Engineering and
Physical Sciences
Research Council

Combining Acoustic, Phonetic, Linguistic and Audiometric data in an Intrusive Intelligibility Metric for Hearing-Impaired Listeners

Mark Huckvale, Gaston Hilkhuysen

Speech, Hearing and Phonetic Sciences,

University College London

m.huckvale@ucl.ac.uk , g.hilkhuysen@ucl.ac.uk

Approach

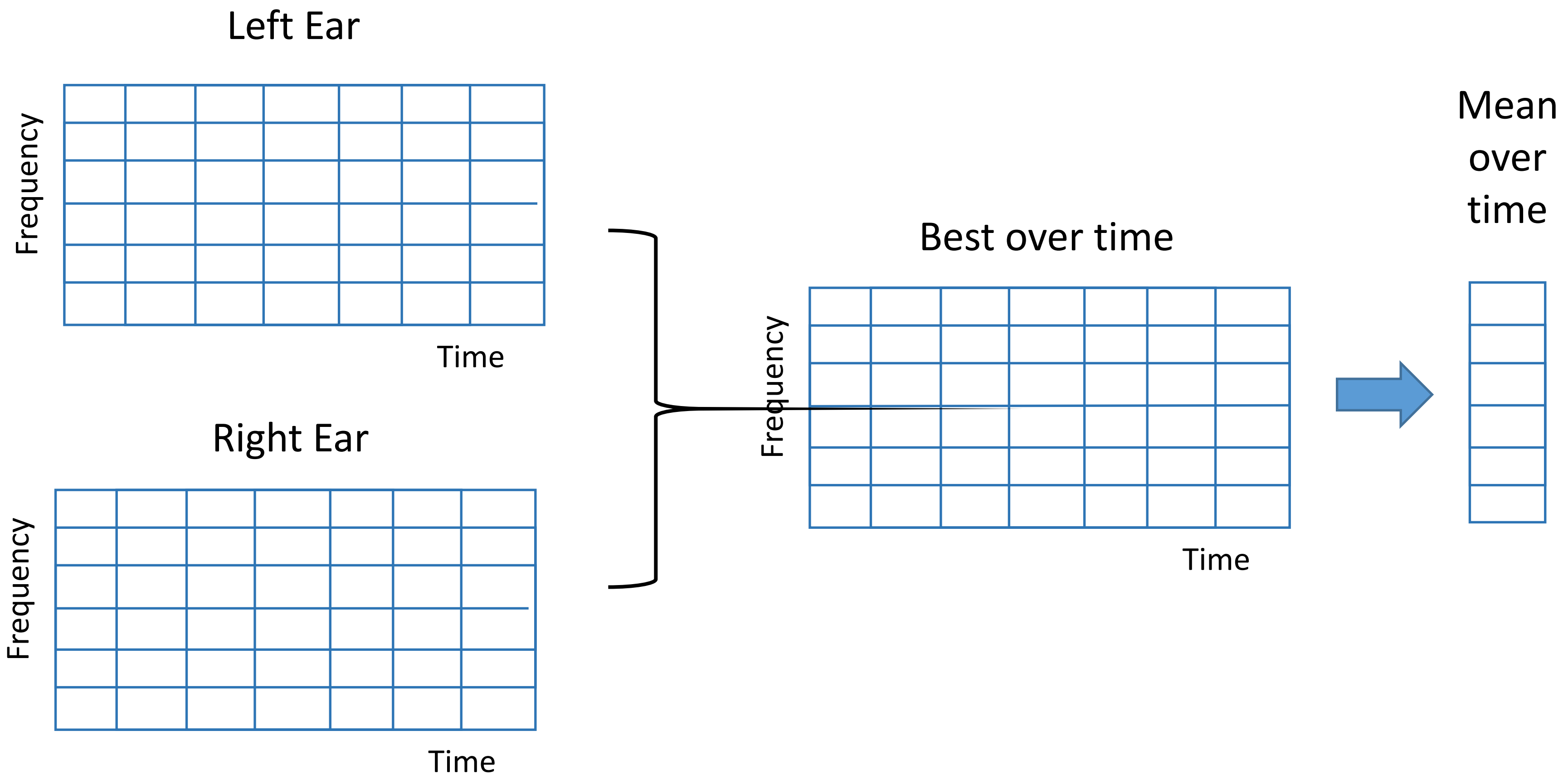
- Build on success of best performing intrusive system in CPC1*
- Determine which signal and metadata features could provide useful information to predict sentence intelligibility to a listener
  - Audio properties: spectrographic changes, enhancement system effects
  - Phonetic properties: phonetic changes, talker identity
  - Linguistic properties: prompt sentence probability
  - Audiometric properties: listener hearing abilities
- Train non-linear regression model to predict intelligibility from features
- Use greedy feature set selection to find best combination of features

*Huckvale & Hilkhuysen, ELO-SPHERES intelligibility prediction model for the Clarity Prediction Challenge 2022, Interspeech 2022

Features used for prediction

Type	Feature set	# Feat.	Description
ACOUSTIC	STOI2EAR	15	STOI correlations between source and processed audio in better ear over time, one correlation per filter channel
ACOUSTIC	SYSTEM	20	Predicted identity of the processing system found by a system classifier, one probability per system
PHONETIC	LATTICE	15	Phone lattice correlations, one correlation per VPM feature
PHONETIC	TALKER	6	Predicted identity of the talker of the sentence found by a scene classifier, one probability per talker
LINGUISTIC	SPROB	11	Sentence probability from language model, and number of words in prompt
AUDIOMETRIC	PTA	8	Average pure-tone thresholds at 8 frequencies

STOI2EAR – Best ear over time

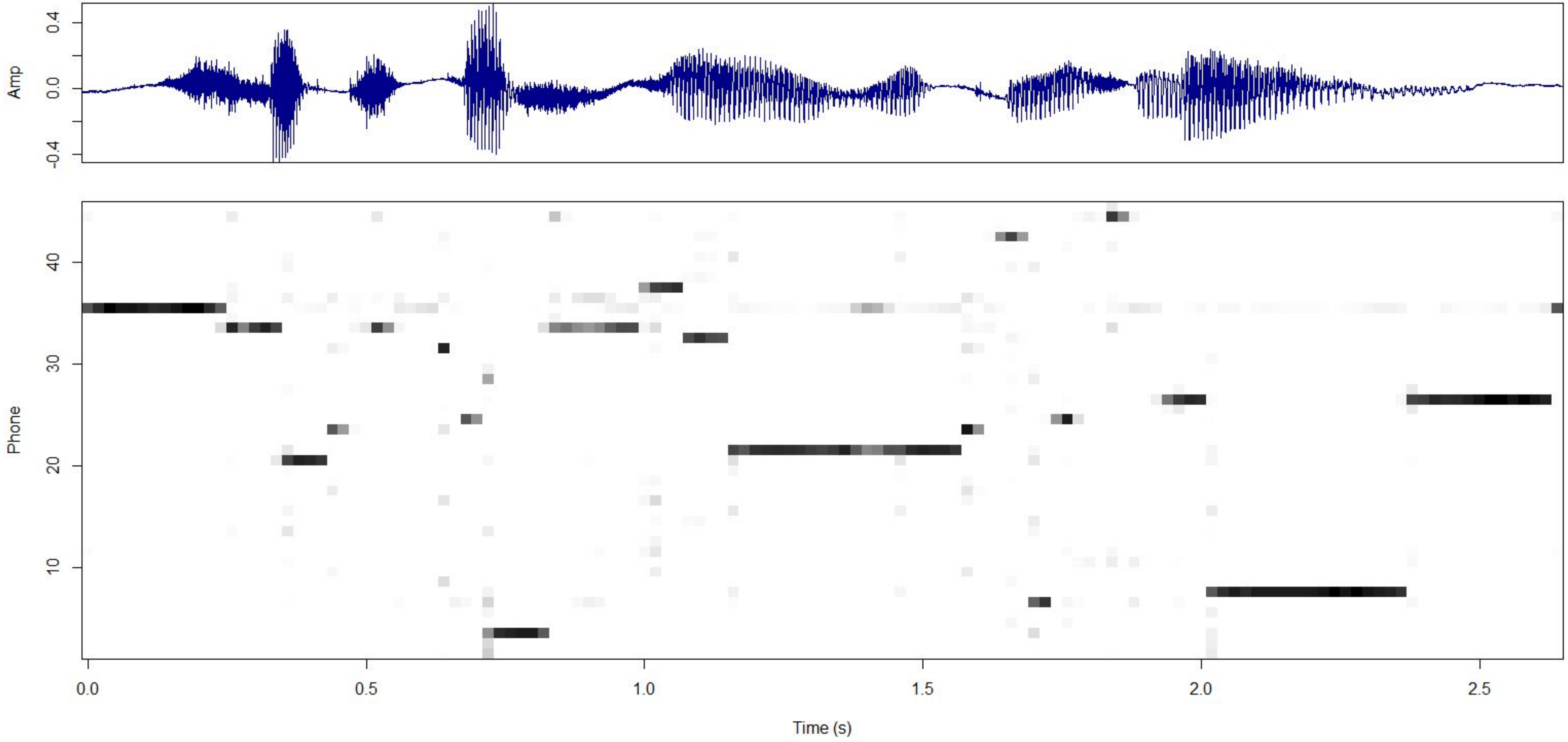


STOI correlations in 15 frequency channels every 12.8ms

Features used for prediction

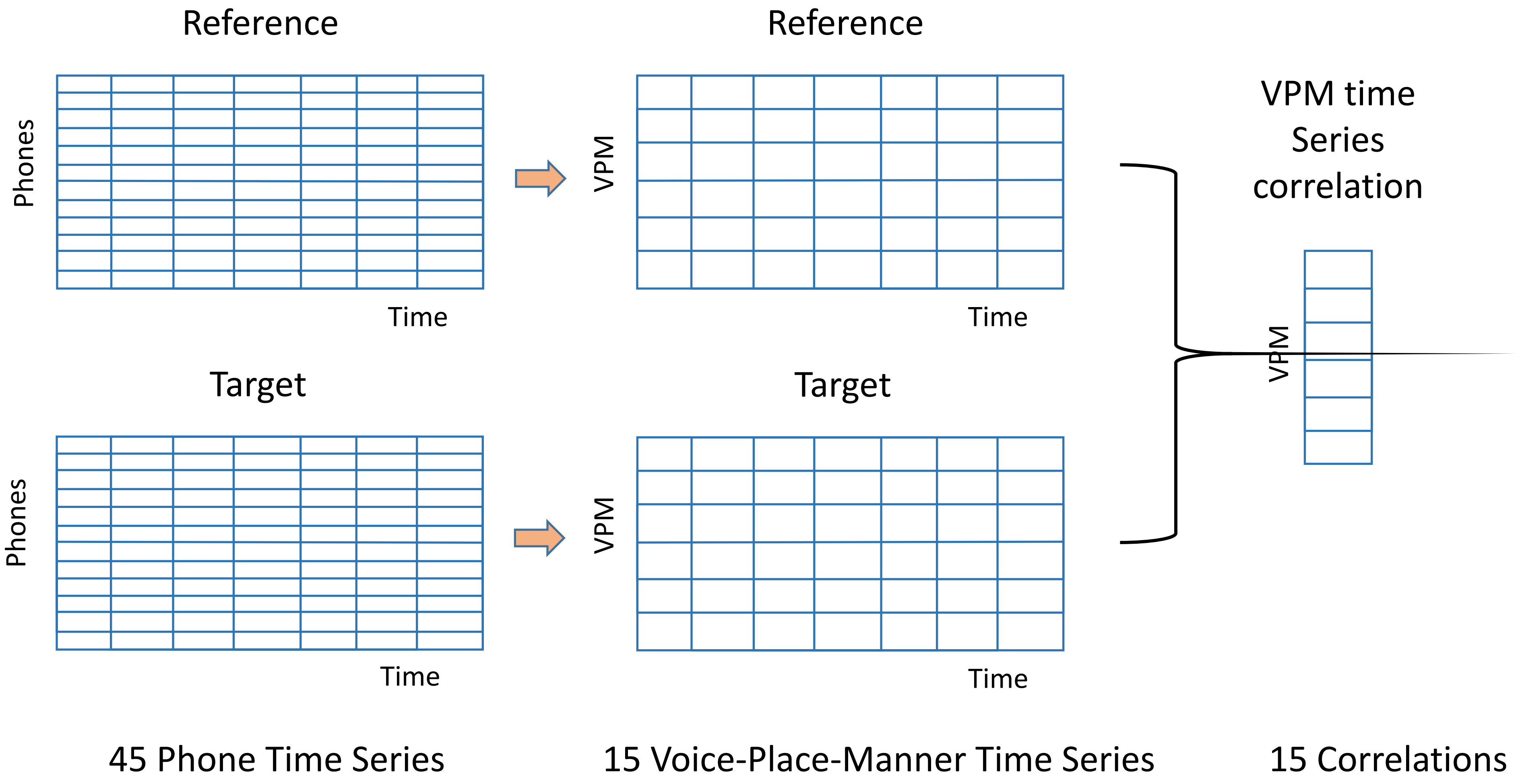
Type	Feature set	# Feat.	Description
ACOUSTIC	STOI2EAR	15	STOI correlations between source and processed audio in better ear over time, one correlation per filter channel
ACOUSTIC	SYSTEM	20	Predicted identity of the processing system found by a system classifier, one probability per system
PHONETIC	LATTICE	15	Phone lattice correlations, one correlation per VPM feature
PHONETIC	TALKER	6	Predicted identity of the talker of the sentence found by a scene classifier, one probability per talker
LINGUISTIC	SPROB	11	Sentence probability from language model, and number of words in prompt
AUDIOMETRIC	PTA	8	Average pure-tone thresholds at 8 frequencies

Phone Lattices: Wav2Vec2 + XLSR + WSJCAM0



['sil s ih k s p l ah s th r iy k w ax l z n ay n sil']

Phone Lattice Correlation

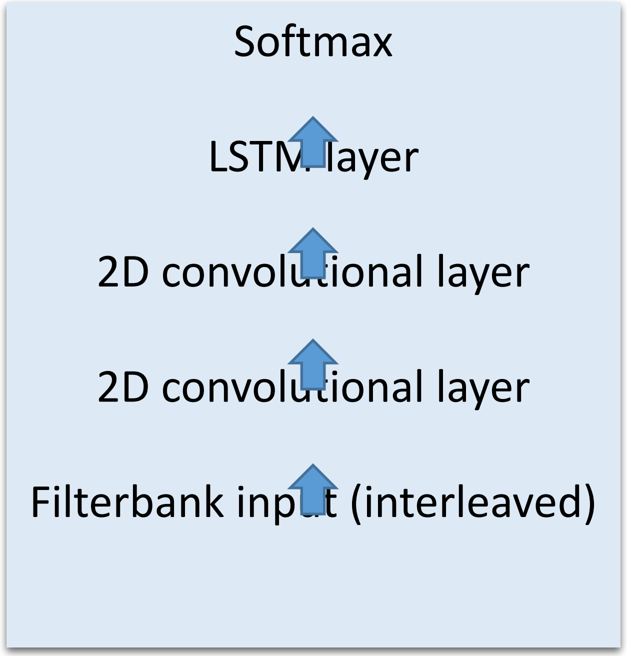


Features used for prediction

Type	Feature set	# Feat.	Description
ACOUSTIC	STOI2EAR	15	STOI correlations between source and processed audio in better ear over time, one correlation per filter channel
ACOUSTIC	SYSTEM	20	Predicted identity of the processing system found by a system classifier, one probability per system
PHONETIC	LATTICE	15	Phone lattice correlations, one correlation per VPM feature
PHONETIC	TALKER	6	Predicted identity of the talker of the sentence found by a scene classifier, one probability per talker
LINGUISTIC	SPROB	11	Sentence probability from language model, and number of words in prompt
AUDIOMETRIC	PTA	8	Average pure-tone thresholds at 8 frequencies

System and Talker classifier

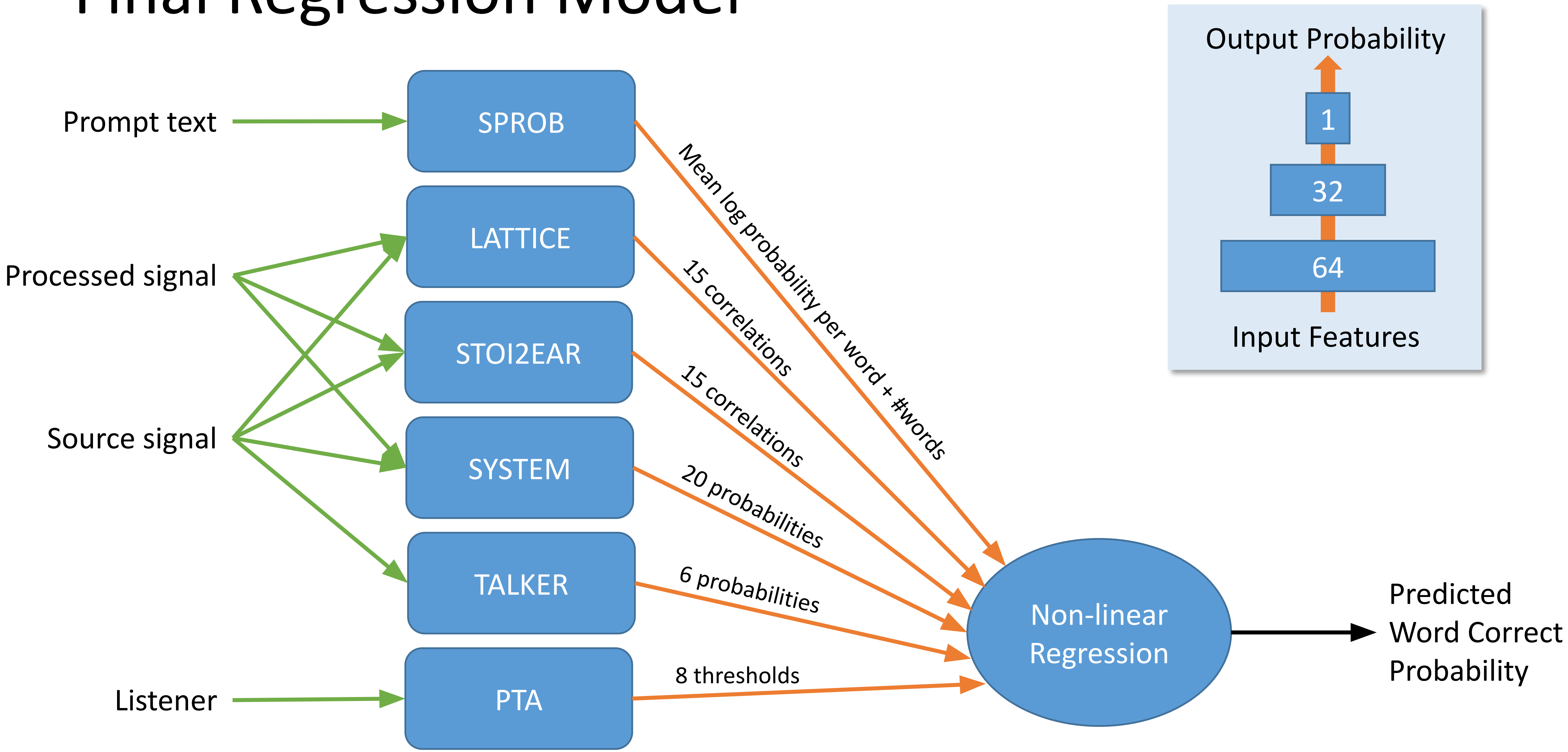
- Aim: to predict scene metadata from audio
- STOI filterbank input
  - Reference and Processed audio
- DNN
  - 2 convolutional layers + LSTM
  - Softmax output
- Train to predict
  - Processing system identity (1 of 20)
  - Talker identity (1 of 6)



Features used for prediction

Type	Feature set	# Feat.	Description
ACOUSTIC	STOI2EAR	15	STOI correlations between source and processed audio in better ear over time, one correlation per filter channel
ACOUSTIC	SYSTEM	20	Predicted identity of the processing system found by a system classifier, one probability per system
PHONETIC	LATTICE	15	Phone lattice correlations, one correlation per VPM feature
PHONETIC	TALKER	6	Predicted identity of the talker of the sentence found by a scene classifier, one probability per talker
LINGUISTIC	SPROB	11	Sentence probability from language model, and number of words in prompt
AUDIOMETRIC	PTA	8	Average pure-tone thresholds at 8 frequencies

Final Regression Model



Results

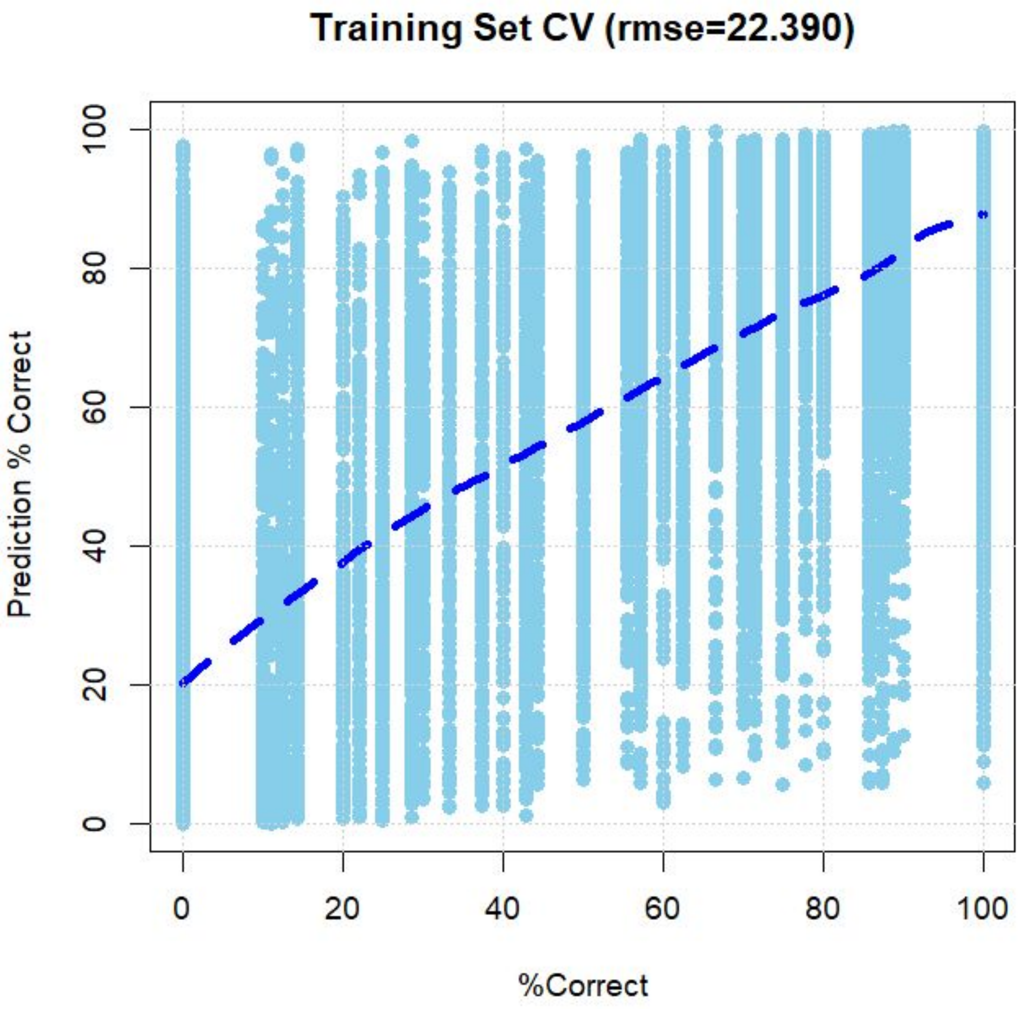
Clarity Baseline with HASPI = 28.584% RMSE
Better Ear STOI = 26.369% RMSE

Feature set	RMS Prediction Error (%)	
	Train (CV)	Test
STOI2EAR alone		
+ LATTICE		
+ SYSTEM		
+ SPROB		
+ PTA		
+ TALKER		

Results

Clarity Baseline with HASPI = 28.584% RMSE
Better Ear STOI = 26.369% RMSE

Feature set	RMS Prediction Error (%)	
	Train (CV)	Test
STOI2EAR alone	25.972	
+ LATTICE	25.344	
+ SYSTEM	23.758	
+ SPROB	23.257	
+ PTA	22.490	
+ TALKER	22.399	



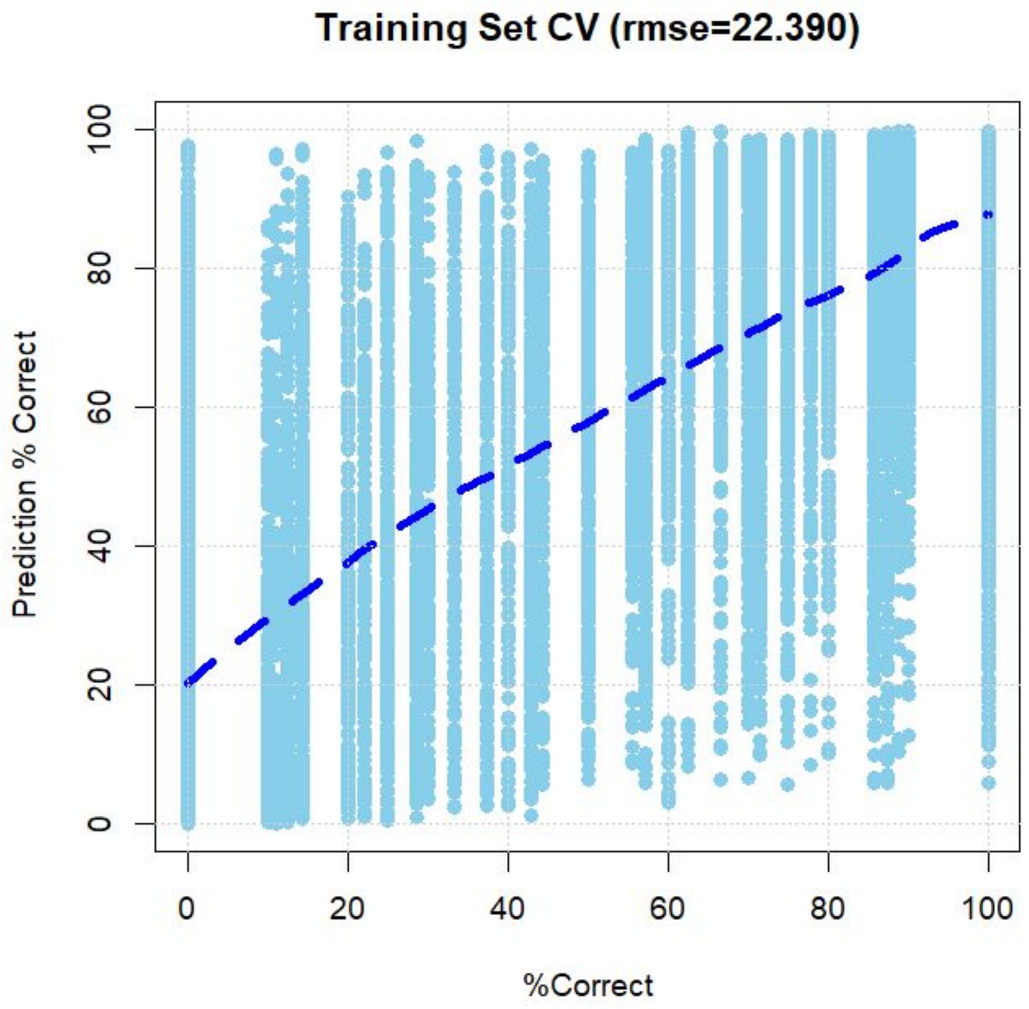
Results

Clarity Baseline with HASPI = 28.584% RMSE
Better Ear STOI = 26.369% RMSE

Feature set	RMS Prediction Error (%)	
	Train (CV)	Test
STOI2EAR alone	25.972	25.793
+ LATTICE	25.344	24.312
+ SYSTEM	23.758	26.570
+ SPROB	23.257	25.895
+ PTA	22.490	24.690
+ TALKER	22.399	24.637

On Train, SYSTEM feature improves RMSE by 1.6%

On Test, SYSTEM feature degrades RMSE by 2.2%



If leave out SYSTEM, Test RMSE=23.133%

Lessons

- SYSTEM feature was key weakness in evaluation containing unseen systems
  - Need to find alternative ways to characterise system behaviour, orthogonal to STOI2EAR and LATTICE
- Still a great deal of unexplained variability
  - Opportunity for investigations into causes of variation
- Listeners could be better characterised
  - Previous work shows audiogram only explains 40% of variability across listeners
  - Need information about listener performance on standardised intelligibility task
- In this method, system and talker were identified separately from audio
  - Systems trained from audio alone could still use this information implicitly