Amplitude Modulation Spectral Analysis: From Conventional Audio Feature Engineering to Speech Foundation Models

Tiago H. Falk
INRS-EMT, MuSAE Lab, University of Québec
INRS-UQO Joint Research Centre on Cybersecurity and Digital Trust

INRS Institut national de la recherche scientifique

Disclaimer:
Material presented here is based on work of dozens of students over the last 15 years, funded by numerous national and international organizations!

What's in the (Speech) Envelope?

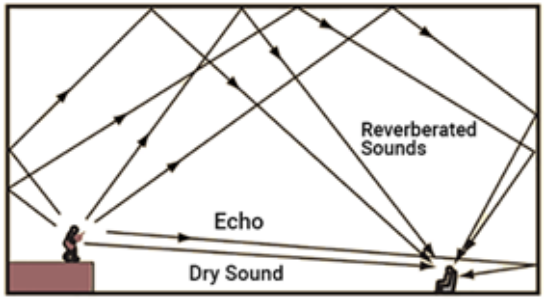
- Slowly varying speech envelopes convey useful information for intelligibility (1939, Dudley)
- Late 70's: Houtgast & Steeneken (room acoustics)
  - Coined the term "modulation spectrum" (MTF)
  - Below 16Hz → important for speech intelligibility
- Mid 90's: Drullman (low/highpass filtering)
  - 2-16Hz → essential for spoken language understanding
- Mid 90's: Hermansky & Morgan
  - RASTA filtering (1-16Hz, emphasis on 4Hz)
  - RASTA-PLP

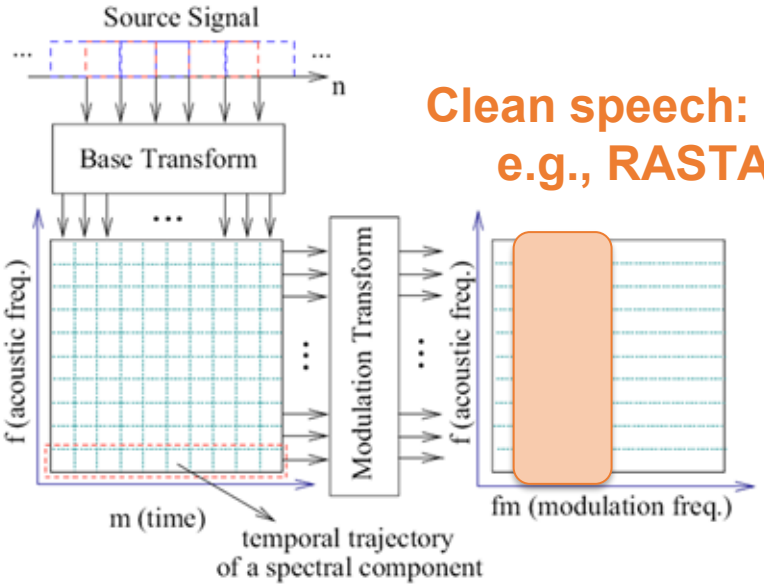
What's in the Envelope (Cont'd)

- Late 90's: Auditory-inspired models, Dau&Kollmeier
  - Modulation "filterbanks"
- Mid 2000s: Spectro-temporal modulation studies on ferrets by Mesgarani & Shamma
- ANIQUE (Kim) and PEMO-Q (Huber & Kollmeier) objective speech quality measurement algorithms
- 2010's-2020's: auditory-inspired models, Falk et al
  - Applications across different fields

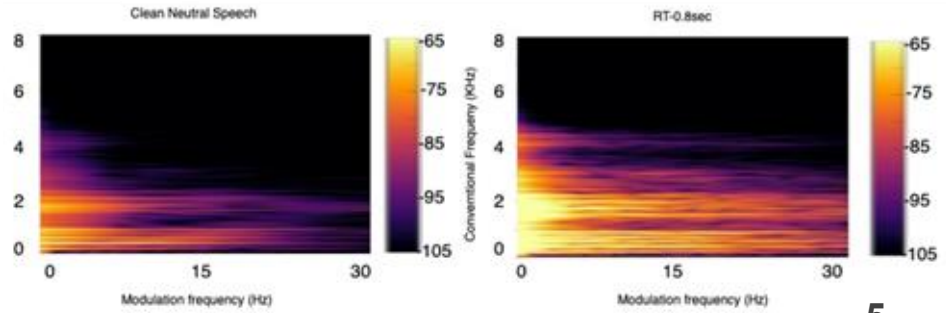
Modulation Spectrum: Analysis-Synthesis

- Spectral analysis of the temporal trajectories of short-time spectral envelopes
  - Analysis and analysis-synthesis models

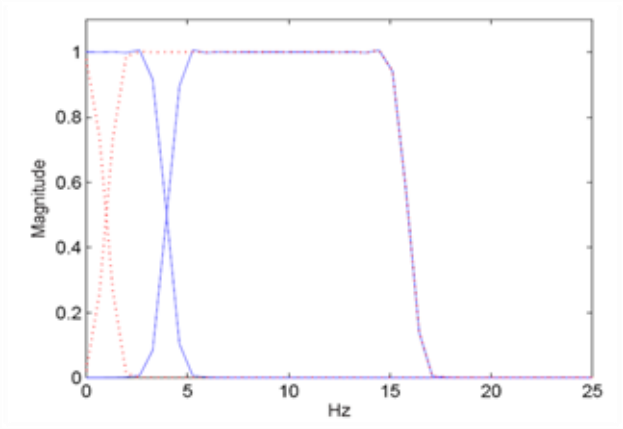


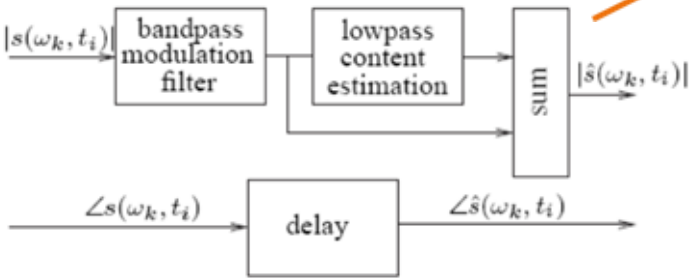


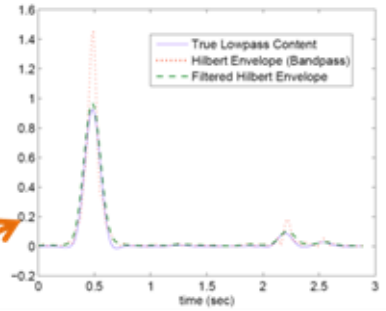
Clean speech: ~ 1-16Hz
e.g., RASTA filter

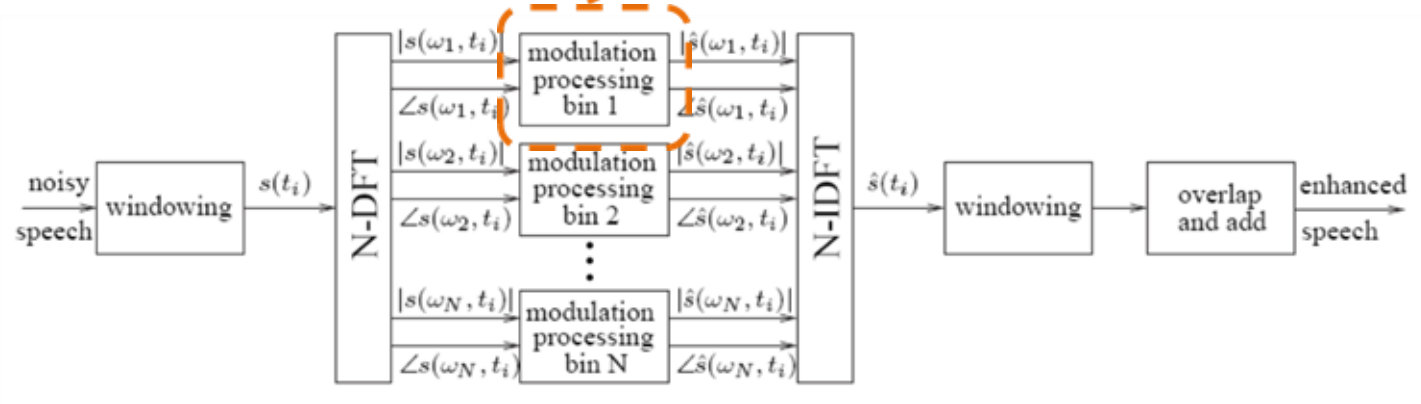


Artificial Bandwidth Expansion









Experiment Results: Noisy/Reverb Speech

Noisy (Plane 5dB)   **Proposed**

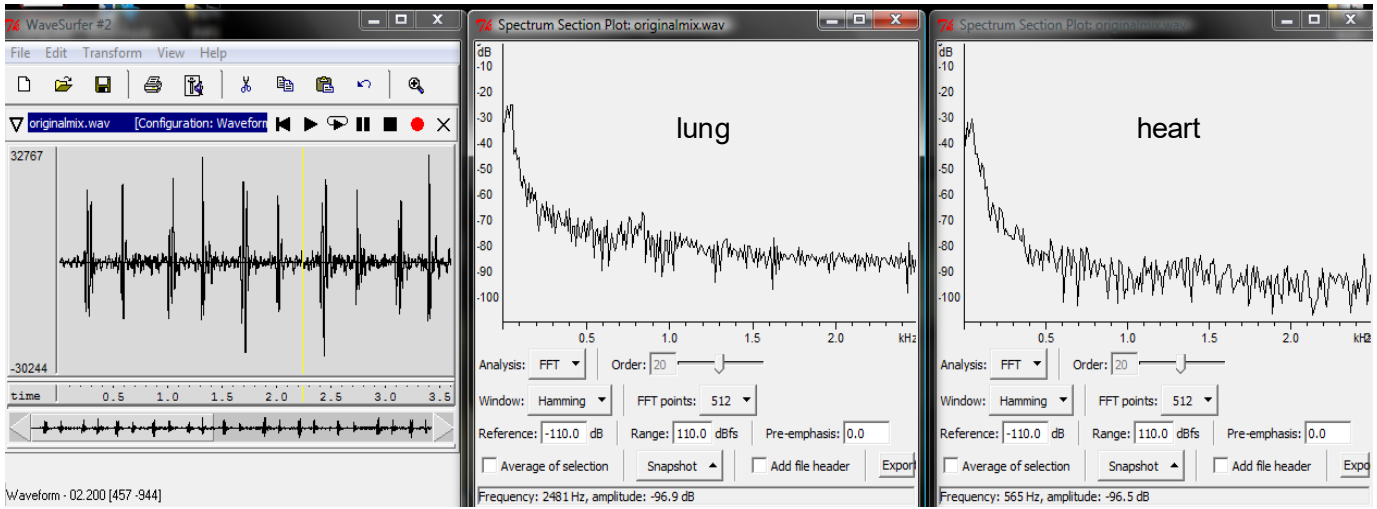
Subjective Test	Noise Type (SNR = 5 dB & 10 dB)		
	Babble (%)	Plane (%)	White (%)
Bandpass	85.00	86.25 	87.50
EVRC	68.75	70.00 	73.25

- Preference percentage of proposed scheme over simple bandpass filtering and EVRC

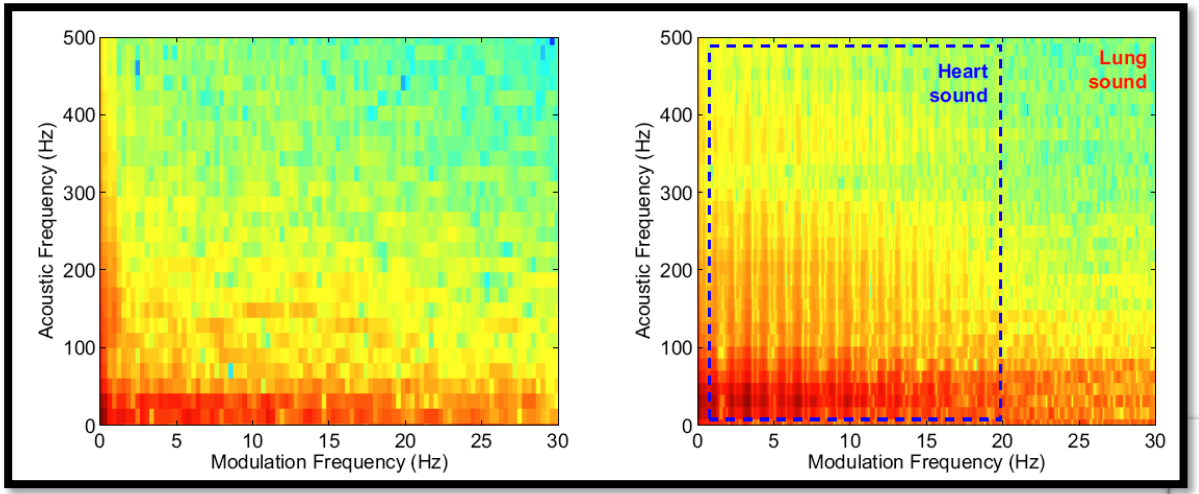
Reverb, RT60 = 1s   **Proposed**

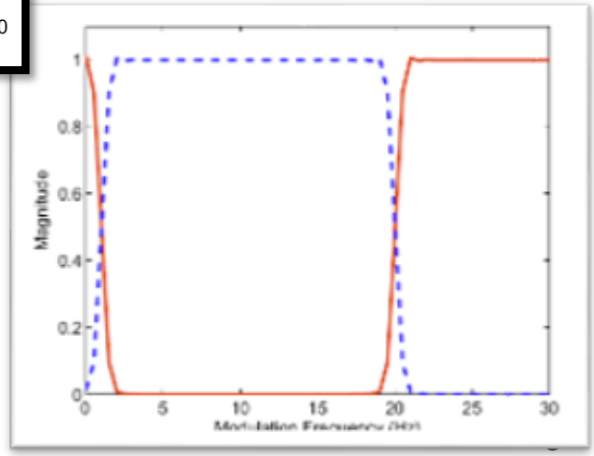
Practical Problem

- Digital Stethoscope
- Measures both heart and lung sounds
  - Sounds overlap in both time and frequency domains

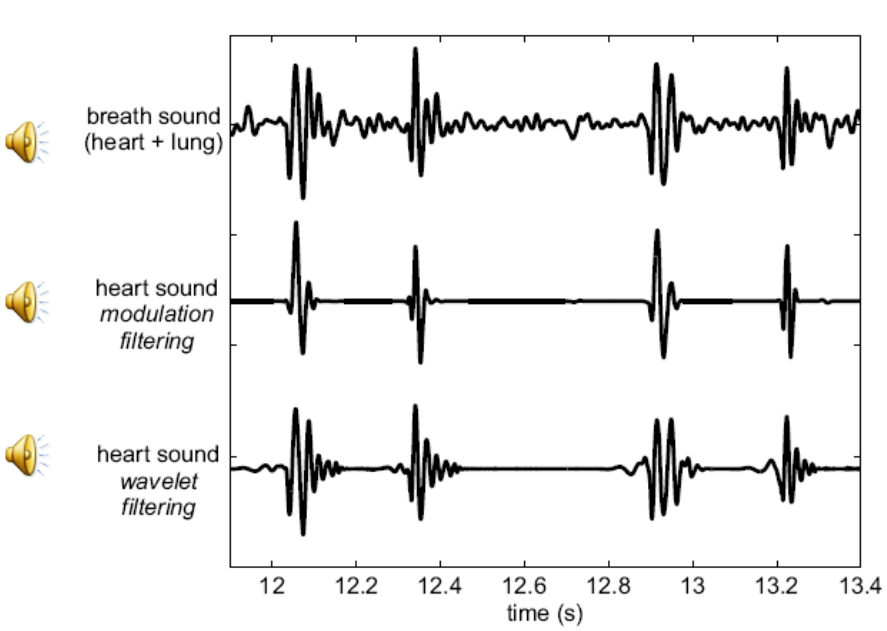


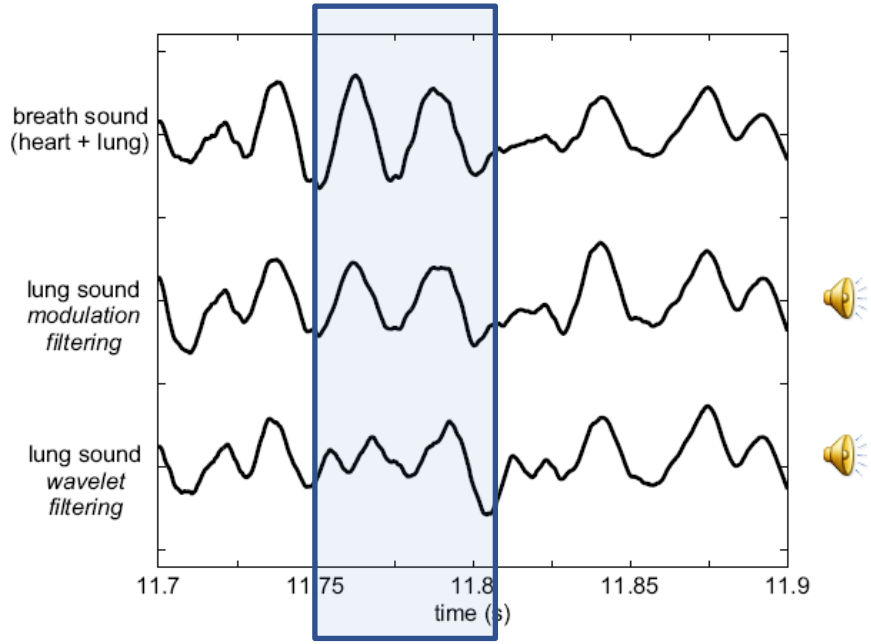
Heart and Lung Sounds





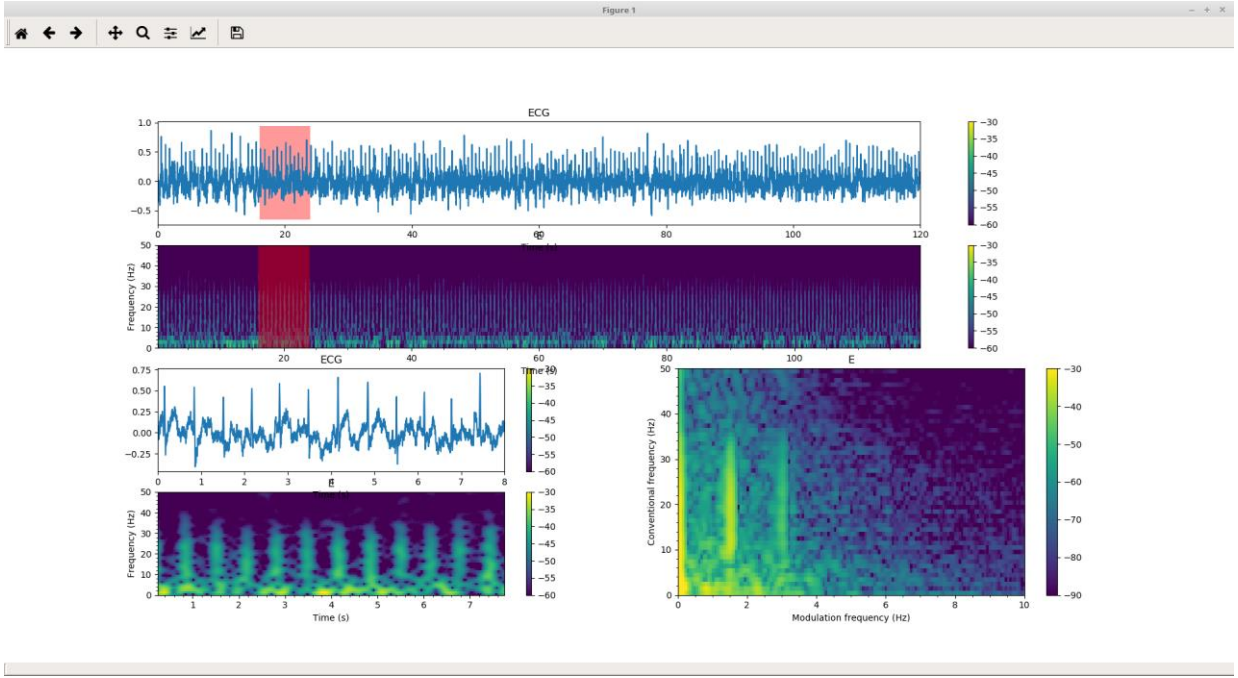
Blind Heart/Lung Sound Separation





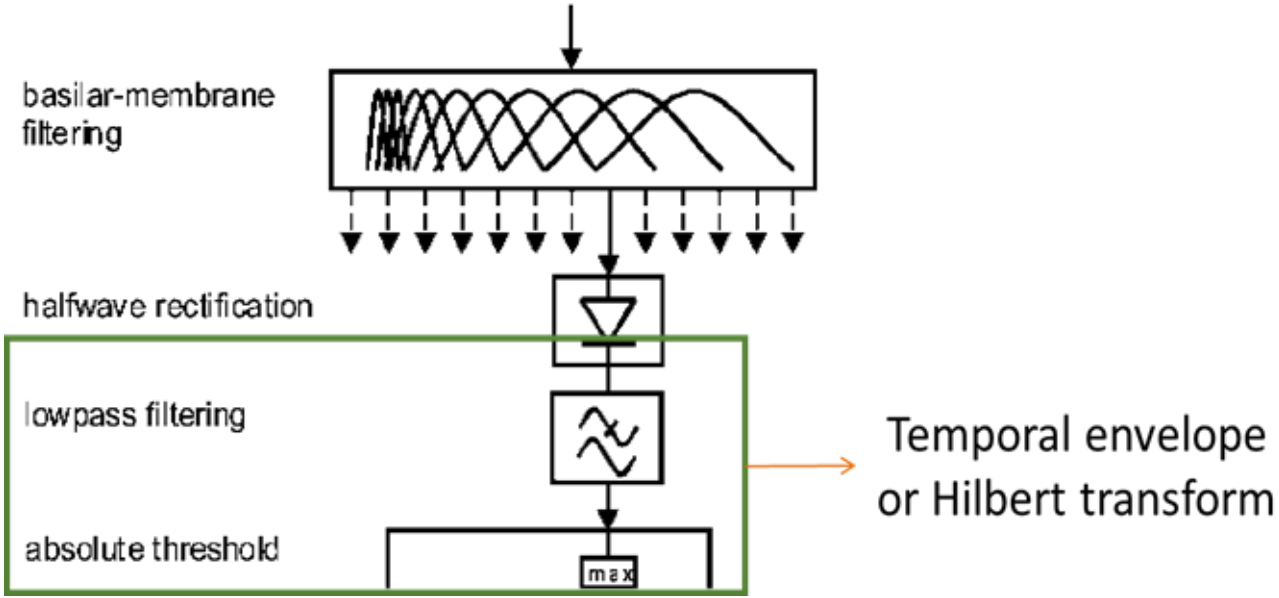
AMA Toolbox

- https://github.com/MuSAELab/amplitude-modulation-analysis-module



Modulation Spectrum - Analysis only

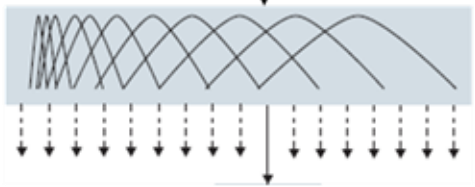
- Spectral analysis of the temporal envelope of the speech signal
  - Commonly involves **analysis** models only



Auditory-Inspired Model (by Tau)

Neutral voice
Pathological voice
Emotional voice
Whispered voice
Non-speech sounds

23-channel gammatone filterbank



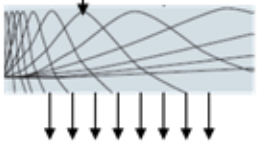
f_c's: 125Hz–3.57kHz
BW's: 38Hz–410Hz (ERB)

Normal hearing
Impaired hearing

Hilbert transform - "temporal envelope"

H(·)

8-channel modulation filterbank

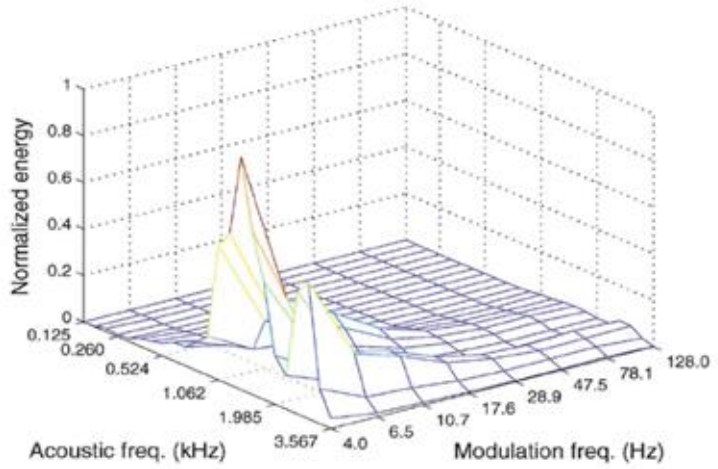


f_c's: 4Hz – 128Hz
BW's: 1.95Hz – 70.8Hz

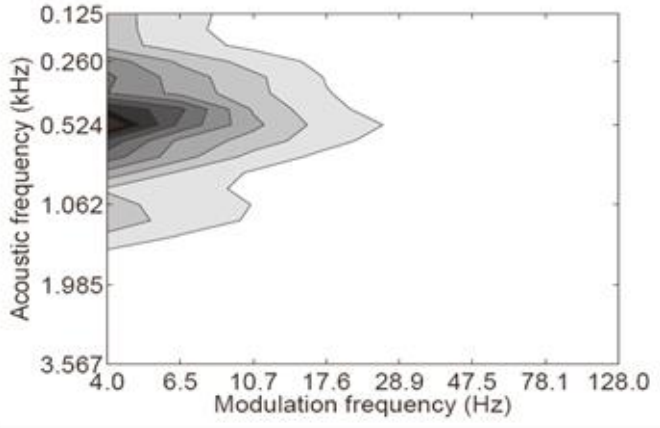
Speech vs non-speech

23 x 8 modulation energy output per 256ms frame

Clean Speech

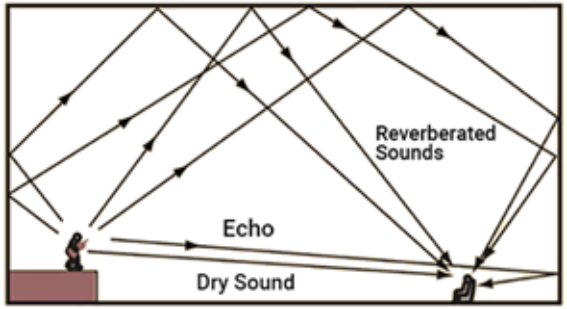


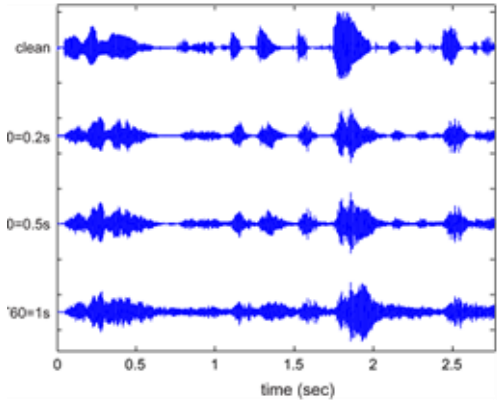
Clean signal 1

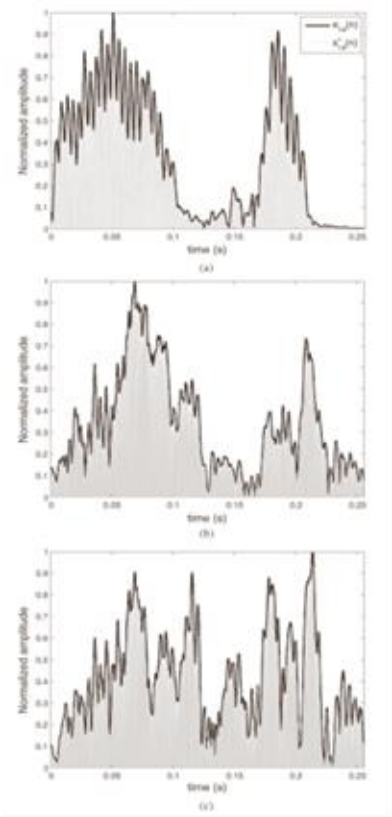


Clean signal 2

Noisy Speech



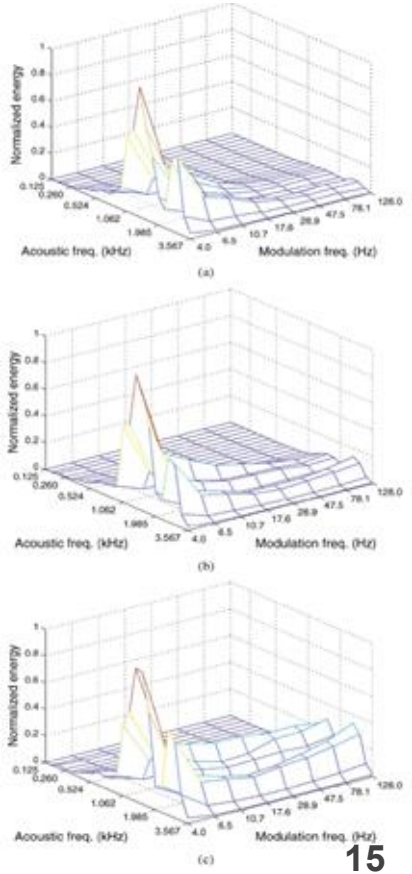




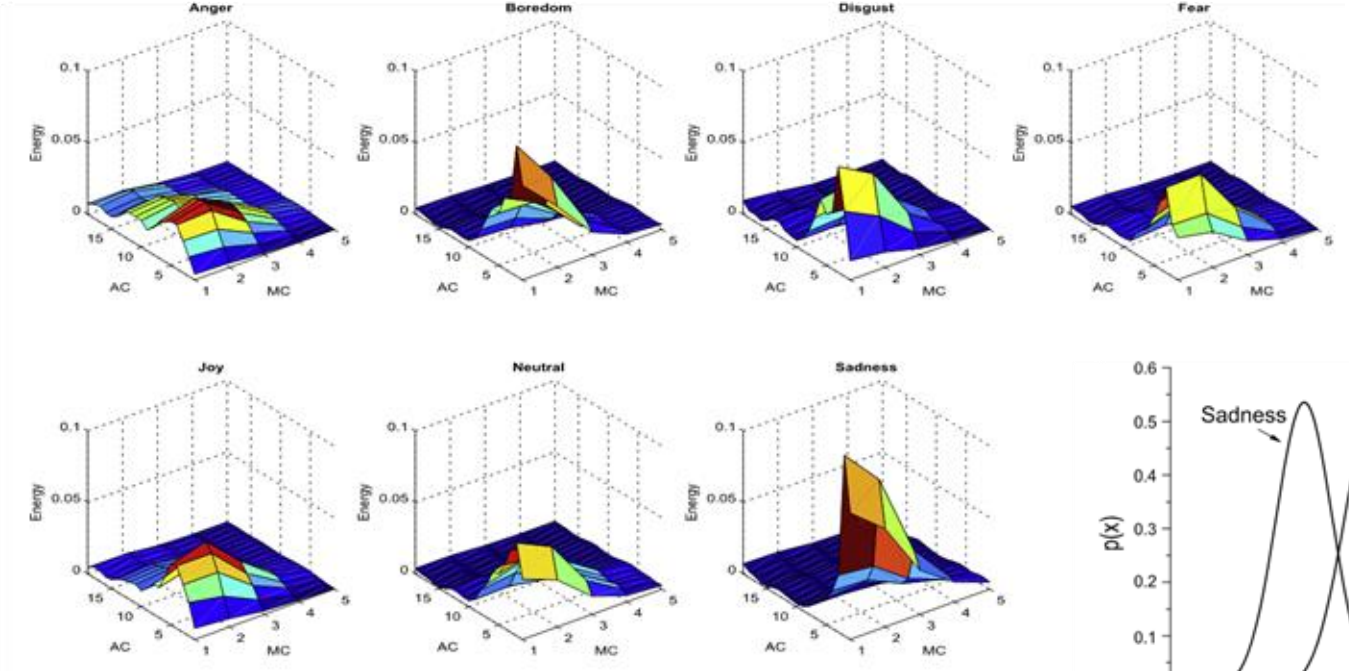
clean

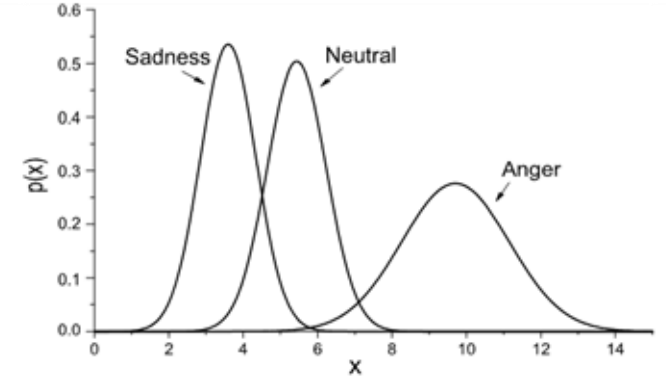
T_{60} = 400 ms

T_{60} = 1000 ms



Emotional Speech





Pathological / Deepfake Speech

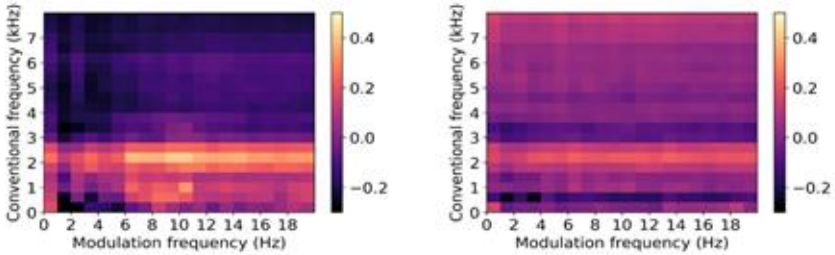
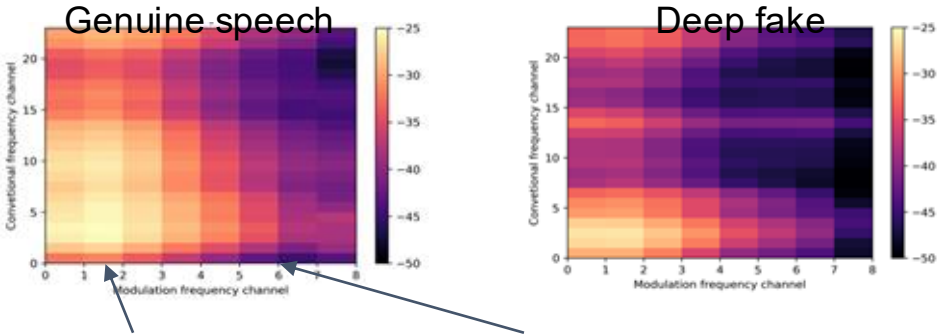


Fig. 2: Average modulation spectrograms for (a) COVID speech, and (b) non-COVID speech



~5Hz: natural speaking rate

>30Hz: Room acoustics

Automatic detection of Parkinson's disease from components of modulators in speech signals

Detección automática de la enfermedad de Parkinson usando componentes moduladoras de señales de voz

DOI: https://doi.org/10.17981/cesta.01.01.2020.05

Artículo de investigación científica. Fecha de recepción: 09/09/2020 Fecha de aceptación: 29/09/2020

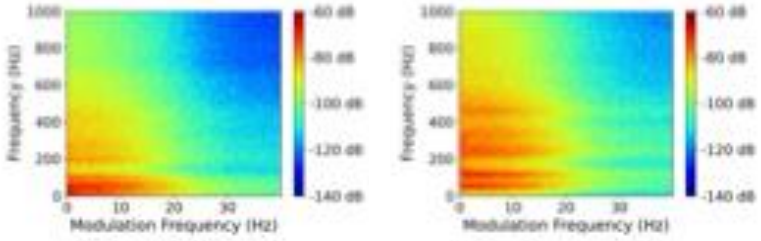
Jhon F. Moofarry
Universidad Santiago de Cali. Cali (Colombia)
jhon.moofarry00@usc.edu.co

Patricia Argüello-Velez
Universidad Santiago de Cali. Cali (Colombia)
patricia.arguello00@usc.edu.co

Milton Sarria-Paja
Universidad Santiago de Cali. Cali (Colombia)
milton.sarria00@usc.edu.co

Feature set	/pataka/		
	NBC	KNN	SVM
MFCC	57.1±2.0	68.1±2.0	77.1±2.0
WIF	54.3±2.9	65.2±4.2	66.1±4.0
MHEC	55.0±3.2	75.1±2.5	78.3±3.1
AAMF	70.1±4.0	71.0±3.1	88.0±2.8

Beehive Recordings



(a) (b)

Fig. 18: Modulation spectrograms of a a) 15-minute clean, and b) noisy beehive audio signal.

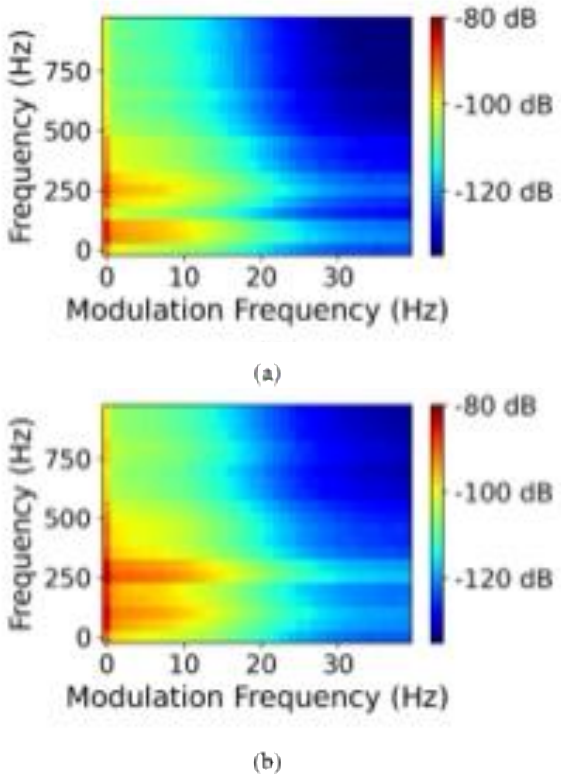
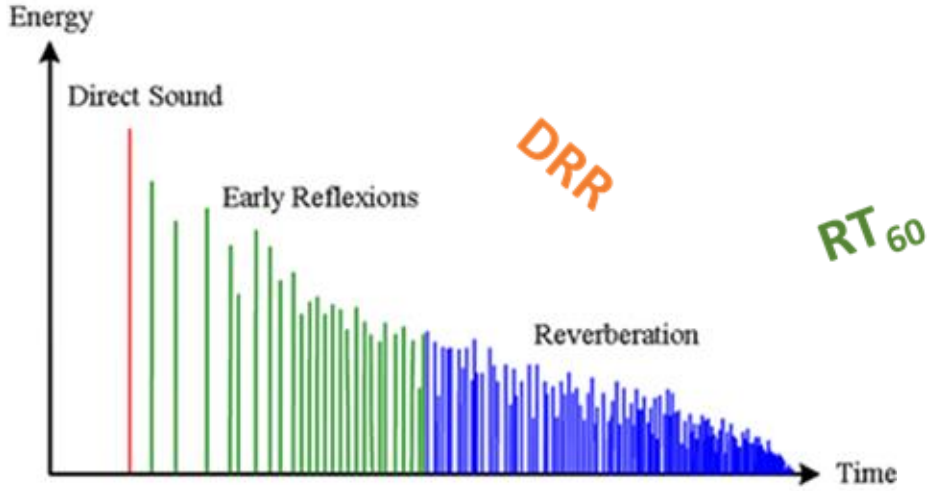


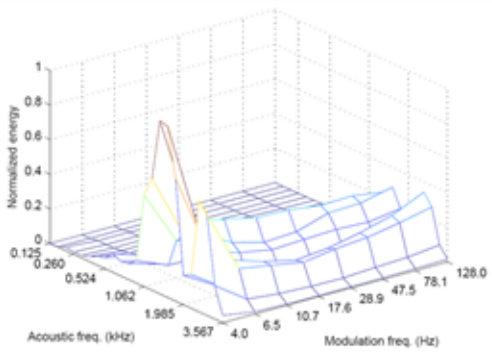
Fig. 4: Modulation spectrograms of a) a Varroa infected beehive, and b) a healthy beehive.

Application #1: Room/Environment acoustics

- Room acoustics characterization
  - Involves analysis of room impulse response (RIR)
  - Recorded or estimated
- Environment: Noise type



Proposed RT_{60} and DRR Estimators



$$\text{RSMR}_k = \frac{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,1}}$$

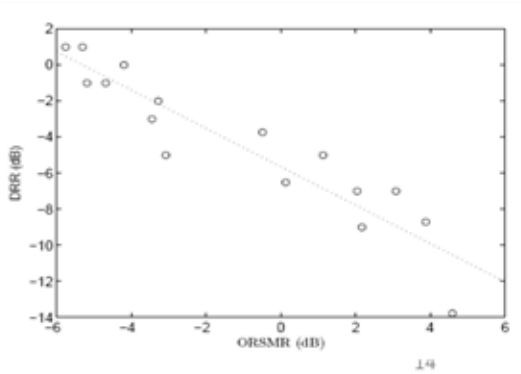
Reverberation-to-speech modulation energy ratio (channel k = 5-8)

$$\hat{RT}_{60} = f(RSMR_5, RSMR_6, RSMR_7, RSMR_8)$$

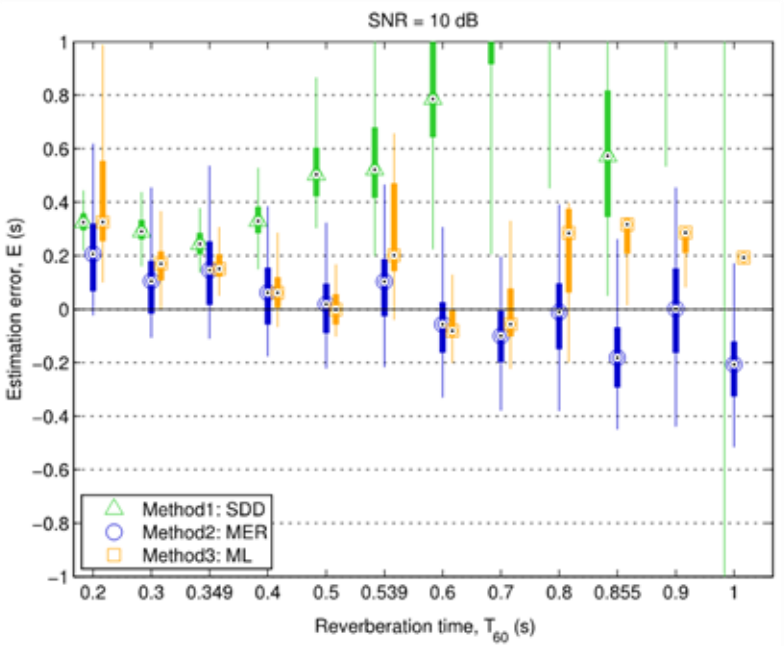
$$\bar{\mathcal{E}}_{j,k} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \mathcal{E}_{j,k}^{act}(i),$$

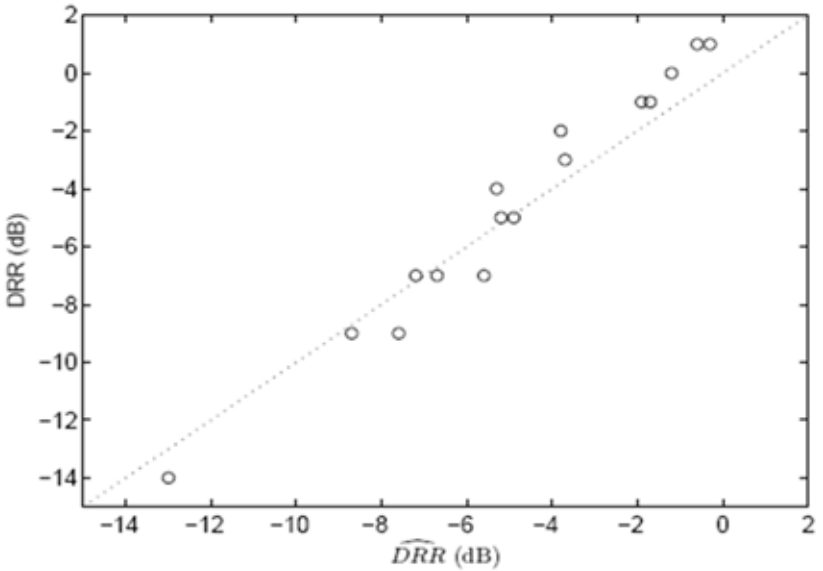
$$\text{ORSMR} = \frac{\sum_{k=5}^{8}\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,1}} = \sum_{i=5}^{8} \text{RSMR}_i.$$

$$\widehat{\text{DRR}} = -5.6467 - 1.0644 \times \text{ORSMR (dB)},$$

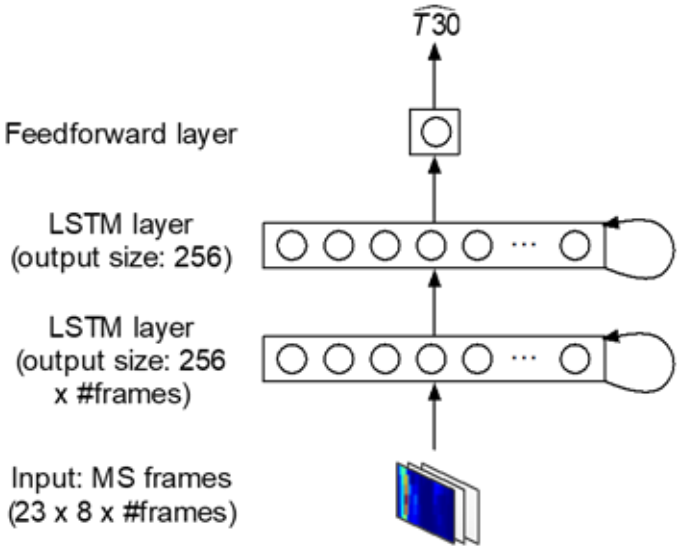


Single-Channel Results (Speech)





Modulation Spectrum as DNN Input



Metric	SNR					
	20 dB		10 dB		0 dB	
	RMSE	MAD	RMSE	MAD	RMSE	MAD
LSTM	**0.306**	**0.212**	**0.311**	**0.218**	0.324	**0.240**
GMR	0.401	0.338	0.402	0.341	0.460	0.392
FNN	0.337	0.262	0.338	0.266	0.355	0.286
RSMR	0.331	0.266	0.325	0.258	**0.304**	0.258
ML-SD	0.394	0.319	0.398	0.322	0.393	0.318

Non-Speech Sounds

Clarity (C)
Definition (D)
Central time (CT)
Early decay time (EDT)
Reverberation time (T)

$$C = 10\log_{10}\left(\frac{\int_0^{t_0}[g(t)]^2\,\mathrm{dt}}{\int_{t_0}^{\infty}[g(t)]^2\,\mathrm{dt}}\right)$$

$$D = \frac{\int_0^{t_0}[g(t)]^2\,\mathrm{dt}}{\int_0^{\infty}[g(t)]^2\,\mathrm{dt}} \times 100\%$$

$$\mathrm{CT} = \frac{\int_0^{\infty}[g(t)]^2 t\,\mathrm{dt}}{\int_0^{\infty}[g(t)]^2\,\mathrm{dt}}$$

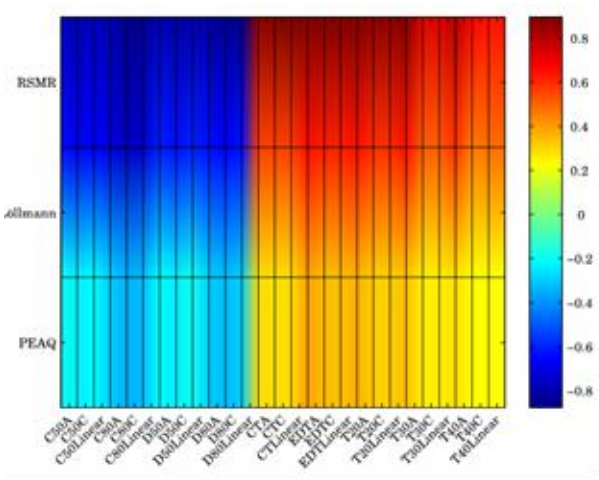


FIGURE 1: Heat map for the ensemble correlations

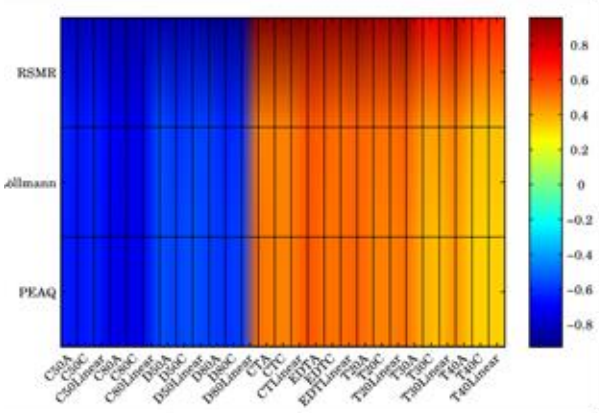
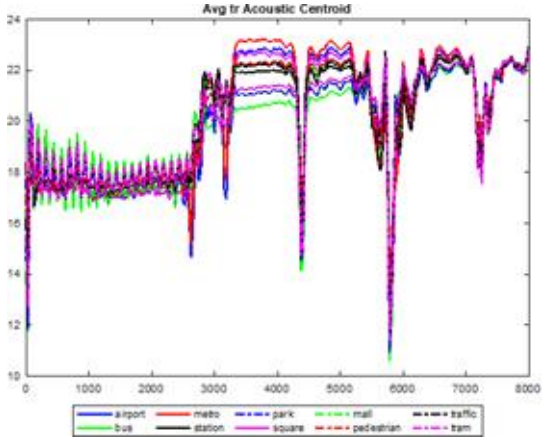


FIGURE 2: Heat map for the solo instrument correlations

Environment Awareness



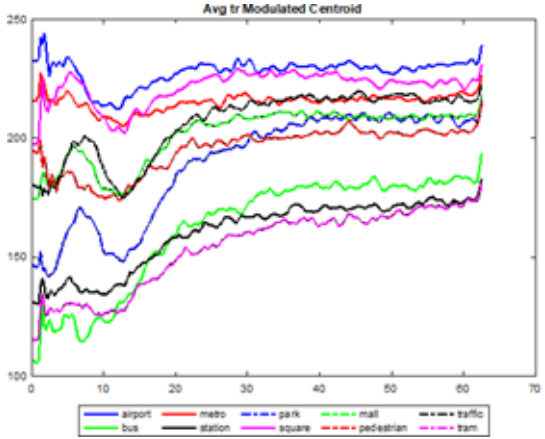


Table 1: Results in terms of accuracy for each class for task 1A.

Class	System			
	S1	S2	S3	Baseline
Airport	60.6%	44.1%	49.2%	45.0%
Public square	39.7%	48.8%	42.8%	44.9%
Bus	84.5%	71.0%	69.7%	62.9%
Metro	59.3%	58.2%	49.2%	53.5%
Metro station	63.6%	48.5%	55.2%	53.0%
Park	78.5%	70.4%	69.7%	71.3%
Shopping mall	62.3%	50.8%	53.9%	48.3%
Street pedestrian	43.4%	37.0%	28.3%	29.8%
Street traffic	85.5%	85.5%	80.8%	79.9%
Tram	68.7%	62.3%	61.6%	52.2%
Average	64.6%	57.7%	56.0%	54.1%

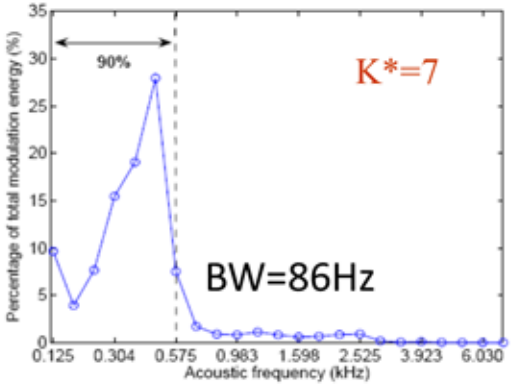
Table 2: Results in terms of accuracy for each class for task 1B.

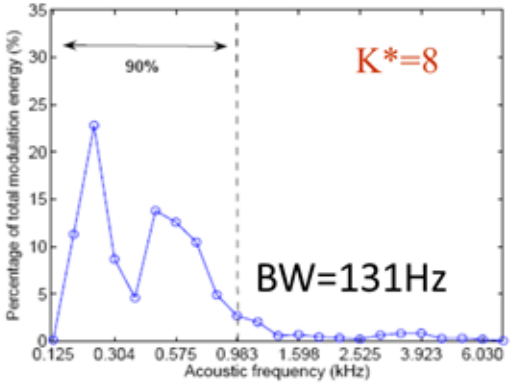
Class	Proposed System	Baseline
Indoor	85.7%	82.0%
Outdoor	83.2%	88.5%
Transportation	93.8%	91.5%
Average	87.6%	87.3%

Application #2: Quality Measurement

Adaptive speech-to-reverberation modulation energy ratio (SRMR)

$$\text{SRMR} = \frac{\sum_{k=1}^{4}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}{\sum_{k=5}^{K^*}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}.$$





Experimental Results

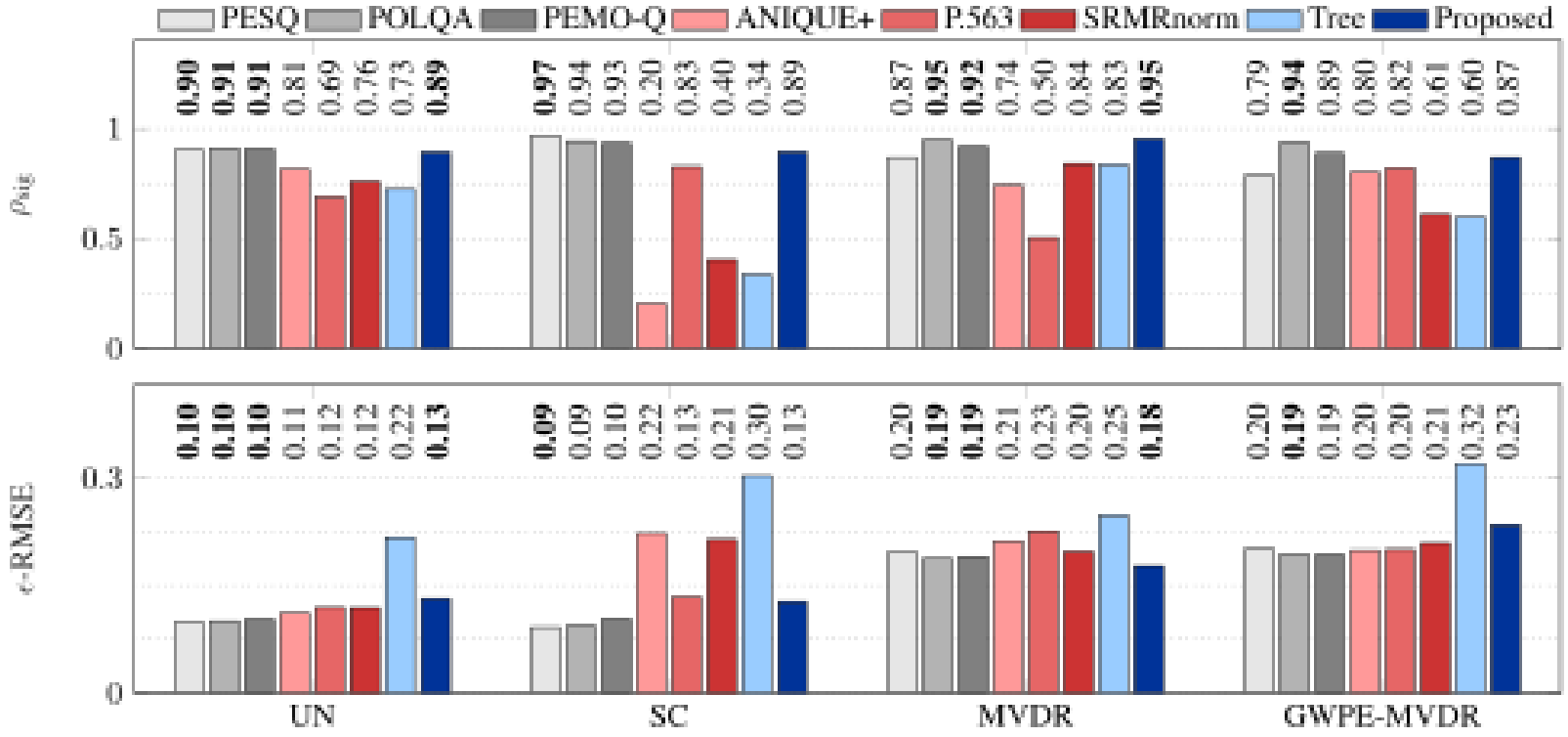
Quality

Algorithm	Overall (reverberant + dereverberated)						Reverberant						Delay-and-sum					
	COL	%↑	RTE	%↑	MOS	%↑	COL	%↑	RTE	%↑	MOS	%↑	COL	%↑	RTE	%↑	MOS	%↑
SRMR	0.84	–	0.82	–	0.79	–	0.81	–	0.82	–	0.79	–	0.87	–	0.83	–	0.80	–
PESQ	0.66	52.9	0.81	5.3	0.72	25.0	0.66	44.1	0.81	5.3	0.70	30.0	0.67	60.6	0.82	5.6	0.78	9.1
P.563	0.44	71.4	0.46	66.7	0.35	67.7	0.38	69.4	0.41	69.5	0.31	69.6	0.54	71.7	0.50	66.0	0.40	66.7
ANIQUE+	0.72	42.9	0.70	40.0	0.77	8.7	0.77	17.4	0.76	25.0	0.84	-31.3	0.67	60.6	0.57	60.5	0.67	39.4
Average	–	55.7	–	37.3	–	33.8	–	43.6	–	33.3	–	22.8	–	64.3	–	44.0	–	38.4

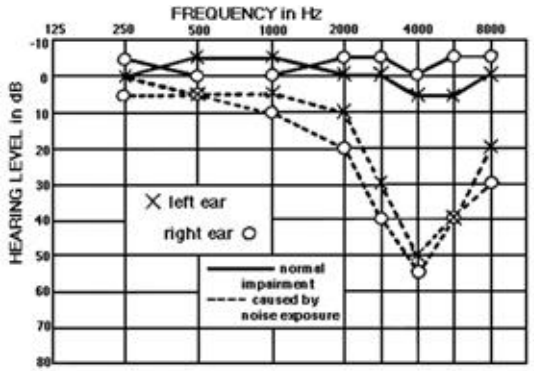
Intelligibility

Algorithm	SRMR			PESQ				P.563				ANIQUE+			
	STI_1	STI_2	STI_3	STI_1	STI_2	STI_3	% ↑	STI_1	STI_2	STI_3	% ↑	STI_1	STI_2	STI_3	% ↑
Reverberation	0.92	0.94	0.96	0.88	0.92	0.92	50.0	0.10	0.11	0.10	95.6	0.42	0.43	0.41	93.2
DSB	0.90	0.92	0.96	0.89	0.93	0.94	33.3	0.12	0.12	0.11	95.5	0.45	0.45	0.46	92.6
Cepstrum	0.90	0.93	0.95	0.86	0.91	0.92	37.5	0.06	0.07	0.06	94.7	0.47	0.48	0.50	90.0
Subspace	0.81	0.86	0.87	0.78	0.85	0.85	13.3	0.20	0.19	0.19	84.0	0.28	0.30	0.34	80.3
Average	–	–	–	–	–	–	33.5	–	–	–	92.4	–	–	–	89.0

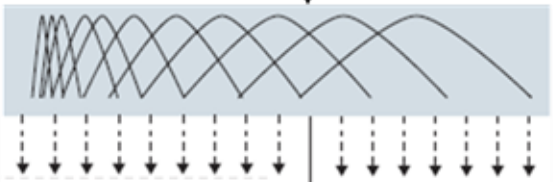
SRMR and LSTMs



Hearing Impaired Listeners



23-channel gammatone filterbank



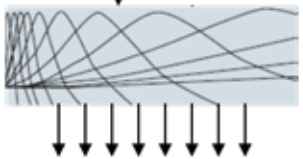
f_c's: 125Hz–3.57kHz
BW's: 38Hz–410Hz (ERB)

Hilbert transform - "temporal envelope"

H(·)

?

8-channel modulation filterbank



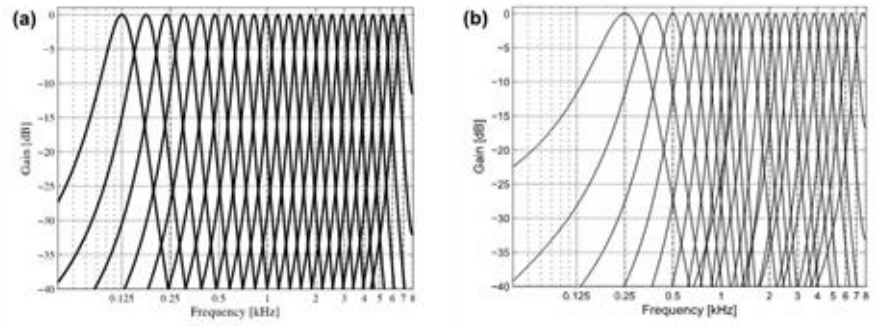
f_c's: 4Hz – 128Hz
BW's: 1.95Hz – 70.8Hz

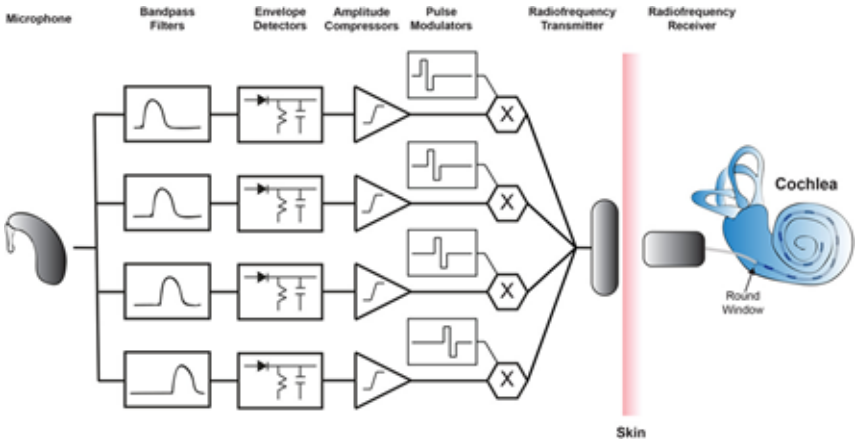
?

23 x 8 modulation energy output per 256ms frame

Cochlear Implants

- Nucleus device (22-channel)





- Modulation channels

Table 1
Filter center frequencies (f_c) and bandwidths (BW), expressed in Hz, of the modulation filters used in the original SRMR and in the proposed SRMR-CI measures.

Channel		1	2	3	4	5	6	7	8
SRMR	f_c	4	6.5	10.7	17.6	28.9	47.5	78.1	128
	BW	1.9	3.4	5.9	9.8	15.9	26.4	43.2	70.8
SRMR-CI	f_c	4	5.94	8.83	13.13	19.5	28.98	43.07	64
	BW	2	3	4.5	6.6	9.8	14.5	21.5	32

Performance: SRMR-CI

RT = 0.3 - 1s
d = 5.5 m
Noise = SSN
-5dB ... 10dB

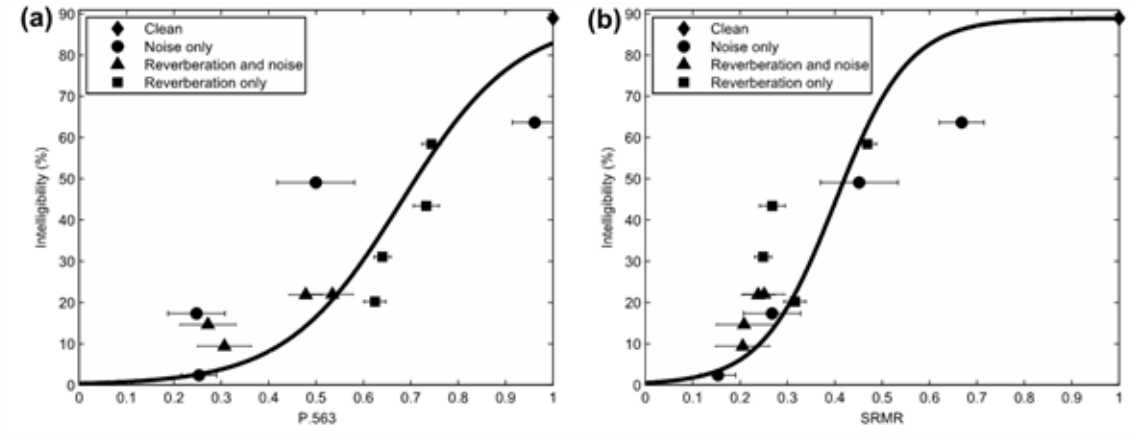
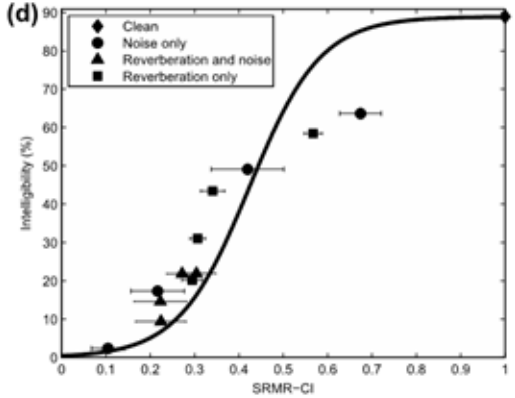
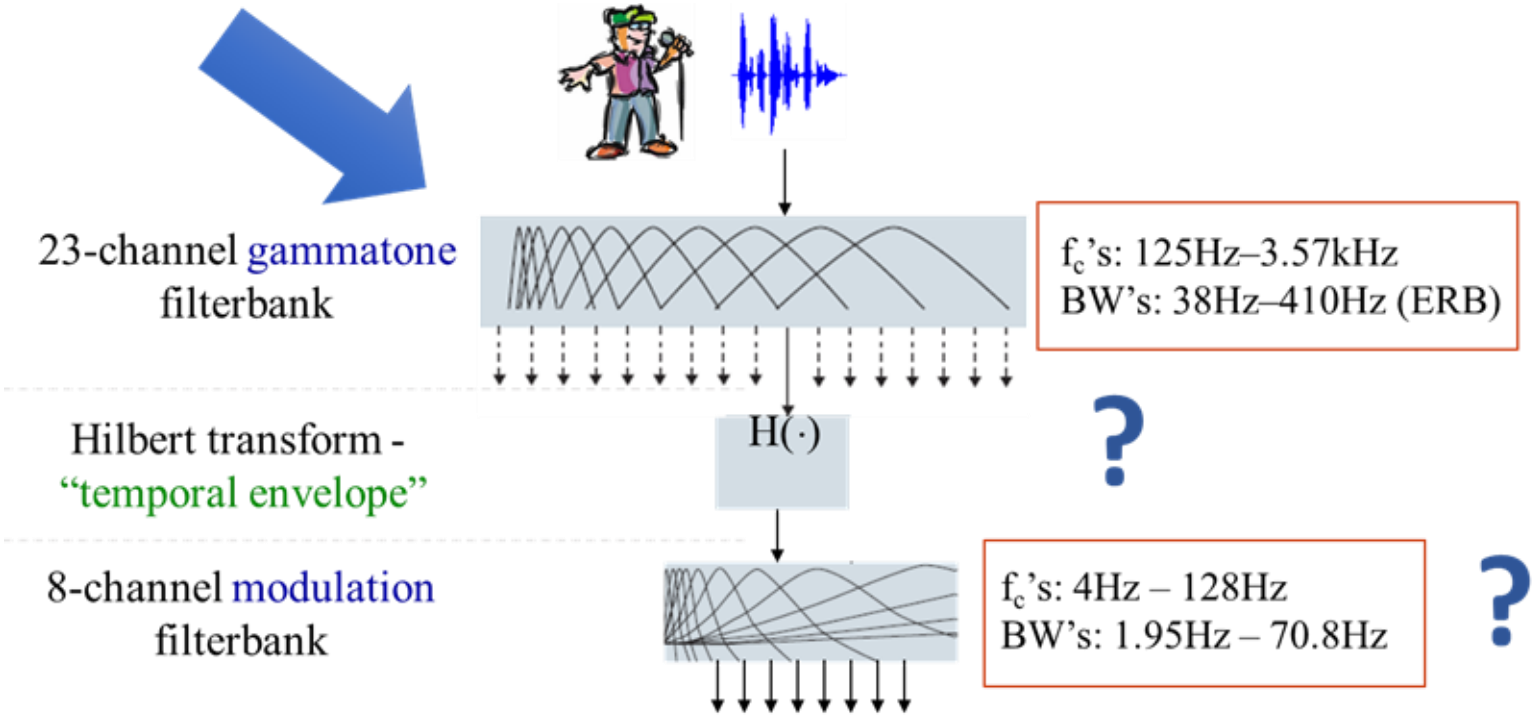


Table 2
Overall per-condition performance criteria of the seven investigated objective measures.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE
NCM	0.96	0.93	0.93	12.4
CSII	0.93	0.91	0.93	10.57
P.563	0.89	0.88	0.89	12.52
SRMR	0.93	0.89	0.92	12.77
ModA	0.82	0.76	0.82	15.70
SRMR-CI	0.96	0.97	0.94	11.29
SRMR – CI_{norm}	0.96	0.97	0.95	10.76

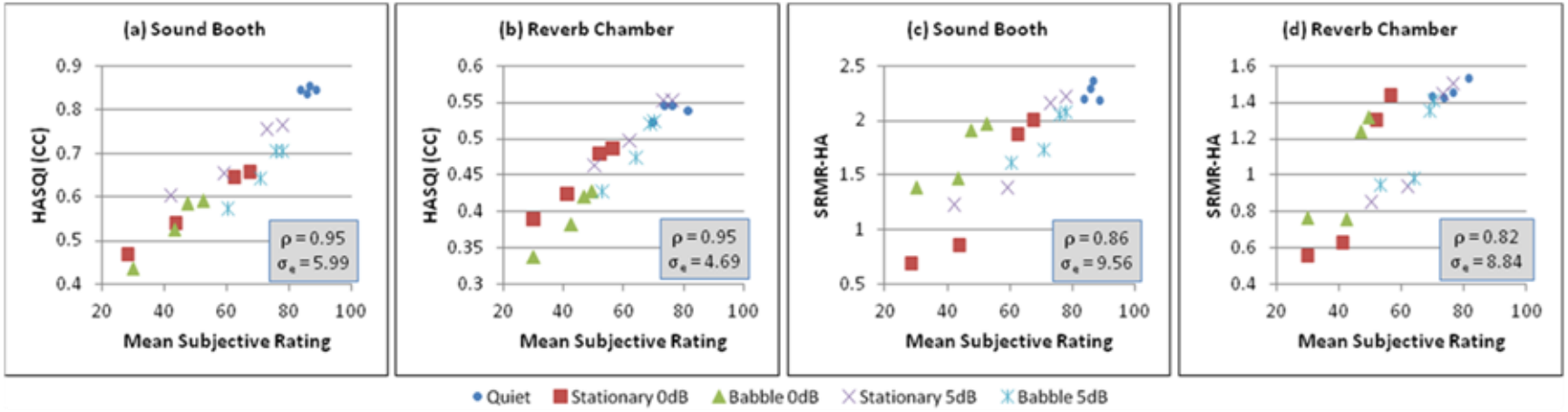


Hearing Aid Users



23 x 8 modulation energy output per 256ms frame

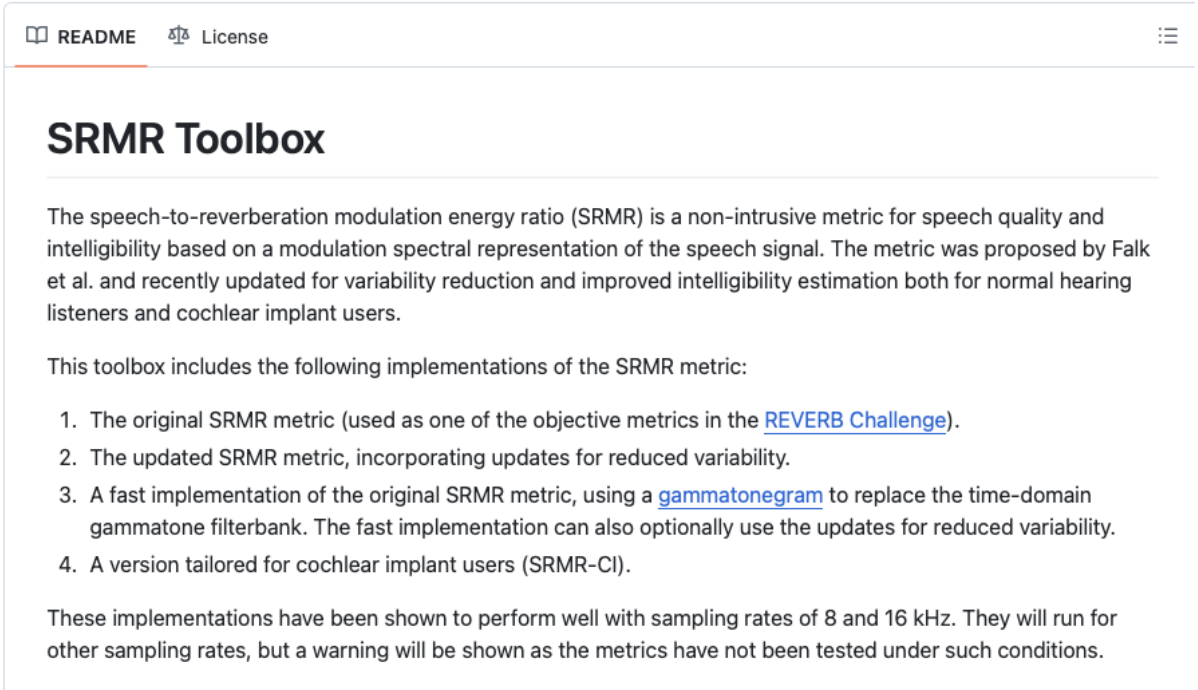
Performance



SRMR: r = 0.75 r = 0.70

Github

- https://github.com/MuSAELab/SRMRToolbox



README License

SRMR Toolbox

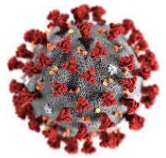
The speech-to-reverberation modulation energy ratio (SRMR) is a non-intrusive metric for speech quality and intelligibility based on a modulation spectral representation of the speech signal. The metric was proposed by Falk et al. and recently updated for variability reduction and improved intelligibility estimation both for normal hearing listeners and cochlear implant users.

This toolbox includes the following implementations of the SRMR metric:

1. The original SRMR metric (used as one of the objective metrics in the REVERB Challenge).
2. The updated SRMR metric, incorporating updates for reduced variability.
3. A fast implementation of the original SRMR metric, using a gammatonegram to replace the time-domain gammatone filterbank. The fast implementation can also optionally use the updates for reduced variability.
4. A version tailored for cochlear implant users (SRMR-CI).

These implementations have been shown to perform well with sampling rates of 8 and 16 kHz. They will run for other sampling rates, but a warning will be shown as the metrics have not been tested under such conditions.

Spectrogram vs Tensorgram



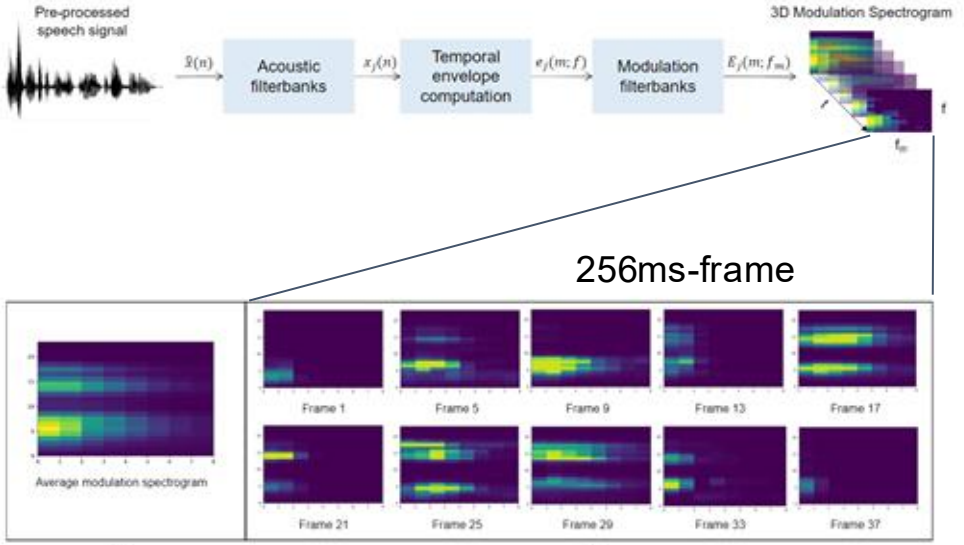


Fig. 2: An example of comparison between (a) a temporally averaged modulation spectrogram, and (b) several frames selected from a 3D modulation spectrogram sequence. Both are computed from the same speech recording.

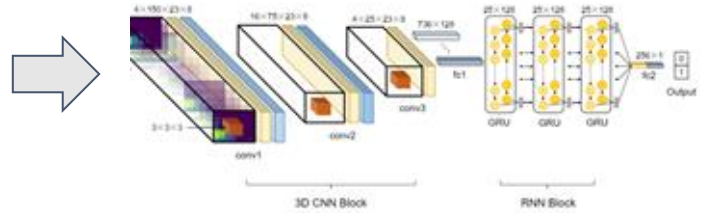
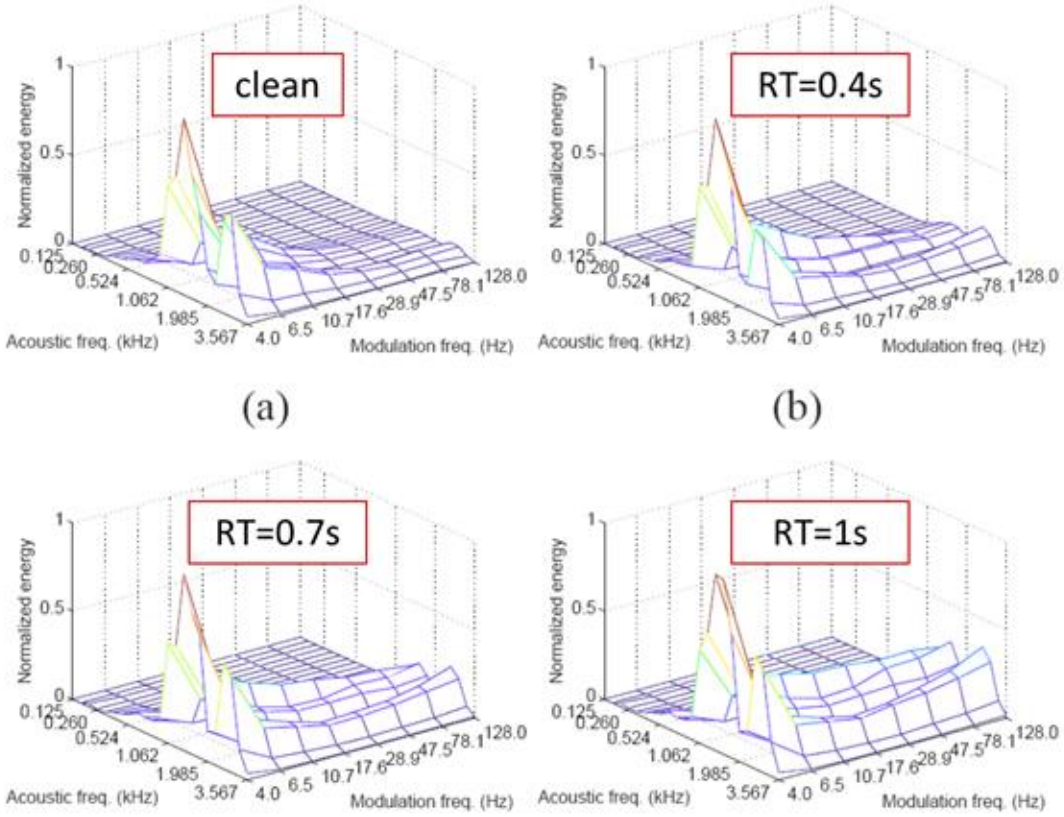


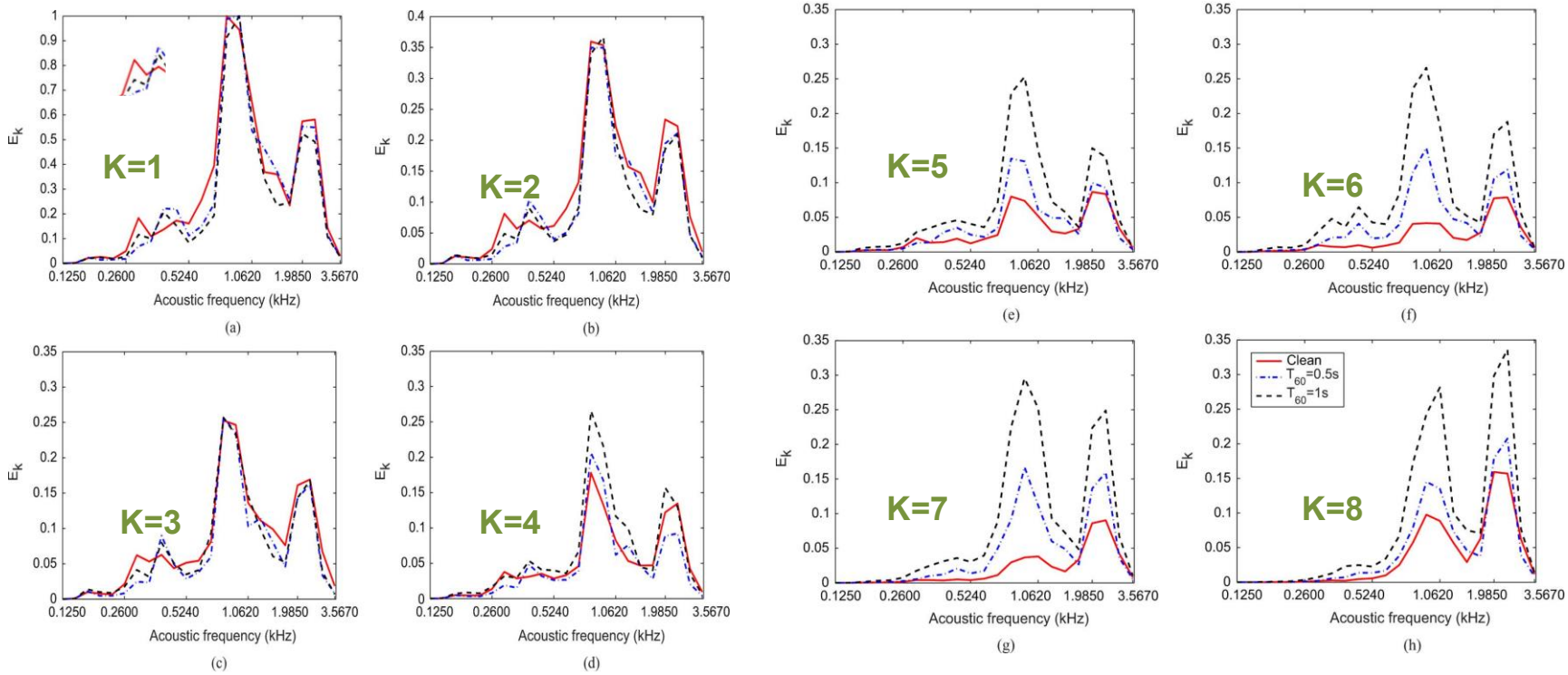
Fig. 3: Model architecture of COVID-CRNN.

Feature	Model	Random-Split			
		No pre-processing		Spectral amplitude subtraction	
		MAE	r	MAE	r
Spectrogram	CNN-2D	3.70 ± 0.17	0.76 ± 0.03	1.71 ± 0.13	0.83 ± 0.02
	CNN-att-2D	3.68 ± 0.08	0.75 ± 0.02	2.17 ± 0.09	0.76 ± 0.01
	CRDNN-2D	3.49 ± 0.24	0.78 ± 0.02	1.77 ± 0.05	0.81 ± 0.01
MFCCs	CNN-2D	3.62 ± 0.13	0.76 ± 0.02	1.85 ± 0.25	0.81 ± 0.03
	CNN-att-2D	3.72 ± 0.09	0.75 ± 0.01	2.15 ± 0.17	0.77 ± 0.01
	CRDNN-2D	3.37 ± 0.05	0.78 ± 0.03	1.78 ± 0.10	0.82 ± 0.01
Modulation spectrogram	CNN-2D	1.54 ± 0.14	0.95 ± 0.01	1.14 ± 0.06	0.96 ± 0.01
	CNN-att-2D	1.52 ± 0.16	0.96 ± 0.03	1.39 ± 0.06	0.97 ± 0.03
Modulation tensorgram	CNN-3D	1.61 ± 0.10	0.94 ± 0.02	1.36 ± 0.09	0.96 ± 0.01
	CNN-att-3D	1.45 ± 0.05	0.96 ± 0.02	1.23 ± 0.06	0.96 ± 0.03
	CRDNN-3D	1.24 ± 0.03	0.95 ± 0.01	1.01 ± 0.08	0.97 ± 0.01

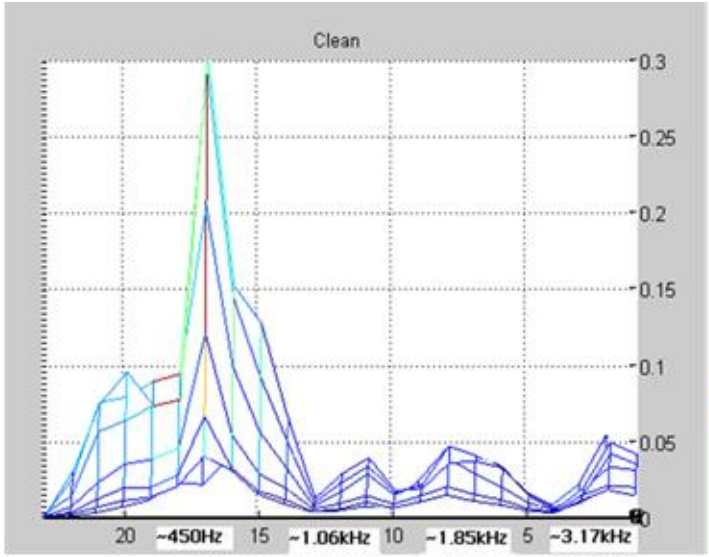
Application #3: Robust Feature Extraction



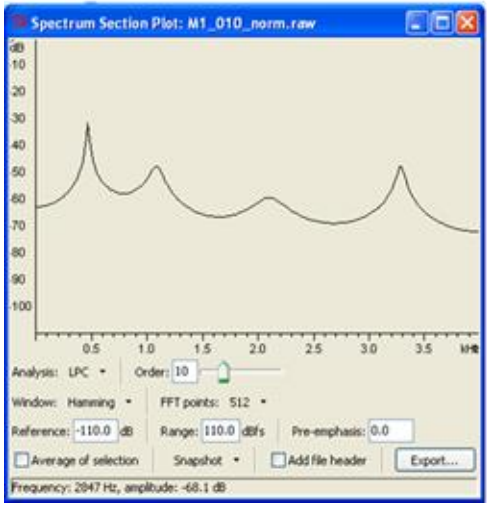
Speech Noise Decoding

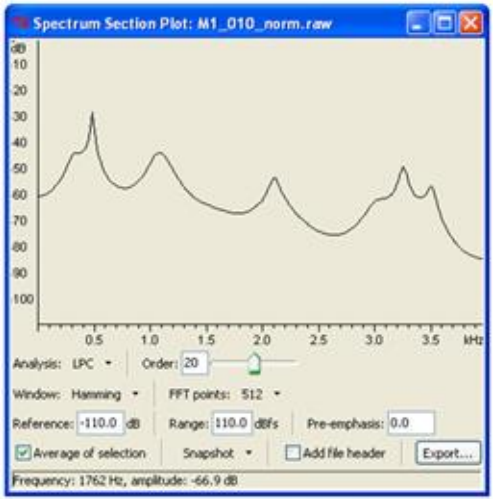


Speaker Information

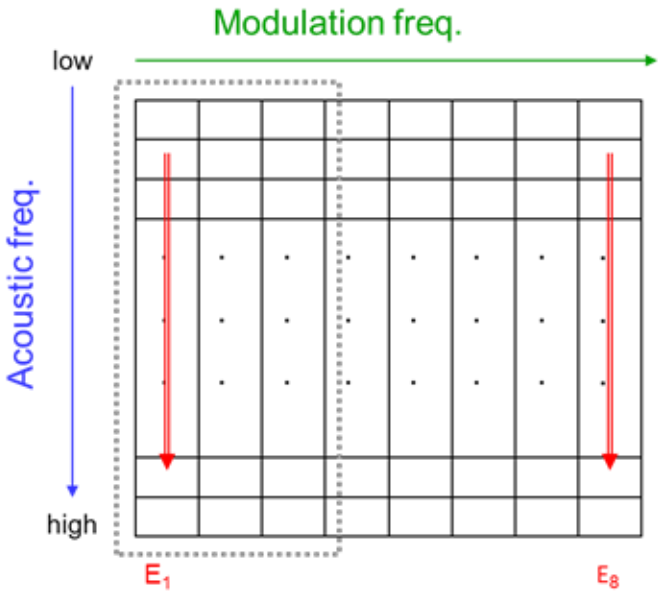


Acoustic (cochlear) frequency





New Speaker ID Features

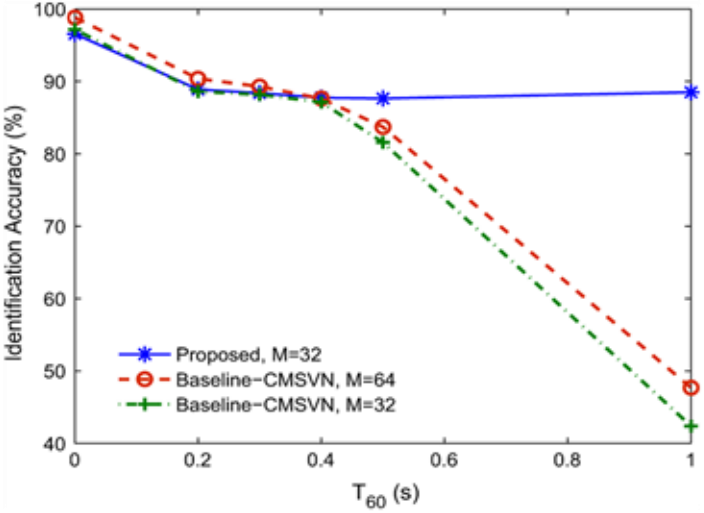


$$LL_s = \frac{1}{N}\sum_{m=1}^{N} \log(p(\mathbf{x}(m)|\mathbf{\Lambda}_s))$$

$$\hat{S} = \underset{1\le s\le N_S}{\operatorname{argmax}} LL_s.$$

$$LL_{k,s} = \frac{1}{N'}\sum_{m=1}^{N'} \log(p_k(\vec{\mathcal{E}}_k(m)|\mathbf{\Lambda}_{k,s})),\ k = 1-3 \qquad \hat{S} = \underset{1\le s\le N_S}{\operatorname{argmax}} \max_{1\le k\le 3} LL_{k,s}.$$

Identification Accuracy



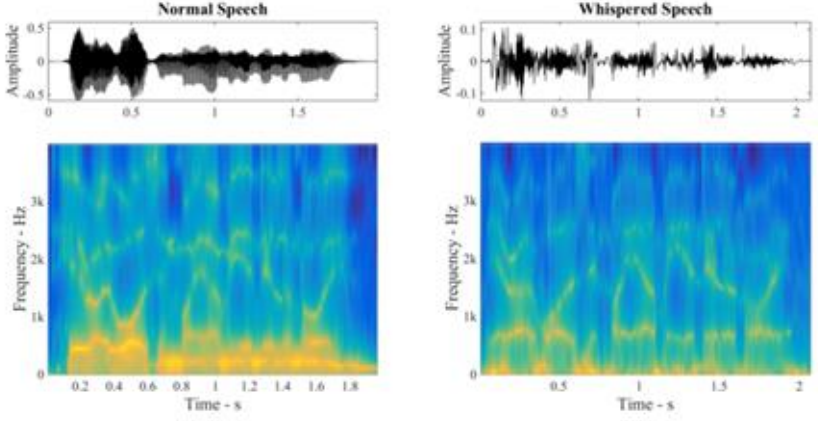
Simulated, single-channel

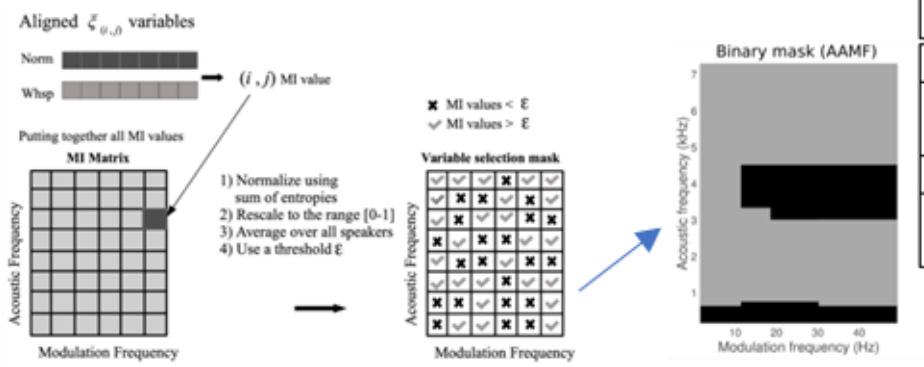
Recorded, multi-channel
(RT60 ~ 600ms)

Condition	Proposed	Baseline, $M = 32$			Baseline, $M = 64$		
	ACC	ACC	INC	ERR	ACC	INC	ERR
Clean	96.8	96.7	0.1	3.0	98.6	-1.8	-128.6
Channel 1	84.6	56.9	48.7	64.3	60	41.0	61.5
Channel 2	82.8	55.8	48.4	61.1	58.9	40.6	58.2
Channel 3	83.2	55.6	49.6	62.2	59.1	40.8	58.9
Channel 4	83.4	54.1	54.2	63.8	58.1	43.5	60.4
Channel 5	81.5	56.6	44.0	57.4	60.8	34.0	52.8
Channel 6	80.7	54.3	48.6	57.8	59.1	36.5	52.8
Average	82.7	55.6	48.9	61.1	59.3	39.4	57.4

Baseline	DSB			Cepstrum			Subspace		
	ACC	INC	ERR	ACC	INC	ERR	ACC	INC	ERR
$M = 32$	63.0	31.3	53.2	46.0	79.8	68.0	61.9	33.6	54.6
$M = 64$	67.4	22.7	46.9	51.6	60.3	64.3	65.6	26.1	49.7

Whispered vs Non-Whispered Speaker ID

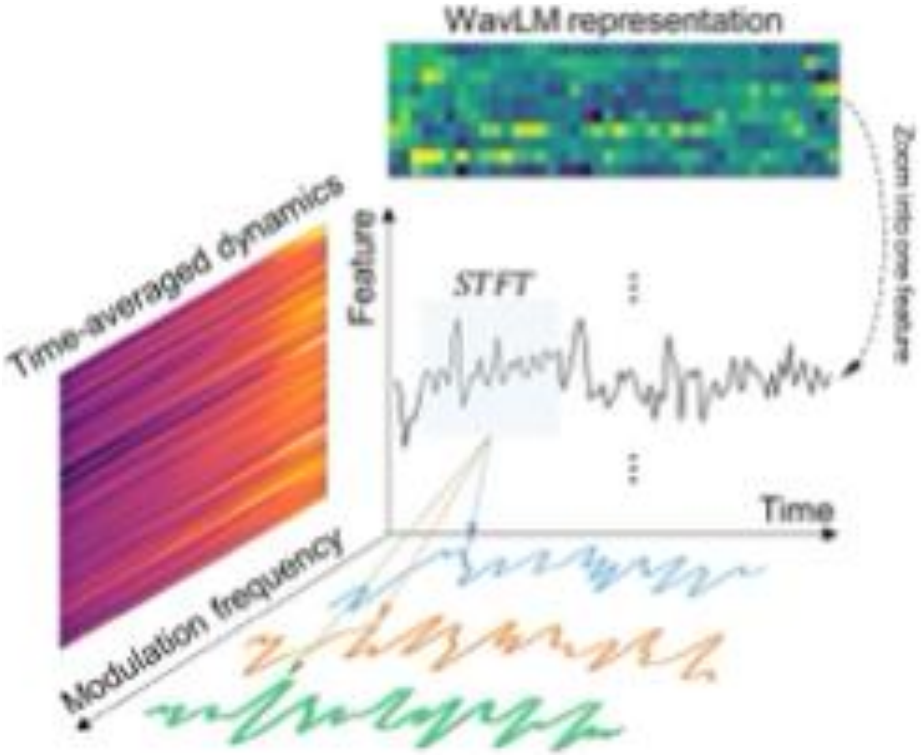


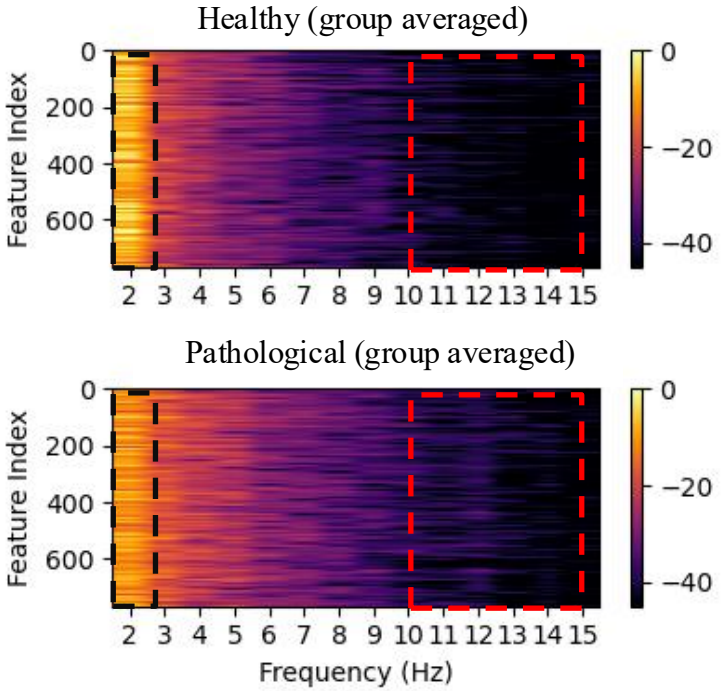


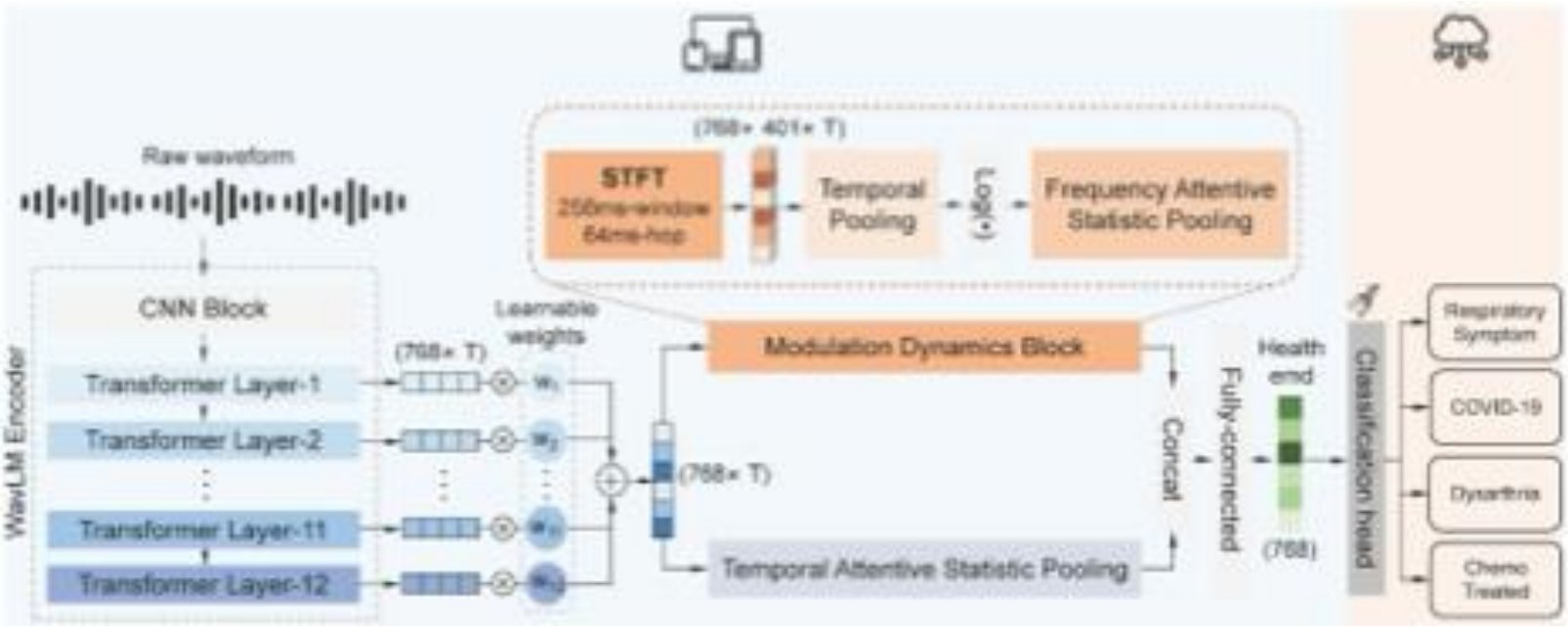
UBM (C)	Normal			Whispered		
	T matrix dimension					
	200	300	400	200	300	400
S1: MFCC						
128	3.78	3.69	3.36	20.99	21.11	20.83
256	3.18	3.44	**3.13**	**20.00**	21.91	20.83
S2: RMFCC						
128	8.44	8.14	7.50	25.83	24.42	25.95
256	8.44	7.50	**6.70**	24.20	25.00	**22.95**
S3: LMFCC						
128	3.13	3.13	3.44	18.13	17.13	17.81
256	3.15	3.44	**3.13**	17.97	18.29	**16.67**
AAMF(FS) - LFCC alignment						
128	1.56	1.58	1.55	21.22	19.82	20.86
256	1.60	1.36	**1.26**	20.51	**19.17**	20.83
S4: AAMF(FS) - AAMF alignment						
128	1.00	1.25	1.18	20.00	20.00	20.57
256	1.56	1.12	**0.94**	18.44	18.29	**14.80**

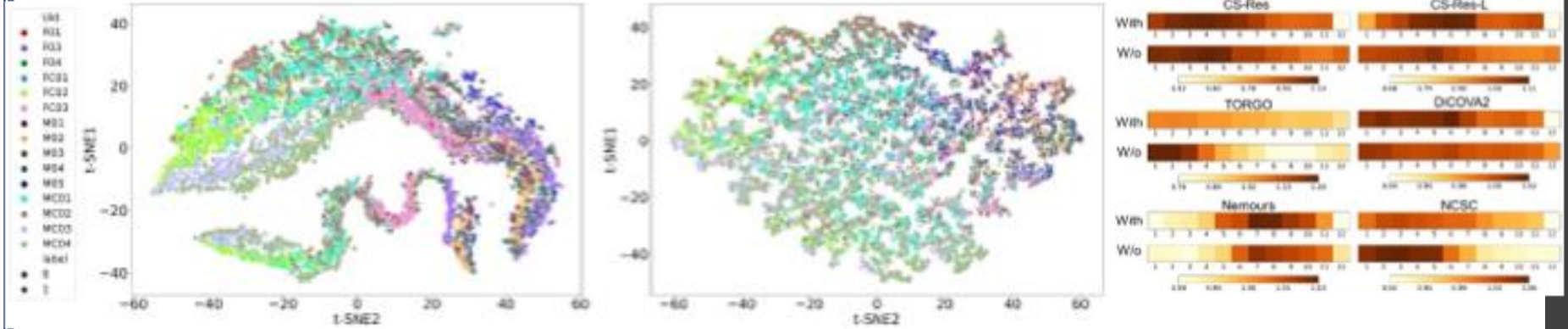

what's
next
?

From Modulation Spectrogram to Modulation Featuregram

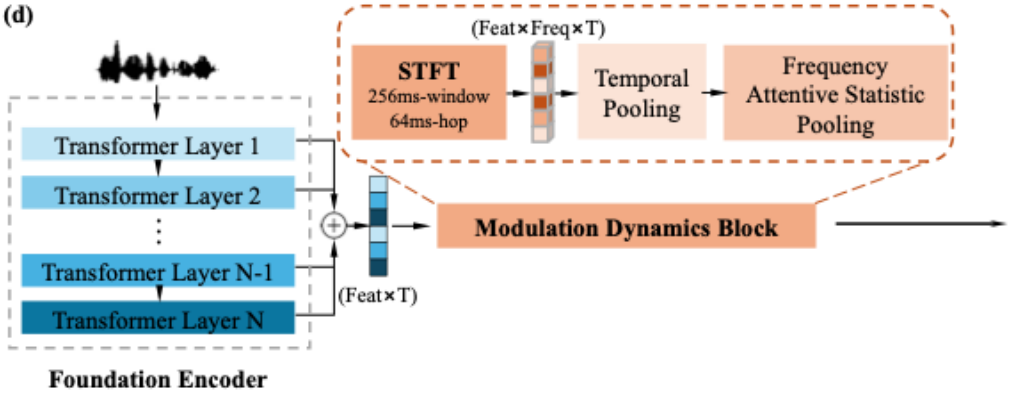


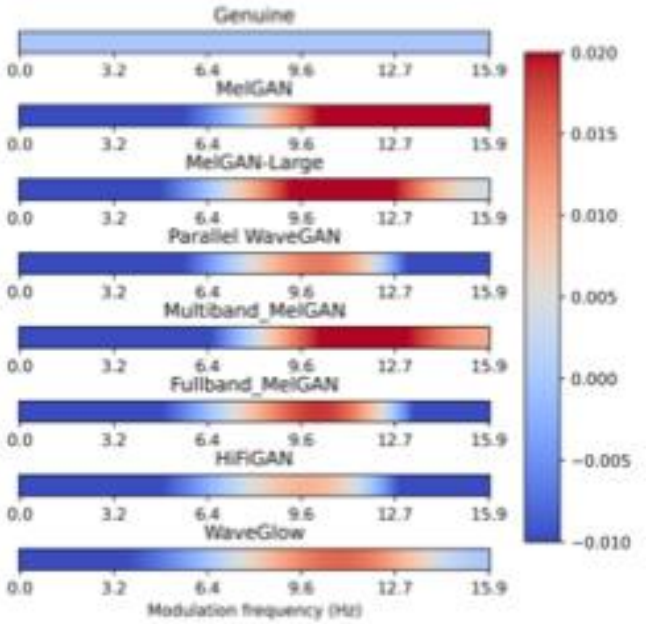



Raw waveform
WavLM Encoder
CNN Block
Transformer Layer-1
Transformer Layer-2
Transformer Layer-11
Transformer Layer-12
Learnable weights
(768× T)
(768× 401× T)
STFT
25ms-window
64ms-hop
Temporal Pooling
Log(·)
Frequency Attentive Statistic Pooling
Modulation Dynamics Block
Temporal Attentive Statistic Pooling
Concat
Fully-connected
Health emd
(768)
Classification head
Respiratory Symptom
COVID-19
Dysarthria
Chemo Treated


t-SNE1
t-SNE2
CS-Res
CS-Res-L
With
W/o
TORGO
DiCOVA2
Nemours
NCSC

Deepfakes


(d)
Transformer Layer 1
Transformer Layer 2
Transformer Layer N-1
Transformer Layer N
Foundation Encoder
(Feat×Freq×T)
STFT
256ms-window
64ms-hop
Temporal Pooling
Frequency Attentive Statistic Pooling
(Feat×T)
Modulation Dynamics Block

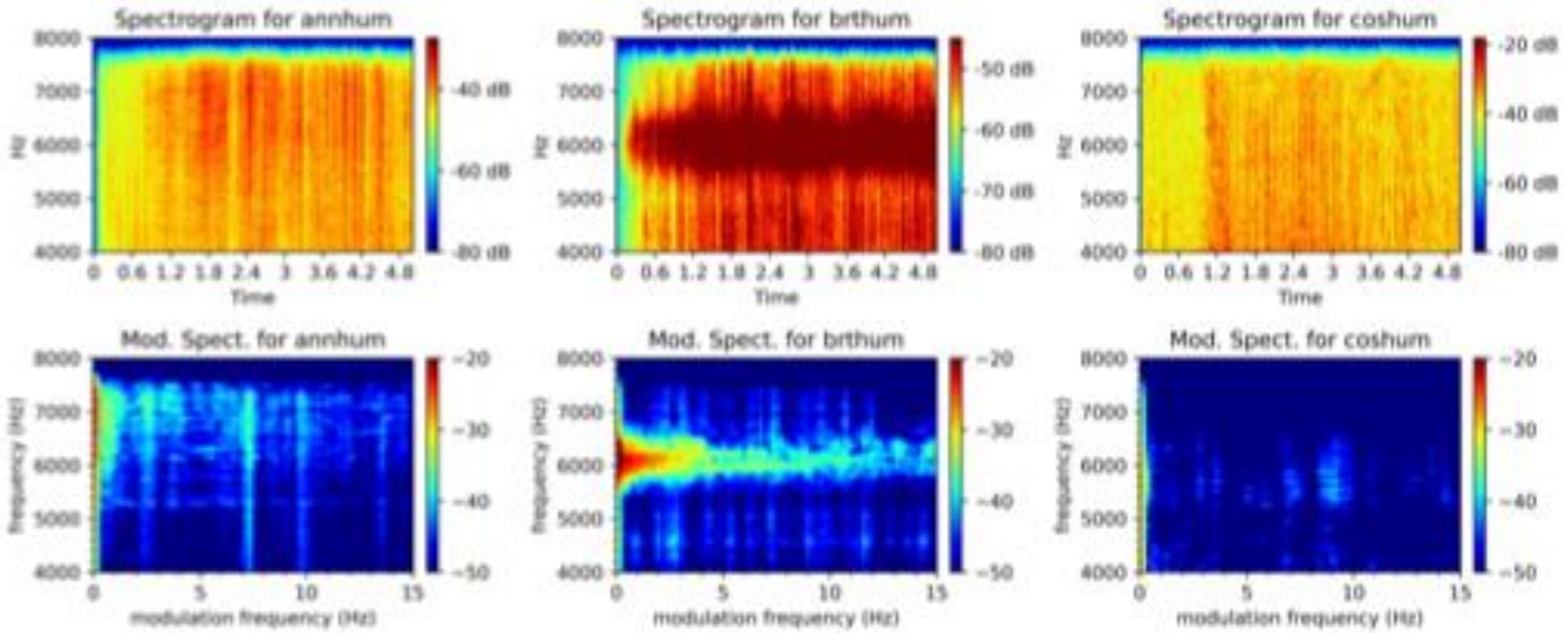

Genuine
MelGAN
MelGAN-Large
Parallel WaveGAN
Multiband_MelGAN
Fullband_MelGAN
HiFiGAN
WaveGlow
0.0
3.2
6.4
9.6
12.7
15.9
Modulation frequency (Hz)
0.020
0.015
0.010
0.005
0.000
−0.005
−0.010

Foundation Models for Edge Bioacoustics

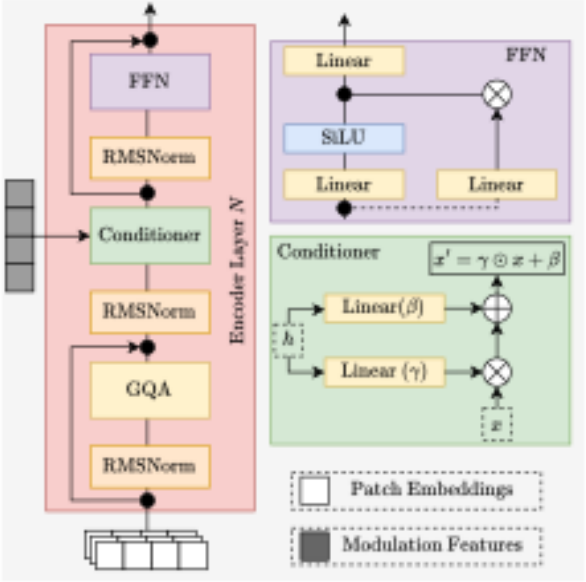
MAIN EVALUATION RESULTS. **BOLD** REPRESENTS THE BEST AND UNDERLINE THE SECOND BEST AVERAGE RESULTS

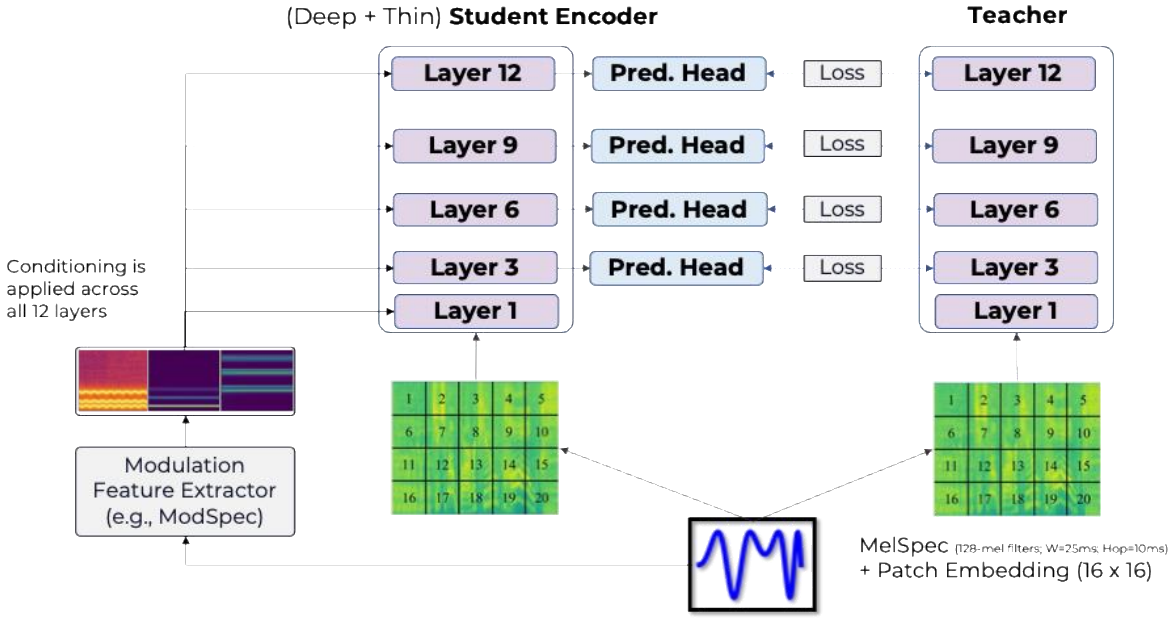
Upstream Information			Downstream Tasks				
Method	Pretraining Data	#P ↓	BSTS ROC-AUC ↑	BSTR MAE ↓	BID ACC ↑	VAD F1-Score ↑	Score
DSP-based Features							
FBanks	—	0	90.04 (± 2.40)	10.84 (± 0.29)	85.39 (± 9.24)	44.53 (± 8.46)	530
Spectrogram	—	0	86.75 (± 2.88)	10.95 (± 0.32)	94.49 (± 5.02)	51.10 (± 7.30)	607
MFCC	—	0	71.42 (± 8.94)	7.22 (± 3.83)	97.95 (± 1.22)	30.28 (± 5.90)	533
General-Purpose Audio Encoders							
BYOL-A	AS-2M	5M	63.81 (± 7.61)	<u>3.14</u> (± 0.21)	98.18 (± 0.91)	63.07 (± 8.83)	733
SS-AST (Tiny)	AS-2M + LS	6M	90.24 (± 2.43)	4.44 (± 0.28)	91.66 (± 2.42)	65.13 (± 1.97)	832
SS-AST (Small)	AS-2M + LS	23M	<u>92.70</u> (± 2.05)	4.45 (± 0.16)	96.61 (± 2.03)	72.43 (± 0.96)	**910**
SS-AST (Base)	AS-2M + LS	89M	**93.91** (± 1.44)	4.11 (± 0.22)	74.22 (± 5.95)	72.19 (± 2.32)	706
MSM-MAE	AS-2M	92M	89.33 (± 1.81)	4.05 (± 0.13)	97.24 (± 0.81)	73.27 (± 1.16)	863
M2D	AS-2M	85M	89.22 (± 3.45)	3.89 (± 0.11)	98.09 (± 0.66)	**74.49** (± 1.37)	<u>883</u>
CAV-MAE	AS-2M	85M	83.37 (± 7.72)	3.42 (± 0.22)	93.60 (± 2.50)	69.26 (± 3.31)	797
BEATs	AS-2M	90M	75.22 (± 4.30)	**3.05** (± 0.12)	<u>99.44</u> (± 0.18)	72.82 (± 1.06)	811
Bioacoustic-based Audio Encoders							
AVES	BIO	94M	90.48 (± 1.80)	3.35 (± 0.16)	**99.46** (± 0.08)	<u>74.42</u> (± 0.95)	**910**
BirdAVES (Base)	BIO + XC	94M	85.32 (± 3.64)	3.39 (± 0.17)	98.22 (± 0.53)	68.59 (± 2.93)	838
BirdAVES (Large)	BIO + XC	315M	88.98 (± 1.04)	3.56 (± 0.10)	99.02 (± 0.20)	72.72 (± 1.61)	743

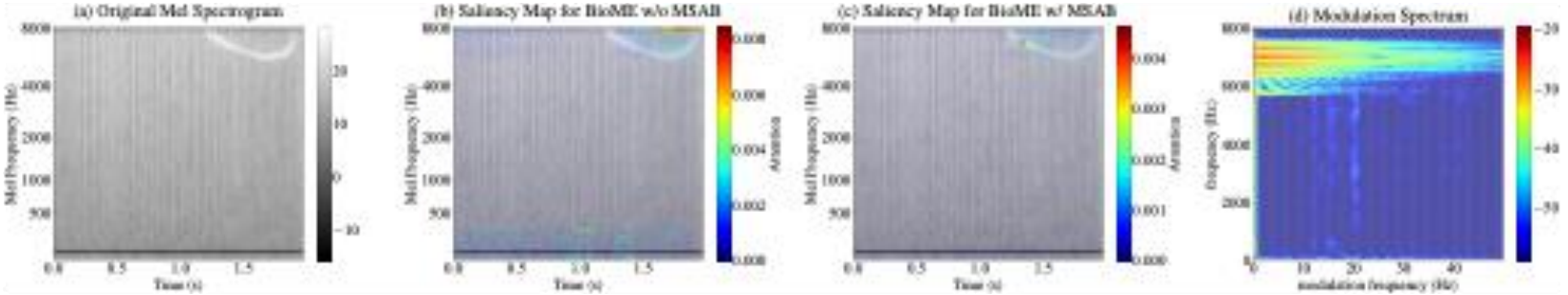
Representative Misclassifications



Inductive Bias to Improve Knowledge Distillation







RESULTS FOR ACOUSTIC BEEHIVE MONITORING TASKS. **BOLD** REPRESENTS THE BEST AND UNDERLINE THE SECOND BEST AVERAGE RESULTS.

Upstream Information		Downstream Tasks				
Method	#P (M) ↓	BSTS ROC-AUC ↑	BSTR MAE ↓	BID ACC ↑	VAD F1-Score ↑	Score
DSP-based Features						
FBanks	0	90.04 (± 2.40)	10.84 (± 0.29)	85.39 (± 9.24)	44.53 (± 8.46)	530
Spectrogram	0	86.75 (± 2.88)	10.95 (± 0.32)	94.49 (± 5.02)	51.10 (± 7.30)	607
MFCC	0	71.42 (± 8.94)	7.22 (± 3.83)	97.95 (± 1.22)	30.28 (± 5.90)	533
General-Purpose Audio Encoders						
BYOL-A	5	63.81 (± 7.61)	3.14 (± 0.21)	98.18 (± 0.91)	63.07 (± 8.83)	733
SS-AST (Tiny)	6	90.24 (± 2.43)	4.44 (± 0.28)	91.66 (± 2.42)	65.13 (± 1.97)	832
SS-AST (Small)	23	92.70 (± 2.05)	4.45 (± 0.16)	96.61 (± 2.03)	72.43 (± 0.96)	910
SS-AST (Base)	89	**93.91** (± 1.44)	4.11 (± 0.22)	74.22 (± 5.95)	72.19 (± 2.32)	706
MSM-MAE	92	89.33 (± 1.81)	4.05 (± 0.13)	97.24 (± 0.81)	73.27 (± 1.16)	863
M2D	85	89.22 (± 3.45)	3.89 (± 0.11)	98.09 (± 0.66)	**74.49** (± 1.37)	883
CAV-MAE	85	83.37 (± 7.72)	3.42 (± 0.22)	93.60 (± 2.50)	69.26 (± 3.31)	797
BEATs	90	75.22 (± 4.30)	**3.05** (± 0.12)	99.44 (± 0.18)	72.82 (± 1.06)	811
Bioacoustic-based Audio Encoders						
AVES	94	90.48 (± 1.80)	3.35 (± 0.16)	**99.46** (± 0.08)	74.42 (± 0.95)	910
BirdAVES (Base)	94	85.32 (± 3.64)	3.39 (± 0.17)	98.22 (± 0.53)	68.59 (± 2.93)	838
BirdAVES (Large)	315	88.98 (± 1.04)	3.56 (± 0.10)	99.02 (± 0.20)	72.72 (± 1.61)	743
[**Ours**] BioME (Bio)	6	92.36 (± 3.03)	4.04 (± 0.26)	97.19 (± 2.62)	68.76 (± 1.16)	**917**
[**Ours**] BioME (Bio)	26	79.62 (± 5.58)	3.63 (± 0.39)	97.96 (± 0.38)	68.03 (± 3.24)	833
[**Ours**] BioME (Bio)	75	70.59 (± 2.69)	3.59 (± 0.11)	98.51 (± 0.42)	73.13 (± 1.88)	770

musaelab.ca

@MuSAELab

Home About Our Team Publications Resources Contact Us EN

Books

[B2] R. Anghinah, W. Paiva, T. Falk and F. Fregni (Eds.), Neurotrama: from emergency room to back to day-by-day life, e-book, Frontiers Media, Lausanne, 2019, 95 pages. doi: 10.3389/978-2-88945-724-3.

[B1] Signal Processing and Machine Learning for Biomedical Big Data, Editors: Ervin Sejdic, Tiago H. Falk, CRC Press, July 2018.

Articles published or accepted in refereed journals

[J136] H. Guimarães, M. Abdohalli, Y. Zhu, S. Maucourt, N. Coallier, P. Giovenazzo, and T. Falk, Benchmarking Self-Supervised Audio Representations for IoT-Enabled Acoustic Beehive Monitoring, IEEE Internet of Things Journal, Vol. 12, No. 21, pp. 45000-45010, Nov. 2025.

[J135] M. Abdohalli, Y. Zhu, H. Guimarães, N. Coallier, S. Maucourt, P. Giovenazzo, and T. Falk, Audio Modulation Spectral Features for Improved Honeybee Colony Population Prediction, IEEE Sensors Journal, Early Access, Oct. 2025.

[J134] Y. Zhu and T. Falk, WavRx: a Disease-Agnostic, Generalizable, and Privacy-Preserving Speech Health Diagnostic Model, IEEE Journal of Biomedical and Health Informatics, Vol. 29, No. 9, pp. 6353-6365, Sept. 2025.

MuSAE Lab

19 followers · Montreal, QC, Canada · http://musaelab.ca · @MuSAELab

Overview · Repositories 17 · Projects · Packages · People 4

Popular repositories

SRMRToolbox — Public
A non-intrusive objective metric for speech quality and intelligibility for normal hearing listeners and cochlear implant users
MATLAB ☆ 73 ⑂ 34

amplitude-modulation-analysis-module — Public
Amplitude Modulation Analysis Module for Python
Python ☆ 42 ⑂ 14

MuLES — Public
MuSAE Lab EEG Server
LabVIEW ☆ 35 ⑂ 7

BLE-Toolkit-LabVIEW — Public
BLE (Bluetooth Low Energy) Toolkit for LabVIEW
LabVIEW ☆ 25 ⑂ 15

AUDDT — Public
A toolkit for benchmarking on a wide variety of audio deepfake datasets.
Python ☆ 25 ⑂ 3

muse_osc — Public
MATLAB ☆ 18 ⑂ 5

Thank you!
Questions?