ELEC5305 Project Proposal

Project Title

Effects of Frequency Band Compression on Speech Intelligibility and Keyword Recognition Accuracy

Student Information

Full Name:Jianxiang Chen

Student ID (SID): 540062964

GitHub Username: cjx259

GitHub Project Link: https://cjx259-au.github.io/elec5305-project-540062964/

Project Overview

This project investigates how frequency band compression and filtering affect both speech intelligibility and the accuracy of automatic keyword recognition. Many practical channels (e.g., telephony) limit speech bandwidth, which can degrade intelligibility and recognition performance. Understanding which frequency regions are most critical helps designers build more robust systems for low-bitrate communications and embedded devices.

Hypothesis. Preserving the 300–3400 Hz “narrowband telephony” region will maintain recognition accuracy relatively well versus aggressive low-pass or high-pass filtering, whereas removing sub-bands inside this region will cause measurable degradation. We will validate this with controlled band-pass and band-stop conditions and two evaluation protocols (robustness and adaptation). (ITU)

Background and Motivation

Classic speech science identifies mid-band energy as central to intelligibility; however, modern keyword spotting (KWS) models are typically trained on full-band 16 kHz audio and may be brittle under bandwidth constraints. By quantifying recognition changes across precisely defined frequency manipulations, this study bridges perception-oriented insights with lightweight machine learning pipelines that are feasible in a single semester.

Proposed Methodology

Dataset

We use Google Speech Commands v0.02, a public KWS dataset of ~1-second utterances sampled at 16 kHz (~105,829 WAV files across core and auxiliary words) under CC-BY-4.0. We focus on the eight core keywords (*yes, no, up, down, left, right, stop, go*) and optionally include *silence/unknown* for robustness. The official validation_list.txt and testing_list.txt files will be used to create train/val/test splits with no speaker overlap. To control confounds, we balance classes by downsampling the over-represented ones, and we avoid any augmentation that changes frequency content (no pitch-shift or time-stretch) in the main experiments.

Pre-processing

All audio is mono at 16 kHz. We RMS-normalize each clip to a consistent loudness range and trim leading/trailing silence with an energy threshold (but keep at least 0.9 s total duration). For time alignment invariance, we allow a small random time shift (±50 ms) during training; this does not affect the spectrum and keeps the study focused on frequency effects

Signal processing conditions (MATLAB). We will synthesize evaluation conditions with:

Low-pass: cutoff 3.4 kHz;

High-pass: cutoff 300 Hz;

Band-pass: 300–3400 Hz;

Band-stop ablations: center frequencies at 1/2/3 kHz with bandwidths 400/600/800 Hz (grid), to test which sub-bands matter most.

Filters will be 8th-order IIR (Butterworth) designed via designfilt and applied using zero-phase filtfilt to avoid phase distortion artifacts. (MathWorks)

Features and models. We will extract MFCCs (mfcc) and/or Mel-spectrograms (melSpectrogram). Two classifier families will be compared: (1) classical ML—SVM/ECOC (fitcecoc) trained on MFCCs; (2) a tiny CNN (3–4 conv layers) operating on Mel-spectrogram “images” for an apples-to-apples feature comparison. (MathWorks)

Evaluation protocols.

Robustness test: Train on original (full-band) data only; evaluate the same model across all filtered conditions to measure bandwidth mismatch robustness.

Adaptation test (optional): Train on a mixed dataset (original + 300–3400 Hz band-pass) and re-evaluate to quantify potential gains under constrained bandwidth deployment.

Metrics and analysis.

Primary: Top-1 accuracy and confusion matrices per condition; report mean ± 95% CI across multiple random seeds.

Secondary: segmental SNR, in-band energy ratios (computed from STFT/Mel spectrograms), and MFCC distances between filtered/unfiltered versions for objective change quantification.

Qualitative: Spectrogram visualizations to illustrate removed/retained bands; error analysis highlighting typical confusions that increase under specific ablations.

Feasibility. All steps use standard MATLAB toolboxes (Signal Processing, Audio, Statistics/ML, and optionally Deep Learning). The pipelines are computationally light (minutes to a few hours on a single workstation) and reproducible in scripts. (MathWorks)

Expected Outcomes

I expect (i) accuracy curves across conditions that isolate the role of specific frequency regions; (ii) clear visual evidence (spectrograms) of what is removed; (iii) a comparative analysis establishing whether 300–3400 Hz retention provides a strong operating point and which internal sub-bands are most critical; and (iv) practical guidance for low-bandwidth KWS deployments. Findings will be tied to long-standing telephony practice and recent trends in robust speech processing. (ITU)

Timeline (Weeks 6–13)

Weeks 6–7: Literature review; implement baseline KWS (original data), establish accuracy and code structure.

Weeks 8–9: Implement all filters and generate processed sets; run robustness tests; produce preliminary plots.

Weeks 10–11: Run adaptation tests; add ablation granularity (multiple band-stop widths); finalize error analysis and confidence intervals.

Weeks 12–13: Final results and writing; prepare figures/tables; complete GitHub repo and enable GitHub Pages with a public project overview.

References (updated with ISBN/DOI/links)

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice (2nd ed.)*. CRC Press. Print ISBN: 978-1466504219; eBook ISBN: 978-0429096181; DOI:

https://doi.org/10.1201/b14529; Publisher page: https://www.taylorfrancis.com/books/mono/10.1201/b14529/speech-enhancement-philipos-loizou (Taylor & Francis)

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of Speech Recognition*. PTR Prentice Hall.
ISBN-10: 0130151572; ISBN-13: 978-0130151575; Catalog page: https://openlibrary.org/books/OL1729705M/Fundamentals_of_speech_recognition (开放图书馆)

Reddy, C. K. A., Dubey, H., Cutler, R., et al. (2021). “INTERSPEECH 2021 Deep Noise Suppression Challenge.” *INTERSPEECH 2021*.
ISCA paper: https://www.isca-archive.org/interspeech_2021/reddy21_interspeech.pdf ; arXiv: https://arxiv.org/abs/2101.01902 (ISCA Archive, arXiv)

Warden, P. (2018). “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition.” *arXiv:1804.03209*.
Paper: https://arxiv.org/abs/1804.03209 ; Dataset card (v0.02, CC-BY-4.0, counts): https://huggingface.co/datasets/google/speech_commands (arXiv, Hugging Face)

ITU-T P.342. “Transmission characteristics for telephone band (300–3400 Hz) terminals.”
Recommendation page: https://www.itu.int/rec/T-REC-P.342 (ITU)