Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA19102 – Language In The Human-Machine Era

Grantee name: Chloe Patman

Details of the STSM

Title: The impact of face coverings on speech processing.

Start and end date: 14/08/2023 to 30/09/2023

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

Phase 1. This phase consisted of learning how to use two automatic speech recognition (ASR) systems (Whisper and wav2vec 2.0). Initially, the STSM outlined that Kaldi would be used as a third system. Given Kaldi's relatively greater complexity and lower performance for English, the present study focused primarily on the performance of Whisper and wav2vec 2.0.

Phase 2. This phase involved better understanding the key components of the scripts used to run the default implementations of Whisper and wav2vec 2.0. Tutorial slides were produced, explaining the key elements of the default models in a manner accessible for linguists.

In producing the tutorial, time was spent investigating what pretrained models and fine-tuning options were available. For example, initially, Whisper was detecting the language of the test cotton mask data as Welsh. Once the script was modified to detect the English language explicitly, accuracy improved.

Phase 3. The final phase involved testing the accuracy of the ASR systems and humans in transcribing face mask data.

The dataset consisted of three Standard Southern British English speakers producing a total of 40 sentences in two mask conditions (Cotton Mask and No Mask). The stimuli were presented in two noise levels crossed with two noise types: 0 dB babble, 0 dB speech shaped noise (SSN), 8 dB babble, and 8 dB SSN. The ASR systems transcribed all the recordings in all the noise conditions. The data was also tested on humans (n = 60) where a between-subjects design was used across the noise level and a within-subjects design was used for the noise type.

The performance of the ASR systems and human listeners was analysed via a word error rate (WER) calculation which is industry standard for ASR technologies. WER was calculated using the word-level normalised Levenshtein distance, dividing the total number of substitutions, insertions and deletions per utterance by the expected number of words per utterance. A mixed effects linear regression model was then run. The best model included four main predictor variables: the transcriber (Whisper, wav2vec 2.0, and humans), mask condition (cotton mask and no mask), noise type (babble vs SSN) and noise level

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

(0 dB vs 8 dB) with interactions between all the variables. A random intercept was included for sentence and for speaker.

Due to space constraints, this summary describes only the significant effects of each main predictor (significance assessed at $p < 0.01$), though significant interactions were also observed. Firstly, the cotton mask condition corresponded to a significant increase in WER. Next, both the babble and the 0 dB noise conditions corresponded to a significant increase in WER compared to the average. Finally, the humans significantly outperformed both ASR systems and Whisper significantly outperformed wav2vec 2.0.

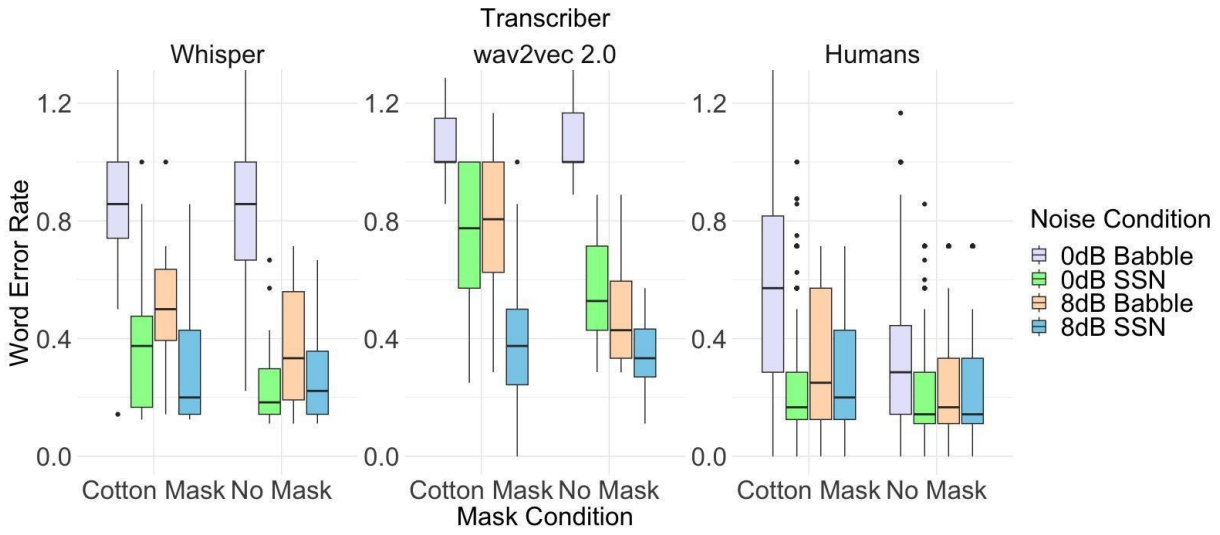

Figure 1. The distribution of WER across stimuli according to mask condition, noise type and noise level and the transcriber.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

The first goal of this STSM was to acquire higher-level knowledge of the methodologies and modelling used in computational linguistics. This goal has been achieved over the STSM as I have learnt how to use two state-of-the-art automatic speech recognition systems and analyse the output data on novel data. In learning how to use these two systems, time was spent breaking down the key components used in default implementation of these scripts. From this, a tutorial presentation was produced which aims to help explain the key components of Whisper and wav2vec 2.0 in a manner accessible for linguists. The aim of this tutorial was to provide other linguists with core knowledge of how the models work and the pretrained data available. This core knowledge is important for achieving efficient processing depending on the data analysed.

The second goal of the STSM was to investigate how face coverings affect speech recognition accuracy for these systems. This goal has been achieved as we designed an experiment to test the systems' performance under different mask and noise conditions. I also learned how to quantify the performance of these systems based on industry standards (e.g., WER). This means conducting a word-by-word comparison (Levenshtein distance) of the output transcript and reference transcript, calculating the number of substitutions, deletions and insertions. The number of deviations from the reference transcript is then divided by the number of words in the utterance, providing the WER. After producing a WER per utterance, a mixed effects linear regression model was performed, comparing the performance of the machine and human systems according to the mask type and noise level/types.

The final part of this STSM was to contribute to two LITHME working groups. This goal will be achieved as I am scheduled to present my work to WG1 and WG2 on the 10th November 2023. Here I will share

the knowledge acquired during the STSM and explain the relevance of this work within the field of forensic speech science and computational linguistics.

Professor Eleanor Chodroff and I plan to write an article outlining the findings of this STSM.