A Study of Salient Modulation Domain Features for Speaker Identification

Simon W. McKnight, Aidan O. T. Hogg, Vincent W. Neo, Patrick A. Naylor

Imperial College London

Speaker Identification

- 1-of-N problem
  - N training speakers, test speaker identified from those
  - closed-set here
- Single frame (e.g. MFCCs) v multiple frames (e.g. modulation spectrum features, x-vectors)

Modulation Spectrum

- Previous research identified 1-16 Hz range of temporal envelope modulation frequencies as containing most useful linguistic information about speech for automatic speech recognition (ASR)
- Temporal envelope v temporal fine structure (instantaneous frequencies)
  - more recent research on cochlear implants showed importance of latter for speaker identification

Imperial College London

Generating Modulation Spectrum Features

- 4-step process:
  - Stage 1: first stage of short-time Fourier transform (STFT)
  - Stage 2: either take amplitudes or use Hilbert transform envelope
  - Stage 3: second stage of STFTs, done in acoustic frequency bands
  - Stage 4: either take amplitudes or use Hilbert transform – latter necessary for instantaneous frequencies
- Reconstruct speech signal using constant overlap-add (COLA) without specific acoustic and/or modulation frequencies to see what effects are: https://swm1718.github.io/ModulationSpectrumAudio/
- See Figure 1 of paper for detailed diagram
- This research focused on 1 second modulation frames with 250 ms steps on top of wideband acoustic frames 3 ms long with 1 ms steps
  - 25 acoustic frequency bands up to 8 kHz
  - 501 modulation frequency bands up to 500 Hz

$$\boldsymbol{\Phi}(l) \in \mathbb{R}^{25 \times 501}$$
$$\boldsymbol{\Phi} = \text{modulation spectrum}$$
$$l = \text{modulation frame}$$

Imperial College London

Data

- TIMIT
  - Designed for ASR, but useful as arranged by speaker and utterance
  - 630 speakers
  - 10 utterances of ~3 seconds per speaker, 2 same for all speakers (SA1 and SA2), sampled at 16 kHz
  - Train set has 468 speakers, test set 162 – used in correlation analysis
  - When machine learning models, rearranged so train and test sets both comprise 630 speakers, with first 7 utterances in train set (inc. SA1, SA2) and last 3 utterances in test set

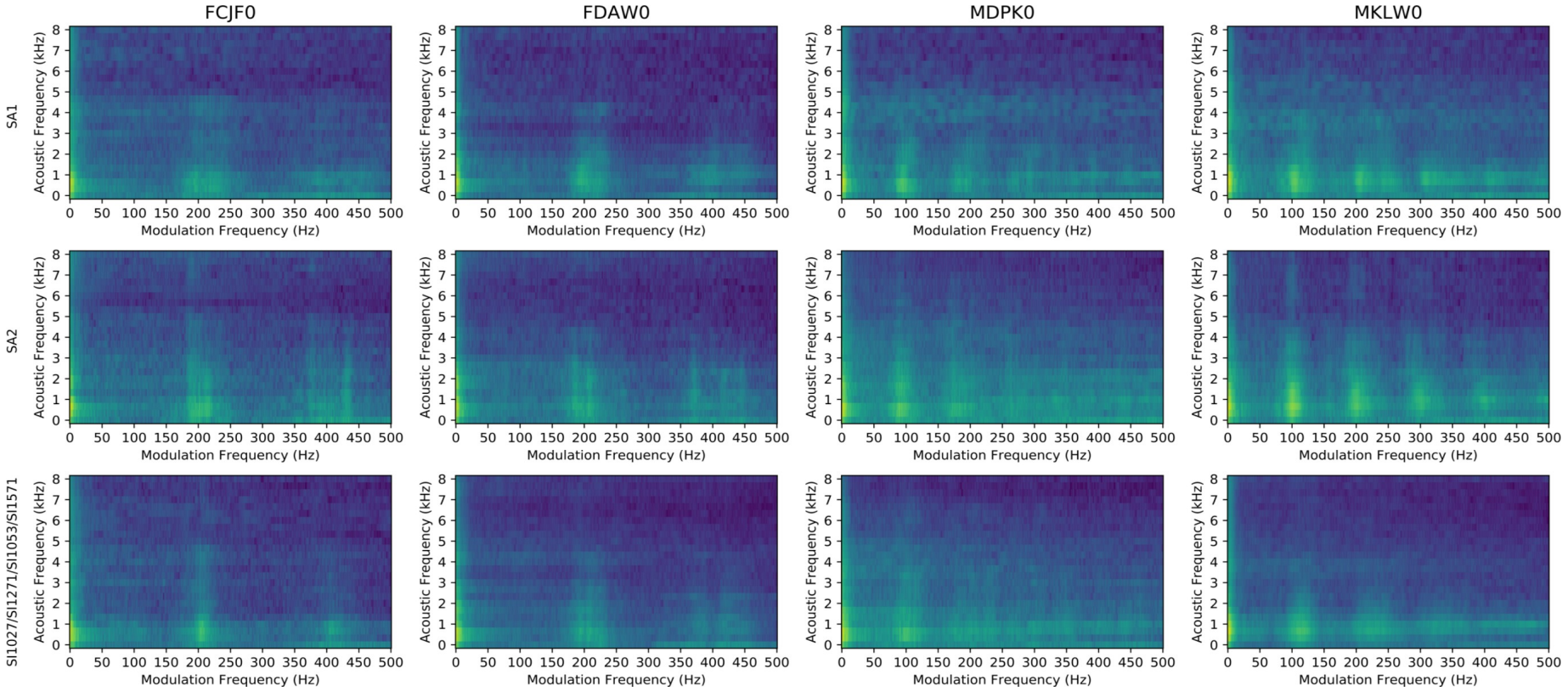
Methods Used

- Correlations
  - Spearman's rank among input features
  - One-way ANOVA for each input feature v output speaker
- Feature importances from random forest models
- Convolutional neural network (CNN) models
- Reconstructed speech signals with specific acoustic and/or modulation frequencies removed

Imperial College London

These are based on the amplitude envelope modulation spectrum

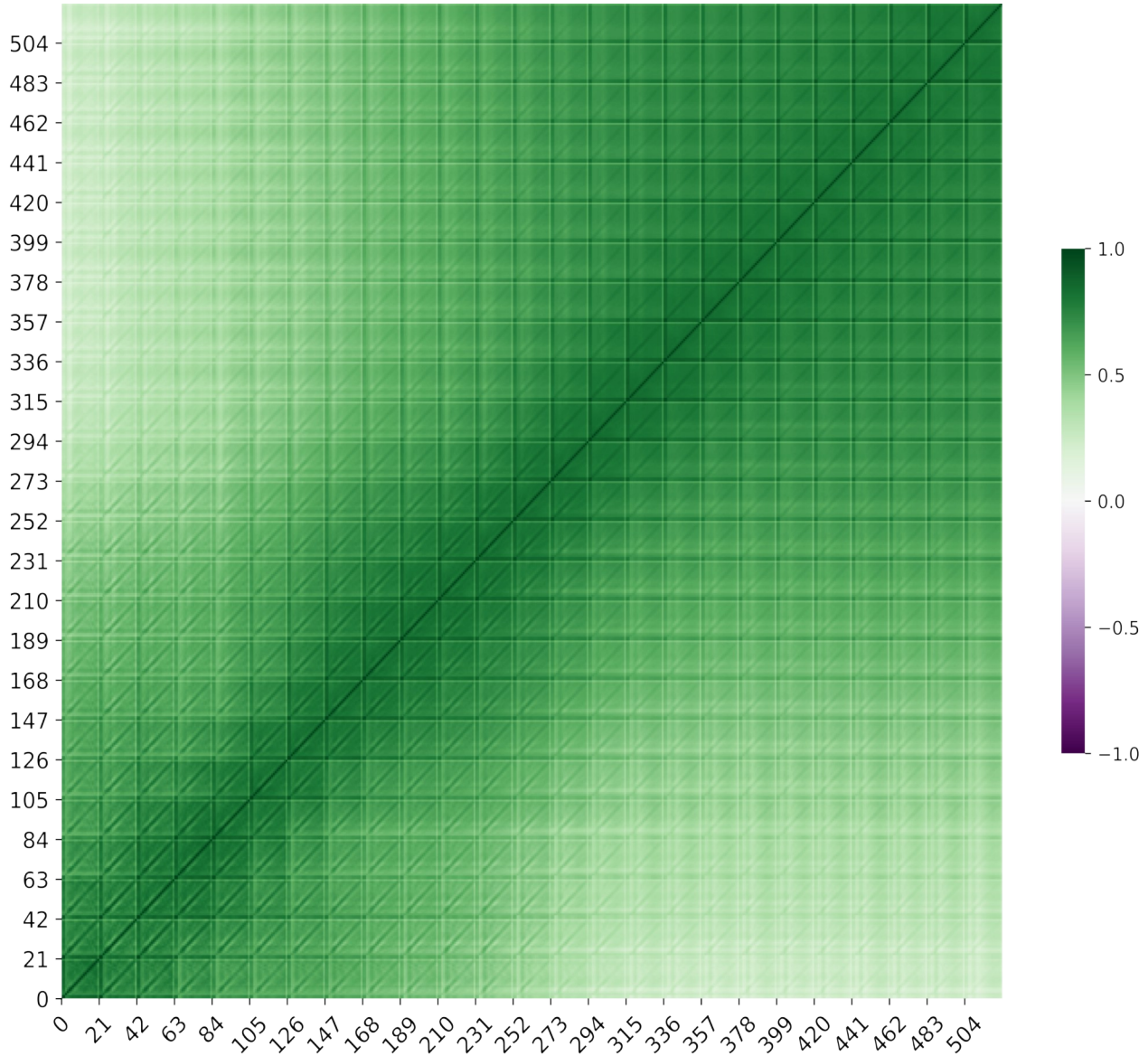
Average Modulation Frame per Speaker/Meeting



Imperial College London

Spearman's Rank Between Feature Correlations

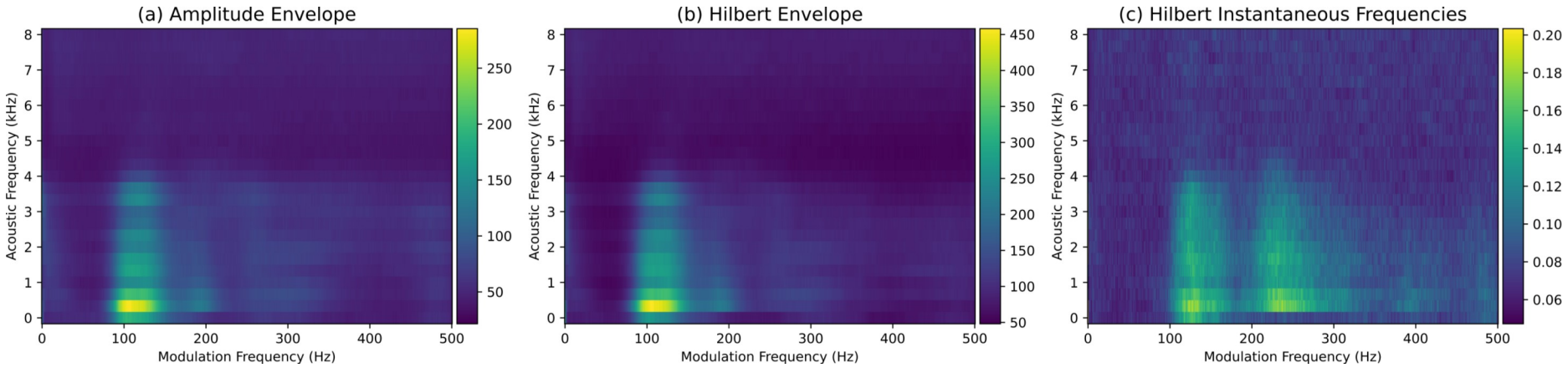
- Wideband acoustic frames
- For 0-20 Hz modulation frequencies
- Flattened to 25 x 21 = 525 features



Imperial College London

$$F_s = \frac{\text{between} - \text{speaker} - \text{means covariance}}{\text{intra} - \text{speaker covariance}}$$

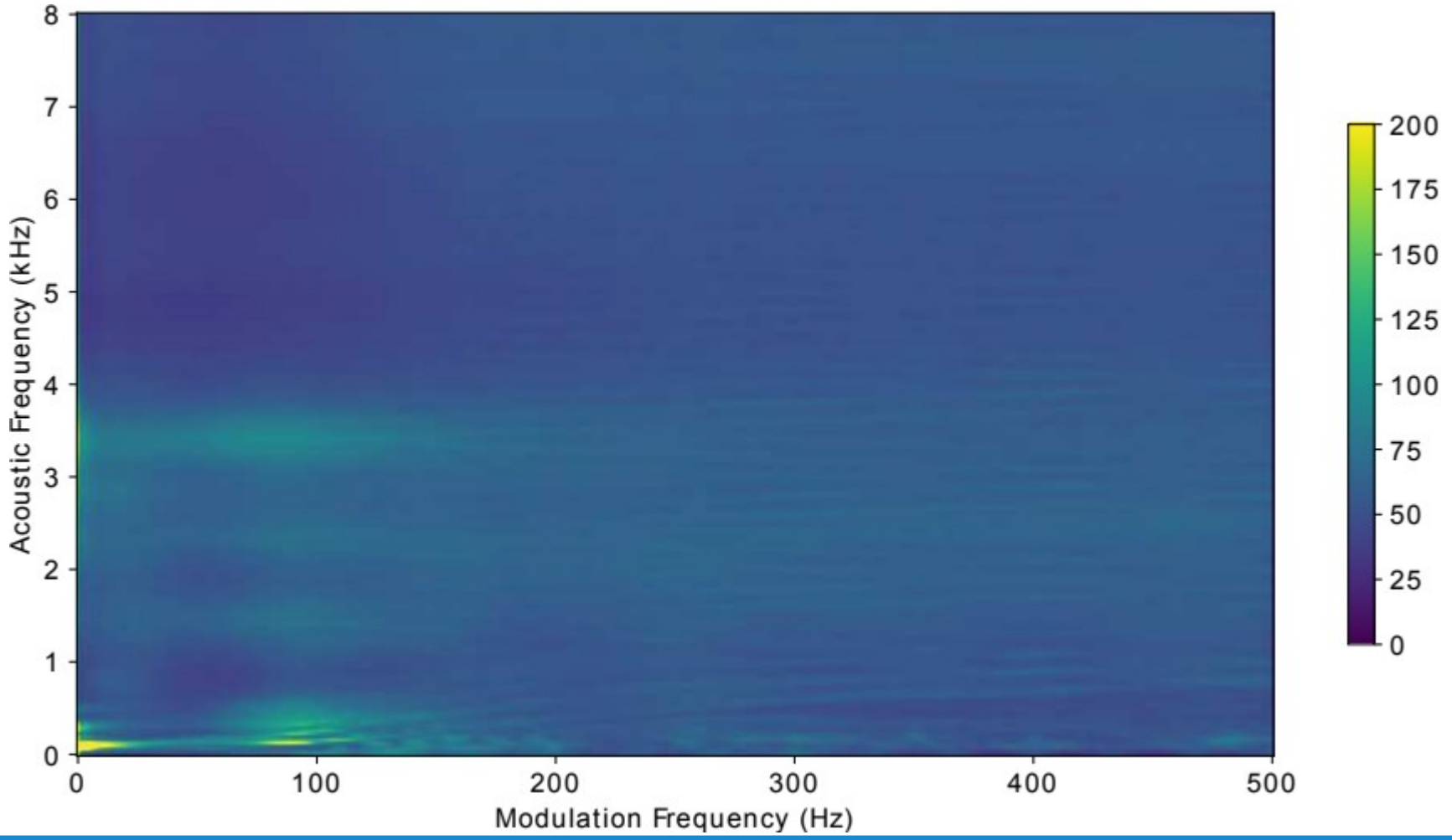
Wideband One-Way ANOVA Correlations



- Based on original TIMIT training set of 462 speakers
- For temporal envelope, first two graphs show strong peak around male fundamental frequencies, with stronger correlation values and slightly lower frequencies for Hilbert envelope
- For temporal fine structure, third graph shows less strong peaks at higher frequencies, so not great on their own but may provide additional information

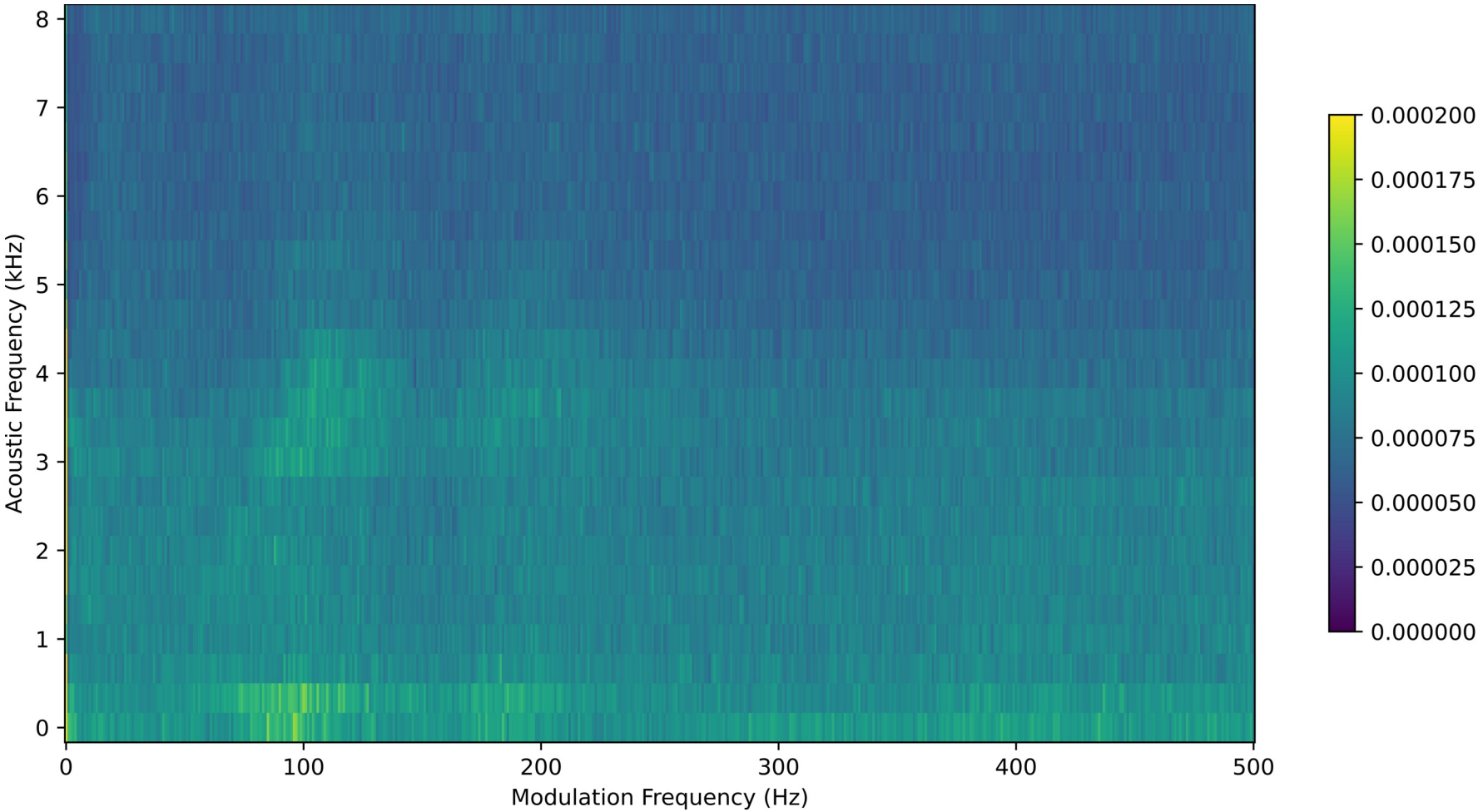
Imperial College London

Narrowband ANOVA Correlations for Amp. Env.



Imperial College London

Feature Importances from Random Forest



Imperial College London

Fitting Random Forest and CNN Models

	Mode	Mean	
	Per MF	**Per Utt.**	**Ave. MF**
RF $\mathbf{\Phi}_{AE}$	12.34	27.63	26.20
CNN $\mathbf{\Phi}_{AE}$	29.03	42.40	26.36
CNN $\mathbf{\Phi}_{HE}$	27.97	48.39	32.77
CNN $\mathbf{\Phi}_{IF}$	5.75	12.20	0.69
CNN $\mathbf{\Phi}_{HE}$ and $\mathbf{\Phi}_{IF}$	31.05	49.26	32.17

MF = modulation frame
Utt. = utterance
Ave. = average
RF = random forest

CNN = convolutional neural network
Φ = modulation spectrum features
AE = amplitude envelope
HE = Hilbert envelope
IF = instantaneous frequency

Imperial College London

Conclusion

- Range of modulation frequencies associated with the fundamental frequency is more important than the 1-16 Hz range most commonly used in automatic speech recognition
- 0 Hz modulation frequency band contains significant speaker information
- Temporal envelope more discriminative among speakers than temporal fine structure, but temporal fine structure still contains useful additional information for speaker identification

Next steps

- See if using filterbanks and discrete cosine transforms (DCTs) in acoustic and modulation domains improve performance
- Test whether modulation spectrum features give as good results as single frame MFCCs and whether using both together improves performance