Interspeech 2025

Preliminary Program

Version: 27/06/2025

Caution: This is a preliminary version of the scientific program and is subject to change at the discretion of the Technical Program Committee Chairs. Requests for changes to time slots or presentation formats will not be considered. The presentation format—oral or poster—does not reflect the scientific quality of the paper; both formats hold equal scientific merit.

Monday 18/08/2025

09:30-10:30 - Keynote speaker - Roger Moore

From Talking and Listening Devices to Intelligent Communicative Machines

11:00-13:00 - Oral - Area 1 - Models of Speech Production

Survey Talk	Speech production, dynamical systems **(Survey Talk, 40 mins)**
2320	Towards a dynamical model of transitions between fluent and stuttered speech
2091	Study of vocal fold vibration using M-mode ultrasound: a proof of concept
565	Articulatory Feature Prediction from Surface EMG during Speech Production
2387	Enhancing Acoustic-to-Articulatory Speech Inversion by Incorporating Nasality

11:00-13:00 - Oral - Area 2 - Speaking Styles, Register and Conversational Speech

968	Modeling Formant Dynamics in Mandarin /ai/: Effects of Speech Style and Speech Rate
1771	Representation of perceived prosodic similarity of conversational feedback
2087	Prolongation in Romanian
1720	Speech Reduction in French: The Relationship Between Vowel Space and Articulation Dynamics
2203	Stress in Spoken and Whistled Greek

11:00-13:00 - Oral - Area 3 - Speech Emotion Recognition 1

2123	Vector Quantized Cross-lingual Unsupervised Domain Adaptation for Speech Emotion Recognition
2191	HYFuse: Aligning Heterogeneous Speech Pre-Trained Representations in Hyperbolic Space for Speech Emotion Recognition
832	Meta-PerSER: Few-Shot Listener Personalized Speech Emotion Recognition via Meta-learning
2778	Breaking Resource Barriers in Speech Emotion Recognition via Data Distillation
418	Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition
1006	Learning More with Less: Self-Supervised Approaches for Low-Resource Speech Emotion Recognition

11:00-13:00 - Oral - Area 5 - Speech Analysis, Detection and Classification 1

357	Temporal Convolutional Network with Smoothed and Weighted Losses for Distant Voice Activity and Overlapped Speech Detection
2466	Attention Is Not Always the Answer Optimizing Voice Activity Detection with Simple Feature Fusion
155	SpeechMLC: Speech Multi-label Classification

193	Fully End-to-end Streaming Open-vocabulary Keyword Spotting with W-CTC Forced Alignment
2034	Comparative Evaluation of Acoustic Feature Extraction Tools for Clinical Speech Analysis
2472	Can We Trust Machine Learning? The Reliability of Features from Open-Source Speech Analysis Tools for Speech Modeling

11:00-13:00 - Oral - Area 6 - Real-time Speech Enhancement

19	Optimized Real-time Speech Enhancement with Deep SSMs on Raw Audio
242	A Two-Stage Hierarchical Deep Filtering Framework for Real-Time Speech Enhancement
2223	Real-Time Audio-Visual Speech Enhancement Using Pre-trained Visual Representations
2642	Lightweight Speech Enhancement Model Based on Harmonic Attention and Phase Estimation with Skin-Attachable Accelerometer
429	TSDT-Net: Ultra-Low-Complexity Two-Stage Model Combining Dual-Path-Transformer and Transform-Average-Concatenate Network for Speech Enhancement
1176	STRUCTURED CODEBOOK BASED HIERARCHICAL FRAMEWORK FOR DNN FOR COMPUTATIONALLY EFFICIENT SPEECH ENHANCEMENT

11:00-13:00 - Oral - Area 7 - Prosody in Speech Synthesis

2189	ProMode: A Speech Prosody Model Conditioned on Acoustic and Textual Inputs
723	Counterfactual Activation Editing for Post-hoc Prosody and Mispronunciation Correction in TTS Models
1940	Investigating Stochastic Methods for Prosody Modeling in Speech Synthesis
1098	GST-BERT-TTS: Prosody prediction without accentual labels for multi-speaker TTS using BERT with global style tokens
2032	ExagTTS: An Approach Towards Controllable Word Stress Incorporated TTS for Exaggerated Synthesized Speech Aiding Second Language Learners
564	Synthetic Data Generation for Phrase Break Prediction with Large Language Model

11:00-13:00 - Oral - Area 8 - Large Language Models in Speech Recognition

1503	LLM-based phoneme-to-grapheme for phoneme-based speech recognition
2003	Pinyin-Guided Chinese Large Language Model-based Speech Recognition
476	Text-Enhanced Audio Encoder for Large Language Model based Speech Recognition via Cross-Modality Pre-training with Unpaired Audio-Text Data
2214	Towards atypical speech transcription using LLM-based ASR
1707	Better Pseudo-labeling with Multi-ASR Fusion and Error Correction by SpeechLLM
1669	Leveraging LLM and Self-Supervised Training Models for Speech Recognition in Chinese Dialects: A Comparative Analysis

11:00-13:00 - Oral - Area 13 - Speech-based Cognitive Assessment 1

2685	HK-GenSpeech: A Generative AI Scene Creation Framework for Speech Based Cognitive Assessment
2461	Enhancing Automated Screening for Neurocognitive Disorders: A Cantonese Phonetic Fuzzy Matching Approach
2097	Leveraging AM and FM Rhythm Spectrograms for Dementia Classification and Assessment
1564	Leveraging Cascaded Binary Classification and Multimodal Fusion for Dementia Detection through Spontaneous Speech
1118	Whisper-Based Multilingual Alzheimer's Disease Detection and Improvements for Low-Resource Language
489	PPGs-BERT: Leveraging Phoneme Sequence and BERT for Alzheimer's Disease Detection from Spontaneous Speech

11:00-13:00 - Poster - Area 2 - Multilinguality, Cross-linguistic Studies, L2 Speech

716	Evaluation of 3 Automatic Alignment Tools for the Processing of Non-native French
2401	CrossPhon: An Auto Phone Mapping Tool to Streamline Cross-language Modeling for Phone Alignment of Low-resource Languages
1827	Multi-lingual and Zero-Shot Speech Recognition by Incorporating Classification of Language-Independent Articulatory Features
1040	Instantaneous changes in acoustic signals reflect syllable progression and cross-linguistic syllable variation
954	Influence of Proficiency and L2 Experience on Dynamic Spectral Cue Utilization in L2 Vowel Perception and Production
152	A Bayesian Approach to L2 Fluency Ratings by Native and Nonnative Listeners
502	Are loan sequences different from foreign sequences? A perception study with Japanese listeners on coronal obstruent – high front vowel sequences
2259	Relative cue weighting in multilingual stop voicing production
664	Variability in intervocalic /t/ and community diversity in Australian English

11:00-13:00 - Poster - Area 2 - Speech and Grammar/Articulatory Analyses

2246	Modeling Probabilistic Reduction using Information Theory and Naive Discriminative Learning
2256	Predictability effects on acoustic distinctiveness of read Polish speech
236	How do both phonological and syntactic complexity influence speech planning?
1121	When focus shapes the flow: prosodic restructuring in Mandarin complex nominals
740	Investigating the Impact of Word Informativeness on Speech Emotion Recognition
1869	Lexical stress affects lenition: The case of Italian palato-alveolar affricates
482	Evaluation of a model for sound radiation from the vocal tract wall
1853	FROST-EMA: Finnish and Russian Oral Speech Dataset of Electromagnetic Articulography Measurements with L1, L2 and Imitated L2 Accents

11:00-13:00 - Poster - Area 12 - Multimodal Resources

190	Hearing from Silence: Reasoning Audio Descriptions from Silent Videos via Vision-Language Model
1593	Mitigating Audiovisual Mismatch in Visual-Guide Audio Captioning
1559	ViCocktail: Automated Multi-Modal Data Collection for Vietnamese Audio-Visual Speech Recognition
776	GALAXY: A Large-Scale Open-Domain Dataset for Multimodal Learning
739	FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems
80	PersonaTAB: Predicting Personality Traits using Textual, Acoustic, and Behavioral Cues in Fully-Duplex Speech Dialogs
1105	FFD: Fine-Finger Diffusion Model for Music to Fine-grained Finger Dance Generation
79	Towards Diverse and Efficient Audio Captioning via Diffusion Models
1664	Pull It Together: Reducing the Modality Gap in Contrastive Learning

11:00-13:00 - Poster - Area 12 - Summarization

2792	Towards Multi-Level Transcript Segmentation: LoRA Fine-Tuning for Table-of-Contents Generation
1341	Pick and Summarize: Integrating Extractive and Abstractive Speech Summarization
1825	Beyond Similarity Scoring: Detecting Entailment and Contradiction in Multilingual and Multimodal Contexts
2771	Comparison-Based Automatic Evaluation for Meeting Summarization

11:00-13:00 - Poster - Area 12 - Spoken Machine Translation 1

1976	Speech transcription from South Tyrolean Dialect to Standard German with Whisper
1166	Length Aware Speech Translation for Video Dubbing
1305	ArticulateX: End-to-End Monolingual Speech Translation in Articulator Space
858	CMSP-ST: Cross-modal Mixup with Speech Purification for End-to-End Speech Translation
317	End-to-End Speech Translation Guided by Robust Translation Capability of Large Language Model
2341	Empowering Large Language Models for End-to-End Speech Translation Leveraging Synthetic Data
1954	Speech-to-Text Translation with Phoneme-Augmented CoT: Enhancing Cross-Lingual Transfer in Low-Resource Scenarios
1595	Scheduled Interleaved Speech-Text Training for Speech-to-Speech Translation with LLMs
2525	End-to-End Speech Translation for Low-Resource Languages Using Weakly Labeled Data

11:00-13:00 - Poster - Area 13 - Depression Detection and Assessment 1

1438	Speech reference intervals: an assessment of feasibility in depression symptom severity prediction
280	DepressGEN: Synthetic Data Generation Framework for Depression

	Detection
1597	Emotion-Guided Graph Attention Networks for Speech-Based Depression Detection under Emotion-Inducting Tasks
1181	Explainable Depression Detection using Masked Hard Instance Mining
2378	Test-Time Training for Speech-based Depression Detection
638	Leveraging Ordinal Information for Speech-based Depression Classification
235	Zero-Shot Speech-Based Depression and Anxiety Assessment with LLMs
2556	Towards the Objective Characterisation of Major Depressive Disorder Using Speech Data from a 12-week Observational Study with Daily Measurements

11:00-13:00 - Poster - Area 14 - Special Session - The 1st SpeechWellness Challenge

2755	Leveraging Text and Speech Processing for Suicide Risk Classification in Chinese Adolescents
124	The 1st SpeechWellness Challenge: Detecting Suicide Risk Among Adolescents
1483	Leveraging Large Language Models for Spontaneous Speech-Based Suicide Risk Detection
2419	Predicting Adolescent Suicidal Risk from Multi-task-based Speech: An Ensemble Learning Approach
2083	In-context learning to detect suicide risk among adolescents
1372	Language-Agnostic Suicidal Risk Detection Using Large Language Models
599	Network of acoustic characteristics for the automatic detection of suicide risk from speech. Contribution to the 2025 SpeechWellbeing challenge by the Semawave team

11:00-13:00 - Special session - Area 14 - Special Session - Interpretability in Audio and Speech Technology

1143	EnvSDD: Benchmarking Environmental Sound Deepfake Detection
918	Echoes of Phonetics: Unveiling Relevant Acoustic Cues for ASR via Feature Attribution
288	Benchmarking Time-localized Explanations for Audio Classification Models
1021	Spectrotemporal Modulation: Efficient and Interpretable Feature Representation for Classifying Speech, Music, and Environmental Sounds
577	Discrete Tokens Exhibit Interlanguage Speech Intelligibility Benefit: an Analytical Study Towards Accent-robust ASR Only with Native Speech Data
945	Analysis of Semantic and Acoustic Token Variability Across Speech, Music, and Audio Domains
46	Is your model big enough? Training and interpreting large-scale monolingual speech foundation models
2574	Semantic-Aware Interpretable Multimodal Music Auto-Tagging
2180	From Words to Waves: Analyzing Concept Formation in Speech and Text-Based Foundation Models

1932	Effective Context in Neural Speech Models
106	Word stress in self-supervised speech models: A cross-linguistic comparison
1526	What do self-supervised speech models know about Dutch? Analyzing advantages of language-specific pre-training
514	Iterative refinement, not training objective, makes HuBERT behave differently from wav2vec 2.0
1911	On the reliability of feature attribution methods for speech classification
2225	An Exploration of Interpretable Deep Learning Models for the Assessment of Mild Cognitive Impairment

11:00-13:00 - Show and Tell - ASR / Tools

2816	Voxplorer: Voice data exploration and projection in an interactive dashboard
2820	ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems
2821	Transcribing Oral History Recordings Using the Transcription Portal
2822	LiRI Corpus Platform: Demonstration of a Web-Based Infrastructure for Multimodal Corpus Analysis
2831	Speech Annotation for A: Accuracy, Access, and Application
2833	LATE: Open Source Toolkit for Latvian and Latgalian Speech Transcription
2838	Scalable Offline ASR for Command-Style Dictation in Courtrooms

14:30-16:30 - Oral - Area 2 - Tone

1173	Neutral Tone Variation in Beijing Mandarin: Is Neutral Tone Toneless?
1673	The Role of Syntactic Structures in Shaping Directionality in Trisyllabic Tone Sandhi: Evidence from Tianjin Mandarin
1835	Acoustic Representation and Realization of Weak Elements Subcategories: In the Case of Tianjin Mandarin
1798	Lexical competition in the process of Cantonese tone merging: Diverse Impact Mechanisms Across Different Individuals and Tone Pairs
284	Tonal Perception in Changde Mandarin
283	Tonal Contrasts in the Malipo Variety of the Mienic Language

14:30-16:30 - Oral - Area 4 - Robust Speaker Verification

183	SSPS: Self-Supervised Positive Sampling for Robust Self-Supervised Speaker Verification
1145	ParaNoise-SV: Integrated Approach for Noise-Robust Speaker Verification with Parallel Joint Learning of Speech Enhancement and Noise Extraction
1865	Disentangling Speaker and Content in Pre-trained Speech Models with Latent Diffusion for Robust Speaker Verification
2167	Evaluating Deep Speaker Embedding Robustness to Domain, Sampling Rate, and Codec Variations
42	Towards Robust Speaker Recognition against Intrinsic Variation with Foundation Model Few-shot Tuning and Effective Speech Synthesis
655	Bayesian Learning for Domain-Invariant Speaker Verification and Anti-Spoofing

14:30-16:30 - Oral - Area 5 - Spatial Audio and Acoustics 1

60	Temporal Modeling of Room Impulse Response Generation via Multi-Scale Autoregressive Learning
588	Effect of noise floor in room impulse response on speech perception under spherical harmonics-based spatial sound reproduction
1912	Direction-Aware Neural Acoustic Fields for Few-Shot Interpolation of Ambisonic Impulse Responses
2478	AuralNet: Hierarchical Attention-based 3D Binaural Localization of Overlapping Speakers
1379	SoundSculpt: Direction and Semantics Driven Ambisonic Target Sound Extraction
9	TF-Mamba: A Time-Frequency Network for Sound Source Localization

14:30-16:30 - Oral - Area 9 - Decoding Algorithms

382	Simultaneous Masked and Unmasked Decoding with Speculative Decoding Masking for Fast ASR without Accuracy Loss
148	WIND: Accelerated RNN-T Decoding with Windowed Inference for Non-blank Detection
955	NGPU-LM: GPU-Accelerated N-Gram Language Model for Context-Biasing in Greedy ASR Decoding
1388	Pushing the Limits of Beam Search Decoding for Transducer-based ASR models
2669	SKIP-SALSA: Skip Synchronous Fusion of ASR LLM Decoders
1290	Efficient Trie-based Biasing using K-step Prediction for Rare Word Recognition

14:30-16:30 - Oral - Area 11 - Conversation, Communication and Interaction 1

167	Gaze-Enhanced Multimodal Turn-Taking Prediction in Triadic Conversations
668	Visual Cues Support Robust Turn-taking Prediction in Noise
1316	Backchannel prediction for natural spoken dialog systems using general speaker and listener information
1117	Rapport-Building Dialogue Strategies for Deeper Connection: Integrating Proactive Behavior, Personalization, and Aizuchi Backchannels
1535	Does effortful speech production indicate communication difficulty caused by noise and hearing aid support?
697	"Dyadosyncrasy", Idiosyncrasy and Demographic Factors in Turn-Taking

14:30-16:30 - Oral - Area 12 - Spoken Machine Translation 2

478	Structured pruning for efficient systolic-array accelerated cascade Speech-to-Text Translation
887	Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language)
380	Multilingual Query-by-Example KWS for Indian Languages using Transliteration
568	Novel Parasitic Dual-Scale Modeling for Efficient and Accurate Multilingual Speech Translation
318	A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard

	Speech Translation
2714	NIRANTAR: Continual Learning with New Languages and Domains on Real-world Speech Data

14:30-16:30 - Oral - Area 13 - Pathological Speech Analysis 1

Survey Talk	Spoken language biomarkers for neurodegenerative diseases **(Survey Talk, 40 mins)**
743	Addressing Task Conflicts in Stuttering Detection via MoE-Based Multi-Task Learning
2767	Comparison of Acoustic and Textual Features for Dysarthria Severity Classification in Amyotrophic Lateral Sclerosis
151	StutterCut: Uncertainty-Guided Normalised Cut for Dysfluency Segmentation
2313	Physiologically-Informed Feature Analysis of Acquired Speech Disorders for Stroke Assessment

14:30-16:30 - Oral - Area 14 - Special Session - Queer and Trans Speech Science and Technology

1050	Web-Based Application for Real-Time Biofeedback of Vocal Resonance in Gender-Affirming Voice Training: Design and Usability Evaluation
2372	On the Production and Perception of a Single Speaker's Gender
86	Conveying gender through speech: insights from trans men
2022	Queer Waves: A German Speech Dataset Capturing Gender and Sexual Diversity from Podcasts and YouTube
2128	Reddit FlairShare: A Human-Annotated Dataset of Gender-Progressive Online Discourse
2229	Voices of 'cyborg awesomeness': Posthuman embodiment of nonbinary gender expression in AI speech technologies

14:30-16:30 - Poster - Area 1 - Articulatory and Vocal Tract Modelling

1820	Articulatory modeling of the S-shaped F2 trajectories observed in Öhman's spectrographic analysis of VCV syllables
349	Random Forest Classifier Modeling of Dynamic Spectral Features in Japanese Whispered Vowels
1716	The Role of Voiced Consonant Duration in Sung Vowel-Consonant and Consonant-Vowel Recognition
125	How sibilant spectra shape gender perception in prepubertal children: A voice morphing study
1650	Constrained LDDMM for Dynamic Vocal Tract Morphing: Integrating Volumetric and Real-Time MRI
1486	2D Immersed Boundary Method in Vocal Tract Acoustics: An Eulerian–Lagrangian Model for Simulation of Diphthongs
963	Reconstruction of the Complete Vocal Tract Contour Through Acoustic to Articulatory Inversion Using Real-Time MRI Data
1151	Co-registration of real-time MRI and respiration for speech research

14:30-16:30 - Poster - Area 1 - Advances in Modelling and Imaging

1751	Theoretical proposal for a unified Bayesian model of adaptation in non-

	interactive and interactive speech production
866	Self-supervised Optimality-Guided Learning of Speech Articulation
1125	Extended High-frequency Cues to Phoneme Recognition: Insights from ASR
700	Decoding Speaker-Normalized Pitch from EEG for Mandarin Perception
1187	SSF-DST: A Spectro-Spatial Features Enhanced Deep Spatiotemporal Network for EEG-Based Auditory Attention Detection
85	Overestimated performance of auditory attention decoding caused by experimental design in EEG recordings
2597	A real-time MRI study on asymmetry in velum dynamics during VCV production with nasal sounds
2444	Exploratory Analysis of Brainstem fMRI Data During Sustained Phonation

14:30-16:30 - Poster - Area 6 - Echo Cancellation, Feedback Control, and Near-end Enhancement

197	Room Impulse Response as a Prompt for Acoustic Echo Cancellation
608	CAGCRN: Real-Time Speech Enhancement with a Lightweight Model for Joint Acoustic Echo Cancellation and Noise Suppression
1572	Exploiting Echo Path Priors for Enhanced Stereo Acoustic Echo Cancellation
2177	Extended Loss: Incorporating Long Context into Training Models when using Short Audio Frames
1496	Analysis and Extension of a Near-End Listening Enhancement Method Based on Long-Term Fractile Noise Statistics
605	A Novel Deep Learning Framework for Efficient Multichannel Acoustic Feedback Control

14:30-16:30 - Poster - Area 6 - Speech Coding and Echo Cancellation

2483	Synonymity-Based Semantic Coding for Efficient Speech Compression
1369	Towards an Ultra-Low-Delay Neural Audio Coding with Computational Efficiency
546	SpecTokenizer: A Lightweight Streaming Codec in the Compressed Spectrum Domain
921	TS3-Codec: Transformer-Based Simple Streaming Single Codec
1253	Towards Bitrate-Efficient and Noise-Robust Speech Coding with Variable Bitrate RVQ
1335	LSPnet: an ultra-low bitrate hybrid neural codec
1409	Vision-Integrated High-Quality Neural Speech Coding
1370	Neural Spectral Band Generation for Audio Coding
647	Multi-Channel Acoustic Echo Cancellation Based on Direction-of-Arrival Estimation

14:30-16:30 - Poster - Area 8 - Multilingual ASR

1359	Switch Conformer with Universal Phonetic Experts for Multilingual ASR
1875	Language-Aware Prompt Tuning for Parameter-Efficient Seamless Language Expansion in Multilingual ASR
1374	Efficient Multilingual ASR Finetuning via LoRA Language Experts
775	Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic

	Speech Recognition
2351	Effects of Speaker Count, Duration, and Accent Diversity on Zero-Shot Accent Robustness in Low-Resource ASR
1839	Leveraging Geographic Metadata for Dialect-Aware Speech Recognition
2260	Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning
398	VietASR: Achieving Industry-level Vietnamese ASR with 50-hour labeled data and Large-Scale Speech Pretraining
184	Open Universal Arabic ASR Leaderboard

14:30-16:30 - Poster - Area 9 - Cross-Lingual and Multilingual Processing

460	Building an Accurate Open-Source Hebrew ASR System through Crowdsourcing
1557	A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic
2296	Automatic Speech Recognition for Low-Resourced Middle Eastern Languages
2626	In-context Language Learning for Endangered Languages in Speech Recognition
2247	CS-FLEURS: A Massively Multilingual and Code-Switched Speech Dataset
1548	Weakly Supervised Data Refinement and Flexible Sequence Compression for Efficient Thai LLM-based ASR
654	Can we train ASR systems on Code-switch without real code-switch data? Case study for Singapore's languages
2646	Swedish Whispers; Leveraging a Massive Speech Corpus for Swedish Speech Recognition

14:30-16:30 - Poster - Area 13 - Hearing Disorders

2448	Robot-assisted Recognition of Vocal Emotions in Pseudospeech for Cochlear Implanted Adolescents
2458	Using Neurogram Similarity Index Measure (NSIM) to Model Hearing Loss and Cochlear Neural Degeneration
1285	Contrastive Learning-based Syllable-Level Mispronunciation Detection and Diagnosis for Speech Audiometry
2111	A Deformable Convolution GAN Approach for Speech Dereverberation in Cochlear Implant Users
1111	L3C-DeepMFC: Low-Latency Low-Complexity Deep Marginal Feedback Cancellation with Closed-Loop Fine Tuning for Hearing Aids
1455	Semantic Processing During Spoken Word Production by Children with Cochlear Implants
72	Linguistic Masking and Its Release in Simulated Electric-acoustic Hearing

14:30-16:30 - Poster - Area 13 - Acoustic Assessment of Respiratory Health

899	SPEAKtoCOPD: a flashmob study to collect COPD speech
861	Developing a LeFF Transformer Model for Exacerbated Speech Detection in COPD and Asthma
84	Towards Pre-training an Effective Respiratory Audio Foundation Model
910	Effect of exercise on voice in people living with COPD

1190	Adaptive Differential Denoising for Respiratory Sounds Classification
1209	Disentangling Dual-Encoder Masked Autoencoder for Respiratory Sound Classification
1477	Patient-Aware Feature Alignment for Robust Lung Sound Classification: Cohesion-Separation and Global Alignment Losses
1295	Improving Respiratory Sound Classification with Architecture-Agnostic Knowledge Distillation from Ensembles

14:30-16:30 - Special session - Area 14 - Special Session - Interspeech 2025 URGENT Challenge

1246	Lessons Learned from the URGENT 2024 Speech Enhancement Challenge
1363	Interspeech 2025 URGENT Speech Enhancement Challenge
734	TS-URGENet: A Three-stage Universal Robust and Generalizable Speech Enhancement Network
749	Multistage Universal Speech Enhancement System for URGENT Challenge
795	Scaling beyond Denoising: Submitted System and Findings in URGENT Challenge 2025
1889	DeepFilterGAN: A Full-band Real-time Speech Enhancement System with GAN-based Stochastic Regeneration
251	FUSE: Universal Speech Enhancement using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge
900	Universal Speech Enhancement with Regression and Generative Mamba

Tuesday 19/08/2025

08:30-10:30 - Oral - Area 1 - Acoustic and Articulatory Cues in Speech Perception

859	Multitalker Babble in English Vowel Perception Training: A Comparison between Humans and Neural Models
2199	Speech stimulus design to study the neural coding of speech and the impact of cochlear synaptopathy
892	Prediction of listening effort ratings for habitual and clear-Lombard speech presented in noise
2088	Language and Accent Familiarity Effects on the Use of Acoustic Cues in Talker Identification
2303	Characterization of voice cue sensitivity and vocal emotion recognition across the adult lifespan
1168	Creaky voice facilitates more efficient phonological processing of Mandarin Tone 3

08:30-10:30 - Oral - Area 4 - Speaker Diarization 1

484	Fine-tune Before Structured Pruning: Towards Compact and Accurate Self-Supervised Models for Speaker Diarization
334	Count Your Speakers! Multitask Learning for Multimodal Speaker Diarization
2030	End-to-End Diarization utilizing Attractor Deep Clustering
2505	SDBench: A Comprehensive Benchmark Suite for Speaker Diarization
1396	Enhancing Serialized Output Training for Multi-Talker ASR with Soft Monotonic Alignment and Utterance-level Timestamp
120	Pretraining Multi-Speaker Identification for Neural Speaker Diarization

08:30-10:30 - Oral - Area 5 - Audio Event Detection and Classification

1642	Training Onset-and-Offset Aware Sound Event Detection on a Heterogeneous Dataset via Probabilistic Sequential Modeling
1613	Multi-view Fusion and Parameter Perturbation for Few-Shot Class-Incremental Audio Classification
1085	Fully Few-shot Class-incremental Audio Classification Using Multi-level Embedding Extractor and Ridge Regression Classifier
1998	Beyond Conventional Metrics: using Entropic Triangles to Explain Balancing Methods in Acoustic Scene Classification
886	Domain Adaptation Method and Modality Gap Impact in Audio-Text Models for Prototypical Sound Classification
1356	Unified Microphone Conversion: Many-to-Many Device Mapping via Feature-wise Linear Modulation

08:30-10:30 - Oral - Area 6 - Multi-channel Speech Enhancement

1462	Mel-McNet: A Mel-Scale Framework for Online Multichannel Speech Enhancement
1171	A Lightweight Hybrid Dual Channel Speech Enhancement System under Low-SNR Conditions

1178	ARiSE: Auto-Regressive Multi-Channel Speech Enhancement
773	WTFormer: A Wavelet Conformer Network for MIMO Speech Enhancement with Spatial Cues Peservation
369	A Three-Stage Beamforming with Harmonic Guidance for Multi-Channel Speech Enhancement
1502	Speech Enhancement with Dual-path Multi-Channel Linear Prediction Filter and Multi-norm Beamforming

08:30-10:30 - Oral - Area 7 - Multilingual Speech Synthesis and Special Applications 1

1344	Parameter-Efficient Fine-Tuning for Low-Resource Text-to-Speech via Cross-Lingual Continual Learning
2520	Accent Normalization Using Self-Supervised Discrete Tokens with Non-Parallel Data
1752	LIST: Language-Independent Speech Token for Multilingual Speech Synthesis with Language Models
469	Developing High-Quality TTS for Punjabi and Urdu: Benchmarking against MMS Models
1443	Synthesizing speech with selected perceptual voice qualities - A case study with creaky voice
762	Intrasentential English in Swedish TTS: perceived English-accentedness

08:30-10:30 - Oral - Area 8 - Self-supervised Learning

1280	Exploring SSL Discrete Speech Features for Zipformer-based Contextual ASR
463	Self-supervised learning of speech representations with Dutch archival data
1616	GigaAM: Efficient Self-Supervised Learner for Speech Recognition
29	DiceHuBERT: Distilling HuBERT with a Self-Supervised Learning Objective
593	Differentiable K-means for Fully-optimized Discrete Token-based ASR
34	Towards Early Prediction of Self-Supervised Speech Model Performance

08:30-10:30 - Oral - Area 12 - Inclusivity

Survey Talk	Towards more fair and inclusive speech technology **(Survey Talk, 40 mins)**
1104	The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages
1102	FaiST: A Benchmark Dataset for Fairness in Speech Technology
481	On the Language and Gender Biases in PSTN, VoIP and Neural Audio Codecs
1760	Evaluating Speech Enhancement Performance Across Demographics and Language

08:30-10:30 - Oral - Area 13 - Speech-based Cognitive Assessment 2

1080	Optimizing Pause Context in Fine-Tuning Pre-trained Large Language Models for Dementia Detection
2099	WhisperD: Dementia Speech Recognition and Filler Word Detection with

	Whisper
91	Acoustic and Linguistic Biomarkers for Cognitive Impairment Detection from Speech
2428	Alzheimer's Dementia Detection Using Perplexity from Paired Large Language Models
871	Understanding Dementia Speech Alignment with Diffusion-Based Image Generation
465	ClaritySpeech: Dementia Obfuscation in Speech

08:30-10:30 - Poster - Area 4 - Language and Accent Identification and Speaker Privacy

2120	Teacher-Free Knowledge Distillation for Improving Short-Utterance Spoken Language Identification
2300	LID Models are Actually Accent Classifiers: Implications and Solutions for LID on Accented Speech
2342	Analyzing the Impact of Accent on English Speech: Acoustic and Articulatory Perspectives
911	A Study of Speech Embedding Similarities Between Australian Aboriginal and High-Resource Languages
2361	An Investigative Study on Recent Sharpness- and Flatness-Based Optimizers for Enhanced Self-Supervised Speaker Verification
1096	Privacy-Preserving Speaker Verification via End-to-End Secure Representation Learning
2290	Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy
364	Federated Learning with Feature Space Separation for Speaker Recognition
1076	Differentially Private Parameter-Efficient Fine-tuning for Large ASR Models

08:30-10:30 - Poster - Area 4 - Characterization and Multimodal Approaches for Speaker Recognition

312	Parameter-Efficient Fine-tuning with Instance-Aware Prompt and Parallel Adapters for Speaker Verification
1984	Unified Text and Speaker Verification using SSL model for Text-Dependent Speaker Verification
1778	Towards Robust Overlapping Speech Detection: A Speaker-Aware Progressive Approach Using WavLM
2029	Towards Secure User Authentication for Headphones via In-Ear or In-Earcup Microphones
1625	Mimic Blocker: Self-Supervised Adversarial Training for Voice Conversion Defense with Pretrained Feature Extractors
1103	A Siamese Network-Based Framework for Voice Mimicry Proficiency Assessment Using X-Vector Embeddings
2171	Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models
2668	Multimodal Zero-Shot Framework for Deepfake Hate Speech Detection in Low-Resource Languages

1299	Joint Target-Speaker ASR and Activity Detection

08:30-10:30 - Poster - Area 5 - Source Separation 1

576	Quadruple Path Modeling with Latent Feature Transfer for Permutation-free Continuous Speech Separation
1552	End-to-End DOA-Guided Speech Extraction in Noisy Multi-Talker Scenarios
49	Speaker Separation for an Unknown Number of Speakers with Encoder-Decoder-Based Contextual Information Module
1764	Attractor-Based Speech Separation of Multiple Utterances by Unknown Number of Speakers
2121	ReSepNet: A Unified-Light Model for Recursive Speech Separation with Unknown Speaker Count
844	Deep-Simplex Multichannel Speech Separation
1315	FLASepformer: Efficient Speech Separation with Gated Focused Linear Attention Transformer
753	Power Spectral Density Estimation for Acoustic Source Separation Using A Spherical Microphone Array
40	Exploring Efficient Directional and Distance Cues for Regional Speech Separation

08:30-10:30 - Poster - Area 5 - Acoustic Analysis and Bioacoustics

1658	Analysis of Avian Biphonic Vocalization Using Computational Modelling
1287	Dog2vec: Self-Supervised Pre-Training for Canine Vocal Representation
2516	Improving Bird Classification with Primary Color Additives
571	Exploring the Power of Empirical Mode Decomposition for Sensing the Sound of Silence: A Pilot Study on Mice Autism Detection via Ultrasonic Vocalisation
1139	Exploring Pre-trained models on Ultrasound Modeling for Mice Autism Detection with Uniform Filter Bank and Attentive Scoring
127	MADUV: The 1st INTERSPEECH Mice Autism Detection via Ultrasound Vocalization Challenge
2175	Significance of Time-Frequency preprocessing for automatic Ultrasonic Vocalization classification in Autism Spectrum Disorder model detection
2311	Robust Vocal Intensity Prediction: Overcoming Dataset Bias with Pretrained Deep Models
1453	SLASH: Self-Supervised Speech Pitch Estimation Leveraging DSP-derived Absolute Pitch

08:30-10:30 - Poster - Area 7 - Singing Voice and Audio Synthesis

1397	VibE-SVC: Vibrato Extraction with High-frequency F0 Contour for Singing Voice Conversion
766	TVC-MusicGen: Time-varying Structure Control for Background Music Generation via Self-supervised Training
1032	Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation
816	Bridging Speech and Singing: Multi-stage Speech-Prompted Singing Voice Conversion with Speaker Embedding Adaptation

154	Neurodyne: Neural Pitch Manipulation with Representation Learning and Cycle-Consistency GAN
1364	VS-Singer: Vision-Guided Stereo Singing Voice Synthesis with Consistency Schrödinger Bridge
305	DAFMSVC: One-Shot Singing Voice Conversion with Dual Attention Mechanism and Flow Matching
1531	Simple and Effective Content Encoder for Singing Voice Conversion via Dimension Reduction
1247	Song Form-aware Full-Song Text-to-Lyrics Generation with Multi-Level Granularity Syllable Count Control

08:30-10:30 - Poster - Area 7 - Voice Conversion 1

815	Towards Better Disentanglement in Non-Autoregressive Zero-Shot Expressive Voice Conversion
383	Voice Conversion for Likability Control via Automated Rating of Speech Synthesis Corpora
1434	REWIND: Speech Time Reversal for Enhancing Speaker Representations in Diffusion-based Voice Conversion
2043	Training-Free Voice Conversion with Factorized Optimal Transport
1229	E2E-BPVC: End-to-End Background-Preserving Voice Conversion via In-Context Learning
2684	Discl-VC: Disentangled Discrete Tokens and In-Context Learning for Controllable Zero-Shot Voice Conversion
1779	ReFlow-VC: Zero-shot Voice Conversion Based on Rectified Flow and Speaker Feature Optimization
2697	"In This Environment, As That Speaker": A Text-Driven Framework for Multi-Attribute Speech Conversion
438	LinearVC: Linear transformations of self-supervised features through the lens of voice conversion
1081	Speaker Normalization and Content Restoration for Zero-Shot Voice Conversion with Attention-Enhanced Discriminator

08:30-10:30 - Special session - Area 14 - Special Session - Source Tracing: The Origins of Synthetic or Manipulated Speech

2001	Audio Deepfake Source Tracing using Multi-Attribute Open-Set Identification and Verification
2079	Unveiling Audio Deepfake Origins: A Deep Metric learning And Conformer Network Approach With Ensemble Fusion
1297	Codec-Based Deepfake Source Tracing via Neural Audio Codec Taxonomy
472	TADA: Training-free Attribution and Out-of-Domain Detection of Audio Deepfakes
1490	Source Verification for Speech Deepfakes
2065	STOPA: A Dataset of Systematic VariaTion Of DeePfake Audio for Open-Set Source Tracing and Attribution
1757	Synthetic Speech Source Tracing using Metric Learning
16	Listen, Analyze, and Adapt to Learn New Attacks: An Exemplar-Free Class Incremental Learning Method for Audio Deepfake Source Tracing
538	VIB-based Real Pre-emphasis Audio Deepfake Source Tracing

1993	Defending Unauthorized Voice Cloning with Watermark-Aware Codecs
1269	Open-Set Source Tracing of Audio Deepfake Systems

11:00-12:00 - Keynote speaker - Alexander Waibel

From Speech Science to Language Transparence

13:30-15:30 - Oral - Area 2 - Tools for Speech Analysis

Survey Talk	Bridging speech science and technology for language varieties and low-resourced languages **(Survey Talk, 40 mins)**
1052	Toolkit for acoustic-phonetic analysis of naturalistic speech data
1037	VoiceNet: Multilingual On-Device Phoneme-To-Audio Alignment
1775	Nosey: Open-source hardware for acoustic nasalance
491	Automatic classification of stop realisation with wav2vec2.0

13:30-15:30 - Oral - Area 4 - Deepfake Detection

172	Bona fide Cross Testing Reveals Weak Spot in Audio Deepfake Detection Systems
527	BiCrossMamba-ST: Speech Deepfake Detection with Bidirectional Mamba Spectro-Temporal Cross-Attention
1442	Few-Shot Speech Deepfake Detection Adaptation with Gaussian Processes
20	Replay Attacks Against Audio Deepfake Detection
422	Enhancing Audio Deepfake Detection by Improving Representation Similarity of Bonafide Speech
1594	Generalizable Audio Deepfake Detection via Hierarchical Structure Learning and Feature Whitening in Poincaré sphere

13:30-15:30 - Oral - Area 5 - Speaker Extraction 1

314	SC-TSE : Speaker Consistency-Aware Target Speaker Extraction
2321	Robust Target Speaker Diarization and Separation via Augmented Speaker Embedding Sampling
1554	Inter-Speaker Relative Cues for Text-Guided Target Speech Extraction
2662	REAL-T: Real Conversational Mixtures for Target Speaker Extraction
43	Online Audio-Visual Autoregressive Speaker Extraction
1371	Plug-and-Play Co-Occurring Face Attention for Robust Audio-Visual Speaker Extraction

13:30-15:30 - Oral - Area 7 - Text Processing and Evaluation for Speech Synthesis 1

308	Acquiring Pronunciation from Speech Audio via Multi-task Learning
779	Intelligibility of Text-to-Speech Systems for Mathematical Expressions
2765	The State Of TTS: A Case Study with Human Fooling Rates
2283	Pairwise Evaluation of Accent Similarity in Speech Synthesis
902	VoiceQualityVC: A Voice Conversion System for Studying the Perceptual Effects of Voice Quality in Speech
2190	Towards Frame-level Quality Predictions of Synthetic Speech

13:30-15:30 - Oral - Area 9 - Novel Architectures for ASR

1701	Weight Factorization and Centralization for Continual Learning in Speech Recognition
2446	Dysfluent WFST: A Framework for Zero-Shot Speech Dysfluency Transcription and Detection
89	Dysarthric Speech Recognition Using Curriculum Learning and Multi-stream Architecture
129	DYNAC: Dynamic Vocabulary based Non-Autoregressive Contextualization for Speech Recognition
643	Beyond Hard Sharing: Efficient Multi-Task Speech-to-Text Modeling with Supervised Mixture of Experts
1062	OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning

13:30-15:30 - Oral - Area 10 - Audio-Visual ASR and Multimodal System

1036	Text Entry for All: Towards Speech-based Multimodal Interaction for Inclusion, Accessibility and the Preservation of the World's Linguistic Heritage **(Blue Sky paper, 40 mins)**
111	Scaling and Enhancing LLM-based AVSR: A Sparse Mixture of Projectors Approach
676	Cocktail-Party Audio-Visual Speech Recognition
1464	Efficient Noise-Robust Hybrid Audiovisual Encoder with Joint Distillation and Pruning for Audiovisual Speech Recognition
451	Unified Audio-Visual Modeling for Recognizing Which Face Spoke When and What in Multi-Talker Overlapped Speech and Video

13:30-15:30 - Oral - Area 12 - Speech Assessment

839	SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information
1823	Continual Speech Learning with Fused Speech Features
1960	Uni-VERSA: Versatile Evaluation of Speech with a Unified Framework
2164	Evaluating ASR robustness to spontaneous speech errors: A study of WhisperX using a Speech Error Database
2693	Is Synthetic Data Truly Effective for Training Speech Language Models?
2245	How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not

13:30-15:30 - Oral - Area 14 - Special Session - Multimodal Information Based Speech Processing (MISP) 2025 Challenge

1382	Pseudo Labels-based Neural Speech Enhancement for the AVSR Task in the MISP-Meeting Challenge
2017	The Multimodal Information Based Speech Processing (MISP) 2025 Challenge: Audio-Visual Diarization and Recognition
875	Cross-attention and Self-attention for Audio-visual Speaker Diarization in MISP-Meeting Challenge
1262	Multi-Channel Sequence-to-Sequence Neural Diarization: Experimental Results for The MISP 2025 Challenge

1717	Leveraging Self-Supervised Learning Based Speaker Diarization for MISP 2025 AVSD Challenge
2648	Overlap-Adaptive Hybrid Speaker Diarization and ASR-Aware Observation Addition for MISP 2025 Challenge

13:30-15:30 - Poster - Area 1 - Segmental and Tonal Units

1043	Perception of long and short vowel contrast in te reo Māori in clean and everyday listening environments
2179	The function of creaky voice in South Korean: A perception study
1857	Talker Normalization in Chinese Bilinguals: A Comparative Study
247	Coping with segmental–prosodic incongruity in spoken word recognition in Japanese
757	What the Filler? Both ASR Systems and Humans Struggle More With Other Kinds of Disfluencies Than With Filler Particles

13:30-15:30 - Poster - Area 1 - Databases and Progress in Methodology

986	Dhvani: A Weakly-supervised Phonemic Error Detection and Personalized Feedback System for Hindi
2400	Evaluating Wav2Vec2-Bert for Computer-Assisted Pronunciation Training for isiZulu
2787	Towards Adaptable and Intelligible Speech Synthesis in Noisy Environments
2595	Harnessing text-to-speech voice cloning models for improved audiological speech assessment
2394	75-Speaker Annot-16: A benchmark dataset for speech articulatory rt-MRI annotation with articulator contours and phonetic alignment
2044	Representing Speech Through Autoregressive Prediction of Cochlear Tokens
1226	Reasoning-Based Approach with Chain-of-Thought for Alzheimer's Detection Using Speech and Large Language Models
947	Finding the Human Voice in AI: Insights on the Perception of AI-Voice Clones from Naturalness and Similarity Ratings
590	Prosodically Enhanced Foreign Accent Simulation by Discrete Token-based Resynthesis Only with Native Speech Corpora

13:30-15:30 - Poster - Area 9 - Computational Resource Constrained ASR

18	Towards One-bit ASR: Extremely Low-bit Conformer Quantization Using Co-training and Stochastic Precision
1777	Unfolding A Few Structures for The Many: Memory-Efficient Compression of Conformer and Speech Foundation Models
503	Ultra-Low Bit Post-Training Quantization of Large Speech Models via K-Means Clustering and Mixed Precision Allocation
353	Effective and Efficient One-pass Compression of Speech Foundation Models Using Sparsity-aware Self-pinching Gates
1193	Context-Driven Dynamic Pruning for Large Multi-Modal Foundation Model
760	Analyzing the Importance of Blank for CTC-Based Knowledge Distillation
764	Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages

13:30-15:30 - Poster - Area 10 - Low Resource Speech Recognition

2615	SardinianVoxes: A Speech Recognition Dataset for the Sardinian Languages
1031	Prompting Whisper for improved verbatim transcription and end-to-end miscue detection
1358	Automated evaluation of children's speech fluency for low-resource languages
1254	Cantonese Punctuation Restoration using LLM Annotated Data
1961	Enhancing Speech Instruction Understanding and Disambiguation in Robotics via Speech Prosody
1874	Beyond Traditional Speech Modifications : Utilizing Self Supervised Features for Enhanced Zero-Shot Children ASR
340	Spoken Language Modeling with Duration-Penalized Self-Supervised Units

13:30-15:30 - Poster - Area 11 - Spoken Dialogue Systems 1

1893	PruneSLU: Efficient On-device Spoken Language Understanding through Vocabulary and Structural Pruning
2681	Leveraging LLMs for Written to Spoken Style Data Transformation to Enhance Spoken Dialog State Tracking
2764	Approaching Dialogue State Tracking via Aligning Speech Encoders and LLMs
595	What Do Humans Hear When Interacting? Experiments on Selective Listening for Evaluating ASR of Spoken Dialogue Systems
2013	SpeechDialogueFactory: A Framework for Natural Speech Dialogue Generation
1433	Who, When, and What: Leveraging the "Three Ws" Concept for Emotion Recognition in Conversation
2607	"Alexa, can you forget me?" Machine Unlearning Benchmark in Spoken Language Understanding
1965	Evaluating Large Language Models in Data Generation for Low-Resource Scenarios: A Case Study on Question Answering
2509	I want a horror -- comedy -- movie: Slips-of-the-Tongue Impact Conversational Recommender System Performance
2564	Towards a Japanese Full-duplex Spoken Dialogue System

13:30-15:30 - Poster - Area 13 - Dysarthric Speech Assessment 1

1152	Regularized Federated Learning for Privacy-Preserving Dysarthric and Elderly Speech Recognition
596	Facilitating Personalized TTS for Dysarthric Speakers Using Knowledge Anchoring and Curriculum Learning
1770	DiffDSR: Dysarthric Speech Reconstruction Using Latent Diffusion Model
2617	Improved Intelligibility of Dysarthric Speech via Conditional Flow Matching
1994	Bridging ASR and LLMs for Dysarthric Speech Recognition: Benchmarking Self-Supervised and Generative Approaches
512	Towards Inclusive ASR: Investigating Voice Conversion for Dysarthric Speech Recognition in Low-Resource Languages
567	Mitigating Overfitting During Speech Foundation Model Fine-tuning:

	Applications to Dysarthric Speech Detection
777	Towards Temporally Explainable Dysarthric Speech Clarity Assessment

13:30-15:30 - Poster - Area 13 - Speech and Voice Disorders 1

2425	Leveraging LLM for Stuttering Speech: A Unified Architecture Bridging Recognition and Event Detection
2496	Seamless Dysfluent Speech Text Alignment for Disordered Speech Analysis
2658	Analysis and Evaluation of Synthetic Data Generation in Speech Dysfluency Detection
2535	Fine-tuning Strategies for Automatic Speech Recognition of Low-Resource Speech with Autism Spectrum Disorder
1011	Identification of Pathological Pronunciation Profiles in ASR Transcription Errors
2333	A simple method for predicting Clinical Scores in Huntington's Disease by leveraging ASR's uncertainty on spontaneous speech
1162	Introducing EMOPARKNZ: the Emotional Speech Database from New Zealand English Speakers with Parkinson's Disease
1362	Revisiting WFST-based Hybrid Japanese Speech Recognition System for Individuals with Organic Speech Disorders

13:30-15:30 - Poster - Area 13 - Speech and Language Technology for Health Applications

358	A Chinese Heart Failure Status Speech Database with Universal and Personalised Classification
1570	Heart Rate as a Proxy Measure to Assess Human Confidence in Spoken Speech
1376	Foundation Model Hidden Representations for Heart Rate Estimation from Auscultation
2206	Towards Fusion of Neural Audio Codec-based Representations with Spectral for Heart Murmur Classification via Bandit-based Cross-Attention Mechanism
113	Perception of Emotional Speech by Individuals with High Borderline Personality Features
979	Visual features of the oral region in Polish sibilants produced by children with various sibilance patterns
2055	Meta-Learning Approaches for Speaker-Dependent Voice Fatigue Models
2596	Decoding Alzheimer's: Interpretable Visual and Logical Attention in Picture Description Tasks

13:30-15:30 - Special session - Area 14 - Special Session - Responsible Speech Foundation Models + SUPERB Challenge

1921	Defending speech-enabled LLMs against adversarial jailbreak threats
662	Mitigating Subgroup Disparities in Multi-Label Speech Emotion Recognition: A Pseudo-Labeling and Unsupervised Learning Approach
1402	Who Gets the Mic? Investigating Gender Bias in the Speaker Assignment of a Speech-LLM
1215	Evaluating Speech Foundation Models for Automatic Speech Recognition

	in the Low-Resource Kanyen'k´eha Language
1228	Benchmarking and Confidence Evaluation of LALMs For Temporal Reasoning
324	Teaching Audio-Aware Large Language Models What Does Not Hear: Mitigating Hallucinations through Synthesized Negative Samples
619	Speech-IFEval: Evaluating Instruction-Following and Quantifying Catastrophic Forgetting in Speech-Aware Language Models
1258	Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models
2377	Improving Multilingual Speech Models on ML-SUPERB 2.0: Fine-tuning with Data Augmentation and LID-Aware CTC
1013	The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties
1797	TalTech Systems for the Interspeech 2025 ML-SUPERB 2.0 Challenge

13:30-15:30 - Show and Tell - Speech Synthesis

2807	Code Mix TTS: An Approach to Infer Human Like Speech for Multi-Lingual Input Texts
2808	Turing's Echo: Investigating Linguistic Sensitivity of Deepfake Voice Detection via Gamification
2810	Unleashing the Inner Monster: Demonstrating High-Fidelity Human to Non-Human Voice Conversion
2811	Accent, Gender, and Audience Perception in Public Speaking: Cognitive Bias and Decision-Making in the Context of Climate Communication
2813	TUNGNAÁ IN LIVE PERFORMANCE: AN IMPLEMENTATION OF INTERACTIVE ARTISTIC TEXT-TO-VOICE
2814	Hear Me Out: Interactive evaluation and bias discovery platform for speech-to-speech conversational AI
2818	Vocal-tract model with two directions: Static design for a dummy head and dynamic design for a speaking machine

16:00-18:00 - Oral - Area 1 - Brain and Cognition

2776	Brain-tuned Speech Models Better Reflect Speech Processing Stages in the Brain
1725	Enhancing Syllabic Recognition via Speech-EEG Phase Analysis and Non-Activity State Modeling
1986	Functional Connectivity and Hilbert-Based Features for Covert Speech EEG Variability Analysis and Classification
1010	Neuro2Semantic: A Transfer Learning Framework for Semantic Reconstruction of Continuous Language from Human Intracranial EEG
2143	Selective Auditory Attention Decoding in Naturalistic Conversations Using EEG-Based Speech Envelope Tracking in Multi-Speaker Environments
1334	MiSTR: Multi-Modal iEEG-to-Speech Synthesis with Transformer-Based Prosody Prediction and Neural Phase Reconstruction

16:00-18:00 - Oral - Area 2 - Regional, Social and Diachronic Variation

1863	Probing prosodic differences between two regional varieties of Brazilian Portuguese

1479	Data-driven approaches to pitch modelling in two Mexican Spanish ethnolects: K-means Clustering & GAMMs
967	Tracking /r/ Deletion: Sociophonetic Perspectives on Post-Obstruent Final Rhotics in French through Forced Alignment with Pronunciation Variants.
924	Agent-based modelling, sound change, and metaphony in Southern Italian varieties of Italo-Romance.
2192	Modeling Vowel System Typology Using Iterated Confusion Minimization
38	Investigating Glottal Stop Coda Loss During Sound Change of Checked Syllables Based on Speech-EGG Voice Offset Alignment

16:00-18:00 - Oral - Area 4 - Dialect Identification in Different Languages

119	Audio-Based Classification and Geographic Regression of Austrian Dialects
421	Jointly Improving Dialect Identification and ASR in Indian Languages using Multimodal Feature Fusion
884	ADI-20: Arabic Dialect Identification dataset and models
1017	Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion
200	Effects of prosodic information on dialect classification using Whisper features
1809	Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification

16:00-18:00 - Oral - Area 5 - Speech Quality Assessment

1728	Non-intrusive Speech Quality Assessment with Diffusion Models Trained on Clean Speech
2683	SQ-AST: A Transformer-Based Model for Speech Quality Prediction
2315	AttentiveMOS: A Lightweight Attention-Only Model for Speech Quality Prediction
1131	Universal Preference-Score-based Pairwise Speech Quality Assessment
2532	FUSE-MOS: Fusion of Speech Embeddings for MOS Prediction with Uncertainty Quantification
1977	SHEET: A Multi-purpose Open-source Speech Human Evaluation Estimation Toolkit

16:00-18:00 - Oral - Area 6 - Speech Enhancement

2335	Investigating Continuous Autoregressive Generative Speech Enhancement
2200	Dynamic Layer Gating for Speech Enhancement
855	Model as Loss: A Self-Consistent Training Paradigm
2725	Test-Time Training for Speech Enhancement
673	Few-step Adversarial Schr{\"o}dinger Bridge for Generative Speech Enhancement
1988	Exploiting Bispectral Features for Single-Channel Speech Enhancement

16:00-18:00 - Oral - Area 10 - General Topics in ASR

731	Running Conventional Automatic Speech Recognition on Memristor Hardware: A Simulated Approach **(Blue Sky paper, 40 mins)**
869	Word Level Timestamp Generation for Automatic Speech Recognition and

	Translation
535	Directional Speech Recognition with Full-Duplex Capability
943	CMT-LLM: Context-Aware Multi-Talker ASR Utilizing Large Language Models
1652	Selective Invocation for Multilingual ASR: A Cost-effective Approach Adapting to Speech Recognition Difficulty

16:00-18:00 - Oral - Area 13 - Dysarthric Speech Assessment 2

1726	Voice Reconstruction through Large-Scale TTS Models: Comparing Zero-Shot and Fine-tuning Approaches to Personalise TTS in Assistive Communication
2711	Data Augmentation using Speech Synthesis for Speaker-Independent Dysarthria Severity Classification
1536	Fairness in Dysarthric Speech Synthesis: Understanding Intrinsic Bias in Dysarthric Speech Cloning using F5-TTS
1174	Synthetic Dysarthric Speech: A Supplement, Not a Substitute for Authentic Data in Dysarthric Speech Recognition
2768	Objective and Subjective Evaluation of Diffusion-Based Speech Enhancement for Dysarthric Speech
2069	Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech

16:00-18:00 - Poster - Area 5 - Acoustic Event Detection and Classification

771	Improving Audio Classification by Transitioning from Zero- to Few-Shot
2469	Zero-Shot Learning for Acoustic Event Classification Using an Attribute Vector and Conditional GAN
474	Leveraging Multi-Level Features of ATST with Conformer-Based Dual-Branch Network for Sound Event Detection
1053	Leveraging Unlabeled Audio for Audio-Text Contrastive Learning via Audio-Composed Text Features
1054	Language-Guided Contrastive Audio-Visual Masked Autoencoder with Automatically Generated Audio-Visual-Text Triplets from Videos
1308	AC/DC: LLM-based Audio Comprehension via Dialogue Continuation
1734	Anomalous Sound Detection Based Feature Fusion and Dual-path Non-linear Independent Components Estimation
1449	An Effective Anomalous Sound Detection Method Based on Global and Local Attribute Mining
1336	Acoustic scattering AI for non-invasive object classifications: A case study on hair assessment

16:00-18:00 - Poster - Area 5 - Spatial Audio and Acoustics 2

2784	SepVAC: Multitask Learning of Speaker Separation, Speaker Localization, Microphone Array Localization, and Room Acoustic Parameter Estimation in Various Acoustic Conditions
1718	TA-RIR: Topology-Aware Neural Modeling of Acoustic Propagation for Room Impulse Response Synthesis
2666	Spatially Weighted Contrastive Learning for Robust Sound Source Localization

2779	Efficient and Microphone-Fault-Tolerant 3D Sound Source Localization
1948	Joint Reference Microphone Selection and Filter Order Determination in Multi-channel Active Noise Control
296	Direct-path Relative Harmonic Coefficients Detection for Multi-source Direction-of-Arrival Estimation in Reverberant Environments
905	D-GAT: Dual Graph Attention Network for Global HRTF Interpolation
746	Deep learning based spatial aliasing reduction in beamforming for audio capture
728	SonarGuard2: Ultrasonic Face Liveness Detection Based on Adaptive Doppler Effect Feature Extraction

16:00-18:00 - Poster - Area 7 - Text Processing and Evaluation for Speech Synthesis 2

661	Grapheme-Coherent Phonemic and Prosodic Annotation of Speech by Implicit and Explicit Grapheme Conditioning
1034	Non-Standard Accent TTS Support via Large Multi-Accent Frontend Pronunciation Knowledge Transfer
1428	Speech-guided Grapheme-to-Phoneme conversion for Cantonese Text-to-Speech
76	Transcript-Prompted Whisper with Dictionary-Enhanced Decoding for Japanese Speech Annotation
401	Enabling the replicability of speech synthesis perceptual evaluations
496	When The MOS Predictor Asks For Training Annotation In Cross Lingual/Domain Adaptation
572	Assessment of the synthetic quality and controllability of laughing onset in speech-laugh synthesis

16:00-18:00 - Poster - Area 7 - Speech Synthesis Paradigms and Methods 1

554	RapFlow-TTS: Rapid and High-Fidelity Text-to-Speech with Improved Consistency Flow Matching
2449	Accelerating Flow-Matching-Based Text-to-Speech via Empirically Pruned Step Sampling
704	Differentiable Reward Optimization for LLM based TTS system
253	Long-Context Speech Synthesis with Context-Aware Memory
551	Monotonic Attention for Robust Text-to-Speech Synthesis in Large Language Model Frameworks
319	Improving Noise Robustness of LLM-based Zero-shot TTS via Discrete Acoustic Token Denoising
854	Bridging the Training–Inference Gap in TTS: Training Strategies for Robust Generative Postprocessing for Low-Resource Speakers

16:00-18:00 - Poster - Area 10 - Language Learning and Assessment

2602	Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian
306	Can ASR generate valid measures of child reading fluency?
2500	SGED-Probe: Probing E2E ASR decoder and aligner for spoken grammar error detection under three speaking practice conditions
1012	Evaluating Logit-Based GOP Scores for Mispronunciation Detection

1497	Towards a Unified Benchmark for Arabic Pronunciation Assessment: Qur'anic Recitation as Case Study
983	OMPAL: Bridging Speech and Learning with an Open-Source Mandarin Pronunciation Assessment Corpus for Global Learners
615	A Perception-Based L2 Speech Intelligibility Indicator: Leveraging A Rater's Shadowing and Sequence-to-Sequence Voice Conversion
248	Multimodal and Multitask Learning for Predicting Multiple Scores in L2 English Speech
2409	Enhancing Generalization of Speech Large Language Models with Multi-Task Behavior Imitation and Speech-Text Interleaving
1375	Mispronunciation Detection Without L2 Pronunciation Dataset in Low-Resource Setting: A Case Study in Finland Swedish

16:00-18:00 - Poster - Area 10 - Multimodal Systems

114	CAMER: Contribution-Aware Multimodal Emotion Recognition
2696	GIA-MIC: Multimodal Emotion Recognition with Gated Interactive Attention and Modality-Invariant Learning Constraints
2251	SNIFR : Boosting Fine-Grained Child Harmful Content Detection Through Audio-Visual Alignment with Cascaded Cross-Transformer
812	CNVSRC 2024: The Second Chinese Continuous Visual Speech Recognition Challenge
268	PAEFF: Precise Alignment and Enhanced Gated Feature Fusion for Face-Voice Association
874	Efficient and Direct Duplex Modeling for Speech-to-Speech Language Model
1524	U-SAM: An audio language Model for Unified Speech, Audio, and Music Understanding
1997	Enhanced Hybrid Transducer and Attention Encoder Decoder with Text Data
1913	The role of audio-visual integration in the time course of phonetic encoding in self-supervised speech models

16:00-18:00 - Poster - Area 12 - Keyword Spotting and Retrieval

2722	Language-Agnostic Speech Tokenizer for Spoken Term Detection with Efficient Retrieval
2631	H-QuEST: Accelerating Query-by-Example Spoken Term Detection with Hierarchical Indexing
159	Vela: Scalable Embeddings with Voice Large Language Models for Multimodal Retrieval
378	Adversarial Deep Metric Learning for Cross-Modal Audio-Text Alignment in Open-Vocabulary Keyword Spotting
908	GTA: Towards Generative Text-To-Audio Retrieval via Multi-Scale Tokenizer
1493	Enhancing Retrieval-Augmented Audio Captioning with Generation-Assisted Multimodal Querying and Progressive Learning
2085	On Retrieval of Long Audios with Complex Text Queries
1607	SIDC-KWS: Efficient Spiking Inception-Dilated Conformer with Self-Attention for Keyword Spotting

285	Multichannel Keyword Spotting for Noisy Conditions
1005	LLM-Synth4KWS: Scalable Automatic Generation and Synthesis of Confusable Data for Custom Keyword Spotting
1038	GraphemeAug: A Systematic Approach to Synthesized Hard Negative Keyword Spotting Examples
2011	SpokenNativQA: A Multilingual and Culturally-Aligned Spoken Queries for LLMs

16:00-18:00 - Special session - Area 14 - Special Session - Connecting Speech Science and Speech Technology for Children's Speech

504	Band-Split Self-supervised Mamba for Infant-centered Audio Analysis
339	Subtyping Speech Errors in Childhood Speech Sound Disorders with Acoustic-to-Articulatory Speech Inversion
2407	PERCEPT-US: A Multimodal American English Child Speech Corpus Specialized for Articulatory Feedback
2117	Children's Voice Privacy: First Steps and Emerging Challenges
1002	FT-Boosted SV: Towards Noise Robust Speaker Verification for English Speaking Classroom Environments
2513	Examining Test-Time Adaptation for Personalized Child Speech Recognition
1946	Employing self-supervised learning models for cross-linguistic child speech maturity classification
273	On Enhancing the Performance of Children's ASR Task in Limited Data Scenario
1216	Egocentric Speaker Classification in Child-Adult Dyadic Interactions: From Sensing to Computational Modeling
1088	Large Language Models based ASR Error Correction for Child Conversations
1962	Challenges in Automated Processing of Speech from Child Wearables: The Case of Voice Type Classifier
658	Improving Child Speech Recognition and Reading Mistake Detection by Using Prompts
936	Improving Automatic Speech Recognition for Children's Reading Assessment with Disfluency-aware Language Models
1245	Oral Reading Errors by Grade 3 Children in Indian Schools: A Hindi-English Perspective
2174	Grammatical Error Detection on Spontaneous Children's Speech Using Iterative Pseudo Labeling
2645	Why is children's ASR so difficult? Analyzing children's phonological error patterns using SSL-based phoneme recognizers
1996	Automatic detection of speech sound disorders in German speaking children: augmenting the data with typically developed speech
2393	Continuous Learning for Children's ASR: Overcoming Catastrophic Forgetting with Elastic Weight Consolidation and Synaptic Intelligence
2733	Exploring Shared-Weight Mechanisms in Transformer and Conformer Architectures for Automatic Speech Recognition
1890	Advancing Pediatric ASR: The Role of Voice Generation in Disordered Speech

1019	CHSER: A Dataset and Case Study on Generative Speech Error Correction for Child ASR
1967	Causal Structure Discovery for Error Diagnostics of Children's ASR

Wednesday 20/08/2025

08:30-10:30 - Oral - Area 2 - Diversity: Age, Sex, Gender, Ethnicity, and More

1136	Age-related changes in multisensory integration of emotions in an audiovisual face-prosody-semantics Stroop task
139	Investigating effects of sex hormones, cycle phases and age on female fundamental frequency
2271	Pre-aspiration in Icelandic Is Conditioned by Gender/Sex
1980	Transcribing Diverse Voices: Using Whisper for ICE corpora
94	Is it all about race?: A cross-examination of /s/ in a multilingual (Nigerian) context
1488	Investigating Gender Bias in Text-to-Audio Generation Models

08:30-10:30 - Oral - Area 3 - Multimodal Emotion Recognition

Survey Talk	Paralinguistic modeling in real-world environments: approaches and applications **(Survey Talk, 40 mins)**
2060	Multi-Modal Multi-Task Affective States Recognition Based on Label Encoder Fusion
548	RA-CLAP: Relation-Augmented Emotional Speaking Style Contrastive Language-Audio Pretraining For Speech Retrieval
1198	EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion via Rank-N-Contrast
1514	Modality-Agnostic Multimodal Emotion Recognition using a Contrastive Masked Autoencoder

08:30-10:30 - Oral - Area 4 - Privacy and Anonymization

2412	Speech Unlearning **(Blue Sky paper, 40 mins)**
287	Unlearning LLM-Based Speech Recognition Models
194	EASY: Emotion-aware Speaker Anonymization via Factorized Distillation
820	Private kNN-VC: Interpretable Anonymization of Converted Speech
1699	Legally validated evaluation framework for voice anonymization

08:30-10:30 - Oral - Area 5 - Anomalous Sound Detection

1563	Dual Orthogonality Sub-center Loss for Enhanced Anomalous Sound Detection
1584	Adaptive Across-Subcenter Representation Learning for Imbalanced Anomalous Sound Detection
467	Towards Few-Shot Training-Free Anomaly Sound Detection
2503	Finetune Large Pre-Trained Model Based on Frequency-Wise Multi-Query Attention Pooling for Anomalous Sound Detection
1142	Acoustic Detection of UAV Abnormality Using One Ground-Based Acoustic Vector Sensor
651	StarGAN-Aug: A Cross-domain Fault Audio Generation Method for High-performance Fault Diagnosis of Power Transformers

08:30-10:30 - Oral - Area 6 - Speaker Extraction 2

970	FlowTSE: Target Speaker Extraction with Flow Matching

2077	MTSE: Multi-Target Speaker Extraction for Conversation Scenarios
1787	Location-Aware Target Speaker Extraction for Hearing Aids
680	ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment
1448	Online AV-CrossNet: A Causal and Efficient Audiovisual System for Speech Enhancement and Target Speaker Extraction
1515	Steering Deep Non-Linear Spatially Selective Filters for Weakly Guided Extraction of Moving Speakers in Dynamic Scenarios

08:30-10:30 - Oral - Area 7 - Speech Synthesis Paradigms and Methods 2

439	Efficient Neural and Numerical Methods for High-Quality Online Speech Spectrogram Inversion via Gradient Theorem
63	Fine-Tuning Text-to-Speech Diffusion Models Using Reinforcement Learning with Human Feedback
1236	Accelerating Diffusion-based Text-to-Speech Model Training with Dual Modality Alignment
1776	SpeechSEC:A Unified Multi-Task Framework for Non-Autoregressive Speech Synthesis, Editing, and Continuation
431	VoiceNoNG: Robust High-Quality Speech Editing Model without Hallucinations
2328	A Watermark for Auto-Regressive Speech Generation Models

08:30-10:30 - Oral - Area 8 - Neural Network Training Methods 1

1399	Cross-lingual Data Selection Using Clip-level Acoustic Similarity for Enhancing Low-resource Automatic Speech Recognition
2692	Spot and Merge: A Hybrid Context Biasing Approach for Rare Word and Out of Vocabulary Recognition
2059	Accurate, fast, cheap: Choose three. Replacing Multi-Head-Attention with Bidirectional Recurrent Attention in ASR
2582	Improving Cross-Attention based on Positional Alignment during Inference for Robust Long-form Speech Recognition
330	Improving End-to-end Mixed-case ASR with Knowledge Distillation and Integration of Voice Activity Cues
1549	Cross-modal Knowledge Transfer Learning as Graph Matching Based on Optimal Transport for ASR

08:30-10:30 - Poster - Area 5 - Music and Audio Analysis

311	Enhancing Lyrics Transcription on Music Mixtures with Consistency Loss
1015	Tonality-Based Accompaniment-Guided Automatic Singing Evaluation
2227	Investigating the Reasonable Effectiveness of Speaker Pre-Trained Models and their Synergistic Power for SingMOS Prediction
930	Focal Modulation Network: A Novel Solution for Polyphonic Music Instrument Recognition without Attention and Aggregation Strategy
1070	A Joint Network for Singing Melody Extraction from Polyphonic Music with Attention Aggregation and Self-Consistency Training
130	Position Also Matters! Separating Same Instruments in String Quartet using Timbral and Positional Cues
847	WhisperMSS: A Two-Stage Framework for Mandarin Singing Transcription

	and Segmentation Using Pretrained Models
1646	Low Complex IIR Adaptive Hear-Through Ambient Filtering for Overcoming Practical Constraints in Earbuds
2533	Sub-band based Adaptive IIR Algorithm with Biquad Filter Stability Constraints for Feedforward Hear-Through Equalization

08:30-10:30 - Poster - Area 5 - Audio Analysis, Generation and Assessment

199	Discrete Audio Representations for Automated Audio Captioning
1573	CLAP-ART: Automated Audio Captioning with Semantic-rich Audio Representation Tokenizer
808	Temp4Cap: Temporally-aligned Automated Audio Captioning
1313	Optimizing CLAP Reward with LLM Feedback for Semantically Aligned and Diverse Automated Audio Captioning
366	Bridging Audio and Vision: Zero-Shot Audiovisual Segmentation by Connecting Pretrained Models
1850	DiffStereo: End-to-End Mono-to-Stereo Audio Generation with Diffusion Transformer
1830	RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio
2138	Crowdsourcing MUSHRA Tests in the Age of Generative Speech Technologies: A Comparative Analysis of Subjective and Objective Testing Methods
2198	SMARTMOS: Modeling Subjective Audio Quality Evaluation for Real-Time Applications

08:30-10:30 - Poster - Area 8 - Other Topics in Speech Recognition

2470	Effect of Loud Speaker Emitted Speech on ASR performance
1200	Self-improvement for Large Speech Models with Unlabeled Data
1785	Contextualized Automatic Speech Recognition with Dynamic Vocabulary Prediction and Activation
1799	Character Error Rate Estimation for Semi-Supervised Training of Speech Recognition for Arabic Dialects
1706	Unified Semi-Supervised Pipeline for Automatic Speech Recognition
2750	Scaling Laws for Synthetic Speech for Model Training
1109	R2S: Real-to-Synthetic Representation Learning for Training Speech Recognition Models on Synthetic Data
1824	Context is all you need? Low-resource conversational ASR profits from context, coming from the same or from the other speaker
1973	Automatic Speech Recognition Biases in Newcastle English: an Error Analysis

08:30-10:30 - Poster - Area 8 - Far-field and Robust Speech Recognition

1513	SuPseudo: A Pseudo-supervised Learning Method for Neural Speech Enhancement in Far-field Speech Recognition
1668	Lightweight Front-end Enhancement for Robust ASR via Frame Resampling and Sub-Band Pruning
201	Calm-Whisper: Reduce Whisper Hallucination On Non-Speech By Calming Crazy Heads Down

397	HuBERT-VIC: Improving Noise-Robust Automatic Speech Recognition of speech foundation model via Variance-Invariance-Covariance Regularization
1614	MOVER: Combining Multiple Meeting Recognition Systems
1360	EmbedAug: An Augmentation Scheme for End-to-End Automatic Speech Recognition
1892	Attention Models and Auditory Transduction Features for Noise Robustness
1415	Lightweight and Robust Multi-Channel End-to-End Speech Recognition with Spherical Harmonic Transform
238	On the Design of a Robust Superdirective Beamformer and Topology Parameter Optimization with Frustum-Shaped Microphone Arrays Featuring Multiple Rings

08:30-10:30 - Poster - Area 11 - Conversation, Communication and Interaction 2

2660	Triadic Multi-party Voice Activity Projection for Turn-taking in Spoken Dialogue Systems
770	Continuous prediction of backchannel timing for human-robot interaction
2038	Impact of Background Noise on Turn-Taking Dynamics in Triadic Conversations
959	Multimodal dynamics of hand gestures and pauses in multiparty interactions
176	TinyClick: Single-Turn Agent for Empowering GUI Automation
408	Improving User Impression of Spoken Dialogue Systems by Controlling Para-linguistic Expression Based on Intimacy
600	Dialogue Response Prefetching Based on Semantic Similarity and Prediction Confidence of Language Model

08:30-10:30 - Poster - Area 11 - Language Modeling for Conversational Systems

409	Analyzing Mitigation Strategies for Catastrophic Forgetting in End-to-End Training of Spoken Language Models
1292	Speechless: Speech Instruction Training Without Speech for Low Resource Languages
2218	LiSTEN: Learning Soft Token Embeddings for Neural Audio LLMs
2109	CRYFISH: On deep audio analysis with Large Language Models
962	Improving Linguistic Diversity of Large Language Models with Possibility Exploration Fine-Tuning
1184	OpusLM: A Family of Open Unified Speech Language Models
834	CAPR: Confidence-Aware Prompt Refinement in Large Language Models

08:30-10:30 - Poster - Area 13 - Multimodal Speech and Language Processing in Healthcare Settings

150	Can Multimodal Foundation Models Help Analyze Child-Inclusive Autism Diagnostic Videos?
726	A Cascaded Multimodal Framework for Automatic Social Communication Severity Assessment in Children with Autism Spectrum Disorder
2322	Accessible Real-time Eye-gaze Tracking for Neurocognitive Health Assessment: A Multimodal Web-based Approach

2332	Multimodal Biomarkers for Schizophrenia: Towards Individual Symptom Severity Estimation
537	Fact-Controlled Diagnosis of Hallucinations in Medical Text Summarization
2430	Evaluating Automatic Speech Recognition Pipelines for Mandarin-English Bilingual Child Language Assessment in Telehealth

08:30-10:30 - Special session - Area 14 - Special Session - Speech Accessibility Project Challenge

566	The Interspeech 2025 Speech Accessibility Project Challenge
2431	Towards Inclusive and Fair ASR: Insights from the SAPC Challenge for Optimizing Disordered Speech Recognition
182	Robust fine-tuning of speech recognition models via weighted average models: application to dysarthric speech
1553	Exploring Generative Error Correction for Dysarthric Speech Recognition
2724	Pathology-Aware Speech Encoding and Data Augmentation for Dysarthric Speech Recognition
2155	Personalized Fine-Tuning with Controllable Synthetic Speech from LLM-Generated Transcripts for Dysarthric Speech Recognition
934	A Self-Training Approach for Whisper to Enhance Long Dysarthric Speech Recognition
1484	Fine-tuning Parakeet-TDT for Dysarthric Speech Recognition in the Speech Accessibility Project Challenge
1705	CBA-Whisper: Curriculum Learning-Based AdaLoRA Fine-Tuning on Whisper for Low-Resource Dysarthric Speech Recognition

11:00-12:00 - Keynote speaker - Carol Espy-Wilson

Speech Kinematic Analysis from Acoustics: Scientific, Clinical and Practical Applications

13:30-15:30 - Oral - Area 1 - Lexicon and Grammar

2369	Processing of grammatical information in cochlear implant simulated speech by German adult listeners
116	A Gradient Effect of Hand Beat Timing on Spoken Word Recognition
1676	The Effect of Word Predictability on Spoken Cross-Language Intelligibility
1191	The Role of Lexicosyntactic and Prosodic Information in Predicting Speaker Turns during Conversation
1675	Sentence-Final Particles in Mandarin Child-Directed Speech: Frequency and Impact on Speech Rate
555	Bilingual Speakers Exhibit Cognitive Fatigue: A Speech Disfluencies Case Study on Research Talks

13:30-15:30 - Oral - Area 2 - Articulatory Analyses

1381	Influence of wall coverings of 3D-printed vocal tract models on measured transfer functions
1733	Supralaryngeal Kinematics of Implosives in Central Vietnamese: An EMA Study
1065	Lateral channel formation in Australian English /l/: insights from

	Magnetic Resonance Imaging
1810	Articulatory variations in Apical Vowels in Southwestern Mandarin
1619	Rhotic Articulation in Australian English: Insights from MRI
666	Articulatory strategy in vowel production as a basis for speaker discrimination

13:30-15:30 - Oral - Area 4 - Disentanglement of Information for Speaker Recognition

2512	LASPA: Language Agnostic Speaker Disentanglement with Prefix-Tuned Cross-Attention
1456	SCD-Conformer: Semantic Content Disentanglement for Text-Independent Speaker Verification
437	Universal Semantic Disentangled Privacy-preserving Speech Representation Learning
1963	Using gender, phonation and age to interpret automatically discovered speech attributes for explainable speaker recognition
2629	Do you read me? - flow of speech effect on speaker recognition systems
57	VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin

13:30-15:30 - Oral - Area 5 - Speech and Audio Analysis and Representation

Survey Talk	Assistive Oral Communication Technology **(Survey Talk, 40 mins)**
669	PAST: Phonetic-Acoustic Speech Tokenizer
2612	Factorized RVQ-GAN For Disentangled Speech Tokenization
506	EnCodecMAE: leveraging neural codecs for universal audio representation learning
426	AxLSTMs: learning self-supervised audio representations with xLSTMs

13:30-15:30 - Oral - Area 7 - Multimodal and Visual Speech Synthesis

1933	MM-MovieDubber: Towards Multi-Modal Learning for Multi-Modal Movie Dubbing
1692	Learning Phonetic Context-Dependent Viseme for Enhancing Speech-Driven 3D Facial Animation
2163	Face2VoiceSync: Lightweight Face-Voice Consistency for Text-Driven Talking Face Generation
47	Revival with Voice: Multi-modal Controllable Text-to-Speech Synthesis
1494	VisualSpeech: Enhancing Prosody Modeling in TTS Using Video
1478	LightL2S: Ultra-Low Complexity Lip-to-Speech Synthesis in the Wild

13:30-15:30 - Oral - Area 9 - Error Correction and Confidence Estimation

649	LLM-based Generative Error Correction for Rare Words with Synthetic Data and Phonetic Context
2016	ASR Confidence Estimation using True Class Lexical Similarity Score
191	Semi-Supervised Learning for Automatic Speech Recognition with Word Error Rate Estimation and Targeted Domain Data Selection
328	Voice Activity-based Text Segmentation for ASR Text Denormalization

1885	Phonetically-Augmented Discriminative Rescoring for Voice Search Error Correction
1432	From Weak Labels to Strong Results: Utilizing 5,000 Hours of Noisy Classroom Transcripts with Minimal Accurate Data

13:30-15:30 - Oral - Area 13 - Pathological Speech Analysis 2

281	Acoustic similarities, articulatory uniqueness: Speech production mechanisms in individuals with congenital lip paralysis
1127	Relationship between objective and subjective perceptual measures of speech in individuals with head and neck cancer
1805	Evaluating the Usefulness of Non-Diagnostic Speech Data for Developing Parkinson's Disease Classifiers
1931	Multimodal Assessment of Speech Impairment in Amyotrophic Lateral Sclerosis Using Audio-Visual and Machine Learning Approaches
2162	Development and Validation of a Wav2Vec 2.0-Based Cross-Language Methodology for Measurement of Articulatory Precision
2700	J-j-j-just Stutter: Benchmarking Whisper's Performance Disparities on Different Stuttering Patterns

13:30-15:30 - Oral - Area 13 - Speech and Voice Disorders 2

1406	Hybrid Expert Knowledge and Self-Supervised Learning for Diagnostic Modeling of Adductor Spasmodic and Primary Myotonic Dysphonia
1868	MVP: Multi-source Voice Pathology detection
2110	Phonetic Posteriorgram-Based Phoneme Selection for Vocal Cord Disorder Classification in Continuous Mandarin Speech
708	Articulatory clarity and variability before and after surgery for tongue cancer
2089	Hybrid HMM-SVM classifier using frication-based features for detection of non-normative sibilant articulation patterns in Polish children’s speech
2373	Fine-Tuning ASR for Stuttered Speech: Personalized vs. Generalized Approaches

13:30-15:30 - Poster - Area 4 - Training and Scoring Methods for Speaker Recognition

2675	Boundary-Conscious Pruning: Hard Set-Aware Model Compression for Efficient Speaker Recognition
715	Pushing the Frontiers of Self-Distillation Prototypes Network with Dimension Regularization and Score Normalization
1512	A Domain Robust Pre-Training Method with Local Prototypes for Speaker Verification
442	Clustering-based hard negative sampling for supervised contrastive speaker verification
82	MASV: Speaker Verification with Global and Local Context Mamba
529	ATMM-SAGA: Alternating Training for Multi-Module with Score-Aware Gated Attention SASV system
628	Rethinking Leveraging Pre-Trained Multi-Layer Representations for Speaker Verification
794	SEED: Speaker Embedding Enhancement Diffusion Model

147	A copula-based generative score-level fusion model for speaker verification

13:30-15:30 - Poster - Area 4 - Evaluation and Forensic Applications of Speaker Recognition

146	Analysis of the ABC classification backends for NIST SRE24
2170	STCON NIST SRE24 System: Composite Speaker Recognition Solution for Challenging Scenarios
591	Vo-Ve: An Explainable Voice-Vector for Speaker Identity Evaluation
494	Variability in performance across four generations of automatic speaker recognition systems
526	On the influence of language similarity in non-target speaker verification trials
782	The Sub-3Sec Problem: From Text-Independent to Text-Dependent Corpus

13:30-15:30 - Poster - Area 6 - Noise Reduction and Dereverberation

483	Boosting StoRM Convergence with Metric Guidance and Non-uniform State-Sampling for Optimal Dereverberation
1942	Unified Variational and Physics-aware Model for Room Impulse Response Estimation
274	MelRe: Vision-Based Mel-Spectrogram Restoration
1581	SpeechRefiner: Towards Perceptual Quality Refinement for Front-End Algorithms
2581	Modality-Specific Speech Enhancement and Noise-Adaptive Fusion for Acoustic and Body-Conduction Microphone Framework
893	Joint Rate Allocation and Sensor Selection for Speech Enhancement in Wireless Acoustic Sensor Networks
1009	Individualized speech enhancement for hearing-impaired listeners
1424	First Analyze Then Enhance: A Task-Aware System for Speech Separation, Denoising, and Dereverberation
65	A Robust Hybrid ACC-PM Approach for Personal Sound Zones

13:30-15:30 - Poster - Area 6 - Bandwidth Expansion and Diffusion-based Speech Enhancement

374	A Semantic Information-based Hierarchical Speech Enhancement Method Using Factorized Codec and Diffusion Model
998	Voice-ENHANCE: Speech Restoration using a Diffusion-based Voice Conversion Framework
1317	Diffusion Buffer: Online Diffusion-based Speech Enhancement with Sub-Second Latency
1721	SNR-Aligned Consistent Diffusion for Adaptive Speech Enhancement
350	MDDM: A Multi-view Discriminative Enhanced Diffusion-based Model for Speech Enhancement
2438	A Neural Codec Approach for Noise-Robust Bandwidth Expansion
692	HWB-Net: A Novel High-Performance and Efficient Hybrid Waveform Bandwidth Extension Method
806	Frequency-Domain Enhanced Extreme Bandwidth Extension Network

	with ICCRN for Superior Speech Quality

13:30-15:30 - Poster - Area 8 - Neural Network Training Methods 2

2746	SiamCTC: Learning Speech Representations through Monotonic Temporal Alignment
1309	Improving Generalization of End-to-End ASR through Diversity and Independence Regularization
2025	Exploring Linear Variant Transformers and k-NN Memory Inference for Long-Form ASR
1079	Attention-Free Dual-Mode ASR with Latency-Controlled Selective State Spaces
128	Thinking Fast and Slow: Robust Speech Recognition via Deep Filter-Tuning
179	Towards Efficiently Whisper Fine-tuning with Monotonic Alignments
1126	Dynamic Acoustic Model Architecture Optimization in Training for ASR
1425	Knowledge Distillation Method for Pruned RNN-T Models via Pruning Bounds Sharing and Losses Confusion
1704	An Effective Training Framework for Light-Weight Automatic Speech Recognition Models

13:30-15:30 - Poster - Area 8 - Neural Network Training Methods and Architectures

747	Distilling a speech and music encoder with task arithmetic
695	MSDA: Combining Pseudo-labeling and Self-Supervision for Unsupervised Domain Adaptation in ASR
1343	REB-former: RWKV-enhanced E-branchformer for Speech Recognition
501	PredTrAD – Prediction-based Transformer for Anomaly Detection in Multivariate Time Series Data
2590	FairASR: Fair Audio Contrastive Learning for Automatic Speech Recognition
1511	Automatic Speech Recognition of African American English: Lexical and Contextual Effects
173	Improving Synthetic Data Training of Contextual Biasing Models with a Keyword-Aware Cost Function
449	SOMSRED-SVC: Sequential Output Modeling with Speaker Vector Constraints for Joint Multi-Talker Overlapped ASR and Speaker Diarization
2388	Thinking in Directivity: Speech Large Language Model for Multi-Talker Directional Speech Recognition

13:30-15:30 - Poster - Area 12 - Language Resources

679	ToxicTone: A Mandarin Audio Dataset Annotated for Toxicity and Toxic Utterance Tonality
1958	ViToSA: Audio-Based Toxic Spans Detection on Vietnamese Speech Utterances
1852	Self-Supervised Models of Speech Processing for Haitian Creole
1437	AfriHuBERT: A self-supervised speech representation model for African languages
1045	The Faetar Speech Recognition Benchmark
2630	LHCP-ASR: An English Speech Corpus of High-Energy Particle Physics

	Talks for Narrow-Domain ASR Benchmarking
452	Towards High-Quality LLM-Based Data for French Spontaneous Speech Simplification: an Exo-Refinement Approach
326	BR-ASR: Efficient and Scalable Bias Retrieval Framework for Contextual Biasing ASR in Speech LLM
1896	SPGISpeech 2.0: Transcribed multi-speaker financial audio for speaker-tagged transcription
720	Loquacious Set: 25,000 Hours of Transcribed and Diverse English Speech Recognition Data for Research and Commercial Use
1934	CEREALES : a new dataset of Quebec French accented speech with applications to speech recognition

13:30-15:30 - Special session - Area 14 - Special Session - Challenges in Speech Data Collection, Curation and Annotation - Part 1

1329	A Study of Audio-Visual Corpus Design and Production: A Perspective from MISP Challenges
1713	VCapAV: A Video-Caption Based Audio-Visual Deepfake Detection Dataset
352	J-SPAW: Japanese speaker verification and spoofing attacks recorded in-the-wild dataset
2490	CommissionsQC: a Québec French speech corpus for automatic speech recognition
540	Granary: Speech Recognition and Translation Dataset in 25 European Languages
2418	Collecting, Curating, and Annotating Good Quality Speech deepfake dataset for Famous Figures: Process and Challenges
2027	Quantifying and Reducing Speaker Heterogeneity within the Common Voice Corpus for Phonetic Analysis
2454	The Speech Accessibility Project: Best Practices for Collection and Curation of Disordered Speech
2774	Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective
1987	Fifteen Years of Child-Centered Long-Form Recordings: Promises, Resources, and Remaining Challenges to Validity
706	Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation
687	Investigating Affect Mining Techniques for Annotation Sample Selection in the Creation of Finnish Affective Speech Corpus
2362	Scalable Spontaneous Speech Dataset (SSSD): Crowdsourcing Data Collection to Promote Dialogue Research
1632	A Multimodal Chinese Dataset for Cross-lingual Sarcasm Detection
2074	Leveraging Large Language Models for Sarcastic Speech Annotation in Sarcasm Detection

13:30-15:30 - Show and Tell - Signal Processing / Multimodel processing

2806	Real-time TSE demonstration via SoundBeam with KD
2809	Real-Time Diffusion Buffer for Speech Enhancement On A Laptop
2817	Co-Speech Motion for Virtual Agents in Dialogue Using LLM-Driven Primitive Action Selection

2826	TargetVoice: Single Channel Low-Latency Target Speaker Extraction
2828	Rollback Speech: Smart Feedback Prompts for Lost Utterances in Unstable Online Calls
2832	Simultaneous Speech Translation Integrated Compact Multiple Sound Spot Synthesis System On A Laptop Carried Out With A Backpack
2834	GenECA: A General-Purpose Framework for Real-Time Adaptive Multimodal Embodied Conversational Agents

16:00-18:00 - Oral - Area 1 - L1 and L2 Acquisition, Perception and Processing

914	Evaluating Progress of CALL System Users on Accentedness and Comprehensibility: An Acoustic and ASR-Based Approach
1748	Does English fish sound like French fiche? Perceptual similarity judgments versus acoustic similarity
1159	Acoustic Features of Mandarin Tone Production in Noise: A Comparison Between Chinese Native Speakers and Korean L2 Learners
1876	The Role of Contextual Variation in Learning Cantonese Tones from Naturalistic Speech
1237	Pitch Target Realization in Putonghua Tone Production of Children from Dialect-Speaking Regions
1847	The Development of Speech Rhythm in Putonghua-Learning Preschool Children in South Xinjiang Uyghur Autonomous Region of China

16:00-18:00 - Oral - Area 4 - Spoofing and Adversarial Attacks

1555	Generalizable Audio Spoofing Detection using Non-Semantic Representations
1560	Adversarial Attacks on Text-dependent Speaker Verification System
1999	Beyond Attacks: Advancing Fake Speech Detection with Attack-Agnostic Methods
920	ASVspoof2019 vs. ASVspoof5: Assessment and Comparison
2618	Evaluating Parameter Sharing for Spoofing-Aware Speaker Verification: A Case Study on the ASVspoof5 Dataset
2701	Can Quantized Audio Language Models Perform Zero-Shot Spoofing Detection?

16:00-18:00 - Oral - Area 5 - Generative Models for Audio

835	Discovering Directions of Uncertainty in Speech Inpainting
209	InfiniteAudio: Infinite-Length Audio Generation with Consistency
300	FoleyMaster: High-Quality Video-to-Audio Synthesis via MLLM-Augmented Prompt Tuning and Joint Semantic-Temporal Adaptation
1119	Video-to-Audio Generation with Fine-grained Temporal Semantics
1516	TTMBA: Towards Text To Multiple Sources Binaural Audio Generation
1137	EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer

16:00-18:00 - Oral - Area 7 - Voice Conversion 2

Survey Talk	Recent Progress on Voice Conversion **(Survey Talk, 40 mins)**
203	ClapFM-EVC: High-Fidelity and Flexible Emotional Voice Conversion with

	Dual Control from Natural Language and Speech
948	PromptEVC: Controllable Emotional Voice Conversion with Natural Language Prompts
1538	StarVC: A Unified Auto-Regressive Framework for Joint Text and Speech Generation in Voice Conversion
1747	FasterVoiceGrad: Faster One-step Diffusion-Based Voice Conversion with Adversarial Diffusion Conversion Distillation

16:00-18:00 - Oral - Area 8 - Streaming ASR

1541	MFLA: Monotonic Finite Look-ahead Attention for Streaming Speech Recognition
1691	Delayed-KD: Delayed Knowledge Distillation based CTC for Low-Latency Streaming ASR
685	Parameter-efficient Fine-tuning of Conformer-based Streaming Speech Recognition into Non-streaming Models
975	On-device Streaming Discrete Speech Units
511	Adapting Whisper for Streaming Speech Recognition via Two-Pass Decoding
1480	Dynamic Context-Aware Streaming Pretrained Language Model For Inverse Text Normalization

16:00-18:00 - Oral - Area 11 - Spoken Language Understanding

532	QUADS: QUAntized Distillation Framework for Efficient Speech Language Understanding
1467	Spoken Language Understanding on Unseen Tasks With In-Context Learning
175	Leveraging Information Retrieval to Enhance Spoken Language Understanding Prompts in Few-Shot Learning
718	Modeling Multi-Turn Spoken Language Understanding with Dynamic Graph Convolutional Networks
1220	DRI-GAN: A Novel Dual Real Input GAN with Triplet Loss for Cross-Lingual and Noisy SLU
1612	“KAN you hear me?” Exploring Kolmogorov-Arnold Networks for Spoken Language Understanding

16:00-18:00 - Oral - Area 13 - Pathological Speech Analysis 3

801	Evaluating the Effectiveness of Pre-Trained Audio Embeddings for Classification of Parkinson's Disease Speech Data
721	On-the-fly Routing for Zero-shot MoE Speaker Adaptation of Speech Foundation Models for Dysarthric Speech Recognition
404	Lightweight Speech Enhancement for Mandarin Esophageal Speech
41	VocalAgent: Large Language Models for Vocal Health Diagnostics with Safety-Aware Evaluation
2261	A Cookbook for Community-driven Data Collection of Impaired Speech in Low-Resource Languages
1133	Voice Quality Dimensions as Interpretable Primitives for Speaking Style for Atypical Speech and Affect

16:00-18:00 - Poster - Area 2 - Prosody and Voice Quality

2135	Tone recognition in low-resource languages of North-East India: peeling the layers of SSL-based speech models
960	Corpus-Based Insights into Mandarin Neutral Tone: Effects of Tonal Context and Structural Patterns in Spontaneous Speech
1325	Tonal variation and word meaning in Taiwanese
783	Sounding Like a Winner? Prosodic Differences in Post-Match Interviews
1938	Exploratory Study of Filled Pauses in Ukrainian Language: Phonetic Properties of Filled Pauses
180	Evaluating the suitability of acoustic parameters for capturing breathy voice in non-pathological female speakers
1949	Robustness of F0 Ratio as a Diagnostic: Comparing Creaky Voice in Danish and Seoul Korean

16:00-18:00 - Poster - Area 3 - Speech Emotion Recognition 2

2326	Towards Machine Unlearning for Paralinguistic Speech Processing
1100	Infant Cry Emotion Recognition Using Improved ECAPA-TDNN with Multi-scale Feature Fusion and Attention Enhancement
1373	Speech Mutil-label Emotion Recognition Using Asymmetric Class Loss Function Based on Effective Samples
1951	EmoDB 2.0: A Database of Emotional Speech in a World that is not Black or White but Grey
2571	Cross-corpus open-set Speech Emotion Recognition Method Based on Spatiotemporal Features with Inverse-Entropy Regularization
1064	CLEP-DG: Contrastive Learning for Speech Emotion Domain Generalization via Soft Prompt Tuning
530	Leveraging Unlabeled Audio-Visual Data in Speech Emotion Recognition using Knowledge Distillation

16:00-18:00 - Poster - Area 3 - Speaker Traits Recognition

1990	Who knows best? Effects of speech disfluencies on incentivized decision-making
2193	Enhancing Transcripts of Open-Source Automatic Speech Recognition Models Through Fine-Tuning with Laughter and Speech-Laugh
988	Investigating the Reasoning Abilities of Large Language Models for Understanding Spoken Language in Interpersonal Interactions
1902	A Naturally Elicited Multimodal Stress Database and Speech Breathing Based Stress Detection
165	From Context to Code-switching: Examining the Interplay of Language Proficiency and Multilingualism in Speech
1801	Extending the Fongbe to French Speech Translation Corpus: resources, models and benchmark
2037	On the Relationship between Accent Strength and Articulatory Features
2195	A Multi-Stream Framework Utilizing 3D Human Reconstruction for Cued Speech Recognition
313	On the cross-modal makeup of charisma: Insights from a field-data analysis

16:00-18:00 - Poster - Area 3 - Speech Emotion Recognition 3

166	Frozen Large Language Models Can Perceive Paralinguistic Aspects of Speech
2233	PARROT: Synergizing Mamba and Attention-based SSL Pre-Trained Models via Parallel Branch Hadamard Optimal Transport for Speech Emotion Recognition
1915	AA-SLLM: An Acoustically Augmented Speech Large Language Model for Speech Emotion Recognition
1439	Speaker-Aware Multi-Task Learning for Speech Emotion Recognition
2009	Multimodal Emotion Diarization: Frame-Wise Integration of Text and Audio Representations
2112	Analysis of Phonetic Level Similarities Across Languages in Emotional Speech
703	Label Semantic-Driven Contrastive Learning for Speech Emotion Recognition
1233	Pitch Contour Model (PCM) with Transformer Cross-Attention for Speech Emotion Recognition

16:00-18:00 - Poster - Area 7 - Multilingual Speech Synthesis and Special Applications 2

2758	RASMALAI: Resources for Adaptive Speech Modeling in Indian Languages with Accents, and Intonations
2031	Kinship in Speech: Leveraging Linguistic Relatedness for Zero-Shot TTS in Indian Languages
2679	Can we reconstruct a dysarthric voice with the large speech model Parler TTS?
432	Voice Adaptation for Swiss German
787	Gradual modeling of the Lombard effect by modifying speaker embeddings from a Text-To-Speech model
433	When Humans Growl and Birds Speak: High-Fidelity Voice Conversion from Human to Animal and Designed Sounds
656	EEG-based Voice Conversion : Hearing the Voice of Your Brain
1403	Streaming Non-Autoregressive Model for Accent Conversion and Pronunciation Improvement
406	Zero-Shot Mono-to-Binaural Speech Synthesis

16:00-18:00 - Poster - Area 7 - Emotion and Expressivity in Speech Synthesis and Voice Conversion

1638	EATS-Speech: Emotion-Adaptive Transformation and Priority Synthesis for Zero-Shot Text-to-Speech
1192	Voice Impression Control in Zero-Shot TTS
754	EME-TTS: Unlocking the Emphasis and Emotion Link in Speech Synthesis
1394	DiEmo-TTS: Disentangled Emotion Representations via Self-Supervised Distillation for Cross-Speaker Emotion Transfer in Text-to-Speech
2586	Spotlight-TTS: Spotlighting the Style via Voiced-Aware Style Extraction and Style Direction Adjustment for Expressive Text-to-Speech
276	Speaker-agnostic Emotion Vector for Cross-speaker Emotion Intensity

	Control
1684	SA-RAS: Speaker-Aware Style Retrieval Augmented Generation for Expressive Zero-shot Text-to-Speech Synthesis
1210	DiffEmotionVC: A Dual-Granularity Disentangled Diffusion Framework for Any-to-Any Emotional Voice Conversion
1101	ZSDEVC: Zero-Shot Diffusion-based Emotional Voice Conversion with Disentangled Mechanism
1115	MPE-TTS: Customized Emotion Zero-Shot Text-To-Speech Using Multi-Modal Prompt

16:00-18:00 - Special session - Area 14 - Special Session - Challenges in Speech Data Collection, Curation and Annotation - Part 2

681	You Are What You Say: Exploiting Linguistic Content for VoicePrivacy Attacks
2734	Recognizing Every Voice: Towards Inclusive ASR for Rural Bhojpuri Women
737	Augment Mandarin to Cantonese Speech Databases via Retrieval-Augmented Generation and Speech Synthesis
1983	An Exploratory Framework for LLM-assisted Human Annotation of Speech Datasets
2433	Automatic Labeling and Correction of Noisy Labels for Robust Self-Supervised Speaker Verification
17	Auto-Landmark: Acoustic Landmark Dataset and Open-Source Toolkit for Landmark Extraction
539	AusKidTalk: The benefits of out-of-domain automatic speech processing tools in corpus building
342	ASR-based segmentation for the analysis of larger child-speech datasets: Performance evaluation on vowels from Australian-English speaking children aged 4 to 11 years
1030	A semi-automatic pipeline for transcribing and segmenting child speech
462	Hybrid Data Sampling for ASR: Integrating Acoustic Diversity and Transcription Uncertainty
160	Whilter: A Whisper-based Data Filter for “In-the-Wild” Speech Corpora Using Utterance-level Multi-Task Classification
447	Adapting Whisper for low-resource Hindi-English Code-Mix speech with on-the-fly Augmentation & LLM-Synthesised Data
1078	Optimizing ASR for Catalan-Spanish Code-Switching: A Comparative Analysis of Methodologies
950	From Scarcity to Sufficiency: Speech Recognition Pipeline for Zero-resource Language
648	MIKU-PAL: An Automated and Standardized Multi-Modal Method for Speech Paralinguistic and Affect Labeling
2451	Multimodal Fusion with Semi-Supervised Learning Minimizes Annotation Quantity for Modeling Videoconference Conversation Experience
1916	Clinical Annotations for Automatic Stuttering Severity Assessment

Thursday 21/08/2025

08:30-10:30 - Oral - Area 1 - Multimodality

2073	Robust Speech-Driven Body Language Generation
178	Beat gestures made by human-like avatars affect speech perception
343	The mutual exclusivity bias of bilingual visually grounded speech models
1773	MultiActor-Audiobook: Zero-Shot Audiobook Generation with Faces and Voices of Multiple Speakers
1321	Incorporating Linguistic Constraints from External Knowledge Source for Audio-Visual Target Speech Extraction
759	Multimodal Silent Recognition of Phonemes Using Radar and Optopalatographic Silent Speech Interfaces

08:30-10:30 - Oral - Area 2 - Segments

2070	French schwa is not acoustically distinct from its two lexical neighbors /ø/ and /œ/
1674	Apical vs. Regular Vowel Duration: A Corpus-based Analysis of Contextual Influences in Standard Mandarin
282	On Apical Vowels in Eastern Zhenjiang Mandarin
1905	Equivalence and differences: Formant patterns of labialization and pharyngealization in Tashlhiyt
1219	Temporal organization of prenuclear glides in Hefei Mandarin
2465	Speaker-specific Patterns of Phonetic Covariation in Korean Word-medial Stops and the Role of Phonological and Morphological Contexts

08:30-10:30 - Oral - Area 5 - Source Separation 2

448	Band-SCNet: A Causal, Lightweight Model for High-Performance Real-Time Music Source Separation
800	CabinSep: IR-Augmented Mask-Based MVDR for Real-Time In-car Speech Separation with Distributed Heterogeneous Arrays
840	DGMO: Training-Free Audio Source Separation through Diffusion-Guided Mask Optimization
257	Cross-Attention-Based Target Sound Extraction by Fully Leveraging Enrollment in a Shared Latent Space
1148	DnR-nonverbal: Cinematic Audio Source Separation Dataset Containing Non-Verbal Sounds
214	Neural Speech Extraction with Human Feedback

08:30-10:30 - Oral - Area 6 - Speech Coding

1289	Unlocking Temporal Flexibility: Neural Speech Codec with Variable Frame Rate
196	SPCODEC: Split and Prediction for Neural Speech Codec
355	Probing the Robustness Properties of Neural Speech Codecs
1106	LSCodec: Low-Bitrate and Speaker-Decoupled Discrete Speech Codec
115	Bringing Interpretability to Neural Audio Codecs
827	NanoCodec: Towards High-Quality Ultra Fast Speech LLM Inference

08:30-10:30 - Oral - Area 8 - Prosody, Phoneme and Stress Modeling in ASR

1475	WHISTRESS: Enriching Transcriptions with Sentence Stress Detection
2068	Prosodic Structure Beyond Lexical Content: A Study of Self-Supervised Learning
1020	Learning Optimal Prosody Embedding Codebook based on F0 and Energy
1918	Pitch Accent Detection improves Pretrained Automatic Speech Recognition
2417	Towards Accurate Phonetic Error Detection Through Phoneme Similarity Modeling
1924	Exploring auditory feedback mechanisms in speech recognition

08:30-10:30 - Oral - Area 10 - Speech Assessment and Language Learning

2359	GoP2Vec: A few shot learning for pronunciation assessment with goodness of pronunciation (GoP) based representations from an i-vector framework and augmentation
829	Enhancing GOP in CTC-Based Mispronunciation Detection with Phonological Knowledge
249	Multilingual Speech Assessment Using Cross-Attention and Multitask Learning
1793	Assessment of L2 Oral Proficiency using Speech Large Language Models
594	Scaling and Prompting for Improved End-to-End Spoken Grammatical Error Correction
1610	Bidirectional Spoken-Written Text Conversion with Large Language Models

08:30-10:30 - Oral - Area 11 - Spoken Dialogue Systems 2

Survey Talk	End-to-end spoken dialog system **(Survey Talk, 40 mins)**
2071	Factors affecting the in-context learning abilities of LLMs for dialog state tracking
843	Spoken question answering for visual queries
1075	Towards Human-like Multimodal Conversational Agent by Generating Engaging Speech
2339	A Chain-of-Thought Reasoning Approach to E2E Spoken Dialogue Systems with an Open-Source Toolkit

08:30-10:30 - Oral - Area 13 - Depression Detection and Assessment 2

1968	An interpretable speech foundation model for depression detection by revealing prediction-relevant acoustic features from long speech
1035	Speech and Text Foundation Models for Depression Detection: Cross-Task and Cross-Language Evaluation
1789	A Study on The Impact of Foundation Models on Automatic Depression Detection from Speech Signals
933	Can Speech Accurately Detect Depression in Patients With Comorbid Dementia? An Approach for Mitigating Confounding Effects of Depression and Dementia
2560	Identifying Vocal and Facial Biomarkers of Depression in Large-Scale Remote Recordings: A Multimodal Study Using Mixed-Effects Modeling

329	M3L: A Multi-Modal and Multi-Lingual Depression Detection Framework

08:30-10:30 - Poster - Area 4 - Speaker Diarization 2

10	Mitigating Non-Target Speaker Bias in Guided Speaker Embedding
862	Speaker Diarization with Overlapping Community Detection Using Graph Attention Networks and Label Propagation Algorithm
1059	DLF-EEND: Dynamic Layer Fusion for End-to-End Speaker Diarization
1150	Pushing the Limits of End-to-End Diarization
1663	Spatio-spectral diarization of meetings by combining TDOA-based segmentation and speaker embedding-based clustering
1749	Selective Channel Attention based Target Speaker Voice Activity Detection for Speaker Diarization under AD-HOC Microphone Array Settings
1807	Diarization-Guided Multi-Speaker Embeddings
2244	Streaming Sortformer: Speaker Cache-Based Online Speaker Diarization with Arrival-Time Ordering
895	A Hybrid Approach to Combining Role Diarization with ASR for Professional Conversations

08:30-10:30 - Poster - Area 4 - Watermarking and Anonymization

642	WAKE: Watermarking Audio with Key Enrichment
1091	Defend for Self-Vocoding: A Novel Enhanced Decoder Network for Watermark Recovery
316	Cross-Modal Watermarking for Authentic Audio Recovery and Tamper Localization in Synthesized Audiovisual Forgeries
575	VoiceMark: Zero-Shot Voice Cloning-Resistant Watermarking Approach Leveraging Speaker-Specific Latents
1530	A Comprehensive Real-World Assessment of Audio Watermarking Algorithms: Will They Survive Neural Codecs?
244	How to Recover Long Audio Sequences Through Gradient Inversion Attack With Dynamic Segment-based Reconstruction
1027	First Steps Towards Voice Anonymization for Code-Switching Speech
2317	Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems
1469	Mitigating Language Mismatch in SSL-Based Speaker Anonymization

08:30-10:30 - Poster - Area 6 - Single-channel Speech Enhancement

2168	MSFNet: A Nested Model for Multi-Sampling-Frequency Speech Enhancement
391	TF-SkiMNet: Speech Enhancement Based on Inplace Modeling and Skipping Memory in Time-Frequency Domain
108	xLSTM-SENet: xLSTM for Single-Channel Speech Enhancement
896	From KAN to GR-KAN: Advancing Speech Enhancement with KAN-Based Methodology
376	Stack Less, Repeat More: A Block Reusing Approach for Progressive Speech Enhancement
1476	Mamba-based Hybrid Model for Speech Enhancement
2012	Restoring Harmonics: Enhancing Speech Quality with Deep Mask and Harmonic Restoration Network

08:30-10:30 - Poster - Area 6 - Speech Enhancement and Representation Learning

392	SaD: A Scenario-Aware Discriminator for Speech Enhancement
1352	Listen through the Sound: Generative Speech Restoration Leveraging Acoustic Context Representation
1270	Efficient Speech Enhancement via Embeddings from Pre-trained Generative Audioencoders
1471	Towards Personalised Audio Visual Speech Enhancement
1745	FlowSE: Efficient and High-Quality Speech Enhancement via Flow Matching
2436	Speech Enhancement based on cascaded two flows
552	X-ARES: A Comprehensive Framework for Assessing Audio Encoder Performance
2528	WavShape: Information-Theoretic Speech Representation Learning for Fair and Privacy-Aware Audio Processing

08:30-10:30 - Poster - Area 7 - Datasets and Tools for Speech Synthesis

989	HiFiTTS-2: A Large-Scale High Bandwidth Speech Dataset
1116	JIS: A Speech Corpus of Japanese Idol Speakers with Various Speaking Styles
2151	FaVC: A Validated, Transcribed, Parallel Farsi Speech Dataset for Voice Conversion
2573	SawtArabi: A Benchmark Corpus for Arabic TTS. Standard, Dialectal and Code-Switching
2536	The text-to-speech in the wild (TITW) dataset
559	Towards Emotionally Consistent Text-Based Speech Editing: Introducing EmoCorrector and The ECD-TSE Dataset
1550	ArVoice: A Multi-Speaker Dataset for Arabic Speech Synthesis
973	A Dataset for Automatic Assessment of TTS Quality in Spanish

08:30-10:30 - Poster - Area 7 - Neural Codecs and Vocoders

1440	FreeCodec: A disentangled neural speech codec with fewer tokens
468	DualCodec: A Low-Frame-Rate, Semantically-Enhanced Neural Audio Codec for Speech Generation
1819	Comparative Analysis of Fast and High-Fidelity Neural Vocoders for Low-Latency Streaming Synthesis in Resource-Constrained Environments
464	Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning
1763	Vocoder-Projected Feature Discriminator
2739	AF-Vocoder: Artifact-Free Neural Vocoder with Global Artifact Filter
1641	Robust neural codec language modeling with phoneme position prediction for zero-shot TTS
2726	DS-Codec: Dual-Stage Training with Mirror-to-NonMirror Architecture Switching for Speech Codec
347	PeriodCodec: A Pitch-Controllable Neural Audio Codec Using Periodic Signals for Singing Voice Synthesis

08:30-10:30 - Poster - Area 8 - Adaptation and Target-speaker ASR

2486	Enhancing Target-speaker Automatic Speech Recognition Using Multiple Speaker Embedding Extractors with Virtual Speaker Embedding
1243	SC-SOT: Conditioning the Decoder on Diarized Speaker Information for End-to-End Overlapped Speech Recognition
2580	Efficient Data Selection for Domain Adaptation of ASR Using Pseudo-Labels and Multi-Stage Filtering
2601	Better Semi-supervised Learning for Multi-domain ASR Through Incremental Retraining and Data Filtering
1263	MOPSA: Mixture of Prompt-Experts Based Speaker Adaptation for Elderly Speech Recognition
606	Visually-Adaptive Guided Robust Speech Recognition with Parameter-Efficient Adaptation
694	Regularizing Learnable Feature Extraction for Automatic Speech Recognition
560	MMLoRA: Multitask Memory Parameter-Efficient Fine-Tuning for Multimodal SER
903	Robust Unsupervised Adaptation of a Speech Recogniser Using Entropy Minimisation and Speaker Codes

08:30-10:30 - Poster - Area 9 - Contextual Biasing and Adaptation

550	GLCLAP: A Novel Contrastive Learning Pre-trained Model for Contextual Biasing in ASR
1300	WCTC-Biasing: Retraining-free Contextual Biasing ASR with Wildcard CTC-based Keyword Spotting and Inter-layer Biasing
646	Ranking and Selection of Bias Words for Contextual Bias Speech Recognition
2621	OWSM-Biasing: Contextualizing Open Whisper-Style Speech Models for Automatic Speech Recognition with Dynamic Vocabulary
1378	Label-Context-Dependent Internal Language Model Estimation for CTC
1624	Assessing the Performance and Efficiency of Mamba ASR in Low-Resource Scenarios
2549	Adapting Whisper for Parameter-efficient Code-Switching Speech Recognition via Soft Prompt Tuning

08:30-10:30 - Special session - Area 14 - Special Session - Speech Emotion Recognition in Naturalistic Conditions Challenge

1112	Towards LLM-Empowered Fine-Grained Speech Descriptors for Explainable Emotion Recognition
1283	From Pretraining to Performance: Benchmarking Self-Supervised Speech Models for Interspeech-25 SER Challenge
1163	Developing A Top-tier Framework in Naturalistic Conditions Challenge for Categorized Emotion Prediction: From Speech Foundation Models and Learning Objective to Data Augmentation and Engineering Choices
1082	Developing a High-performance Framework for Speech Emotion Recognition in Naturalistic Conditions Challenge for Emotional Attribute Prediction
1391	EmoSphere-SER: Enhancing Speech Emotion Recognition through

	Spherical Representation with Auxiliary Classification
1841	Explainable Speech Emotion Recognition Through Attentive Pooling: Insights from Attention-Based Temporal Localization
368	ABHINAYA - A System for Speech Emotion Recognition In Naturalistic Conditions Challenge
1972	The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions
1041	MATER: Multi-level Acoustic and Textual Emotion Representation for Interpretable Speech Emotion Recognition
2033	Multi-task learning for speech emotion recognition in naturalistic conditions
2636	Medusa: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions
1662	Interactive Fusion of Multi-View Speech Embeddings via Pretrained Large-Scale Speech Models for Speech Emotional Attribute Prediction in Naturalistic Conditions
1445	Advancing Emotion Recognition via Ensemble Learning: Integrating Speech, Context, and Text Representations
2014	Improving Speech Emotion Recognition Through Cross Modal Attention Alignment and Balanced Stacking Model
1141	EmoJudge: LLM Based Post-Hoc Refinement for Multimodal Speech Emotion Recognition
1357	Lessons Learnt: Revisit Key Training Strategies for Effective Speech Emotion Recognition in the Wild
1380	Enhancing Speech Emotion Recognition with Multi-Task Learning and Dynamic Feature Fusion

08:30-10:30 - Show and Tell - Education / Assistive Technology

2812	SCRIBAL: A Digital Transcription Tool in Higher Education
2815	From Static to Dynamic: Enhancing AAC with Generative Imagery and Zero-Shot TTS
2819	Concurrent Speech and Auditory Tag Clouds for Non-Visual Web Interaction
2823	Towards Domain-Specific Spoken Language Understanding for a Catalan Voice-Controlled Video Game
2827	Accessible Delivery of Visual-Acoustic Biofeedback for Speech Sound Disorder
2835	End-to-End Indian Language Dubbing with Zero-Shot Speaker Preservation
2839	SLP Sidekick: An Open-Source, Multilingual Speech Therapy Platform

11:00-12:00 - Keynote speaker - Judith Holler

Using and comprehending language in face-to-face conversation

13:30-15:30 - Oral - Area 2 - Prosody

863	The prosodic characteristics of Standard Chinese rhetorical questions in naturalistic settings
2028	ProBiEM: Acoustic and Lexical Correlates of Prosodic Prominence in

	English-Malayalam Bilingual Speech
2161	Are you being sarcastic? Prosodic cues to irony perception in German
1873	Can AI Understand Mandarin Speech Prosody? A Framework and Benchmark Showcase
2159	Generating Consistent Prosodic Patterns from Open-Source TTS Systems
143	Multimodal Prosody Modeling: A Use Case for Multilingual Sentence Mode Prediction

13:30-15:30 - Oral - Area 3 - Emotions and Foundational Models

1232	EAA: Emotion-Aware Audio Large Language Models with Dual Cross-Attention and Context-Aware Instruction Tuning
1979	Chain-of-Thought Distillation with Fine-Grained Acoustic Cues for Speech Emotion Recognition
2093	Exploring the Limits of Conformer CTC-Encoder for Speech Emotion Recognition using Large Language Models
261	Token-Level Logits Matter: A Closer Look at Speech Foundation Models for Ambiguous Emotion Recognition
149	Assessing the feasibility of large language models for detecting micro-behaviors in team interactions during space missions
756	A-SMiLE: Affective Sparse Mixture-of-Experts Adapter with Multi-Task Learning for Spoken Dialogue Models

13:30-15:30 - Oral - Area 4 - Speaker Recognition

2018	The 2024 NIST Speaker Recognition Evaluation
515	EmoSpeechAuth: Emotion-Aware Speaker Verification
2345	TELVID: A Multilingual Multi-modal Corpus for Speaker Recognition
2545	A Simple-Yet-Effective Data Augmentation Method for Speaker Identification in Novels
736	IDIR: Identifying and Distilling Informative Relations for Speaker Verification
2737	Analysis of ABC Frontend Audio Systems for the NIST-SRE24

13:30-15:30 - Oral - Area 6 - Prediction and Evaluation of Speech Quality and Intelligibility

1377	Non-Intrusive Binaural Speech Intelligibility Prediction Using Mamba for Hearing-Impaired Listeners
1599	No Audiogram: Leveraging Existing Scores for Personalized Speech Intelligibility Prediction
1756	Feature Importance across Domains for Improving Non-Intrusive Speech Intelligibility Prediction in Hearing Aids
991	Intelligibility Prediction for Time-Modified Speech Signals Using Spectro-Temporal Modulation Features
1947	French Listening Tests for the Assessment of Intelligibility, Quality, and Identity of Body-Conducted Speech Enhancement
984	Benchmarking Neural Speech Codec Intelligibility with SITool

13:30-15:30 - Oral - Area 7 - Speech Synthesis Paradigms and Methods 3

1084	Efficient Streaming TTS Acoustic Model with Depthwise RVQ Decoding Strategies in a Mamba Framework
455	APTTS: Adversarial Post-training in Latent Flow Matching for Fast and High-fidelity Text-to-Speech
277	Eigenvoice Synthesis based on Model Editing for Speaker Generation
1066	Score-Based Training for Energy-Based TTS Models
2447	Accelerating Autoregressive Speech Synthesis Inference With Speech Speculative Decoding
1122	BitTTS: Highly Compact Text-to-Speech Using 1.58-bit Quantization and Weight Indexing

13:30-15:30 - Oral - Area 8 - Multi-Talker ASR

Survey Talk	Advances in Conversational Speech Recognition **(Survey Talk, 40 mins)**
1886	AISHELL-5: The First Open-Source In-Car Multi-Channel Multi-Speaker Speech Dataset for Automatic Speech Diarization and Recognition
2142	Speaker Targeting via Self-Speaker Adaptation for Streaming Multi-talker ASR
2572	Speaker-Distinguishable CTC: Learning Speaker Distinction using CTC for Multi-Talker Speech Recognition
2414	Improving Practical Aspects of End-to-End Multi-Talker Speech Recognition for Online and Offline Scenarios

13:30-15:30 - Oral - Area 12 - ASR Assessment and Foundational Models

891	Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches
1950	SOVA-Bench: Benchmarking the Speech Conversation Ability for LLM-based Voice Assistant
2488	An approach to measuring the performance of Automatic Speech Recognition(ASR) models in the context of Large Language Model(LLM) powered applications
246	DC-Spin: A Speaker-invariant Speech Tokenizer for Spoken Language Models
310	Exploring the Effect of Segmentation and Vocabulary Size on Speech Tokenization for Speech Language Models
2741	Hearing deficits of transformer-based ASR for anechoic and spatial signals

13:30-15:30 - Poster - Area 4 - Speech Deepfakes

717	Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection
1703	Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection
1419	A Comparative Study on Proactive and Passive Detection of Deepfake Speech
2583	PhonemeFake: Redefining Deepfake Realism with Language-Driven Segmental Manipulation and Adaptive Bilevel Detection
1095	From Sharpness to Better Generalization for Speech Deepfake Detection

100	Unmasking real-world audio deepfakes: A data-centric approach
2105	A Data-Driven Diffusion-based Approach for Audio Deepfake Explanations
942	PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing
2298	Rehearsal with Auxiliary-Informed Sampling for Audio Deepfake Detection

13:30-15:30 - Poster - Area 4 - Speech Deepfakes, Antispoofing and Backdoor Attacks

1543	Layer-Wise Decision Fusion for Fake Audio Detection Using XLS-R
2659	SynHate: Detecting Hate Speech in Synthetic Deepfake Audio
1234	Can Emotion Fool Anti-spoofing?
1411	Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models
362	Amplifying Artifacts with Speech Enhancement in Voice Anti-spoofing
1895	Thai Speech Spoofing Detection Dataset with Variations in Speaking Styles
372	CBA: Backdoor Attack on Deep Speech Classification via Audio Compression
1872	LRBA: Stealthy Backdoor Attacks on Speech Classification via Latent Rearrangement in VITS
2677	LitMAS: A Lightweight and Generalized Multi-Modal Anti-Spoofing Framework for Biometric Security

13:30-15:30 - Poster - Area 5 - Speech Analysis and Quality Assessment

2063	HASRD: Hierarchical Acoustic and Semantic Representation Disentanglement
1882	Performance of Montreal Forced Aligner on Cantonese Spontaneous Speech
2075	Segmentation-Variant Codebooks for Preservation of Paralinguistic and Prosodic Information
15	AdaKWS: Towards Robust Keyword Spotting with Test-Time Adaptation
518	Multivariate Probabilistic Assessment of Speech Quality
1992	A Study on Speech Assessment with Visual Cues
2269	Efficient Streaming Speech Quality Prediction with Spiking Neural Networks
1435	Unifying Listener Scoring Scales: Comparison Learning Framework for Speech Quality Assessment and Continuous Speech Emotion Recognition

13:30-15:30 - Poster - Area 5 - Speech Analysis, Detection and Classification 2

205	Identifying Primary Stress Across Related Languages and Dialects with Transformer-based Speech Encoder Models
2519	SupraDoRAL: Automatic Word Prominence Detection Using Suprasegmental Dependencies of Representations with Acoustic and Linguistic Context
1639	LombardTokenizer: Disentanglement and Control of Vocal Effort in a Neural Speech Codec
98	Robust Personal Voice Activity Detection for Mitigating Domain Mismatch and False Acceptance Scenarios
30	Adaptive Knowledge Distillation for Device-Directed Speech Detection

322	Flexible VAD-PVAD Transition: A Detachable PVAD Module for Dynamic Encoder RNN VAD
2178	Speaker Conditioning of Voice Activity Detection via Implicit Separation
730	ASDA: Audio Spectrogram Differential Attention Mechanism for Self-Supervised Representation Learning
1242	DuRep: Dual-Mode Speech Representation Learning via ASR-Aware Distillation

13:30-15:30 - Poster - Area 13 - Pathological Speech Analysis 4

793	On the Relevance of Clinical Assessment Tasks for the Automatic Detection of Parkinson's Disease Medication State from Speech
738	Speech power spectra: a window into neural oscillations in Parkinson's disease
521	Synchronous analysis of abnormal acoustic and linguistic production in Parkinson's speech
2056	Automatic Detection and Sub-typing of Primary Progressive Aphasia from Speech: Integrating Task-Specific Features and Spatio-Semantic Graphs
964	Towards Classification of Typical and Atypical Disfluencies: A Self Supervised Representation Approach
1124	Stuttering Detection Based on Self-Attention Weights of Temporal Acoustic Vector Sequence
587	Speech-Based Chronic Kidney Disease Diagnosis via Transformer Fusion of Glottal and Spectrogram Features
1851	Influence of Room Acoustics on Objective Voice Assessment Methods in the Context of Speech and Language Therapy
2307	Multimodal Speech-Based Biomarkers Outperform the ALS Functional Rating Scale in Predicting Individual Disease Progression in ALS

13:30-15:30 - Poster - Area 13 - Pathological Speech Analysis 5

2280	Pitfalls and Limits in Automatic Dementia Assessment
2751	On the Within-class Variation Issue in Alzheimer's Disease Detection
219	Alzheimer's Disease Detection Using Co-Attention Mechanism for Acoustic and ASR-Transcribed Text Features
1598	Beyond Manual Transcripts: The Potential of Automated Speech Recognition Errors in Improving Alzheimer's Disease Detection
761	Voice-Based Dysphagia Detection: Leveraging Self-Supervised Speech Representation
523	ADCeleb: A Longitudinal Speech Dataset from Public Figures for Early Detection of Alzheimer's Disease
2000	Anne Rowling Neurological Speech Corpus: clinically annotated longitudinal dataset for developing speech biomarkers in neurodegenerative disorders
259	Multitask Learning with Fused Attention for Improved ASR and Mispronunciation Detection in Children's Speech Sound Disorders
2272	Multimodal Speech, Language and Orofacial Analysis for Remote Assessment of Positive, Negative and Cognitive Symptoms in Schizophrenia

13:30-15:30 - Special session - Area 14 - Special Session - Biosignal-enabled Spoken Communication

1887	GTAnet: Geometry-Guided Temporal Attention for EEG-Based Sound Source Tracking in Cocktail Party Scenarios
1637	Decoding Listener's Identity: Person Identification from EEG Signals Using a Lightweight Spiking Transformer
1400	Recreating Neural Activity During Speech Production with Language and Speech Model Embeddings
2752	Towards Sentence Level Imagined Speech Generation from EEG signals
2550	Word-Level Error Analysis in Decoding Systems: From Speech Recognition to Brain-Computer Interfaces
1304	NeuroSpex+: Dual-Task Training of Neuro-Guided Speaker Extraction with Speech Envelope and Waveform
1914	DiffMV-ETS: Diffusion-based Multi-Voice Electromyography-to-Speech Conversion using Speaker-Independent Speech Training Targets
2147	Conformer-based Ultrasound-to-Speech Conversion
860	Training Articulatory Inversion Models for Interspeaker Consistency
1519	Enhancing Acoustic-to-Articulatory Inversion with Multi-Target Pretraining for Low-Resource Settings
1727	Articulatory Vowel Distinctiveness in Spanish
769	EEG-based Speech Decoding Based on Multi-mode Joint Modeling
1183	A Silent Speech Decoding System from EEG and EMG with Heterogenous Electrode Configurations
1537	NAM-to-Speech Conversion with Multitask-Enhanced Autoregressive Models
1160	RESOUND: Speech Reconstruction from Silent Videos via Acoustic-Semantic Decomposed Modeling