POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Towards Intent Recognition from Bone Conduction Speech Signals

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Tomaz Maia Suller

Advisor: Prof. Paolo Bestagini

Co-advisor: Prof. Mirco Pezzoli

Academic year: 2025-2026

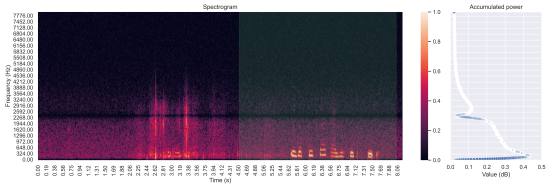
1. Introduction

Bone conduction (BC) microphones are better suited for wearable speech applications than standard aerial conduction (AC) ones due to the smaller power consumption and greater resistance to environmental disturbances of the former. Their low bandwidth and strong dependence on sensor positioning however makes has meant their use most common as an auxiliary, rather than a primary, sensor for natural language understanding tasks. Figure 1 illustrates in two spectrograms how BC signals are most concentrated in frequencies below 1 kHz while AC signals cover the entire 8 kHz bandwidth given the 16 kHz sampling frequency.

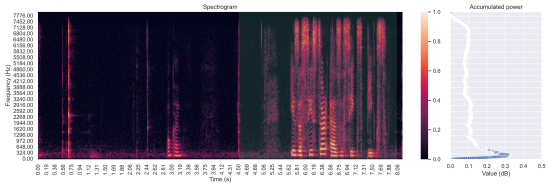
In this work, the exclusive use of BC is explored, with a focus on voice command detection, also known as intent recognition or audio-to-intent (A2I).

The main contributions of this work are:

1. introducing a data collection protocol and of a privately-held dataset collected following the protocol, constituting the first English-language BC intent recognition dataset to the best of the author's knowledge; and
2. reporting the performance of A2I and speech transcription, or automatic speech recognition (ASR), models designed and trained for AC data no BC data, showing



(a) Bone Conduction



(b) Aerial Conduction

Figure 1: Comparison between bone and aerial conduction signals collected simultaneously.

2. Problem statement

The computational method at hand must extract structured speech commands from bone conduction time series data captured using an in-ear, STMicroelectronics 3-axis accelerometer in a speaker- and accent-invariant manner, ideally with no per-user adjustment required.

In this work, a three *field* intent structure is

adopted, with each intent containing: object (the target of the command), action (what is performed to the object), and location (the place in which the object is located).

3. Methodology

Intent classification is modelled as a supervised learning problem, for which a dataset of BC audio recordings and the intents such recordings express is required; its collection methodology is reported in Section 3.1.
A machine learning algorithm is also required. Following the recent success of end-to-end deep learning approaches in AC [3], a single neural network is employed; its architecture is described in Section 3.2.

3.1. Data collection

Before collection, a set of *prompt* sentences for participants to speak had to be determined. This set was taken from two sources: first, from a subset of the prompts in the Fluent Speech Commands (FSC) benchmark [2], a reference voice command dataset; the subset focused on commands related to house controls and music playback. second, from a subset of phonetically representative sentences from a speech manual [1]; the subset was balanced using a custom greedy algorithm such that the selected subset would have a more balanced phoneme distribution than the original sentence set while also restricting the number of sentences to keep per-participant data collection time reasonable. In total, 231 prompts – 116 from FSC and 115 from the speech manual – were selected.
Collection was performed using a custom data acquisition setup built using commercial STMicroelectronics components. The system allows participants to control the progression of the prompts presented to them while recording bone- and aerial- conduction microphone speech at a sampling frequency of 16 kHz.
After collection, samples were manually analysed for inconsistencies and labelling of speech segments. A total of 11.975 utterances spanning 2.95 h were collected from 38 volunteers (24 reported male at birth, 14 female) from 12 different nationalities (15 Italian).

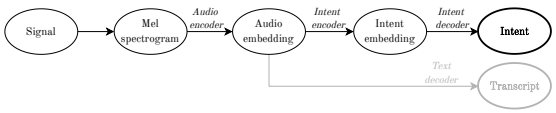


Figure 2: High-level representation of the model architecture.

3.2. Model architecture

A high-level component view of the model architecture is illustrated in Figure 2. It was introduced in [4].
A mel filter bank with 40 triangular filters spanning frequencies up to 8 kHz is computed over a short-time Fourier transform with Hann windows of 25 ms with 50% overlap, serving as input features to the audio encoder. The audio encoder employs a first set of convolutional layers, followed by two BiLSTM recurrent layers, and two final fully connected ones. Its embeddings enable both text transcription through the text decoder (employed for model pre-training only) and A2I through the intent classifier. The intent classifier adds a further intent encoder with an additional two-layer BiLSTM and a fully connected layer.
Decoding in both cases is performed by beam search, with scores assigned by a recurrent neural network with attention. The text encoder employs location-aware attention, a variant specifically designed to assist in monotonic sequence-to-sequence tasks, complemented by an attention coverage penalty and by shallow fusion with a pre-trained language model. The intent decoder on the other hand employs key-value attention.

4. Experiments and Results

4.1. Data pre-processing

The use of a model designed for and pre-trained on AC data for BC signal classifications opens the question of how to best pre-process BC data, which is measured in different physical units and is more susceptible to noise in lower frequency bands. Pre-processing experiments therefore focused on how to normalise the data before providing it to the model, and on which cutoff frequency to use for removing noise from participant body movements.
Normalisation was performed by computing mean and variance from each sample in the Lib-

riSpeech dataset over which the model was pre-trained, and then by taking the mean of these values. BC data was thus normalised to a mean of $-1.393 \cdot 10^{-5}$ and a standard deviation of $8.830 \cdot 10^{-2}$. Normalisation resulted in minor improvements in F1-score, with a slight reduction in accuracy generally outweighed by a great increase in recall.

Filter cutoff frequency analysis was performed with band- and high-pass order 3 Butterworth filters. Bandpass filters are defined by two frequencies, here named the low cutoff frequency f^l and the high cutoff frequency f^h, while the high-pass filter is ony defined by f^l. All combinations of $f^l = 10\,\text{Hz}, 20\,\text{Hz}, \ldots, 200\,\text{Hz}$ and $f^h = 1000\,\text{Hz}, 2000\,\text{Hz}, 5000\,\text{Hz}$, and $7999\,\text{Hz}$ were experimented with.

No significant variation was present between experiments sharing the same f^h but with different f^l; variation was however present in experiments with different f^h, with a sharp increase in pre-trained model performance on BC data when $f^h \geq 5\,\text{kHz}$. This result is surprising, as signals in higher frequency bands suffer attenuation not only from the bone conduction phenomenon itself, but also from the sensor frequency response; nevertheless, enough information remained to support the model.

Based on these results, to reduce the potential for mechanical noise interference, and to maximise signal availability to the model, a high-pass filter with $f^l = 100\,\text{Hz}$ was selected.

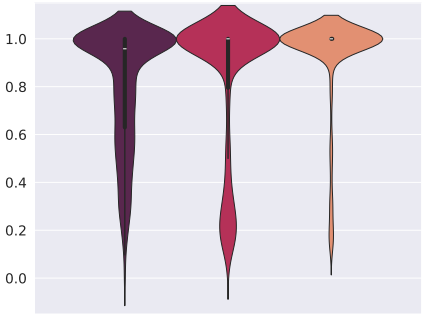
4.2. Model training

As described in Section 3.2, the intent classification model is composed of an audio encoder trained on text transcription and on a downstream intent classifier. Each of these modules is trained separately, so that, during classifier training, the audio encoder is frozen.

Given the high number of parameters and extensive dataset used to train the audio encoder – LibriSpeech contains 960h of audio recordings –, this model component was fine-tuned with BC data from the original AC trained model. Fine-tuning took place in two stages, both using the same training parameters as the ones for the provided model, except for a reduced initial learning rate (from 1.0 to 0.1): first, the audio encoder was fixed, and the text decoder was trained, in an attempt to maintain the original embedding distribution; then, both components were jointly optimised. Figure 3 illustrates how fine-tuning improved model performance in both domains, bringing BC performance close to that of AC.

The intent classifier instead is trained from scratch, as FSC provides a more comparable total duration and speaker diversity to the collected dataset than LibriSpeech. In this case, the initial learning rate is kept at its original value, but learning curve analysis motivates training the model beyond its provided 6 training epochs. Figure 4 illustrates performance of the model architecture pre-trained on AC data, and of two checkpoints form training for each dataset: one after 6 epochs and one after 30; it also represents performance both for prompts the model had access to during training, and prompts held-out for evaluation purposes.

(a) Bone Conduction

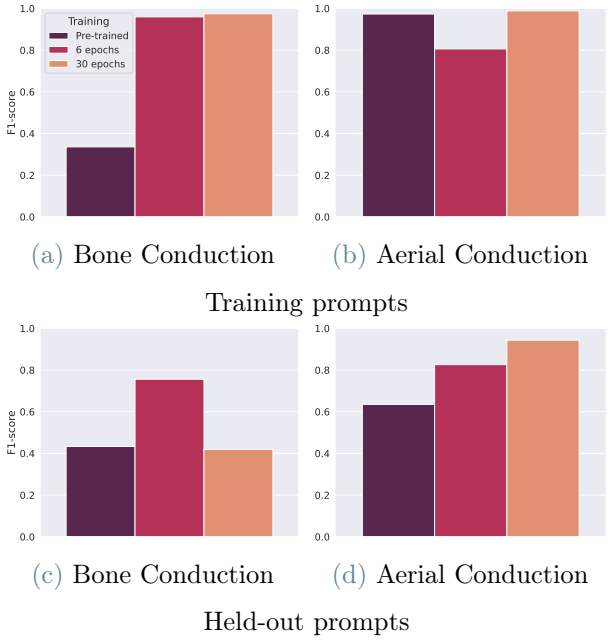


(b) Aerial Conduction

Figure 3: Normalised Levenshtein similarity between model predictions and ground-truth phonetic transcript from CMUdict.

These results show that, despite an initial dip in performance after some training, AC performance in training and held-out prompts benefited from longer training. This is not the case in BC however, in which performance in the training prompts increased slightly while held-out performance decreased sharply.

It is therefore hypothesised that pre-training in-



(a) Bone Conduction (b) Aerial Conduction

Training prompts

(c) Bone Conduction (d) Aerial Conduction

Held-out prompts

Figure 4: Intent classification F1-score across model checkpoints and datasets.

troduces a semantic structure to the embeddings of the audio encoder while supports the model in generalising to never-before-seen sentences, and that the small quantity and diversity of BC data from the collected dataset is not able to adequately represent such semantic concepts. It should be noted this situation does not configure a typical case of overfitting, since validation performance (computed over the same prompt pool as that of the training set) still increases, suggesting the model is able to generalise to new scenarios and speaker voices and accents.

5. Conclusion

Executed experiments focused on modifying aspects of the training procedure of the employed neural network model to increase its performance for bone conduction signals. Normalising the dataset provided some performance improvements, and filtering showed the unexpected presence of important BC speech signal components above 5 kHz despite high expected attenuation, even if it did not show significant performance differences between cutoff frequencies for the high-pass filter.

Model classification results provide strong evidence for the feasibility of employing existing AC techniques for BC data, despite issues around how the A2I model generalises to new prompts over BC. This is attributed to the lower diversity and size of the BC dataset, which does not allow the model to map semantically similar sentences close to each other in its embedding space.

The ability of the model to generalise in the BC case to different speakers, and in the AC case to different speakers and prompts, nevertheless suggests training over a larger and more diverse BC dataset should allow the model to achieve comparable results across both domains

5.1. Future work

Dataset Work should be done on how to best scale the dataset without the need for manual collection and extraction of new samples, given the difficulties of crowdsourcing BC data. While text-to-speech systems would seem a good option, preliminary analysis suggests state-of-the-art models still struggle to generate non-aerial speech profiles. Another possible avenue involves purposefully degrading AC signals to make them more similar to BC, and using these generated signals to further train the audio encoder to increase its generalisation capabilities.

Model training Alternative training regimes may improve performance further than simple full-model fine tuning followed by training from scratch. Low-rank fine-tuning, as well as more advanced techniques such as knowledge distillation from a more capable model (including potentially from a textual language model) could unlock further performance benefits without harming the generalisation ability of the model.

Model analysis Understanding the inner workings of the model may support its development and debugging. For example, the application of interpretability techniques to understand which frequency bands are more relevant could indicate whether the model prioritises higher frequencies, as hypothesised; and the importance of specific components inside the network could show points for saving compute or storage, if implemented.

6. Acknowledgements

This work has been developed as part of an extracurricular internship at the MEMS Software

Solutions Team at STMicroelectronics, who the author gratefully thanks for their support

References

[1] Ph D. Grant Fairbanks. *Voice And Articulation Drillbook*. 1940.

[2] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Interspeech 2019*, pages 814–818. ISCA, September 2019.

[3] Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-End Speech Recognition: A Survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351, 2024.

[4] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A General-Purpose Speech Toolkit, June 2021.