Recognizing Speech with no Microphone: using Hidden Markov Models on Intraoral Sensor Data

Nathan Klumpenaar (M.J.N.Klumpenaar@student.tudelft.nl)
Supervisors: Przemysław Pawełczak, Vivian Dsouza
Part of the 2025 CSE3000 Research Project

TU Delft

Introduction

The *Densor* [1] is an intraoral sensor platform designed to capture unique data inside the human mouth. While primarily focused on sleep analytics, this thesis explores a different use case: its use for voice activity detection (VAD) and keyword recognition.

An initial literature survey and data exploration showed the success and ease of use of Hidden Markov Models (HMMs), leading to the research question:

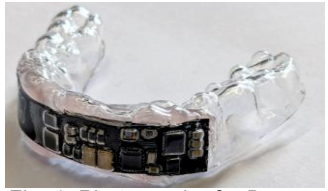


Fig. 1: Photograph of a Densor, by Dsouza et al., 2024 [1].

Is it possible to detect and recognize speech with Hidden Markov Models using data gathered by the intraoral Densor sensing platform?

Feature Extraction Pipeline

Sensor values → Overlapping sliding windows (125 ms) → Z-Score normalization → Extracting feature values

Five sensor channels
- Barometer
- Ambient Light Sensor (ALS)
- Inertial Measurement Unit (IMU)
  - Reports X, Y, Z acceleration

Four feature values per channel
- Mean value
- First differential mean value
- Standard deviation
- Amount of zero crossings

Fig. 2: Feature extraction

Limitations & Future Work

- **Limited variety in dataset**: Prevented most session/user independent testing, leading to limited reliability of results.
- **Controlled position & environment:** All recordings were done with a stationary user. Results would presumably be worse in a dynamic environment.
- **Future work:** A more (varied) dataset and exploring more advanced models.

Conclusions

- **An F1-Score of 0.73 and a keyword recognition accuracy of up to 72% show that Densor-recorded data contains enough speech information for basic speech detection and recognition.**
- **These results are limited by the small and unvaried dataset. It is not expected that these models perform well on real world data.**
- **A bigger and more varied dataset can improve the models and possibly support more advanced models as well.**

[1] Dsouza, V., Pronk, J., Peppelman, C., Madariaga, V. I., Pereira-Cenci, T., Loomans, B., & Pawełczak, P. (2024). Densor: An Intraoral Battery-Free Sensing Platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(4), 1–30. https://doi.org/10.1145/3699746

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Methodology

Voice Activity Detection

1. Split data into a train and test set.
2. Partition the training data into 1-second segments labelled either speech or no-speech.
3. Automatically select the highest-performing features, based on *F1-Score*.
4. Train the final HMMs: one speech, one no-speech, using only the selected features.
5. Evaluate the performance based on *recall*, *precision* and *F1-Score*.

Keyword Recognition

1. Extract all keyword segments from the data. Each segment is labelled with the spoken keyword.
2. Split these segments into a test and train set.
3. Automatically select the highest-performing features, based on *accuracy*.
4. Train the final HMMs: one for each keyword, using only the selected features.
5. Evaluate the performance based on the *accuracy*.

3 Automated Feature Selection Loop

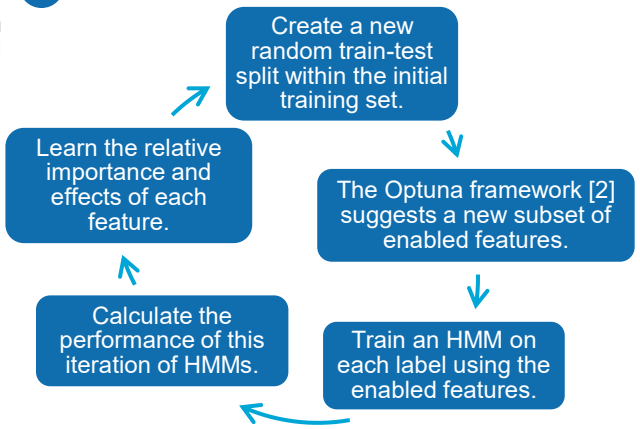
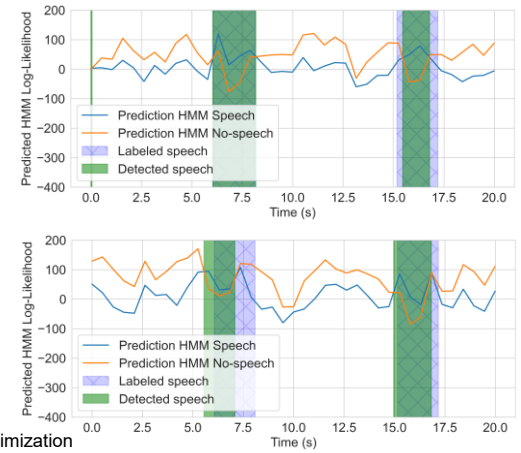


Fig. 3: Automated feature selection loop

Results

Voice Activity Detection

The best performing result, on a session-dependent dataset, using all sensors:
- **Precision: 0.74**
- **Recall: 0.73**
- **F1-Score: 0.73**



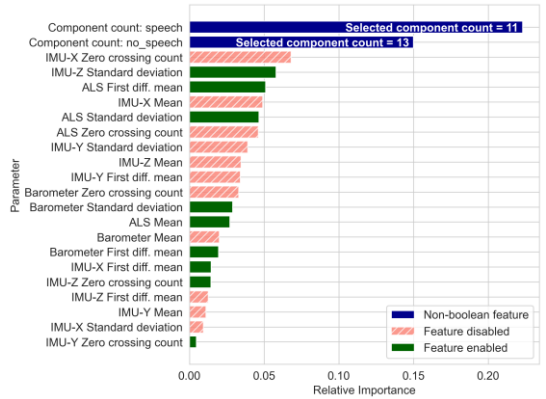
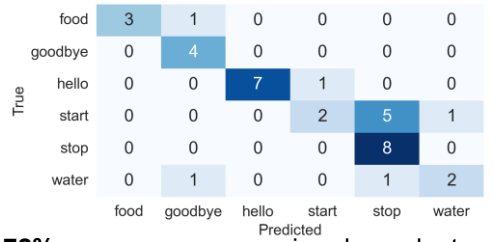


Fig 5. VAD automated feature selection results

Fig. 4: Subset of VAD results

Keyword recognition



72% accuracy on a session-dependent dataset using all sensors.

67% accuracy on a session-independent, user-dependent dataset using all sensors.

Fig. 6: Subset of keyword recognition results