Saplings: a session by budding Research Scholars

PhD Scholar	Title	Description
Arshdeep Singh, IIT Mandi boparaiarshdeep@gmail.com	An Ensemble framework using Deep Hidden Features for Audio Scene Classification	Understanding of the physical environment (commonly refers as scene) using audio cues is an alternate way as that using visual cues. The light weight, low power consumption and no light constraint of the audio sensor makes it an attractive choice in place of cameras. However, the unstructured nature of sound, high interclass variability and no prior knowledge of sound sources present in the environment etc. makes it a challenging task. In this talk, I will discuss an ensemble framework, which utilizes the hidden information learned by a 1D convolution neural network (SoundNet). SoundNet is a pre-trained neural network, trained on raw audio signals directly. We show that various layers of SoundNet give complementary information. Thus, the complementary information can be used to enhance the overall performance. The proposed framework operates on raw-audio signals only without any need of time-frequency transformation as most of the current studies do. Under low data condition, the proposed strategy can be utilized to increase the performance in place of data augmentation based techniques, by maximally utilizing the underlying model.
Rini Sharon, IIT Madras ee15d210@smail.iitm.ac.in	Study of temporal syllabic structure percieved by the brain	Clinical applicability of electroencephalography (EEG) is well established, however the use of EEG as a choice for constructing brain computer interfaces to develop communication platforms is relatively recent. To provide more natural means of communication, there is an increasing focus on bringing together speech and EEG signal processing. Quantifying the way our brain processes speech is one way of approaching the problem of speech recognition using brain waves. This paper analyses the feasibility of recognizing syllable level units by studying the temporal structure of speech reflected in the EEG signals. The slowly varying component of the delta band is present in all other EEG frequency bands. Analysis shows that removing the delta trend in EEG signals results in signals that reveals syllable like structure. Using a 25 syllable framework, classification of EEG data obtained from 13 subjects yields promising results, thus encouraging the potential of revealing speech related temporal structure in EEG.
Mari Ganesh Kumar, IIT Madras mari@cse.iitm.ac.in	Subspace techniques for task independent EEG person identification	There has been a growing interest in studying electroencephalography signals (EEG) as a possible biometric. The brain signals captured by EEG are rich and carry information related to the individual, tasks being performed, mental state, and other channel/measurement noise due to session variability and artifacts. To effectively extract person-specific signatures present in EEG, it is necessary to define a subspace that enhances the biometric information and suppresses other nuisance factors. i-vector and x-vector are state-of-art subspace techniques used in speaker recognition. In this paper, novel modifications are proposed for both frameworks to project person-specific signatures from multi-channel EEG into a subspace. The modified i-vector and x-vector systems outperform baseline i-vector and x-vector systems with an absolute improvement of 10.5% and 15.9%, respectively.
Ch Srikanth Raj, IISc Bangalore sraj@iisc.ac.in	Neural network constraint for better dereverberation in multi channel linear prediction	Reverberation is the dominant source of signal distortion in distance speech acquisition. The diffuse (late reverb) part of the room response affects spectro-temporal properties and affects speech intelligibity. We consider cancellation of late reverb component using multi-channel linear prediction (MCLP) in short time Fourier transform (STFT) domain. The late reverb part of the reverberant signal is modeled using a linear predictor with a delay, and the prediction residual is the desired early reverb signal, in each STFT frequency bin. Prediction filters are estimated assuming a time-varying complex Gaussian source model for the residual. Maximum likelihood estimation leads to an iterative speech power spectral density (PSD) weighted prediction error (WPE) minimization problem. The method is sensitive to the estimate of the desired signal PSD. In this work, we propose a deep neural network (DNN) based non-linear constraint for the desired signal PSD. An auto encoder trained on clean speech STFT coefficients is used as the desired signal prior. The estimate for the desired signal obtained in each iteration of WPE is used to predict the desired signal PSD using the DNN. We explore two different architectures based on (i) fully-connected (FC) feed-forward, and (ii) recurrent long short-term memory (LSTM) layers. Experiments using measured room impulse responses show that the LSTM-DNN based PSD constraint results in improved FwSNR, PESQ and STOI scores compared to the traditional methods for late reverb cancellation.

Purvi Agrawal,
IISc Bangalore
purvia@iisc.ac.in

Unsupervised Representation Learning For Noise Robust Speech Recognition

The performance of an automatic speech recognition (ASR) system deteriorates in the presence of background noise. If the conventional ASR system is seen as broadly divided into two blocks: feature extraction and system training, our focus is on "how to obtain robust speech representations (features)". In particular, we are trying to benefit the machine understanding of speech by designing a data-driven representation learning paradigm. The modulation filtering method is one such approach to obtain robust representations, based on enhancing perceptually relevant regions of the modulation spectrum while suppressing the regions susceptible to noise.
We explore learning of modulation filters purely from a data-driven perspective using deep generative models, namely, restricted Boltzmann machine, convolutional autoencoder, generative adversarial network, and variational autoencoder. Our research in this direction has shown that the speech representations obtained from the proposed method improve the ASR performance in noisy conditions. In addition, they are found to be resilient to semi-supervised training of ASR. In the talk, I will be discussing about the motivation of the work and some key findings from it.

Arthi S,
IISc Bangalore
arthi.ramaniam@gmail.com

Perception of Auditory Source Width Expansion

As a part of my research work, we are exploring the spatial aspects of source localization and auditory source width expansion (ASW). Localization of sources has been classically studied as perception of point source. In my research, we look at perception of sources with finite width with respect to human perception. Source width has been classically studied in reverberant conditions of auditoria, while the perceptual auditory process for the perception of these parameters is not clearly understood. We conduct human listening experiments to quantify these parameters on a perceptual scale in anechoic conditions. We have built a multi-channel linear array of loudspeakers and its associated amplifiers and D/A converters to study ASW. We simulate sources of different physical width by energizing different number of loudspeakers and try to quantify the perceptual width of the source on a relative scale, analogous to loudness scale and pitch scale. Inter-aural time difference and inter-aural level differences are understood to be significant cues for localization of point sources. We analyze the binaural phase differences to understand the perceptual cues for source width.