaggle dataset, as our focus is on measuring generalization capability rather than relative improvement.

3.4 End-to-End Pipeline Integration

Our system works like a two-stage pipeline. First, Whisper turns incoming audio into text. Then, a fine-tuned BERT classifier uses the text to guess which of three surgical intent categories it belongs to. We can put the performance of each stage in context by using both ASR- and intent-level baselines. Testing under different acoustic conditions also reveals how front-end transcription errors propagate throughout the entire pipeline. This integrated design enables us to conduct a controlled and comprehensible analysis of robustness across the entire speech-to-intent workflow.

4 Data

Our system uses two sources of speech data: a custom domain-specific corpus and a publicly available medical speech dataset.

4.1 Custom surgical command dataset

We constructed a domain-specific dataset of 120 spoken surgical commands, grouped into three intent categories: **Request Instrument**, **Adjust Device**, and **Request Information**. The data set contains 60 manually recorded samples and 60 acoustically enhanced versions by Audacity, designed to simulate common operating room distortions. We used a smartphone microphone to record everything in a quiet room. We talk about the four types of audio in this data set below.

(a) Normal-speed clear speech:
We recorded 30 commands at a normal volume and speaking speed. These recordings will represent pristine or optimal environments with little or no ambient noise. Some commands are: "Hand me scalpel", "Give me forceps", "Increase camera brightness", "Turn off the cauterizer", and "Show blood pressure". Each recording is paired with its canonical reference transcription and corresponding intent label.

(b) Fast-and-quiet speech:
We made 30 more recordings by saying each command quickly and quietly. This method mimics low-volume or stressed communication in operating rooms, where doctors may need to give instructions quickly or with minimal voice projection.

(c) Noise augmentation:
We used Audacity to add white noise to the original recordings, making them sound like the noise generated by operating-room equipment, such as suction devices and ventilators. Table 1 lists the settings that were used to make noise-augmented audio. These additions help us understand how Whisper's ASR performance deteriorates in the presence of acoustic interference.

Table 1: White noise augmentation parameters

Parameter	Value	Description
Amplitude (0–1)	0.02	Loudness of added noise
Duration	Equal to original	Matches command duration
Mixing	Overlay	Keeps speech audible

(d) Reverb–Bass–Treble augmentation:
We used Audacity's reverberation and equalization filters to simulate the sound in the room as it would be in an operating room, taking into account factors such as reflective surfaces and microphone coloration. Table 2 shows the reverb settings, and Table 3 shows the Bass and Treble settings. These additions test Whisper's strength against echo, color, and spectral distortion.

Table 2: Reverb parameters

Setting	Value	Description
Room Size (%)	40	Simulated reflection space
Pre-delay (ms)	10	Delay before first reflection
Reverberance (%)	50	Controls reflection density
Damping (%)	50	Controls decay rate
Tone Low (%)	100	Maintains low frequencies
Tone High (%)	100	Preserves high clarity
Wet Gain (dB)	+10	Strength of reflected signal
Dry Gain (dB)	0	Keeps original loudness
Stereo Width (%)	100	Preserves spatial width

Table 3: Bass and Treble parameters

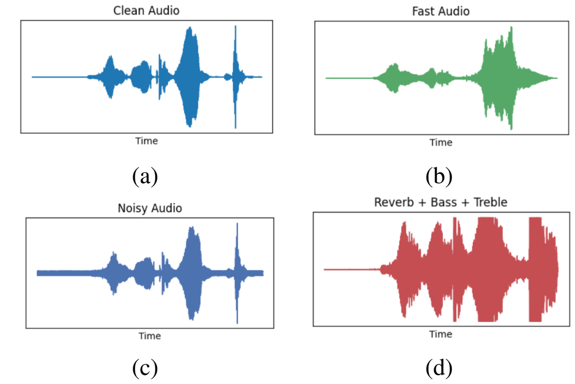
Setting	Value (dB)	Description
Bass	+15	Enhances low-frequency strength
Treble	+11	Improves brightness and sharpness
Volume	+5	Balances overall loudness

(e) Waveform visualizations:
Figure 3 shows the four audio conditions in our data set: clear normal-speed speech, fast-and-quiet speech, noise-augmented speech, and Reverb–Bass–Treble–augmented speech. These visualizations show how each augmentation changes the signal's structure and energy profile. They provide a qualitative picture of the acoustic variability that Whisper must deal with during ASR.
needspace
5

4.2 Kaggle medical speech dataset

To determine if our models extend beyond the limitations of our small custom dataset, we also utilize the publicly accessible *Medical Speech, Transcription, and Intent* dataset from Kaggle.[3] This dataset

[3] https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent



(a) (b) (c) (d)

Figure 3: Waveform examples for (left to right): normal-speed clear speech, fast-and-quiet speech, noise-augmented speech, and Reverb–Bass–Treble–augmented speech.

contains hundreds of medical voice commands, spoken by different individuals. This means that the speaker's identity, accent, microphone quality, and recording conditions are all much more varied than in our 120-command corpus.

The Kaggle dataset differs from ours in that it includes a broader range of general medical instructions. This provides us with a different perspective on real clinical speech.

Each audio file has a transcription and one of several intent labels that are related to common actions of medical devices.

The Kaggle samples have two roles in our research:

- The Kaggle samples provide a significant, clean baseline to compare Whisper's ASR accuracy with our manually recorded and augmented commands.
- They enable evaluation of whether our BERT classifier, trained on a small controlled dataset, can generalize to real-world medical speech scenarios.

The Kaggle dataset has more than just transcripts and intent labels. It also contains metadata that provides information about the quality of the recordings, including background noise, audio clipping, and overall audio clarity. The average subjective audio quality rating is 3.68 on a 1–5 scale, and most samples have n *no noise* or *light noise*. Figure 4 shows how the overall audio quality scores are spread out over the 6,661 recordings. These differences in quality make the dataset a good way to test how well Whisper works and whether our classifier can work with a broader range of real-world clinical speech conditions.

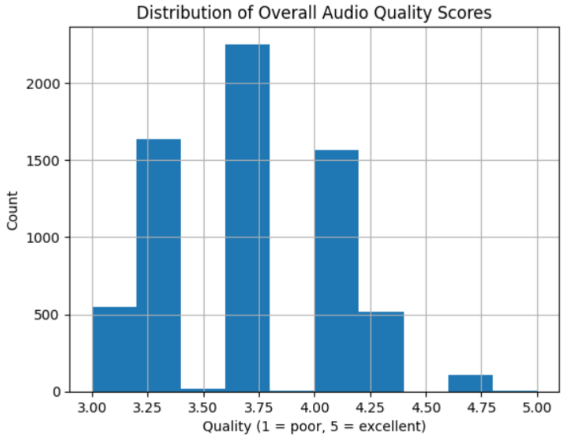


Figure 4: Distribution of overall audio-quality scores (1 = poor, 5 = excellent) in the Kaggle medical speech dataset.

5 Experiments and Evaluation

5.1 ASR Results

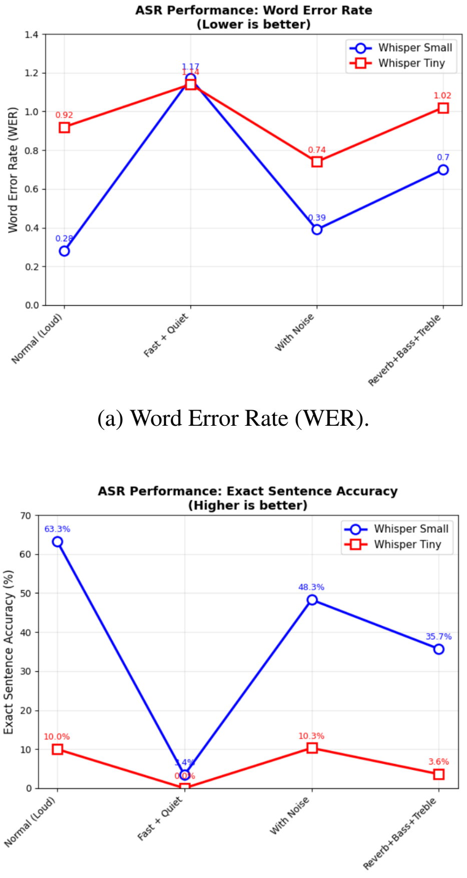
We evaluate Whisper's transcription quality across five acoustic settings: *(i) normal clean speech*, *(ii) fast-and-quiet speech*, *(iii) noise-augmented speech*, *(iv) reverb–bass–treble–augmented speech*, and *(v) a large clean dataset from Kaggle*. Figure 5 shows Word Error Rate (WER) in subfigure 5a and Exact Sentence Accuracy (ESA) in subfigure 5b for Whisper-small (our main ASR model) and Whisper-tiny (baseline).

5.1.1 Overall trends

As shown in Figure 5a, Whisper-small achieves strong performance on normal clean speech, with a low WER of 0.28, and a corresponding ESA of 63.3% (Figure 5b). Performance degrades substantially on fast-and-quiet speech (WER = 1.17; ESA = 3.4%), indicating that reduced volume and rapid articulation are the most difficult conditions for Whisper.

5.1.2 Effects of acoustic distortions

Adding noise raises the WER by a small amount (0.39), but ESA remains relatively high (48.3%), indicating that Whisper performs well with background noise. Reverberation and equalization exacerbate spectral smearing, which increases WER to 0.70 and reduces ESA to 35.7%. This supports the idea that reverberation affects ASR performance more than additional noise.



(a) Word Error Rate (WER).

(b) Exact Sentence Accuracy (ESA).

Figure 5: Whisper ASR performance across four acoustic conditions for Whisper-small (main) and Whisper-tiny (baseline). Lower WER and higher ESA indicate better transcription quality.

5.1.3 Performance on the Kaggle dataset

Whisper-small achieves a WER of 0.234 on the large Kaggle dataset (Figure 5a), demonstrating robust performance on diverse medical speech. The ESA of 14% reflects the vocabulary mismatch between Kaggle's medical symptoms and our surgical command references. **Note:** We report only Whisper-small performance on Kaggle, as our focus is on cross-domain generalization rather than model comparison.

5.1.4 Comparison to the baseline ASR

Figure 5a shows that Whisper-tiny always has a higher WER, especially for fast-and-quiet speech (1.25). The significant difference between tiny models highlights the importance of model capacity in handling changes in sound.

5.1.5 Error patterns

Whisper exhibits several systematic failure modes that directly affect downstream intent classification.

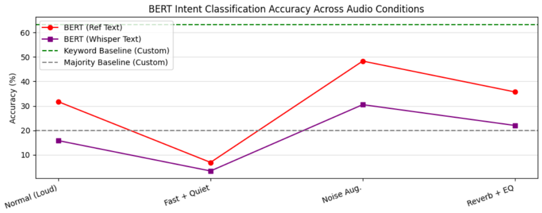


Figure 6: BERT intent classification accuracy across acoustic conditions using reference transcriptions and Whisper-generated transcriptions. Keyword and majority baselines are shown for comparison.

Common failure cases include:

- **Phonetic drift:** GT: “pass needle holder” Whisper: “past middle older”
- **Cross-language hallucination:** GT: “hand me scalpel” Whisper: “hàn mié escapo”
- **Word deletion:** GT: “increase suction power” Whisper: “increase power”
- **Invention of unseen words:** GT: “prepare scissors” Whisper: “repair siskors”

These errors illustrate how acoustic degradation produces unpredictable transcription variants, which then propagate into the intent classifier and cause significant accuracy loss in the cascade pipeline.

5.2 BERT Classification Results

We evaluate the downstream intent classifier using both *reference transcriptions* (i.e., clean canonical text) and *Whisper-generated transcriptions*. Figure 6 shows BERT's accuracy across the four acoustic conditions and compares performance against two baselines: a keyword-matching classifier and a majority-class model.

5.2.1 Reference vs. Whisper transcriptions

As shown in Figure 6, BERT achieves its highest accuracy with noise-augmented reference text (48.3%), followed by reverb+EQ (35.7%) and normal speech (31.7%). However, when using Whisper transcriptions, accuracy decreases significantly across all conditions—dropping from 48.3% to 30.5% with noise, and from 31.7% to 15.8% with normal speech. Fast and quiet speech yields the poorest outcomes for both input types (6.9% for reference text and 3.4% for Whisper text), confirming the severe ASR degradation detailed in Section 5.1.

5.2.2 Comparison to keyword and majority baselines

The keyword-based classifier achieves 63.3% accuracy on our custom dataset, surpassing BERT in every aspect. This is what we expect because the custom dataset has short, formulaic commands with keywords that are very good at predicting what someone wants. The majority-class baseline, on the other hand, remains low (20%), indicating that the dataset is not heavily biased toward a dominant class.

5.2.3 Effects of Whisper transcription errors

Our results indicate that errors propagate significantly from the ASR stage to the intent classifier. Because BERT was trained on clean reference text, any changes made by Whisper, such as substitutions, dropped words, or cross-language hallucinations, immediately impact classification accuracy. This demonstrates that making ASR more reliable is crucial for utilizing a cascaded Whisper–BERT pipeline in real-life surgical settings.

5.2.4 Generalization to the Kaggle dataset

On the Kaggle dataset, BERT achieves near-perfect accuracy with reference text (99.6%) and maintains strong performance with Whisper transcriptions (86.1%). This demonstrates excellent generalization to diverse medical speech, though the 13.5% performance drop reveals the impact of ASR errors even on high-quality transcriptions (WER: 0.234). The high accuracy across both conditions indicates that BERT effectively handles the broader medical vocabulary present in Kaggle's data, confirming the robustness of our intent classification approach beyond surgical commands.

5.3 Cross-Dataset Comparison and Unified Analysis

To better understand the overall behavior of our speech-to-intent system, we compare all datasets and acoustic conditions together. This subsection provides a unified view of (i) ASR robustness across clean, distorted, and real-world speech, and (ii) the resulting impact on downstream intent classification.

Figure 7 shows the Whisper-small ASR performance across all five datasets, while Figure 8 summarizes BERT classification accuracy for both reference text and Whisper-generated transcriptions.

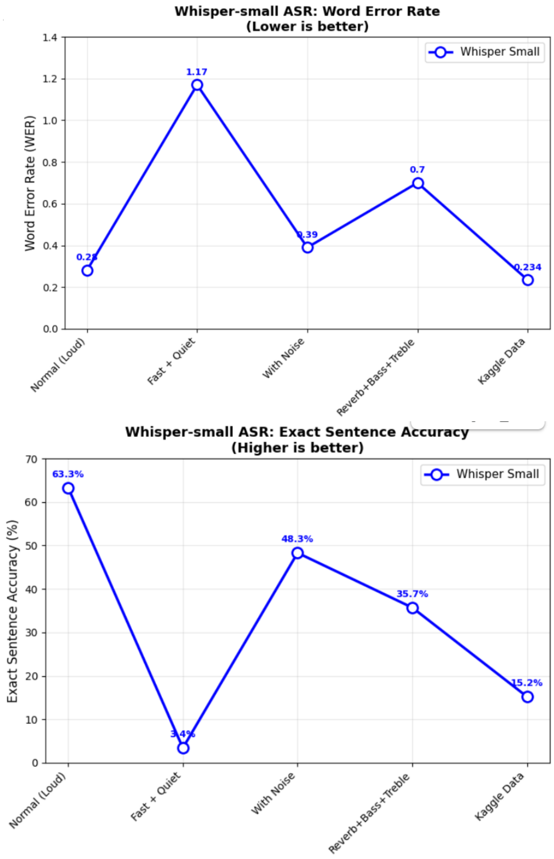


Figure 7: Unified Whisper-small performance across all datasets: (top) Word Error Rate (WER), (bottom) Exact Sentence Accuracy (ESA). Normal clean speech and Kaggle data yield the highest ASR accuracy, while fast-and-quiet speech produces the most severe degradation.

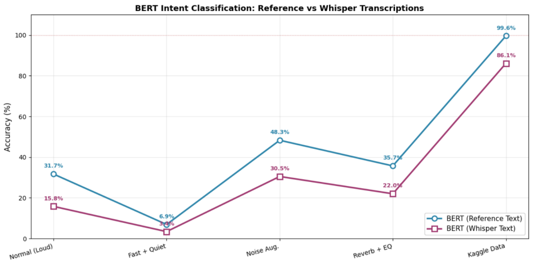


Figure 8: Unified BERT classifier accuracy across all datasets, comparing reference text vs. Whisper transcriptions. Accuracy correlates strongly with upstream ASR quality, with Kaggle data yielding the best cross-domain performance.

Key Findings

- **Clean speech vs. distorted speech:** Normal clean speech achieves low WER (0.28) and high ESA (63.3%), while fast-and-quiet speech severely degrades performance (WER = 1.17; ESA = 3.4%). Acoustic clarity is the single strongest factor influencing ASR success.

- **Effect of augmentation type:** Noise augmentation increases WER only moderately, while reverb and equalization induce more severe transcription drift. This confirms that spectral smearing is harder for Whisper to model than additive noise.

- **Downstream BERT sensitivity:** BERT accuracy drops sharply when using Whisper transcriptions instead of reference text. Errors in ASR propagate directly to the classifier, reducing accuracy by 15–25 percentage points across most conditions.

- **Kaggle dataset behavior:** Despite its larger speaker and acoustic diversity, Kaggle speech yields the *best overall results*: 86.1% accuracy using Whisper transcriptions, and 99.6% using reference text. This highlights the classifier's strong generalization to real-world data, and Whisper's stability on clean professional recordings.

- **Cross-dataset robustness patterns:** The combined results reveal a nearly linear relationship between WER and intent accuracy: conditions with higher ASR errors consistently produce lower classification scores. This validates the necessity of noise-robust ASR for reliable medical command understanding.

Overall, this unified comparison shows that the Whisper–BERT pipeline is effective on clean or moderately noisy medical speech but becomes fragile with low-volume or reverberant commands. Improving ASR robustness is therefore crucial for deployment in real surgical environments.

6 Conclusion and Future Work

In our project, we developed and tested an end-to-end speech-to-intent pipeline for detailed understanding of surgical commands using OpenAI's Whisper ASR model and a fine-tuned BERT classifier. We assembled a domain-related dataset of 120 speech commands covering three categories of intent and extensively used fast-and-quiet to strengthen the audio speech, additive noise, and Reverb–Bass–Treble effects to simulate the conditions. In our experiments, Whisper-small transcribed regular speech correctly. Moderately noisy signals with relatively low word error rates, but performance degrades sharply under fast, low-volume, and reverberant speech conditions. For downstream tasks, BERT performs accurately on an example text, but deduces drastically when presented with noisy data Whisper Transcriptions, confirming the propagation of ASR Errors directly into the intent ASR Errors classification. Even when applied to the extensive medical Kaggle dataset, evaluation demonstrated that our classifier can generalize effectively to other speakers and commands when there is good text. Readily accessible, thereby emphasizing the need for robust ASR performance and large-scale datasets.

Future work will involve bolstering both the data and the automatic speech recognition (ASR) front-end. Regarding data, the aim will be to build a larger, multi-speaker corpus including a wider range of surgery scenarios and evaluate further data augmentation strategies in addition to multi-condition training, based on previous work in robustness to noise and SpecAugment-inspired transformations (Ko et al., 2015; Park et al., 2019; Balam et al., 2020). Regarding the ASR front-end, the most important next step would be to fine-tune or adapt the Whisper model and other related ASR architectures (Radford et al., 2023; [identifier 10447520]) in the medical speech domain, ensuring speech recognition performance in the face of challenging noises and reverberations. At the same time, other related ideas include exploring other transformer architectures relevant to intent classification tasks (Devlin et al., 2019; Sanh et al., 2019; [bioinformatics reference]) and closer integration of ASR and intent classification.

References

S. Alharbi, M. Alrazgan, A. Alrashed, T. AlNomasi, R. Almojel, R. AlHarbi, F. AlShehri, and M. Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:133517–133541.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

Jagadeesh Balam, Jocelyn Huang, Vitaly Lavrukhin, Slyne Deng, Somshubra Majumdar, and Boris Ginsburg. 2020. Improving noise robustness of an end-to-end neural model for automatic speech recognition. *arXiv: Audio and Speech Processing*.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

P. Deepa. 2022. Speech technology in healthcare. *Elsevier Health Sciences*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John S. Garofolo, Lori F. Lamel, William M. Fisher, David S. Pallett, Nancy L. Dahlgren, Victor Zue, and Jonathan G. Fiscus. 1993. Timit acoustic-phonetic continuous speech corpus.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proceedings of Interspeech*, pages 3586–3589.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: A survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Francesca Sasanelli, Khang Duy Ricky Le, Samuel Boon Ping Tay, Phong Tran, and Johan W. Verjans. 2023. Applications of natural language processing tools in orthopaedic surgery: A scoping review. *Applied Sciences*, 13(20).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *ArXiv*, abs/2306.09968.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Yalta, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech*.

Kateryna Zinchenko, Chien-Yu Wu, and Kai-Tai Song. 2017. A study on speech recognition control for a surgical robot. *IEEE Transactions on Industrial Informatics*, 13(2):607–615.