Closure Report

7th Edition of ISCA Supported Summer School on Speech Signal Processing (S4P), July 05–09, 2025

Organized by

Speech Research Lab,
Dhirubhai Ambani University (DAU)
(formerly DA-IICT), Gandhinagar
Website: https://sites.google.com/view/s4p2025/home

Organizing Chair
Prof. (Dr.) Hemant A. Patil
Room No. 4103, Faculty Block-4,
DAU, Gandhinagar–382 007, India
Mail: hemant_patil@dau.ac.in
Telephone: +91-79-68261650
Website: https://sites.google.com/site/hemantpatildaiict/

Sponsors

Dhirubhai Ambani University

IndSCA

elra
Language Resources Association

tcs Research

Scientech
www.ScientechWorld.com

Technical Co-sponsor

इलेक्ट्रॉनिकी एवं
सूचना प्रौद्योगिकी मंत्रालय
MINISTRY OF
ELECTRONICS AND
INFORMATION TECHNOLOGY
सत्यमेव जयते

IEEE
Signal
Processing
Society
GUJARAT CHAPTER

1 Message from organising chair

On behalf of the Organizing Committee, we record our appreciation for the valuable contribution made by eminent world-class invited speakers, participants, international program committee, DAU faculty colleagues, staff, administration and student volunteers towards conducting the 7th edition of ISCA supported summer school with the theme *'Automatic Speech Recognition (ASR)'* during July 05-09, 2025 at DAU Gandhinagar, India. This summer school gave a platform to interact with distinguished invited speakers, to discover novel methods and broaden our knowledge in the broad area of voice biometrics. Furthermore, to encourage young talent, the school presented the 6th edition of the *5 Minute Ph.D. Thesis (5MPT)* contest with four ISCA endorsed cash prizes.

We were honored to host a remarkable lineup of experts and researchers from leading institutions worldwide. From Japan, we welcomed IEEE SP DLs by **Prof. (Dr.) Akihiko K. Sugiyama** (Founder, Damascus Corporation, Tokyo, and Professor, Kansai University, Japan), **Prof. (Dr.) Tatsuya Kawahara** (Kyoto University, Japan), and **Prof. (Dr.) Yu Tsao** (Research Fellow and Deputy Director, Research Center for Information Technology Innovation, Academia Sinica, Taiwan). From the UK, **Prof. (Dr.) Thomas Hain** (The University of Sheffield, UK) shared his insights, while Switzerland was represented by **Dr. Petr Motlicek** and **Dr. Mathew Magimai Doss** (both Senior Researchers, IDIAP Research Institute, Martigny, Switzerland). Singapore was represented by **Dr. Nancy F. Chen** (Generative AI Group Leader, A*STAR – Agency for Science, Technology and Research, Singapore). From India, we had the privilege of hosting **Prof. (Dr.) B. Yegnanarayana** (Emeritus Professor, IIIT Hyderabad), **Prof. (Dr.) Hema A. Murthy** (Emeritus Professor, IIT Madras), **Prof. (Dr.) S. Umesh** (IIT Madras), **Prof. (Dr.) Sriram Ganapathy** (Indian Institute of Science, Bengaluru, and Google Research, Bengaluru), **Prof. (Dr.) K. Sri Rama Murty** (IIT Hyderabad), **Prof. (Dr.) Vipul Arora** (IIT Kanpur), **Prof. (Dr.) Anil Kumar Vuppala** (IIIT Hyderabad), **Prof. (Dr.) Vinayak Abrol** (IIIT Delhi), and **Prof. (Dr.) Hemant A. Patil** (DAU, Gandhinagar). Their collective expertise and contributions enriched the discussions on speech processing, machine learning, and AI technologies.

At the Summer School, motivated from INTERSPEECH 2018, we organized a special session on *Industry Perspective Talks* in which senior industry personnel, namely, **Dr. Sri Garimella** (Director, Applied Science, Amazon, Bengaluru), **Dr. Sunil Kumar Kopparapu** (Principal Scientist, TCS Research & Innovations Labs, Mumbai), **Dr. Nagaraj Adiga** (Senior Principal Research Scientist, Outcomes.ai, Bengaluru), **Dr. Debmalya Chakrabarty** (Senior Applied Scientist, Amazon Alexa, Bengaluru), **Dr. Premjeet Singh** (Lead Engineer, Samsung Research Institute, Bengaluru), **Dr. Nirmesh J. Shah** (Senior Research Scientist, Sony Research India, Bengaluru), **Dr. Bidisha Sharma** (Senior AI Scientist, Uniphore, Bengaluru), **Dipesh K. Singh** (Speech Recognition Engineer, Augnito, Mumbai), and **Thoshith S.** (Speech Solutions Architect, Gnani.ai, Bengaluru) shared their insights and experiences.Bengaluru).

Events of this kind cannot happen without generous financial support from sponsors. In this regard, we express our deep gratitude and appreciation to our **Gold Sponsors**, namely, **Dhirubhai Ambani University (formerly DA-IICT), Gandhinagar** and **International Speech Communication Association (ISCA)**, as well as our **Bronze Sponsors**, **Indian Speech Communication Association (IndSCA)**, **European Language Resources Association**, **Tata Consultancy Services**, and **Scientech Technologies**. We also extend our thanks to and **The Ministry of Electronics and Information Technology (MeitY)**, for their invaluable support. Our technical co-sponsors, **IEEE Signal Processing Society, Gujarat Section Chapter**, which played a crucial role in making this event possible. In addition, we would like to thank **Prof. Phil D. Green** (University of Sheffield, UK) for their valuable feedback on our proposal for possible ISCA support to S4P 2025.

The members of the Organizing Committee hope that the participants and invited speakers had a memorable experience and pleasant stay at Gandhinagar, and hope that you will continue to visit DAU in future and participate in such ISCA supported events.

Prof. (Dr.) Hemant A. Patil
(Member if IEEE, ISCA, APSIPA, EUSIPCO)
Dhirubhai Ambani University (formerly
DA-IICT), Gandhinagar
ISCA Distinguished Lecturer (2020–2022)
APSIPA Distinguished Lecturer (2018–2019)

2 About S4P 2025

S4P-Summer School on Speech Signal Processing 2025 is being organized as a part of educational outreach activities at Speech Research Lab, Dhirubhai Ambani University (DAU) (Formerly DA-IICT), Gandhinagar, India. It will provide opportunities to students, faculty, researchers, and professionals to enhance their fundamentals and get exposed to research areas in the field of speech signal processing. The school consists of a theme topic and tutorials surrounding it.

This school is the 7^{th} **summer school** at DAU, following the successful earlier six summer schools, namely, S4P 2024, S4P 2019, S4P 2018, S4P 2017, S4P 2016, and ASAP 2016, focusing on different topics in the broad area of speech signal processing.

The theme for **S4P 2025** is **Automatic Speech Recognition (ASR)**. ASR system is now a key component of commercially successful Intelligent Personal Assistants (IPAs), such as Apple's Siri, Microsoft's Cortana, Google Assistant, Amazon's Echo/Alexa, Samsung's Bixby, IBM's Watson, etc. Design of ASR system depends upon various factors, such as near-field *vs.* far-field speech, recording and transmission channel conditions, acoustic model, language model, signal degradation conditions (acoustic noise), etc. Understanding these technological challenges is the major goal of **S4P 2025**. This school is targeted mainly towards graduate/post-graduate students, college teachers/faculty in educational institutions, and scientists/researchers in research laboratories/industries, who are interested in this area, and would like to learn more about this area from some of the best researchers in the world.

The summer school greatly benefits from invited talks from eminent and world-class researchers from academia, industry, and research laboratories from India and abroad. In addition, **S4P 2025** is highly benefited by world-renowned international program committee (PC) members from multiple countries. Furthermore, to encourage young talent, the school presents the 6^{th} edition of the **5 Minutes PhD Thesis Contest (5MPT)** with four cash prizes, the second edition of the **Poster Presentation Contest (PPC)** and the 1^{st} edition of **Deepfake Challenge**. Finally, **S4P 2025** also features the 4^{st} **edition of industry perspective talks** by experts from the speech technology industry.

3 Committee

Patron

1. Tathagata Bandyopadhyay, Director General, DAU Gandhinagar

International Program Committee

1. Shrikanth (Shri) Narayanan, University of Southern California, USA.
2. Hervé BOURLARD, EPFL Honorary Professor and former Director of Idiap Research Institute, Switzerland.

3. Nicholas Evans, EURECOM, France.
4. Dilek Hakkani-Tur, University of Illinois Urbana-Champaign, USA.
5. Kong Aik Lee, The Hong Kong Polytechnic University, Hong Kong.
6. Dong Yu, Tencent AI Lab, USA.
7. Elmar Nöth, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany.
8. Tiago H. Falk, MuSAE Lab, Montreal/Gatineau, Canada.
9. Tatsuya Kawahara, School of Informatics, Kyoto University, Japan.
10. Hexin Liu, Nanyang Technological University, Singapore.
11. Lori Lamel, CNRS, France.
12. Nancy Zlatintsi, NTUA, Greece.
13. Torbjørn Karl Svendsen, Norwegian University of Science and Technology, Norway
14. Akihiko Sugiyama, Founder, Damascus Corporation, Japan.
15. Khalid Choukri, Secretary General, European Language Resources Association (ELRA), France.

4 Invited Speakers

Akihiko K. Sugiyama
Damascus Corporation,
Japan

Thomas Hain
University of Sheffield,
UK

Petr Motlicek
IDIAP, Switzerland

Mathew Magimai Doss
IDIAP, Switzerland

Yu Tsao
Academia Sinica, Taiwan

Nancy F. Chen
Institute for Infocomm
Research (I2R),
Singapore

Tatsuya Kawahara
Kyoto University

Bayya Yegnanarayana
IIIT Hyderabad

Hema A. Murthy
IIT Madras

Srinivasan Umesh
IIT Madras

Sriram Ganapathy
IISc Bengaluru

K. Sri Rama Murty
IIT Hyderabad

Vipul Arora
IIT Kanpur

Anil Kumar Vuppala
IIIT Hyderabad

Vinayak Abrol
IIIT Delhi

Hemant A. Patil
DAU Gandhinagar

5 Industry Perspective Talks

Sri Garimella
Amazon AGI, Bengaluru

Sunil Kumar Kopparapu
TCS Research, Mumbai

Nagaraj Adiga
Outcomes.ai, Bengaluru

Debmalaya Chakroborty
Amazon AGI, Bengaluru

Premjit Singh
Samsung R&D Institute, Bengaluru

Nirmesh J. Shah
Sony Research, Bengaluru

Bidisha Sharma
Uniphore, Bengaluru

Dipesh K. Singh
Augnito, Bengaluru

Thoshith S
Gnani.ai, Bengaluru

6 Organizing Committee

Hemant A. Patil
DAU Gandhinagar, India

Eng-Siong Chng
Nanyang Technological University (NTU), Singapore

Mathew Magimai Doss
IDIAP Research Institute, Switzerland

Hardik B. Sailor
Institute for Infocomm Research (I2R), Singapore

Rodrigo Capobianco Guido
São Paulo State University (UNESP), Brazil

7 Arrangement Committee

No.	Volunteers	Admin and Helpdesk Team
1	Ravindrakumar M. Purohit	Siddharth Swaminarayan
2	Priya A. Damdar	Krutika Raval
3	Dharmendra Vaghela	Kirit Pandya
4	Aniket Pandey	Divyam Mishra
5	Manish Manojkumar Prajapati	Jaydeep Panchal
6	Vishnu Vardhan G V S	Sanjay Bariya
7	Arth J. Shah	Bhavesh Shah
8	Hiya Chaudhari	Rajendra Shah
9	Nisarg Trivedi	Jitendra Parmar
10	Siddharth Kumar	Sawan Kumar Sachaniya
11	Kausthubh Wade	Nimesh Patel
12	Saroj Pandit	Rajesh Patel
13	Nischay Agrawal	Gyanesh Pandya
14	Sahil Sadarangani	Prabhunath Sharma
15	Devanshi Trivedi	Souvik Sarkar
16	Himanshi Borad	Anuradha Srivastava
17	Manav Gaikwad	Deepali Sharma
18	Jayraj Lakkad	Shirish Varma
19	Rohith	Niketa Raval
20	Lakshman	Geeta Nair
21	Kunjan Gajre	Ramesh Prajapati
22	Rajnidhi Gupta	Abhilash Bhaskaran
23	Satyam Tiwari	Jainik Patel
24	Satyam Rana	Mahendrabhai (Housekeeping)
25	Vijay Hothi	Vipul Makwana
26	Ami Gandhi	Ashvin Chaudhari
27		Darshan Prajapati
28		Priyank Santola
29		Chaitanya Bhamare
30		Dhruti Joshi
31		Saurabh Nayee
32		Juhi Patel
33		Jaydeep Panchal
34		Jhalabhai (Electrician, Security Guard Team)

Table 3: List of Volunteers and Admin & Helpdesk Team

8 Sponsors

Gold Sponsors

Dhirubhai Ambani University (formerly DA-IICT), Gandhinagar

International Speech Communication Association (ISCA)

Bronze Sponsors

Indian Speech Communication Association (IndSCA)

European Language Resources Association (ELRA)

Tata Consultancy Services (TCS)

Scientech Technologies

Technical Co-sponsors

Ministry of Electronics and Information Technology (MEITy)

IEEE Signal Processing Society – IEEE Gujarat Section

9 Announcement and Publicity of S4P 2025

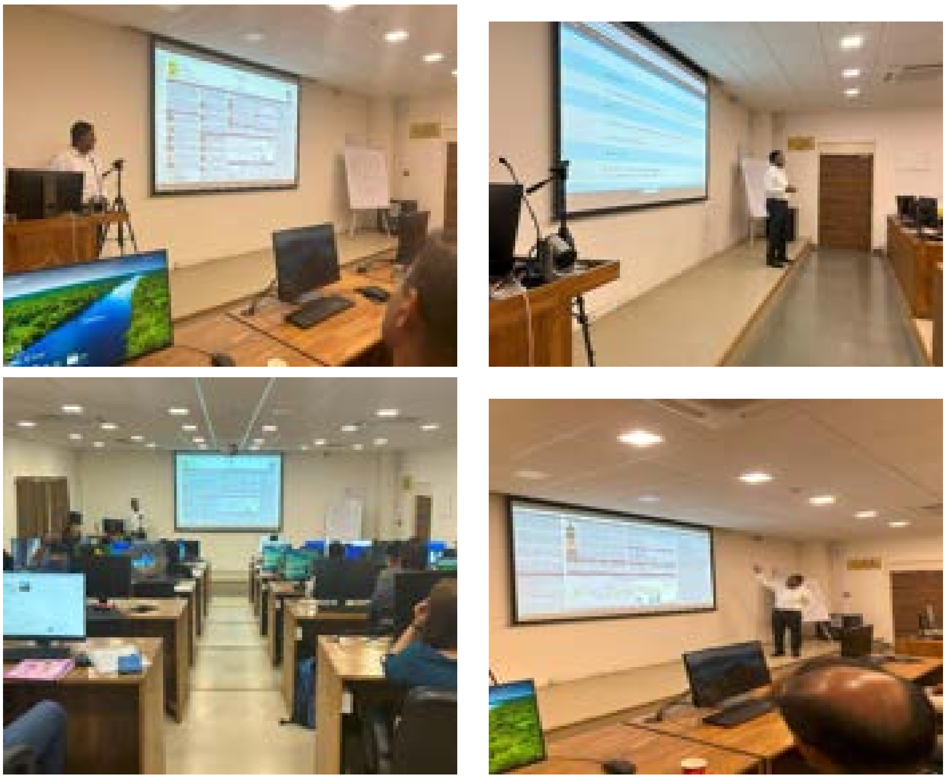


Figure 1: Captured moments from the publicity event

Further the poster of this grand event was sent to about 200 companies. In addition, hard copies of the poster were distributed during ICASSP 2025, a flagship event by IEEE SPS.

Figure 2: Prof. Patil has done wide publicity of S4P 2025 via poster distribution to the former ISCA President, Prof. (Dr.) Tanja Schultz (University of Bremen, Germany)

10 ISCA Supported Summer School on Speech Signal Processing

This event witnessed **81** participants and **25** invited speakers.

Figure 3: Group photographs of the participants and speakers of S4P 2025

11 Participation in S4P 2025

Around **81** participants from India and **25** speakers across the world have attended S4P 2025. The brief detail of the participants is shown in Figure 4. The participants represented 31 institutes/colleges/universities across India. The distribution of the participants by state-wise is given in Figure 2.

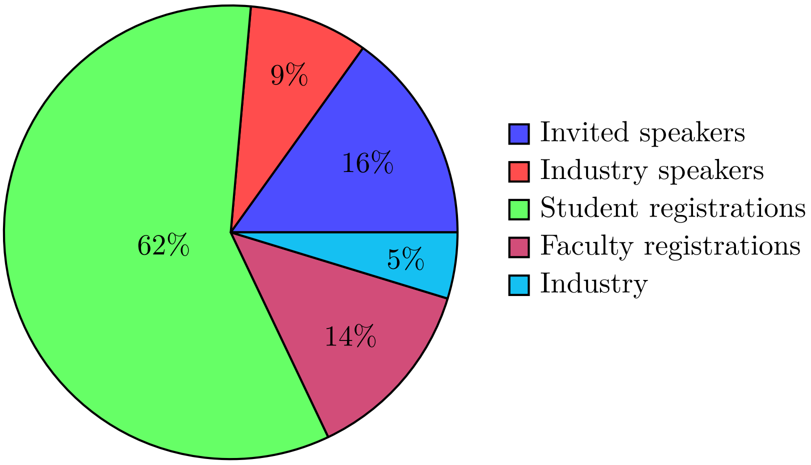


Figure 4: Participation in the S4P 2025

State	No. of Participants	Institutes / Universities Represented
Gujarat	35+	DAU (Gandhinagar), SCET (Surat), Sarvajanik College of Engg. & Tech (Surat), LD College (Ahmedabad), Nirma University (Ahmedabad), PDEU (Gandhinagar), Adani University, VGEC, GEC Gandhinagar, MSU Vadodara
Karnataka	10+	IIT Dharwad, IIIT Dharwad, PES University, RV College, BMSCE Bengaluru
Maharashtra	6+	Pimpri Chinchwad College (Pune), MET BKC (Nashik), COEP Pune
Telangana	5+	IIIT Hyderabad, KL University Hyderabad, Vasavi College of Engg.
Industry / Others	5+	TCS (Mumbai, Bengaluru), Sprinklr, Research Labs
Andhra Pradesh	3+	KL University Vaddeswaram, Koneru Lakshmaiah Education Foundation (KLEF), VIT-AP Amaravati
Odisha	3+	KIIT Bhubaneswar, CET Bhubaneswar
Tamil Nadu	3+	Amrita Vishwa Vidyapeetham (Coimbatore), PSG Tech Coimbatore
Delhi (NCR)	2+	IIIT Delhi, NSUT Delhi
Assam	2+	IIT Guwahati, Tezpur University
Rajasthan	2+	Jodhpur Institute of Engg. & Tech, LNMIIT Jaipur
Kerala	2+	SCTIMST (Trivandrum), NIT Calicut
Goa	1	Goa College of Engineering
Himachal Pradesh	1	IIT Mandi
West Bengal	1	Jadavpur University (Kolkata)

Table 5: Statewise distribution of participants in S4P 2025

12 Student Travel and Participation Grants

12.1 TCS Travel Grants

A total of **15 students** were announced under the TCS Travel Grants, allocated on a *first-come, first-served* basis. Each selected student was eligible for travel support of up to INR 1,000. The procedure required students to first pay the full registration fee; upon successful attendance at the Summer School, the eligible amount was reimbursed to 05 students.

12.2 IndSCA Travel Grants

Similarly, **15 students** received the IndSCA Travel Grant, also awarded on a *first-come, first-served* basis. The grant provided travel support up to INR 1,000. As with the TCS Travel Grant, reimbursement was processed after the Summer School for those students who attended.

#	Name	Institute
1	Devarakonda S. Charan	IIIT Hyderabad
2	Ravi Sastry Kolluru	IIIT Hyderabad
3	Pasupuleti Venkatesh	IIT Dharwad
4	Sujit Pangeni	KIIT University
5	Shruthi BS	IIIT Dharwad

Table 6: Awardees of the TCS Grants (B.Tech and M.Tech Students)

#	Name	Institute
1	Suraj B Madagaonkar	IIT Dharwad
2	Narottam Bhattacharjee	Tezpur University
3	Sujeet Kumar	IIT Dharwad
4	Akansha Tyagi	IIT Mandi
5	Shraddha Revankar	IIIT Dharwad
6	Olivia Babi	IIIT Dharwad

Table 7: Awardees of the IndISCA Travel Grants (Ph.D. Students)

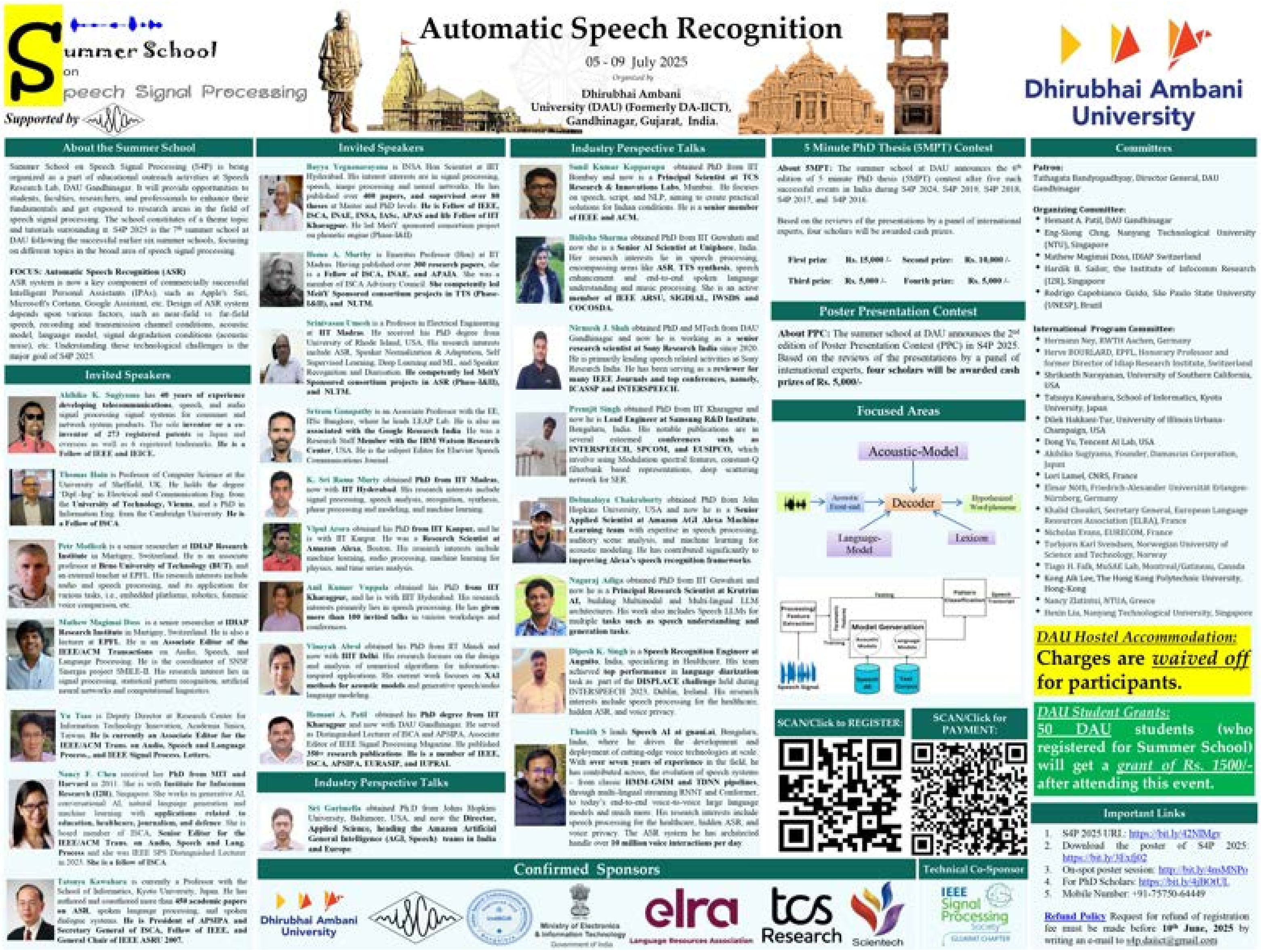
12.3 DAU Student Grants

To encourage strong participation from the host institution, **50 DAU Student Grants** were introduced in order to support students by DAU (as part of its sponsorship). Each DAU student participant was provided a grant of INR 1,500, allocated on a *first-come, first-served* basis. Eligibility was restricted to students officially enrolled at DAU at the time of registration. The reimbursement was processed post-event, conditional upon attendance.

#	Name	Course
1	Ravindrakumar M. Purohit	Ph.D
2	Priya Damdar	Ph.D
3	Vishal Kumar Yadav	Ph.D
4	Dharmendra Harish Vaghera	M.Tech
5	Vishnu Vardhan G V S	M.Tech
6	Daksh Arvindbhai Patel	M.Tech
7	Manish Manojkumar Prajapati	M.Tech
8	Sahil	M.Tech
9	Nischay	M.Tech
10	Aniket Pandey	M.Tech
11	Saroj Kumar	M.Sc
12	Siddharth Kumar	M.Sc
13	Kaustubh Wade	M.Sc
14	Devansh Modi	B.Tech
15	Rudra Bhatt	B.Tech
16	Hari Sharma	B.Tech
17	Ayush Patel	B.Tech
18	Arth Shah	B.Tech
19	Hiya	B.Tech
20	Nisarg Trivedi	B.Tech

Table 8: Awardees of the DAU Grants (B.Tech, M.Tech, M.Sc, and Ph.D Students)

13 Poster of the S4P 2025



Summer School on Speech Signal Processing

Supported by

Automatic Speech Recognition

05 - 09 July 2025

Organized by

Dhirubhai Ambani University (DAU) (Formerly DA-IICT), Gandhinagar, Gujarat, India.

Dhirubhai Ambani University

About the Summer School

Summer School on Speech Signal Processing (S4P) is being organized as a part of educational outreach activities at Speech Research Lab, DAU Gandhinagar. It will provide opportunities to students, faculties, researchers, and professionals to enhance their fundamentals and get exposed to research areas in the field of speech signal processing. The school constitutes of a theme topic and tutorials surrounding it. S4P 2025 is the 7th summer school at DAU following the successful earlier six summer schools, focusing on different topics in the broad area of speech signal processing.

FOCUS: Automatic Speech Recognition (ASR)
ASR system is now a key component of commercially successful Intelligent Personal Assistants (IPAs), such as Apple's Siri, Microsoft's Cortana, Google Assistant, etc. Design of ASR systems depends upon various factors, such as near-field vs. far-field speech, recording and transmission channel conditions, acoustic model, language model, signal degradation conditions (acoustic noise), etc. Understanding these technological challenges is the major goal of S4P 2025.

Invited Speakers

Akihiko K. Sugiyama has **40 years of experience developing telecommunications**, speech, and audio signal processing signal systems for consumer and network system products. The sole **inventor or a co-inventor of 273 registered patents** in Japan and overseas as well as 6 registered trademarks. **He is a Fellow of IEEE and IEICE.**

Thomas Hain is Professor of Computer Science at the University of Sheffield, UK. He holds the degree 'Dipl.-Ing' in Electrical and Communication Eng. from the **University of Technology, Vienna**, and a PhD in Information Eng. from the Cambridge University. **He is a Fellow of ISCA.**

Petr Motlicek is a senior researcher at **IDIAP Research Institute** in Martigny, Switzerland. He is an associate professor at **Brno University of Technology (BUT)**, and an external teacher at EPFL. His research interests include audio and speech processing, and its application for various tasks, i.e., embedded platforms, robotics, forensic voice comparison, etc.

Mathew Magimai Doss is a senior researcher at **IDIAP Research Institute** in Martigny, Switzerland. He is also a lecturer at **EPFL**. He is an **Associate Editor of the IEEE/ACM Transactions** on Audio, Speech, and Language Processing. He is the coordinator of SNSF Sinergia project SMILE-II. His research interest lies in signal processing, statistical pattern recognition, artificial neural networks and computational linguistics.

Yu Tsao is Deputy Director at Research Center for Information Technology Innovation, Academia Sinica, Taiwan. **He is currently an Associate Editor for the IEEE/ACM Trans. on Audio, Speech and Language Process., and IEEE Signal Process. Letters.**

Nancy F. Chen received her **PhD from MIT and Harvard** in 2011. She is with **Institute for Infocomm Research (I2R)**, Singapore. She works in generative AI, conversational AI, natural language generation and machine learning with **applications related to education, healthcare, journalism, and defence**. She is board member of ISCA, **Senior Editor for the IEEE/ACM Trans. on Audio, Speech and Lang. Process** and she was IEEE SPS Distinguished Lecturer in 2023. **She is a fellow of ISCA.**

Tatsuya Kawahara is currently a Professor with the School of Informatics, Kyoto University, Japan. He has authored and coauthored more than **450 academic papers on ASR**, spoken language processing, and spoken dialogue systems. **He is President of APSIPA and Secretary General of ISCA, Fellow of IEEE, and General Chair of IEEE ASRU 2007.**

Invited Speakers

Bayya Yegnanarayana is INSA Hon Scientist at IIIT Hyderabad. His interest interests are in signal processing, speech, image processing and neural networks. He has published over **400 papers, and supervised over 80 theses** at Master and PhD levels. **He is Fellow of IEEE, ISCA, INAE, INSA, IASc, APAS and life Fellow of IIT Kharagpur.** He led MeitY sponsored consortium project on phonetic engine (Phase-I&II).

Hema A. Murthy is Emeritus Professor (Hon) at IIT Madras. Having published over **300 research papers**, she is a **Fellow of ISCA, INAE, and APAIA**. She was a member of ISCA Advisory Council. **She competently led MeitY Sponsored consortium projects in TTS (Phase-I&II), and NLTM.**

Srinivasan Umesh is a Professor in Electrical Engineering at **IIT Madras**. He received his PhD degree from University of Rhode Island, USA. His research interests include ASR, Speaker Normalization & Adaptation, Self Supervised Learning, Deep Learning and ML, and Speaker Recognition and Diarization. **He competently led MeitY Sponsored consortium projects in ASR (Phase-I&II), and NLTM.**

Sriram Ganapathy is an Associate Professor with the EE, IISc Bangalore, where he leads LEAP Lab. He is also an **associated with the Google Research India**. He was a Research Staff **Member with the IBM Watson Research Center**, USA. He is the subject Editor for Elsevier Speech Communications Journal.

K. Sri Rama Murty obtained **PhD from IIT Madras**, now with **IIT Hyderabad**. His research interests include signal processing, speech analysis, recognition, synthesis, phase processing and modeling, and machine learning.

Vipul Arora obtained his PhD **from IIT Kanpur**, and he is with IIT Kanpur. He was a **Research Scientist at Amazon Alexa**, Boston. His research interests include machine learning, audio processing, machine learning for physics, and time series analysis.

Anil Kumar Vuppala obtained his PhD **from IIT Kharagpur**, and he is with IIIT Hyderabad. His research interests primarily lies in speech processing. He has **given more than 100 invited talks** in various workshops and conferences.

Vinayak Abrol obtained his PhD from IIT Mandi and now with **IIIT Delhi**. His research focuses on the design and analysis of numerical algorithms for information-inspired applications. His current work focuses on **XAI methods for acoustic models** and generative speech/audio language modeling.

Hemant A. Patil obtained his **PhD degree from IIT Kharagpur** and now with DAU Gandhinagar. He served as Distinguished Lecturer of ISCA and APSIPA, Associate Editor of IEEE Signal Processing Magazine. He published **350+ research publications. He is a member of IEEE, ISCA, APSIPA, EURASIP, and IUPRAI.**

Industry Perspective Talks

Sri Garimella obtained Ph.D from Johns Hopkins University, Baltimore, USA, and now the **Director, Applied Science, heading the Amazon Artificial General Intelligence (AGI, Speech) teams in India and Europe.**

Industry Perspective Talks

Sunil Kumar Kopparapu obtained PhD from IIT Bombay and now is a **Principal Scientist at TCS Research & Innovations Labs**, Mumbai. He focuses on speech, script, and NLP, aiming to create practical solutions for Indian conditions. He is a **senior member of IEEE and ACM.**

Bidisha Sharma obtained PhD from IIT Guwahati and now she is a **Senior AI Scientist at Uniphore**, India. Her research interests lie in speech processing, encompassing areas like **ASR, TTS synthesis**, speech enhancement and end-to-end spoken language understanding and music processing. She is an active **member of IEEE ARSU, SIGDIAL, IWSDS and COCOSDA.**

Nirmesh J. Shah obtained PhD and MTech from DAU Gandhinagar and now he is working as a **senior research scientist at Sony Research India** since 2020. He is primarily leading speech related activities at Sony Research India. He has been serving as a **reviewer for many IEEE Journals and top conferences, namely, ICASSP and INTERSPEECH.**

Premjit Singh obtained PhD from IIT Kharagpur and now he is **Lead Engineer at Samsung R&D Institute**, Bengaluru, India. His notable publications are in several esteemed **conferences such as INTERSPEECH, SPCOM, and EUSIPCO**, which involve using Modulation spectral features, constant-Q filterbank based representations, deep scattering network for SER.

Debmalya Chakrabarty obtained PhD from John Hopkins University, USA and now he is a **Senior Applied Scientist at Amazon AGI Alexa Machine Learning team** with expertise in speech processing, auditory scene analysis, and machine learning for acoustic modeling. He has contributed significantly to **improving Alexa's speech recognition frameworks.**

Nagaraj Adiga obtained PhD from IIT Guwahati and now he is a **Principal Research Scientist at Krutrim AI**, building Multimodal and Multi-lingual LLM architectures. His work also includes Speech LLMs for multiple **tasks such as speech understanding and generation tasks.**

Dipesh K. Singh is a **Speech Recognition Engineer at Augnito**, India, specializing in Healthcare. His team achieved **top performance** in **language diarization** task as part of the **DISPLACE challenge** held during INTERSPEECH 2023, Dublin, Ireland. His research interests include speech processing for the healthcare, hidden ASR, and voice privacy.

Thejith S leads **Speech AI at gnani.ai**, Bengaluru, India, where he drives the development and deployment of cutting-edge voice technologies at scale. With **over seven years of experience** in the field, he has contributed across, the evolution of speech systems – from classic **HMM-GMM and TDNN pipelines**, through multi-lingual streaming RNNT and Conformer, to today's end-to-end voice-to-voice large language models and much more. His research interests include speech processing for the healthcare, hidden ASR, and voice privacy. The ASR system he has architected handle over **10 million voice interactions per day**

5 Minute PhD Thesis (5MPT) Contest

About 5MPT: The summer school at DAU announces the 6th edition of 5 minute PhD thesis (5MPT) contest after five such successful events in India during S4P 2024, S4P 2019, S4P 2018, S4P 2017, and S4P 2016.

Based on the reviews of the presentations by a panel of international experts, four scholars will be awarded cash prizes.

First prize:	**Rs. 15,000 /-**	**Second prize:**	**Rs. 10,000 /-**
Third prize:	**Rs. 5,000 /-**	**Fourth prize:**	**Rs. 5,000 /-**

Poster Presentation Contest

About PPC: The summer school at DAU announces the 2nd edition of Poster Presentation Contest (PPC) in S4P 2025. Based on the reviews of the presentations by a panel of international experts, **four scholars will be awarded cash prizes of Rs. 5,000/-**

Focused Areas

SCAN/Click to REGISTER:

SCAN/Click for PAYMENT:

Committees

Patron:
Tathagata Bandyopadhyay, Director General, DAU Gandhinagar

Organizing Committee:
- Hemant A. Patil, DAU Gandhinagar
- Eng-Siong Chng, Nanyang Technological University (NTU), Singapore
- Mathew Magimai Doss, IDIAP Switzerland
- Hardik B. Sailor, the Institute of Infocomm Research (I2R), Singapore
- Rodrigo Capobianco Guido, São Paulo State University (UNESP), Brazil

International Program Committee:
- Hermann Ney, RWTH Aachen, Germany
- Herve BOURLARD, EPFL, Honorary Professor and former Director of Idiap Research Institute, Switzerland
- Shrikanth Narayanan, University of Southern California, USA
- Tatsuya Kawahara, School of Informatics, Kyoto University, Japan
- Dilek Hakkani-Tur, University of Illinois Urbana-Champaign, USA
- Dong Yu, Tencent AI Lab, USA
- Akihiko Sugiyama, Founder, Damascus Corporation, Japan
- Lori Lamel, CNRS, France
- Elmar Nöth, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
- Khalid Choukri, Secretary General, European Language Resources Association (ELRA), France
- Nicholas Evans, EURECOM, France
- Torbjørn Karl Svendsen, Norwegian University of Science and Technology, Norway
- Tiago H. Falk, MuSAE Lab, Montreal/Gatineau, Canada
- Kong Aik Lee, The Hong Kong Polytechnic University, Hong-Kong
- Nancy Zlatintsi, NTUA, Greece
- Haizhou Li, Nanyang Technological University, Singapore

DAU Hostel Accommodation:
Charges are *waived off* for participants.

DAU Student Grants:
50 DAU students (who registered for Summer School) will get a *grant of Rs. 1500/-* after attending this event.

Important Links

1. S4P 2025 URL: https://bit.ly/42NlMgr
2. Download the poster of S4P 2025: https://bit.ly/3Exj0Z
3. On-spot poster session: http://bit.ly/4nsMNPo
4. For PhD Scholars: https://bit.ly/4jBOOUL
5. Mobile Number: +91-75750-64449

Refund Policy Request for refund of registration fee must be made before **10th June, 2025** by writing an e-mail to s4p.daiict@gmail.com

Confirmed Sponsors

Dhirubhai Ambani University

Ministry of Electronics & Information Technology, Government of India

elra Language Resources Association

tcs Research

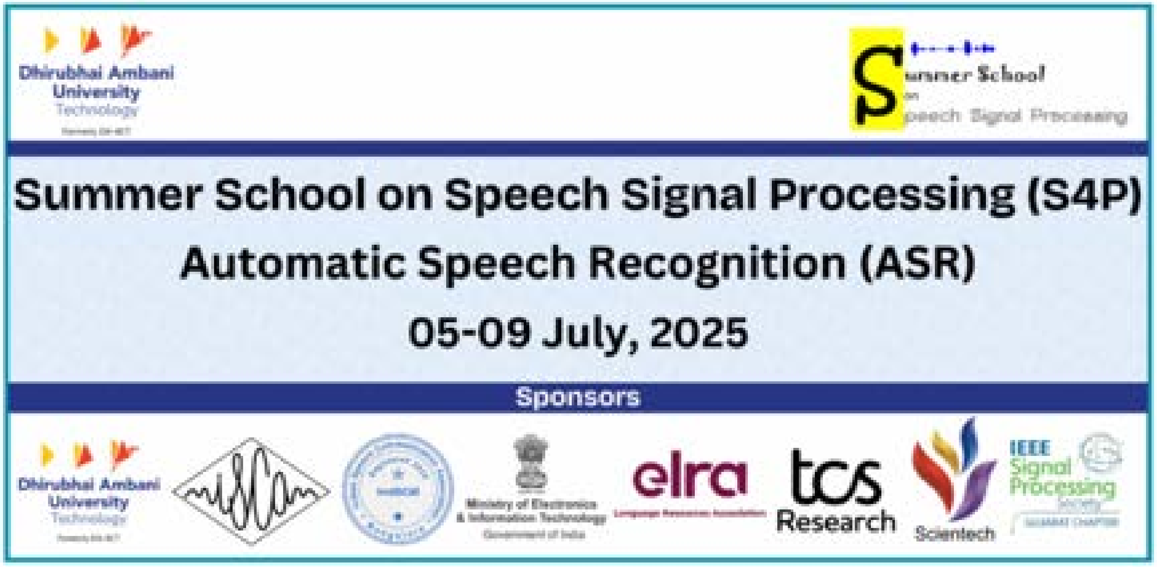
Scientech

Technical Co-Sponsor

IEEE Signal Processing Society, GUJARAT CHAPTER

Figure 5: Official Poster of S4P 2025.

14 Banner and Proceedings of the S4P 2025



Summer School on Speech Signal Processing (S4P)

Automatic Speech Recognition (ASR)

05-09 July, 2025

Sponsors

Dhirubhai Ambani University Technology
Summer School on Speech Signal Processing
Ministry of Electronics & Information Technology Government of India
elra
tcs Research
Scientech
IEEE Signal Processing Society

Figure 6: Official Banner of S4P 2025.

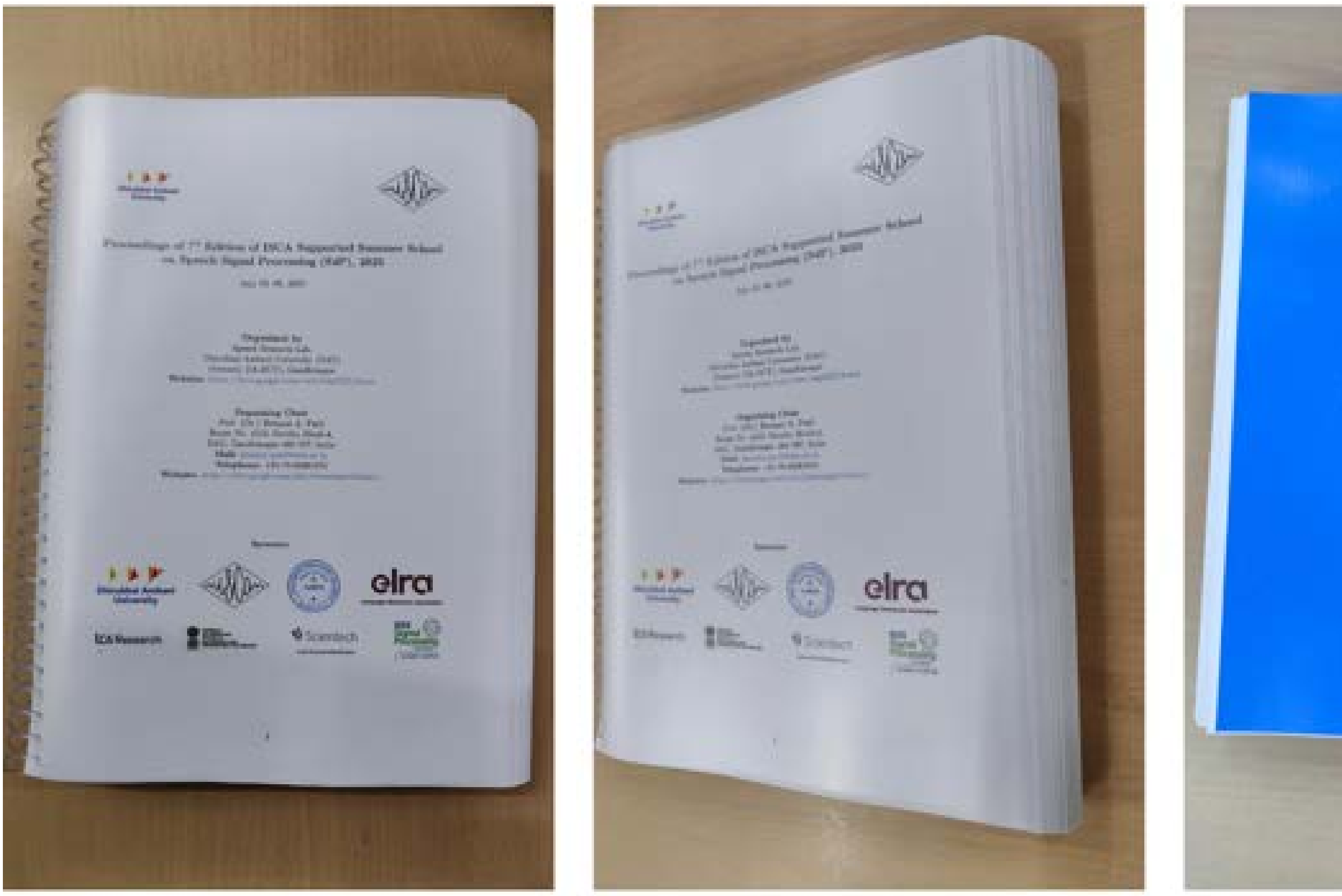
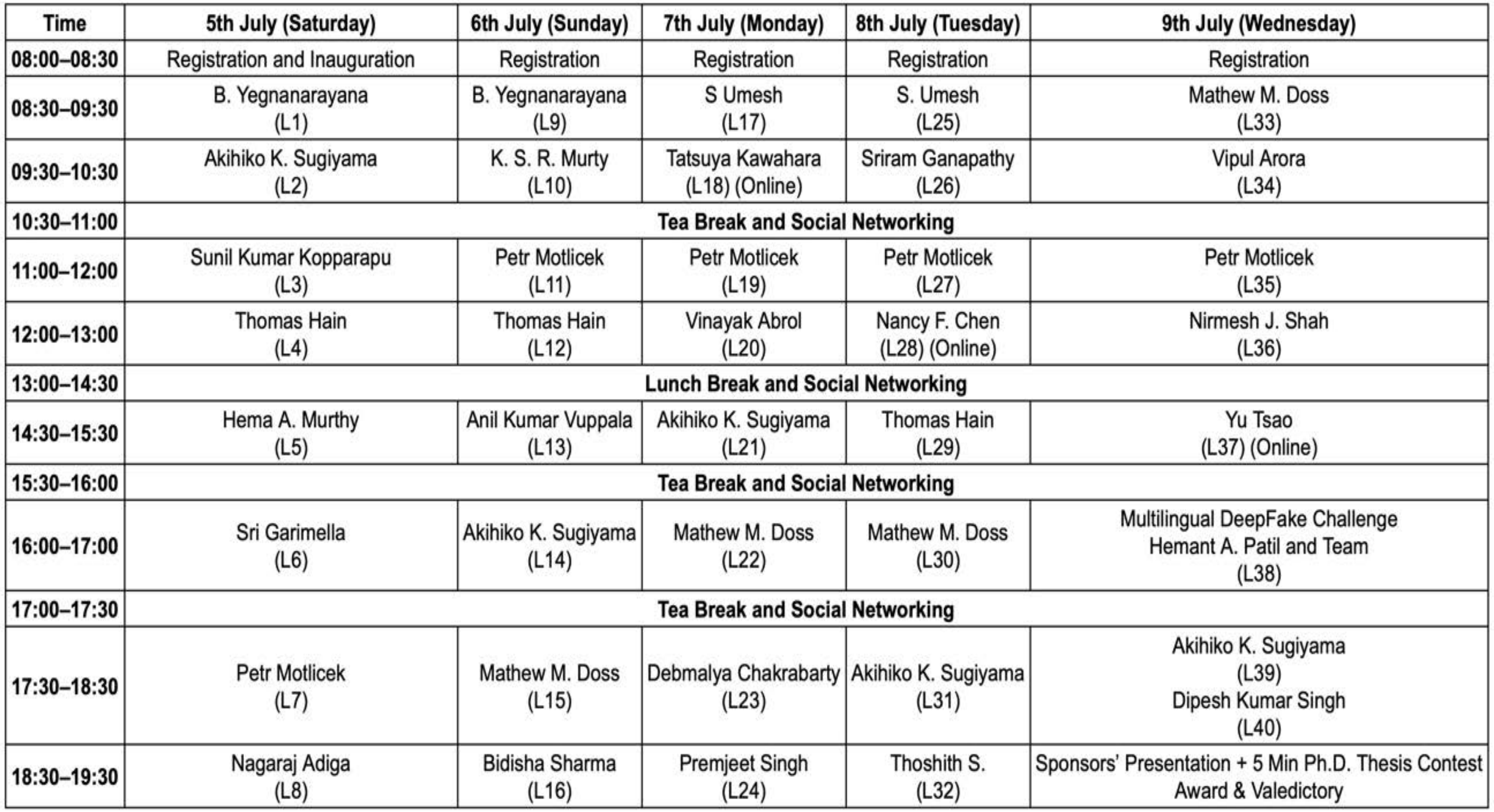


Figure 7: Proceedings of the S4P 2025.

15 Technical Program Schedule



Time	5th July (Saturday)	6th July (Sunday)	7th July (Monday)	8th July (Tuesday)	9th July (Wednesday)
08:00–08:30	Registration and Inauguration	Registration	Registration	Registration	Registration
08:30–09:30	B. Yegnanarayana (L1)	B. Yegnanarayana (L9)	S Umesh (L17)	S. Umesh (L25)	Mathew M. Doss (L33)
09:30–10:30	Akihiko K. Sugiyama (L2)	K. S. R. Murty (L10)	Tatsuya Kawahara (L18) (Online)	Sriram Ganapathy (L26)	Vipul Arora (L34)
10:30–11:00	Tea Break and Social Networking				
11:00–12:00	Sunil Kumar Kopparapu (L3)	Petr Motlicek (L11)	Petr Motlicek (L19)	Petr Motlicek (L27)	Petr Motlicek (L35)
12:00–13:00	Thomas Hain (L4)	Thomas Hain (L12)	Vinayak Abrol (L20)	Nancy F. Chen (L28) (Online)	Nirmesh J. Shah (L36)
13:00–14:30	Lunch Break and Social Networking				
14:30–15:30	Hema A. Murthy (L5)	Anil Kumar Vuppala (L13)	Akihiko K. Sugiyama (L21)	Thomas Hain (L29)	Yu Tsao (L37) (Online)
15:30–16:00	Tea Break and Social Networking				
16:00–17:00	Sri Garimella (L6)	Akihiko K. Sugiyama (L14)	Mathew M. Doss (L22)	Mathew M. Doss (L30)	Multilingual DeepFake Challenge Hemant A. Patil and Team (L38)
17:00–17:30	Tea Break and Social Networking				
17:30–18:30	Petr Motlicek (L7)	Mathew M. Doss (L15)	Debmalya Chakrabarty (L23)	Akihiko K. Sugiyama (L31)	Akihiko K. Sugiyama (L39) Dipesh Kumar Singh (L40)
18:30–19:30	Nagaraj Adiga (L8)	Bidisha Sharma (L16)	Premjeet Singh (L24)	Thoshith S. (L32)	Sponsors' Presentation + 5 Min Ph.D. Thesis Contest Award & Valedictory

Figure 8: Detailed Day-wise Technical Program of S4P 2025.

Label	Lecture Topic
L1	Challenges in Processing Natural Signals Like Speech
L2	Mechanical Noise Suppression: Debutant Of Phase In Signal Enhancement After 30 Years of Silence
L3	Audio & Speech Processing — What have we been doing?
L4	Selecting Data for Semi-Supervised ASR
L5	Signal Processing Guided Machine Learning In Various Domains
L6	Representation Learning of Speech and its Applications
L7	ASR - from input data to industrial applications
L8	Advances in Speech Large Language Models for Recognition and Translation
L9	Speech Signal Processing Using Single Frequency Filtering (SFF)
L10	Phase Processing of Speech Signals
L11	ASR - from HMM/GMMs to LLM-based engines
L12	Multilingual speech recognition - Modelling the relationship between languages
L13	ASR and SLT in Indian Languages
L14	Personal Information Devices: Portable To Wearable, Stand-alone To Connected, Players To Sensors
L15	From Spectral Feature Representations to Supervised Learning-Based Feature Representations
L16	Beyond Supervision: Leveraging Pseudo-Labels and LLMs for Domain-Specific ASR
L17	Recent Advances in ASR of Indian Languages
L18	Universal Speech Recognition Using IPA And Articulatory Features
L19	ASR - contextuallisation of ASR systems
L20	On Quantization of Neural Models for Speech Tasks
L21	Phase-Based Time-Frequency Filtering as an Alternative to the Classical Beamforming
L22	Self-Supervised Learning (SSL)-Based Representation Learning for Speech Processing
L23	Towards Universal Audio Understanding: A Unified Encoder for Speech and Audio Tasks
L24	Automatic Speech Recognition: From Problem to Research to Productisation
L25	Speech Foundation Models for ASR in Indian Languages
L26	Demystifying the Black Box: Explainability and Trust in Modern AI
L27	ASR - recognition of apriori unknown words, detection of rare word entities
L28	Multimodal, Multilingual Generative AI: From Multicultural Contextualization to Empathetic Reasoning
L29	Self-supervised Models for Robust Speech Content Representations
L30	From Spectral Feature Representations to Supervised Learning-Based Feature Representations
L31	Linear Microphone Array Parallel to the Driving Direction for In-Car Speech Enhancement
L32	Building Real-Time Voice-to-Voice LLMs: Toward Expressive, Multilingual, and Task-Aware AI
L33	Statistical Interpretation of the SSL-Based Representation Learning
L34	Towards Multilingual Speech Tokenization
L35	ASR - data selection and learning using weakly labeled data, performance monitoring
L36	Generative Modeling for Emotional Speech Synthesis: Progress, Pitfalls, and Possibilities
L37	Neural Speech Enhancement and Assessment and Their Applications in Assistive Oral Communication Technologies
L38	Multi-Lingual Audio DeepFake Detection Challenge
L39	IEEE Fellow Elevation: Keys to Success
L40	Next-Generation Speech Recognition: Scaling Self-Supervised Learning, Multimodal Fusion, and LLM-Augmented ASR for Real-World Deployment

Table 9: Lecture Program

16 Inauguration Ceremony and Welcome Address

The Summer School commenced with an Inauguration Ceremony that began with a prayer to the Almighty, followed by the traditional lighting of the lamp. The ceremony was graced by the presence of distinguished dignitaries: Prof. (Dr.) B. Yegnanarayana, Prof. (Dr.) Hema A. Murthy, Prof. (Dr.) Akihiko K. Sugiyama, Prof. (Dr.) Thomas Hain, Dr. Bhaskar Chaudhury (Dean AP).

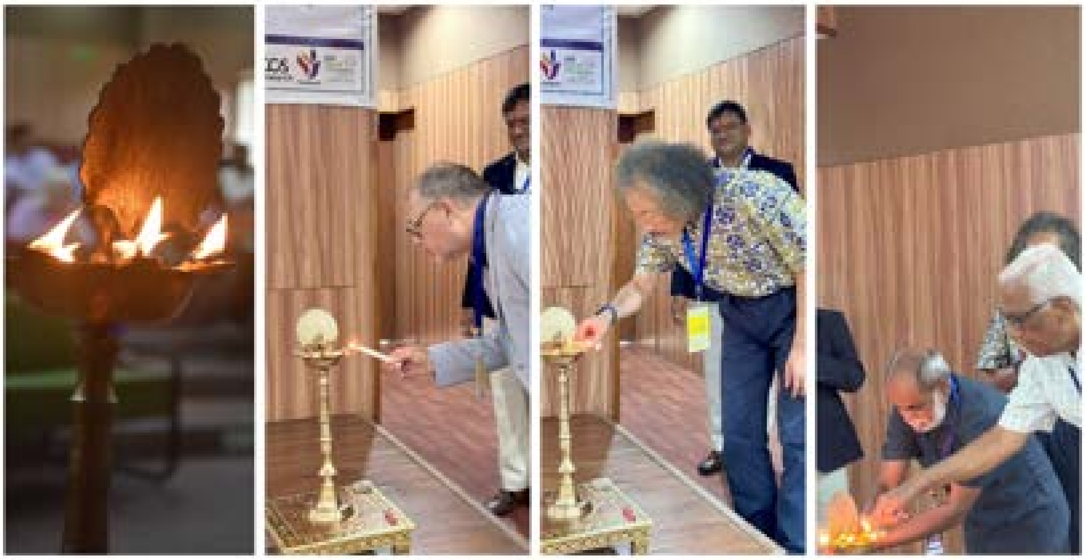


Figure 9: Inauguration ceremony of S4P 2025, where Prof. Thomas Hain, Prof. Akihiko Sugiyama, and Prof. B. Yegnanarayana formally commenced the event.

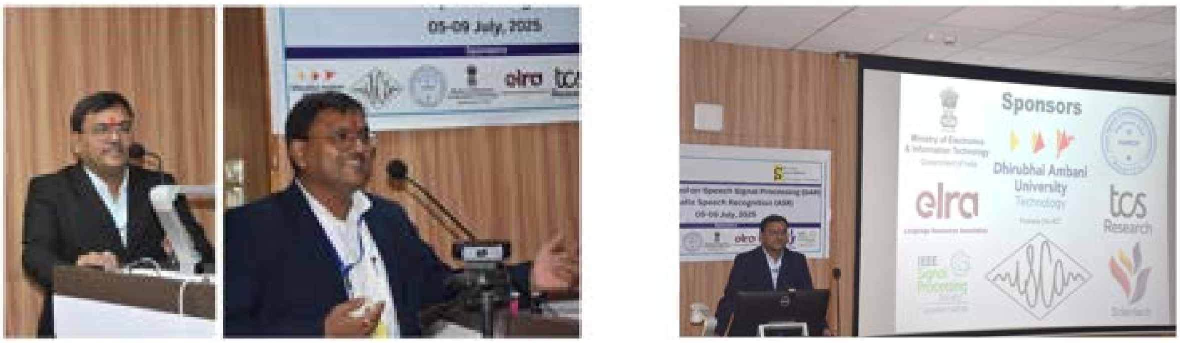


Figure 10: Inaugural and Welcome Address by Prof. Hemant A. Patil, Organizing Chair, S4P 2025

Figure 11: Appreciation to sponsors by Prof. Hemant A. Patil, Organizing Chair of S4P 2025.

17 Invited Speaker's Talks

17.1 Prof. (Dr.) Akihiko K. Sugiyama (IEEE Fellow)

Figure 12: Lecture by Prof. Akihiko Sugiyama, invited speaker at S4P 2025.

17.1.1 Talk 2: Mechanical Noise Suppression: Debutant of Phase In Signal Enhancement After 30 Years of Silence

Abstract: This talk presents challenges, solutions, and applications in commercial products of mechanical noise suppression. The topic has become more important as dissemination of consumer products that process environmental signals in addition to human speech. Three typical types of mechanical noise signals with small, medium, and large signal power, represented by feature phones and camcorders, digital cameras, and standard and tablet PCs, respectively, are covered. Mechanical noise suppression for small power signals is performed by continuous spectral template subtraction with a noise template dictionary. Medium power mechanical noise is suppressed in a similar manner only when its presence is notified by the parent system such as the digital camera. When the power is large, explicit detection of the mechanical noise based on phase information determines suppression timings. In the all three scenarios, the phase information of the input noisy signal is randomized for making the residual noise inaudible in frequency bins where noise is dominant. The phase has been unaltered in the past 30 years after Lim, thus, these suppression algorithms opened the door to a new signal enhancement era. Sound demonstrations before and after suppression highlight the effect of the algorithms and make the talk engaging

17.1.2 Talk 14: Personal Information Devices: Portable To Wearable, Stand-alone To Connected, Players To Sensors

Abstract: This talk presents a history of personal information devices. The origin is an audio player dated back to the 1990s which was born at an intersection of audio coding algorithms to provide sufficient subjective audio quality and a sufficient memory size on a single chip. LSI technology was indispensable to its birth which had a revolutionary impact on the hardware business. The audio-only device was naturally extended to include video signals to cover multimedia applications commonly encountered today in our daily life. Integration with a mobile phone brought us continuous extensions to wearables, connected operations, and sensing functions.

17.1.3 Talk 21: Phase-Based Time-Frequency Filtering as an Alternative to the Classical Beamforming

Abstract: This talk presents phase-based time-frequency filtering as an alternative to the classical beamforming. The classical beamforming is decomposed into direction-of-arrival estimation and direction-based attenuation. This decomposition makes the design of directivity pattern free from the sensor arrangement, enabling a sharp beam with a small number of sensors. Audio beamforming for PC applications is presented as an example with a design technique for a constant beam-width across the frequency in multiple channels. Successful evaluation results confirm the constant beam-width design.

17.1.4 Talk 31: Linear Microphone Array Parallel to the Driving Direction for In-Car Speech Enhancement

Abstract: This talk presents a linear microphone array parallel to the driving direction for in-car speech enhancement. In contrast to other linear microphone arrays in the car cabin reported in a literature or implemented as a commercial product, the array axis is arranged in parallel to the driving direction. Thanks to the 90o-rotated array axis with the constraints on the microphone position specific to the car environment, a mirror image of the directivity toward the talker with respect to the array axis is no longer projected in the direction of interference but redirected to a direction with no interference. As a result, the talker speech can be discriminated from the interference by directivity, leading to good interference reduction with little speech distortion. Simulation results confirms this position.

17.1.5 Talk 39: IEEE Fellow Elevation: Keys to Success

Abstract: This talk presents how to prepare the nomination when one is nominated for IEEE Fellow. There are some key considerations to maximize the chances of success when a nomination is prepared. IEEE has a clear guideline to write an effective nomination which most nominators do not refer to. Nominations in line with the guide makes the nominator/nominee confident about the nomination and simultaneously makes the evaluators comfortable in the process of evaluation through easy understanding and comparison. The talk is based on the presenter's experiences as a member of the IEEE Fellow Committee to make the final decision and a Society Fellow Evaluation Committee to perform the initial evaluation as well as a nominator/reference/endorser. The considerations in the talk such as items to be included, descriptions of the accomplishments, and the order of presentation are useful for other occasions when one would like to appeal accomplishments for Senior-Member elevation, award nomination, and promotion in the affiliation.

17.2 Prof. (Dr.) Thomas Hain (ISCA Fellow)

Figure 13: Lecture by Prof. Thomas Hain, invited speaker at S4P 2025.

17.2.1 Talk 4: Selecting Data for Semi-Supervised ASR

Abstract: Training of ASR models has long followed the path of multi-style training, i.e. more diverse data is better data. Labelled data is still hard to come by, hence semi-supervised training is often used. In contrast the amounts of unlabeled data available now can be vast - and the question of data selection may be important again. In this talk we briefly review standard strategies for semi-supervised training and data selection. We then move on to present recent work on data selection using new methods for word error rate estimation and present results on ASR training.

17.2.2 Talk 12: Multilingual speech recognition - Modelling the relationship between languages

Abstract: Multilingual speech recognition is now commonly used in ASR systems such as Whisper or foundation models such as XLSR. Models are simply trained using data from many languages, and possibly joint tokenization of different writing scripts. In this talk we briefly review the history of modelling followed by recent work on trying to understand the relationship between languages with the aim to make progress towards representing the 7000 languages of the world, where most are under-resourced. Simple mapping models are shown to better understand and model relationships between languages and thus allow better generalisation to new low resource languages.

17.2.3 Talk 29: Self-supervised Models for Robust Speech Content Representations

Abstract: Self supervised models have revolutionised language and speech processing. The fact that unlabelled data can be used to inform and bootstrap models for a vast range of tasks has given rise to a completely different view of speech technology. However even though most models are trained on large amounts of data, domain generalisation can be poor. Model training is very typically very costly and fine-tuning a task may not lead to good results because of domain mismatches. In this presentation some properties of SSL derived methods were explored, leading to novel ways to fine tune models to a domain and content oriented task types. Instead of using model specific loss functions generic alignment loss allows for fast fine tuning with much lower computational cost.

17.3 Dr. Petr Motlicek

Figure 14: Lecture by Dr. Petr Motlicek, invited speaker at S4P 2025

17.3.1 Talk 7: ASR - from input data to industrial applications

Abstract: This lecture examines the key challenges involved in deploying Automated Speech Recognition (ASR) systems, covering the entire pipeline from data collection to model training, execution and iterative assessment. It will address critical considerations such as data transcription strategies, commonly used tools in both research and industry, data privacy concerns, licensing of available resources, and the trade-offs between offline and streaming solutions. The lecture will also explore ASR performance metrics and discuss scalability challenges in real-world applications.

17.3.2 Talk 11: ASR - from HMM/GMMs to LLM-based engines

Abstract: This lecture offers a high-level overview of the key challenges in achieving high-accuracy Automatic Speech Recognition (ASR) systems. It will begin by introducing foundational ASR concepts widely used in recent decades, such as HMM/GMM-based approaches, and progress toward the latest advancements involving the integration of ASR with large language models (LLMs). The session will also highlight recent developments in speech pre-processing, including voice activity detection, handling multi-speaker scenarios, speaker diarization or mapping of the ASR output with additional information available for given use-case. Finally, the lecture will showcase specific industrial applications where ASR technologies play a critical role.

17.3.3 Talk 19: ASR - contextuallisation of ASR systems

Abstract: This lecture will focus on current approaches for contextualizing ASR output using prior knowledge. ASR systems are often tailored for specific applications where auxiliary data—containing relevant context or domain-specific information—is available to enhance recognition accuracy. The session will conclude with a demonstration of how ASR can be integrated into real-world applications, with a particular focus on air traffic management.

17.3.4 Talk 27: ASR - recognition of apriori unknown words, detection of rare word entities

Abstract: This lecture will address a common requirement from ASR users: how to incorporate new words or named entities into ASR output without retraining the entire system. These terms are typically not present in the original training data and were unknown during the initial training phase. The lecture will explore earlier methods used in traditional hybrid ASR systems, as well as more recent techniques developed for end-to-end architectures. The primary goal is to improve recognition accuracy, particularly for rare or out-of-vocabulary words.

17.3.5 Talk 35: ASR - data selection and learning using weakly labeled data, performance monitoring

Abstract: This lecture will explore methods for iteratively training ASR systems using data drawn from large, readily available sources. While the volume of data is typically not a limiting factor, its quality can vary significantly and it in most cases lacks manual annotations. The lecture will cover strategies for effective data selection, techniques for iterative learning that mitigate catastrophic forgetting, and approaches for training with weakly labeled or noisy data. Eventually the lecture will also consider performance monitoring, including generation of reliable confidence scores as relevant ASR output.

17.4 Dr. Mathew Magimai Doss

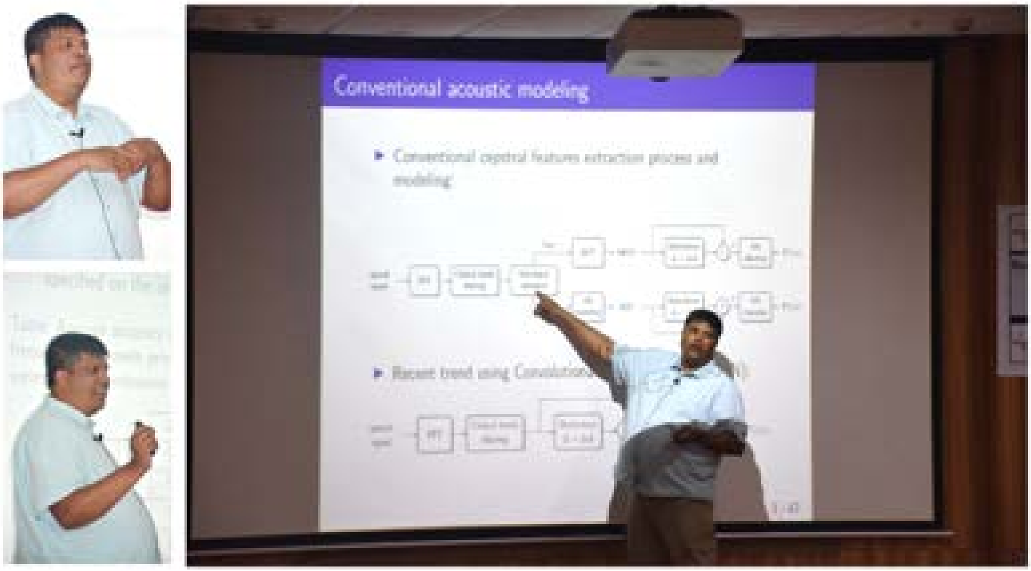

Conventional acoustic modeling
Conventional cepstral features extraction process and modeling
Recent trend using Convolutional

Figure 15: Lecture by Dr. Mathew Magimai Doss, invited speaker at S4P 2025.

17.4.1 Talk 15: From spectral feature representations to supervised learning-based feature representations

Abstract: In the first part, I will start with an overview of extraction of features using signal processing techniques and their modeling by different distribution modeling methods for automatic speech recognition, and show how this led to different supervised learning based feature representations such as, tandem feature and auto-association/auto-encoder features.

17.4.2 Talk 22: End-to-end acoustic modeling

Abstract: In the second part, I will present an end-to-end acoustic modeling method pioneered at Idiap, where raw waveform is directly modeled by neural network in a task dependent manner. I will provide links to conventional signal processing techniques and show how these kind of neural networks can be analyzed to gain insight into the information captured by them.

17.4.3 Talk 30: Self-supervised learning (SSL) based representation learning for speech processing

Abstract: In this part, I will start with an overview of self-supervised learning based feature representation learning methods. I will then present recent works at Idiap on self supervised feature representation based speech synthesis and voice conversion to demonstrate how this leads to new directions where speech synthesis and speech recognition/assessment can be put in loop. Specifically, the talk will focus on (a) multispeaker speech synthesis, (b) unsupervised rhythm and voice conversion for improving dysarthric speech recognition and (c) children voice privacy.

17.4.4 Talk 33: Statistical interpretation of the SSL-based representation learning

Abstract: In this talk, I will present an on-going work at Idiap to show the link between classical approaches to model feature distributions and self-supervised learning based models. Through this link, I will provide a statistical interpretation of SSL models and show how different pre-trained models like wav2vec2, HuBERT, wavLM and Whisper can be analyzed and distinguished, and how the information learned by them could be interpreted. This talk will conclude by drawing parallels between past methods and current methods from statistical pattern recognition point of view and providing suggestions for future research.

17.5 Prof. (Dr.) Yu Tsao

Figure 16: View from the back of the auditorium during Prof. Yu Tsao's online talk.

17.5.1 Talk 37: Neural Speech Enhancement and Assessment and Their Applications in Assistive Oral Communication Technologies

Abstract: This presentation is divided into three parts. Firstly, we will discuss our recent advancements in neural speech enhancement (SE), a critical element in various speech-related applications. The primary objective of SE is to enhance speech signals by mitigating distortions caused by additive and convoluted noises, thereby improving human-human and human-machine communication efficacy. We'll delve into the system architecture and fundamental theories behind neural SE approaches, as well as explore important directions aimed at achieving better performance. Moving on to the second part, we will focus on our recent progress in neural speech assessment (SA), which aims to effectively evaluate the quality and intelligibility of spoken audio—a crucial aspect in numerous speech-related applications. Traditionally, the evaluation process often relies on listening tests involving human participants, which can be both resource-intensive and impractical due to the need for a large number of listeners. To address this challenge, neural SA metrics have garnered notable attention. We will discuss the fundamental systems of neural SA, highlight several factors influencing performance, and explore emerging trends in this domain. Finally, we will present some applications of neural SE and SA in assistive oral communication technologies. These applications include impaired speech transformation and noise reduction for assistive hearing and speaking devices. Through these discussions, our aim is to illustrate the potential impact of neural-based approaches in improving communication accessibility for individuals with oral communication disorders.

17.6 Dr. Nancy F. Chen (ISCA Fellow)

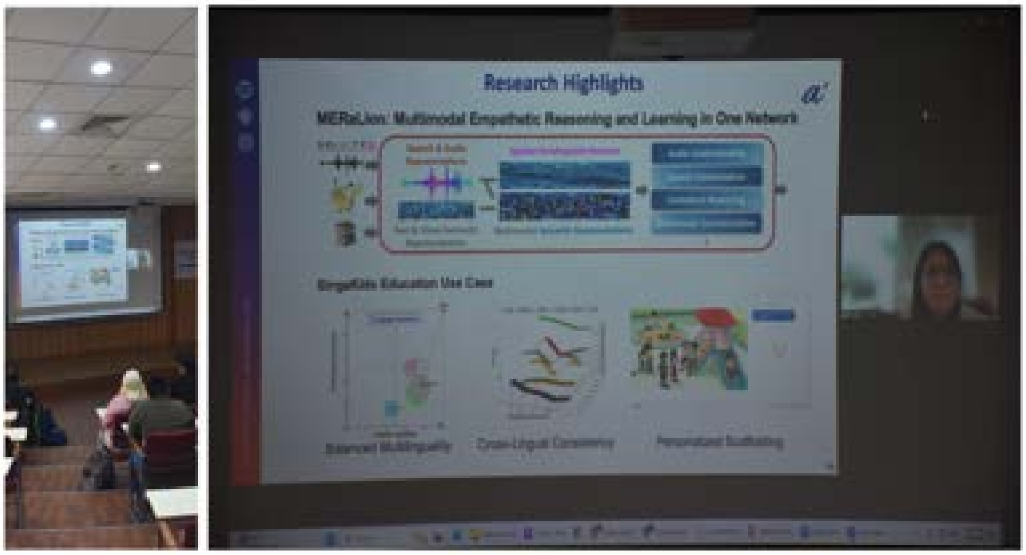

Research Highlights
MERaLion: Multimodal Empathetic Reasoning and Learning in One Network
SingaKids Education Use Case
Balanced Multilinguality
Cross-Lingual Consistency
Personalized Scaffolding

Figure 17: View from the back of the auditorium during Dr. Nancy F. Chen's online talk, along with a snapshot of her presentation.

17.6.1 Talk L28: Multimodal, Multilingual Generative AI: From Multicultural Contextualization to Empathetic Reasoning

Abstract: In this seminar, we will take a historical view on large language models from a speech technology lens and draw R&D examples from initiatives such as MeraLion (Multimodal Empathetic Reasoning and Learning In One Network), our generative AI efforts in Singapore's National Multimodal Large Language Model Programme. Speech and audio information is rich in providing more comprehensive understanding of spatial and temporal reasoning in addition to social dynamics that goes beyond semantics derived from text alone. Cultural nuances and multilingual peculiarities add another layer of complexity in understanding human interactions. In addition, we will draw use cases in education to highlight research endeavors, technology deployment experience and application opportunities.

17.7 Prof. (Dr.) Tatsuya Kawahara (APSIPA President, Fellow of IEEE, Secretary General of ISCA)

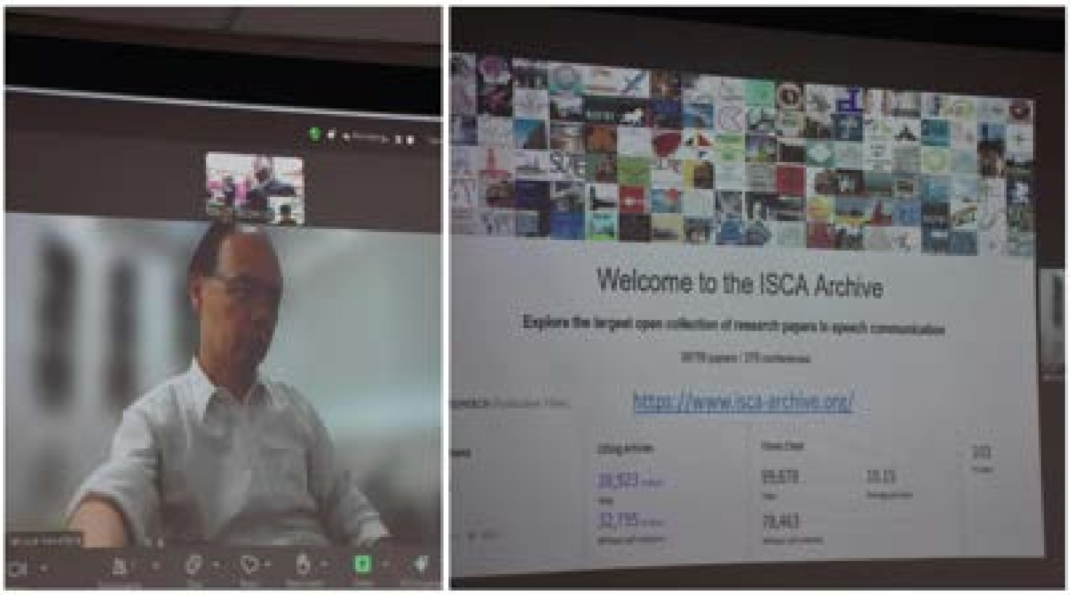

Welcome to the ISCA Archive
Explore the largest open collection of research papers in speech communication
https://www.isca-archive.org/

Figure 18: Prof. Tatsuya Kawahara (Fellow ISCA), presenting ISCA Publications and Archive, serving as ISCA Board Member.

17.7.1 Talk L18: Universal Speech Recognition Using IPA And Articulatory Features

Abstract: While the end-to-end framework has achieved remarkable advancements in automatic speech recognition (ASR), it is heavily optimized for the training dataset and lacks flexibility for multi-lingual ASR, particularly in low-resource languages. The problem is significantly mitigated by the SSL-pretrained models, such as X-LSR, but not completely solved. An alternative solution to universal ASR is to adopt language-independent tokens such as IPA (International Phonetic Alphabet). Since IPA is defined by the articulatory features, it is possible to incorporate the knowledge of articulatory features during the training. This talk addresses several approaches in this direction.

17.8 Prof. (Dr.) B. Yegnanarayana (Fellow of IEEE, ISCA, INAE, INSA, IASc, APAS)

recognition, audio search, enhancement, etc
NLP: Speech to text, and then text processing, after removing all naturalness
Speech
vs speech signal processing
(SSP)
The output y[n]
for an input x[n] is given by

Figure 19: Discussing phase and magnitude, fundamental concepts in signal processing by Prof. Bayya Yegnanarayana

17.8.1 Talk 1: Challenges in Processing Natural Signals Like Speech

17.8.2 Talk 9: Speech signal processing using single frequency filtering (SFF)

Abstract: Signal processing in general, and speech signal processing in particular, is normally performed using block processing methods, like discrete Fourier transform. Frame- based block processing of signals has some disadvantages, especially in processing the phase spectral component. Filtering-based methods can be explored as an alternative for processing speech signals. In this presentation, we will discuss single frequency filtering (SFF) method for speech signal processing, especially for extracting speech production information from the phase component. Starting with the basics of signals and systems for discrete time signals, this talk presents the main ideas of SFF that are useful in extracting the time-varying formants and pitch harmonics contours from speech signals. The results will be demonstrated for speech signals from different types of voices

17.9 Prof. (Dr.) Hema A. Murthy (Fellow of ISCA, INAI, APAIA)

Figure 20: Prof. (Dr.) Hema A. Murthy explaining the role of signal processing in preparing data for machine learning in speech, music, and EEG applications.

17.9.1 Talk 5: Signal Processing Guided Machine Learning in Various Domains

Abstract: The buzzword for building applications of practical relevance is “big data” today. This has led to a separate field called “Data Science” being offered by arious universities, and Institutes. The field of “data science” has grown to accommodate for the variability in the underlying statistical structure that exists natural signals. Both classical machine learning and deep learning rely on the availability of large amount of a wide variety of data. Deep learning models which are massive neural networks ultimately learn the underlying structure of data.

While Deep Learning has revolutionized machine learning, in this talk we focus on the use of signal processing to preprocess or mine existing data, so that accurate data is presented to machine learning models. Domain specific signal processing has the capability to identify events in a signal. The event itself may have varying statistical characteristics. Presentation of detected events to machine learning enables faster convergence, and a smaller data footprint. We draw examples from Speech, Music and EEG signals to show that “signal processing and machine learning” must work together to build systems of relevance for a given domain.

17.10 Prof. (Dr.) S. Umesh

Figure 21: Prof. S. Umesh presented an insightful talk on recent advances in ASR and Speech Foundation Models for Indian languages.

17.10.1 Talk 17: Recent Advances in ASR of Indian Languages

Abstract: In this talk, I will give an overview of the current work on ASR in Indian Languages at SPRING LAB, IIT Madras. I will give an overview of the three broad architectures including encoder-decoder, CTC and transducer based approaches to ASR. This will be followed by details of our efforts to collect speech data and build ASR models in various Indian languages. All of our models and data are available in open source, and I will give a demo (https://asr.iitm.ac.in/demo/home) of the ASR systems as well as speech-2-speech translation system by pipelining our ASR and MT systems with a TTS.

17.10.2 Talk 25: Speech Foundation Models for ASR in Indian languages

Abstract: In this talk, I will give an overview of Speech Foundation Models. While the motivation for these models come from text language models, unlike text, the discretisation of speech signal is not straightforward. I will start with contrastive predictive coding ideas, followed by some popular models like wav2vec2.0 and HuBERT. This will be followed by details of recent work from SPRING LAB, where we have proposed two speech foundation models – ccc-wav2vec2.0 and data2vec-aqc. These models have done exceedingly well in SUPERB challenge and also in a study that did a large scale evaluation of Speech Foundation Models (Yang et. Al IEEE TASLP vol.32, pg. 2884-2899). We are particularly excited since these were just built on 960-hours of data, and yet were competing with bigger models built on 60,000 or 94K hour models. Motivated by the success on American English, we have pretrained ccc-wav2vec2.0 and data2vec-aqc models based on 30,000 hours of Indian languages. These models when fine-tuned for ASR tasks give state-of-art performance for Indian languages. I will wrap the talk with demos of systems developed at SPRING LAB.

17.11 Prof. (Dr.) Sriram Ganapathy

Figure 22: A snapshot from the talk shows Dr. Sriram Ganapathy's humble smile while explaining the concept.

17.11.1 Talk 26: Demystifying the Black Box: Explainability and Trust in Modern AI

Abstract: As artificial intelligence systems in speech, text and vision, become more complex and opaque, ensuring their interpretability and trustworthiness is essential—especially when users only have black-box access. In this talk, I will detail two recent advancements from our work, that tackle these challenges across vision, audio, and language tasks. First, I will introduce Distillation-Aided Explainability (DAX), a gradient-free framework that generates saliency-based explanations using a learnable mask generation network and a student distillation network. DAX outperforms existing methods across modalities using both objective and human-centric evaluation metrics. This part of the talk will based on the work detailed in IEEE JSTSP24 Second, I will present our recent work on Trust Assessment of LLMs - FESTA (Functionally Equivalent Sampling for Trust Assessment), an unsupervised, black-box technique that estimates model uncertainty by probing input consistency and sensitivity through equivalent and complementary samples. Together, these methods show how we can peek inside the black box—using distillation and input sampling approximations—to build approaches that inspire confidence and understanding of deep learning models and LLMs, as they become ubiquitous in several safety-critical domains.

17.12 Prof. (Dr.) K. Sri Rama Murty

Summer School on Speech Signal
Automatic Speech Recognition (ASR)
05-09
tcs Research

Figure 23: Dr. K. Sri Rama Murty explaining the significance of phase processing in speech signals.

17.12.1 Talk 10: Phase Processing of Speech Signals

Abstract: Phase information, long regarded as secondary to magnitude in speech signal processing, has emerged as a powerful cue for analyzing and interpreting speech. This talk highlights key contributions of phase-based methods, particularly those leveraging the Short-Time Fourier Transform (STFT), in uncovering fine temporal and spectral structures of speech. Techniques based on group delay and instantaneous frequency enable high-resolution representations that are sensitive to vocal tract dynamics and source characteristics. Modified group delay functions, product spectrum analysis, and phase modeling approaches have shown remarkable utility in applications such as formant estimation, voice activity detection, speaker and speech recognition, and glottal event analysis. Despite challenges like phase wrapping and windowing artifacts, phase processing continues to provide complementary and sometimes superior information compared to magnitude-based methods, underscoring its growing importance in modern speech technology.

17.13 Prof. (Dr.) Vipul Arora

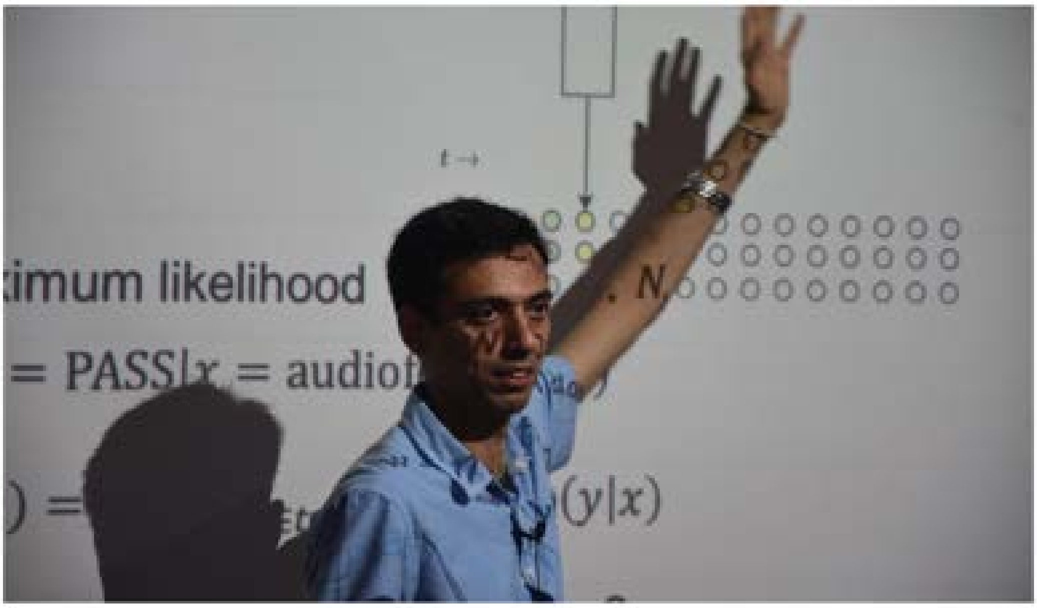


Figure 24: Prof. Vipul Arora explaining the working mechanism of a language-agnostic tokenizer for multiple Indian languages.

17.13.1 Talk 34: Towards Multilingual Speech Tokenization

Abstract: Modern NLP tools, including LLMs, depend on tokens derived from text orthography (written form), which is the conventional written form specific to a language. We present alternative ways to derive tokens directly from speech audio, bypassing orthography. The goal is to obtain universal multilingual tokenization that extracts language-independent features akin to IPA. We will discuss popular tokenization methods, such as speech-to-text ASR and self-supervised learning, and alternative approaches such as audio fingerprinting, wav2tok, and BEST-STD. We will introduce a pairwise training paradigm that circumvents the need for a written form of a language. Finally, we will present our language-agnostic tokenizer, tested across multiple Indian languages, for the word search task.

17.14 Prof. (Dr.) Anil Kumar Vuppala

Figure 25: Prof. Anil Kumar Vuppala presenting on ASR and SLT tasks.

17.14.1 Talk 13: ASR and SLT in Indian Languages

Abstract: This talk focuses on advancements in Automatic Speech Recognition (ASR) in Indian languages and Spoken Language Translation (SLT). The research highlights the creation of the IIITH-CSTD corpus, a large-scale Telugu speech dataset collected through crowd-sourced strategies, and evaluates different ASR architectures on this corpus. The presentation also delves into SLT, outlining both cascaded and end-to-end models, and introduces ”Shruthilipi Anuvaad,” a dataset creation pipeline for low-resource Indic-to-Indic speech translation using weakly labeled data. Furthermore, it details the IIITH-BUT system for low-resource Bhojpuri to Hindi speech translation, discussing hyperparameter optimization, data augmentation, and cross-lingual transfer learning techniques.

17.15 Prof. (Dr.) Vinayak Abrol

Summer School on Speech Signal Pro
Automatic Speech Recognition
05-09 July, 2025
elra

Figure 26: Prof. Vinayak Abrol explaining quantization in modern deep learning–based speech tasks.

17.15.1 Talk 20: On Quantization of Neural Models for Speech Tasks

Abstract: As deep learning models for speech tasks grow in size and complexity, reducing their computational and memory demands becomes critical for efficient deployment, especially on edge devices. Two key strategies to achieve this are model compression and quantization. While model compression focuses on reducing the structural complexity through methods like pruning or distillation, quantization tackles the numerical precision of model parameters, activations, and/or gradients, enabling models to operate with lower bit-widths (e.g., 8-bit instead of 32-bit). This talk will introduce the fundamentals of quantization and discuss why popular methods like post-training quantization (PTQ) and quantization-aware training (QAT) often fall short when applied to modern speech models that include complex components such as channel aggregation, squeeze-and-excitation or attention modules. I will present recent work that addresses these limitations, offering more robust quantization strategies tailored for state-of-the-art speech architectures. The session aims to provide beginner students with a clear understanding of the practical challenges and emerging solutions in making speech models lightweight without compromising accuracy.

17.16 Prof. (Dr.) Hemant A. Patil

Figure 27: Prof. (Dr.) Hemant A. Patil is presenting on multilingual audio deepfake detection and the MLADDC dataset.

17.16.1 Talk L38: Multi-Lingual Audio DeepFake Detection Corpus

Abstract: Deepfakes are artificially generated fake media using deep learning (DL) methods. Recent study found that deepfakes are challenging to detect even for human listeners, however, machines can do better job in their detection. This talk present development of recent Multi-Lingual Audio Deepfake Detection Corpus (MLADDC) to boost the Audio DeepFake Detection (ADD) research. Existing datasets for ADD suffer from several limitations; in particular, they are limited to one or two languages. Proposed dataset contains 20 languages, which have been released in 4 Tracks (6 - Indian languages, 14 - International languages, 20 languages half-truth data, and combined data). Moreover, the proposed dataset has 400 K files (1,125+ hours) of data, which makes it one of the largest datasets. Deepfakes in MLADDC have been produced using advanced Deep Learning (DL) methods, such as HiFi- GAN and BigVGAN. Another novelty of this corpus lies in its sub-dataset, that has partial deepfakes (Half-Truth). We compared our dataset with various existing datasets, using cross-database method. For comparison, we also proposed baseline accuracy of 68.44%, and EER of 40.9% with MFCC features and CNN classifier (14 languages track only) indicating technological challenges associated with ADD task on proposed dataset. The talk will also discuss some of the open research challenges in this ADD research, more so, in the multilingual context.

18 Industry Perspective Talks

18.1 Dr. Sri Garimella

Figure 28: Dr. Sri Garimella delivering a fruitful talk on representation learning for multimodal LLMs..

18.1.1 Talk 6: Representation Learning of Speech and its Applications

Abstract: Speech representation learning is a crucial component in various speech processing applications, including speech recognition (ASR), speaker identification, emotion recognition, and language identification. This field focuses on encoding speech signals into compact, informative representations—either as dense embeddings or discrete tokens—which serve as input for downstream tasks. In this presentation, we explore the evolution of speech encoding techniques, beginning with a widely-adopted unsupervised or self-supervised learning approach. We then delve into a series of advancements that have significantly improved the performance of speech representation models (or speech encoders), specifically, task-aware training, knowledge distillation, dual-mode encoding capabilities, and language-aware encoding through attention mechanisms. These innovations have collectively enhanced the accuracy of speech-based applications. Furthermore, we examine the effects of integrating these advanced speech encoders with multimodal large language models (LLMs). The talk concludes with a comparative analysis of our developed models against other external ASR models.

18.2 Dr. Sunil Kumar Kopparapu

ool on Speech Signal Processing (S4P)
atic Speech Recognition (ASR)
05-09 July, 2025
Summer School on Speech Signal Processin
Automatic Speech Recognition (ASR)
tcs Research

Figure 29: Dr. Sunil Kumar Kopparapu presenting recent work from TCS Research Lab, Mumbai.

18.2.1 Talk 3: Audio & Speech Processing — What have we been doing?

Abstract: In this talk, we will explore some of the more recent work that has been happening in TCS in the area of audio and speech signal processing. We will look at aspects that are required to enable voice user interface for the emergent user. Some of the things that we will cover are (a) how the knowledge of the microphone location impacts the performance of an ASR, (b) a novel data argumentation method for enabling robust ASR, and (c) the importance of choosing an appropriate vocabulary size hyper parameter in an e2e ASR. Time permitting, we will look at a spoken grammar assessment tool, the need for a new metric for audio captioning and some experiments around text to speech.

18.3 Dr. Nagaraj Adiga

Figure 30: Dr. Nagaraj Adiga illustrating advances in Speech Large Language Models for recognition and translation.

18.3.1 Talk 8: Advances in Speech Large Language Models for Recognition and Translation

Abstract: This talk explores the latest progress in Speech Large Language Models (Speech LLMs), with a focus on their use in automatic speech recognition (ASR) and automatic speech translation (AST). Unlike traditional cascaded systems, Speech LLMs enable end-to-end modeling of spoken language, offering improved contextual understanding and more natural human-like interactions. We examine recent architectural innovations integrating speech and language through unified tokenization and adaptation mechanisms. While ASR datasets are increasingly available, there remains a critical lack of high-quality AST resources for Indian languages. To bridge this gap, we introduce IndicST, a new dataset for training and evaluating Speech LLMs in the Indian linguistic context. Additionally, we analyze how component-level changes within Speech LLMs impact ASR and AST outcomes across Indian languages. The session concludes by addressing key challenges and future opportunities for real-world deployment of these models.

18.4 Dr. Debmalaya Chakroborty

Figure 31: Dr. Debmalaya Chakraborty introducing team members during the talk at S4P 2025, with a friendly expression.

18.4.1 Talk 23: Towards Universal Audio Understanding: A Unified Encoder for Speech and Audio Tasks

Abstract: Recent advancements in speech and audio encoders have drawn significant attention due to their integration with Large Language Models for diverse acoustic tasks. While most research has focused on developing specialized encoders for either speech or audio domains, with limited solutions addressing streaming constraints, there remains a critical gap in unified approaches. This presentation introduces a novel universal audio-speech encoder designed to process the complete spectrum of acoustic inputs, from human speech to environmental sounds. Our encoder generates robust representations that seamlessly interface with large language models for multiple downstream tasks, including automatic speech recognition, speech translation, audio captioning, and event detection. We address the fundamental challenges of unifying traditionally separate speech and audio encoding paradigms while effectively handling both streaming and non-streaming applications. Through our analysis of existing foundation models, we identify their limitations and present innovative techniques to bridge these gaps. Experimental results demonstrate that our universal encoder achieves comparable or superior performance to specialized models across various benchmarks, marking a significant step toward a truly unified audio processing framework.

18.5 Dr. Premjeet Singh

Figure 32: Explaining the workflow of ASR, from initial problem formulation to deployment in a production-ready environment by Dr. Premjeet Singh

18.5.1 Talk 24: Automatic Speech Recognition: From Problem to Research to Productisation

Abstract: Automatic speech recognition (ASR) is an essential part of Samsung Galaxy AI features such as, transcript assist, live translate, and call transcript. Real-time usefulness of such features require the ASR system to be robust against various practical challenges like, background noises, room reverberation, effects of recorded/playback sound, latency, on-device memory, etc. This talk discusses how Galaxy AI team aims towards recognizing such challenges, overcoming them, and providing product-based solutions over speech from network calls and listening mode applications.

18.6 Dr. Nirmesh J. Shah

Figure 33: Dr. Nirmesh Shah highlighting observations on emotion similarity across state-of-the-art results while discussing their paper.

18.6.1 Talk 36: Generative Modeling for Emotional Speech Synthesis: Progress, Pitfalls, and Possibilities

Abstract: Emotional speech synthesis aims to generate human-like speech that conveys not only linguistic content but also expressive emotional states. As emotionally rich speech becomes critical for applications in entertainment, virtual storytelling, and immersive user experiences, this domain holds the potential to bridge the emotional gap between humans and machines. Understanding and replicating emotions in speech is not just a technical challenge, but a step toward more natural and empathetic human-computer interaction. This talk presents a comprehensive overview of the emotional speech synthesis field, highlighting key developments and challenges in modeling emotional prosody and style. We will explore the role of generative modeling techniques in advancing emotional Text-to-Speech (TTS), covering how these methods enable more controllable, diverse, and realistic synthesis. The talk will also discuss fundamental issues such as emotion representation, emotion controllability, and the evaluation of emotional expressiveness. Emphasis will be placed on current trends and open research problems, offering insights into future directions for building emotionally aware and socially intelligent speech systems. Finally, I will briefly share insights from our recent paper, which introduces a method for emotion intensity regularization in emotional voice conversion, contributing toward finer emotional control in synthesized speech.

18.7 Dr. Bidisha Sharmra

Figure 34: Dr. Bidisha Sharma explaining semi-supervised learning in low-resource and low-data constraints.

18.7.1 Talk 16: Beyond Supervision: Leveraging Pseudo-Labels and LLMs for Domain-Specific ASR

Abstract: Training high-quality automatic speech recognition (ASR) models typically requires extensive transcribed data, posing a significant challenge for domain adaptation in scenarios with limited or no labeled audio. In this talk, I will introduce a unified approach for training transducer-based ASR models in a semi-supervised setting, leveraging large unlabeled corpora and minimal or zero manual annotation. This work explores the generation and refinement of pseudo-labels using outputs from multiple ASR models, enhanced through consensus mechanisms, large language models (LLMs), and speech-based LLMs (SpeechLLMs). We propose a flexible framework that combines model prompting, multi-system alignment, and filtering techniques based on consensus voting, named entity recognition (NER), and error-rate prediction. Our experiments across diverse datasets, including call centre and conversational corpora, demonstrate that these strategies not only improve the quality of pseudo-labels but also enable scalable training of ASR models with significantly reduced reliance on human-annotated data. The findings demonstrate the potential of semi-supervised pipelines to democratize ASR development, especially in low-resource or domain-specific settings.

18.8 Dipesh K. Singh

Figure 35: Mr. Dipesh K. Singh on advances in ASR using self-supervised learning, large language models, and multimodal foundation models.

18.8.1 Talk 40: Next-Generation Speech Recognition: Scaling Self-Supervised Learning, Multimodal Fusion, and LLM-Augmented ASR for Real-World Deployment

Abstract: The field of automatic speech recognition (ASR) has undergone transformative changes with the rise of self-supervised learning (SSL), large language models (LLMs), and multimodal foundation models. In this talk, we explore how these advancements address persistent challenges in multi-accent generalization, noise robustness, and real-time ambient speech processing. We begin by examining how self-supervised pretraining (e.g., wav2vec 3.0, Whisper-v3) has reduced reliance on labeled data while improving cross-accent generalization. Next, we discuss innovations in noise-robust ASR, including dynamic acoustic adaptation and neural dereverberation techniques that leverage visual or contextual cues for improved performance in chaotic environments. A key focus is the integration of LLMs into end-to-end ASR systems, enabling not just transcription but semantic disambiguation, speaker-adaptive correction, and task-aware contextualization (e.g., for medical or legal domains). We also highlight emerging work on ”ASR as a sensor”—using speech recognition for ambient intelligence in healthcare, education, and human-computer interaction.

18.9 Thoshith S.

hool on Speech Signal Processing (SSP)
matic Speech Recognition (ASR)
05-09 July, 2025
elra
tcs

Figure 36: Mr. Thoshith S. on real-time, multilingual, task-aware voice-to-voice AI systems.

18.9.1 Talk 32: Building Real-Time Voice-to-Voice LLMs: Toward Expressive, Multilingual, and Task-Aware AI

Abstract: Voice-to-voice systems are redefining the boundaries of conversational AI by enabling direct, real-time speech interactions—where input speech is understood, reasoned over, and responded to with expressive, human-like output speech. These systems go beyond simple transcription or translation, aiming to capture the richness of human communication across language, emotion, and intent. This talk explores the core components and design considerations behind such systems: modeling multilingual speech, preserving speaker affect and prosody, and enabling intelligent, task-aware responses. Multitask training approaches that combine speech recognition, language identification, and emotion modeling help develop representations that are robust across speakers, dialects, and use cases. Speech generation models are guided not only by linguistic content but also by cues from the speaker's tone, emotion, and conversational rhythm. A critical part of enabling natural interaction is low-latency end-of-utterance detection, which determines when a system should respond. This timing mechanism plays a vital role in creating responsive, turn-based dialogues, especially in real-time environments where naturalness is key. Use cases such as real-time voice-to-voice translation, spoken task execution, and interactive AI agents illustrate the promise of voice-native systems that can understand and act through speech alone. Rather than treating speech simply as a carrier for text, these systems view it as the primary interface for reasoning, action, and expression. This talk offers a look at the current capabilities and limitations of voice-to-voice systems, and the emerging research directions driving them toward more natural, multilingual, and emotionally intelligent communication.

19 Some memories of the S4P 2025

19.1 Felicitation of Invited Speakers and Tea Break Discussions

The invited speakers spared their valuable time and shared their rich and wide research experience and expertise with the participants during the Summer School. They were felicitated at a special function organized as part of the Summer School.

Figure 37: The felicitation of the invited speaker during lecture breaks.

Figure 38: Social networking between participants and speakers.

20 Interaction of the Invited Speakers with the Director General, DAU, Gandhinagar

Figure 39: Invited speakers from academia and industry engaged in an interactive session with Director General, Prof. Tathagata Bandopadhyay.

21 5-Minutes Ph.D. Thesis Contest

The Summer School at DA-IICT hosted the **sixth edition of the 5 Minutes Ph.D. Thesis (5MPT) Contest**, following the success of previous editions held during S4P 2024, S4P 2019, S4P 2018, S4P 2017, and S4P 2016. This unique event provided doctoral students from across India and abroad with a platform to present their research in the broad areas of speech and audio signal processing within a concise **5-minute format**. The contest aimed to enhance research communication skills, provide visibility to emerging research, and facilitate meaningful interactions between students and eminent researchers from academia and industry. Participation was open to all Ph.D. students registered for S4P 2025. Applications were reviewed by an internationally renowned expert committee, and shortlisted candidates delivered their presentations during the school. Each talk was evaluated based on clarity, conciseness, and impact. To recognize excellence, **four scholars were awarded cash prizes**, endorsed by ISCA, IndSCA, Google, and DAU:

- **First Prize:** INR 15,000
- **Second Prize:** INR 10,000
- **Third Prize:** INR 5,000
- **Fourth Prize:** INR 5,000

The event drew inspiration from the globally acclaimed **3 Minute Thesis (3MT) competition**, initiated at the University of Queensland in 2008, and subsequently adopted by numerous universities and international conferences, including the European Signal Processing Conference (EUSIPCO).

21.1 Winners of the 5MPT

1. Ravindrakumar M. Purohit

Figure 40: Captured moments from S4P 2025 5MPT event.

Figure 41: Prize Winner Mr. Ravindrakumar Purohit (DAU, Gandhinagar).

22 On-Spot Poster Session

We organised an on-spot poster presentation session during Tea/Lunch break for all five days of the S4P 2024. The key motivation of the session is to encourage the participants of summer school to present their ongoing research/published papers and get review feedback from experts and attendees of the event.

22.1 Winners of the On-Spot Poster Session

1. Aniket Pandey, Arth J. Shah, Hemant A. Patil
2. Kunjan Gajre, Rajnidhi Gupta, Ravindrakumar Purohit, Hemant A. Patil
3. Arth J. Shah, Dharmendra H. Vaghera, Ravindrakumar M. Purohit, Hemant A. Patil

Figure 42: Awarding certificates to the winners of the S4P 2025 Poster Presentation event.

Cash prizes of INR 5,000 each were awarded to the four best presentations, and certificates were provided to all participants. Winners were announced during the S4P 2025 award ceremony.

23 Deepfake Challenge

The session focused on introducing the **MLADDC dataset** and the need for multilingual approaches in audio DeepFake detection. It highlighted advancements in **DeepFox 2.0**, a state-of-the-art corpus featuring text-symmetric pairs, 'half-truth' audio for misinformation simulation, coverage across Indian and global languages, 51 generation models, and novel source-tracing capabilities. The discussion addressed the **growing DeepFake threat landscape**, differentiating it from traditional spoofing, showcasing realistic AI-generated speech like 'Fake Obama', and stressing the urgency of multilingual solutions since many public figures and applications operate in regional languages. Both humans and machines face challenges in reliable detection, especially under noisy conditions.

The session also featured the **S4P 2025 DeepFake Detection Challenge**, with 9 registered teams (7 submissions) from India and the USA. Using a baseline BiLSTM classifier on linear filterbank features, participants competed for prizes, with Ms. Priyanshi K. Patel's team securing first place, followed by Ms. Sneha Ilame's and Mr. Parth Patel's teams. The event encouraged global collaboration, drawing participants from 18 countries, and promoted practical experimentation in ADD research.

Key achievements included demonstrating the feasibility of multilingual datasets, raising awareness of audio security risks, releasing an advanced benchmark corpus, and fostering international cooperation. Limitations were acknowledged, such as restricted language and model diversity, GPU resource constraints, and the need for unseen test sets for half-truth detection. Looking ahead, plans involve **dataset expansion**, new challenges like **DeepFox Challenge 2** and **SingFox Challenge**, and broader collaborations to build the world's largest and most diverse audio DeepFake corpus.

23.1 Winners of the DeepFake Challenge

1. Priyanshi Ka Patel (M.Tech ML, DAU, Gandhinagar)
2. Sneha Vijay Ilame (NFSU, Gandhinagar)
3. Patel Parth Rajeshbhai (M.Tech ML, DAU, Gandhinagar)

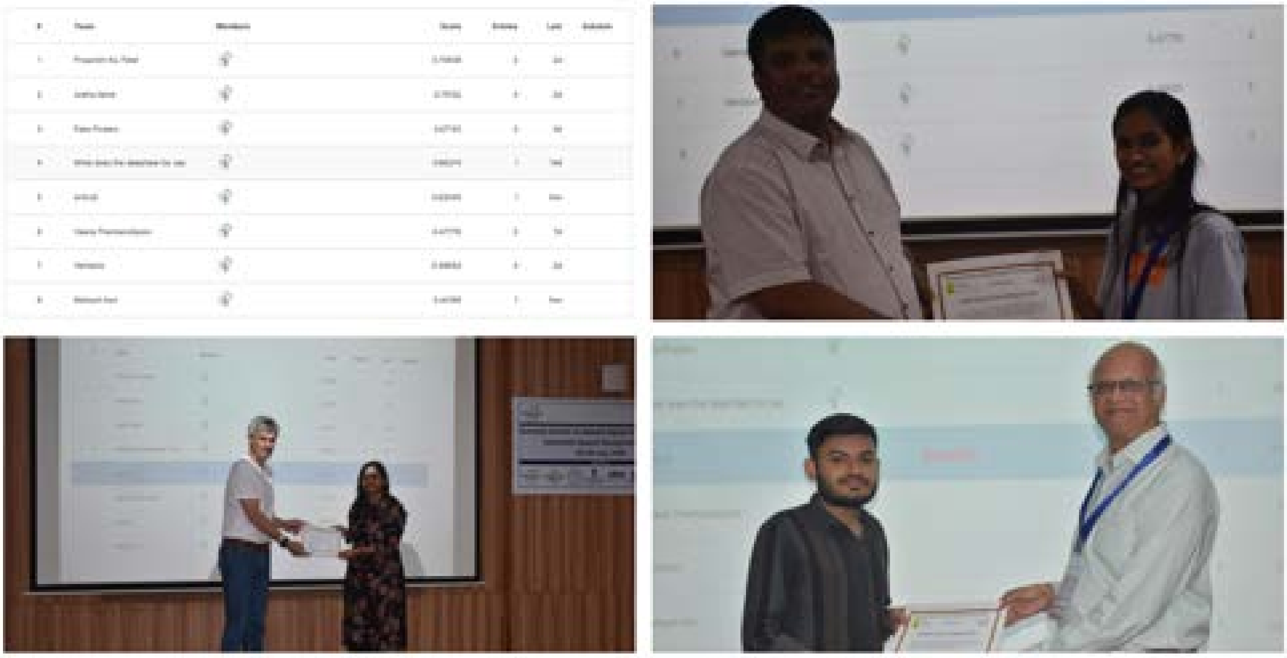


Figure 43: Leaderboard result and three challenge winners of cash price.

24 Volunteers of S4P 2025, DAU Gandhinagar

Thus, summer school activity strengthens a great bond of team spirit and interpersonal skills within members of Speech Group @ DAU.

Figure 44: Volunteers of the S4P 2025.

25 Appreciation of S4P 2025 on Social Media

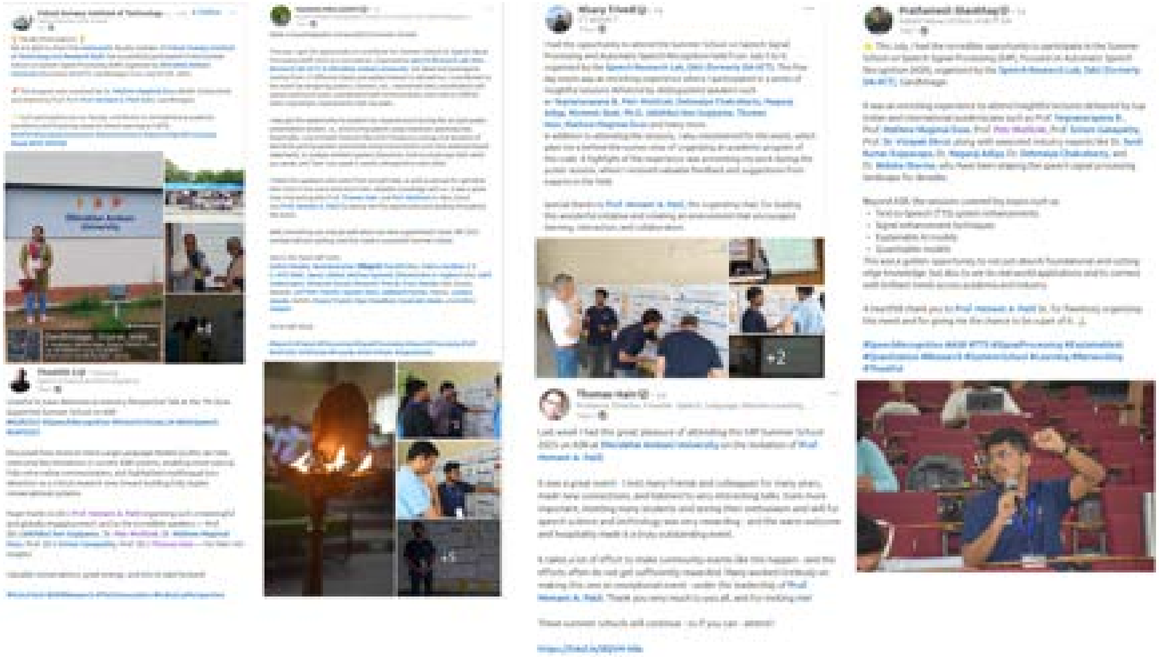


Figure 45: Participants and speakers shared their experiences from S4P on LinkedIn.

26 Feedback from the Participants

It was a valuable learning experience, and I gained many new insights.

– Joel S (IIT Gandhinagar)

A memorable and enriching program. Grateful for the chance to meet global experts in Speech and be part of this platform.

– Patel Vinaben Shankarbhai (GTU Ahmedabad)

Well-organised summer school with warm hospitality and motivated students. Suggested shorter tasks and more time for discussions and student presentations.

– Petr Motlíček (Idiap Research Institute Switzerland)