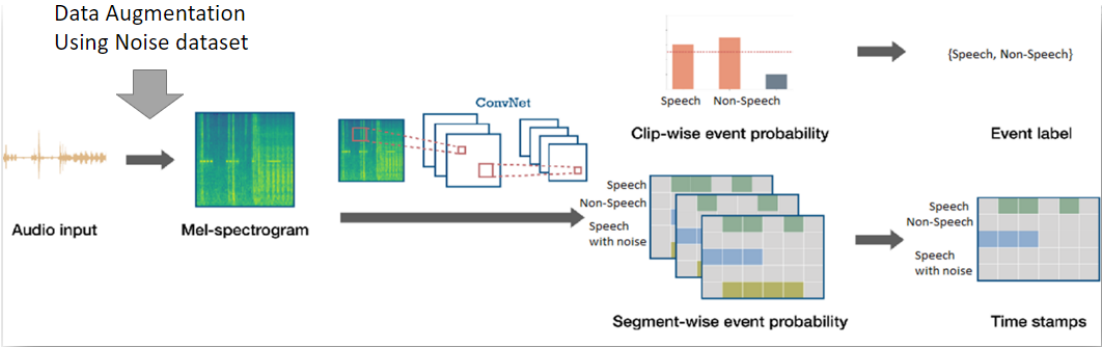
Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarize the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.



The challenge expects us to detect speech and no speech in noisy indoor environment with approximately 95% accuracy. In order to solve this challenge, we plan to use pretrained convolutional neural network-based method. With this model, segment-wise speech probabilities will be generated from the given audio data. For generating correct speech probabilities from different domains and noisy audios, data augmentation with selected noise datasets will be used. Therefore, maintaining fast inference times with low resource-based CNN architectures will be achieved.

2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

For selecting a successful speech activity detection model, after a literature research, we examined CNN14 model from PANN's work[1]. In this method, classifier is trained to classify 527 classes on audioset. Also, CNN14 model is converted to a detection model (CNN14 segmentwise) to be able to produce time intervals of classified classes. With this model, classifying speech class on audioset[2] with more than 95% precision is achieved. However, this CNN14 model is very large for low resource (edge) devices with 80M parameters. In PLSA[3] work, it is proved that with using smaller models like Mobilenet[4] and Efficientnet[5], achieving similar mAP (Mean Average Precision: Average precision on each class in Audioset) with CNN14 model is possible. Also, a better training pipeline for audioset is proposed.

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981

Mobilenet model is designed for fast interface on CPU and edge devices. It has very less floating-point instructions compared to other CNN models.

As show in the below table, mobilenet is able to get nearly the same precision with CNN14 on speech class.

Table 1: Related work results on Audioset data

Method	**Mean Average Precision on AudioSet**	**Precision on Speech**	**Parameter Count**
CNN14	0.44	0.981	80M
CNN14 segment-wise	0.42	*	80M
PSLA Mobilenet	0.41 (0.30)	0.980	4M
PSLA Efficientnet	0.46 (0.37)	*	10M

Methodology:

We plan to convert Mobilenet to detection model like Cnn14 segmentwise. We will pretrain it on this model on Audioset but only using classes as speech or non-speech. Then, this model will be finetuned on REACH data set and other selected Speech Activity Detection datasets (these datasets are chosen by considering the challenge's needs) mentioned below. Data augmentation will be applied to generate noisy and different domain data, both to pretraining and finetuning parts.

These datasets will be used as speech activity dataset:

- Musan[6]
- Avaspeech[7]: Ava speech contains labels as "Speech", "Noisy Speech", "Speech with Music", "Non Speech"
- CommonLanguage[8]: Commonlanguage Dataset has speech data from 45 different language.
- QUT-NOISE-TIMIT[9]: QUT-NOISE-TIMIT dataset contains large domain noises and will be used as noise datasets for data augmentation.
- Room Impulsive Responses[10]: This dataset contains indoor noises.

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981

References:

[1] Kong, Qiuqiang, et al. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 2880-2894.

[2] https://research.google.com/audioset

[3] Gong, Yuan, Yu-An Chung, and James Glass. "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3292-3306.

[4] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

[5] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[6] https://www.openslr.org/resources/17

[7] https://research.google.com/ava

[8] https://zenodo.org/record/5036977/

[9] D. Dean, S. Sridharan, R. Vogt, M. Mason (2010) "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms", in Proceedings of Interspeech 2010, Makuhari Messe International Convention Complex, Makuhari, Japan.

[10] https://github.com/RoyJames/room-impulse-responses

3. SCALABILITY AND FLEXIBILITY OF THE SOLUTION: Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains

Since we will pretrain our model in audioset dataset by Google[2], which contains 5.5 Million-hours of data, it will already cover a wide variety of domains in terms of the content of the speech data. Pretraining on this dataset and then finetuning on the given dataset domain, will make the performance and adaptation of the model very easy. Also, our training pipeline depends heavily on data augmentation. With data augmentation we will generate different noise levels and different domain data.

To show generalization of audioset domain, we compare CNN14 segmentwise model with open source Gaussian Mixture Model based Webrtc Vad algorithm.

Table 2: Preliminary results

Set	VAD	FA (False Alarm Rate)	Miss Rate	Detection Error Rate
AMI	Google Webrtc Vad	2.18	17.17	19.25
	cnn14_att_10sec	5.43	4.74	10.17
	cnn14_att_1sec	3.71	7.45	11.16
VoxConverse	Google Webrtc Vad	2.75	1.71	4.46
	cnn14_att_10sec	3.77	0.59	4.36

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981

	cnn14_att_1sec	3.13	0.18	3.32

We used 2 datasets: Ami and VoxConserve. Cnn14 segment wise model already performs better than Webrtc algorithm on Ami dataset. Although it is a very challenging noisy meeting dataset, it performs 2 times better than Webrtc. This Cnn14 model is only pretrained on Audioset.

As a proof of concept we used REACH sample health dataset too.

REACH dataset has 140 non speech samples. While Webtrc algoritm confuses speech and non speech on this set by telling 97 of samples has speech where CNN14 model only wrongly classified 2 samples.

Table 3: Preliminary results with REACH sample data

Reach Samples (140 non-speech samples)	**Speech**	**Non-speech**
Google Webrtc Vad	97	43
Cnn14_att	2	138

With our proposed training pipeline and with the use of Mobilenet, we will train a fast and accurate speech activity detection model.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planed for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment…) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

As our proposed solution includes working with sensitive and personal data such as health and audio data, data governance and legal compliance are critical. To prevent any actions that can violate data security, the following actions are and will be taken:

1. All the model trainings and data usage will be done parallel with General Data Protection Regulations. SESTEK has 6 European Union funded projects where SESTEK has entered and passed very strict Ethics Committee checks in terms of data protection policies. In addition, SESTEK's 29 nationally funded projects were inspected and passed the national data protection policies, KVKK.
2. For the model trainings, there will be no additional voice data collection, only open-source data will be used. Open-source data proposed in the proposal have either Creative Commons Attribution 4.0 International or CC-BY-SA licensing where if the source is declared, these datasets can be used commercially. Only QUT-Noise-Timit dataset's license includes only research usage, therefore, this dataset will only be used in tests.
3. As the challenge deals with sensitive data and the challenge specifically addresses to no usage of cloud, the proposed model will work on device so there will be no cloud-based or device-based storage.
4. In case the data provider requests to store data, an asymmetric encryption that complies with Advanced Encryption Standards can be applied so that all the data will be encrypted.

SESTEK's Speech Recognition and Voice Activity Detection modules have a success rate of approximately 95% with call centre data. As SESTEK's main target group are call centres, huge amounts of data flow and the ability to test the models for different data sets enables this success rate. This challenge has some certain risks as the speech detection environment and domain differs and this creates the following problems. First of all, although call centre records are not that different from indoor data, the

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981

audio source is close source and data is collected with specific recorders. Secondly, in the regular call centre records, there is a certain amount of noise level but the challenge requires a speech recognition model that can work with very noisy data. Lastly, health domain is not a very active domain in SESTEK' customer portfolio.

SESTEK will fulfil the challenge's requirements as the risks are known beforehand. As SESTEK works in Agile Methodology, each step of the project from design to deployment will be taken carefully.

The most major risk of the project is the lack of data variety. To prevent the risks, for the design stage, the datasets are chosen according to challenge's needs. The datasets include a wide variety of speech data from different speakers in different languages where the indoor data is selected with different voice parameters. In addition, the whole dataset to be shared by the data provider will increase the data variety. For the development stage, there are minor risks as the currently used voice activity detection algorithms will be converted to speech activity detection models. The performance checks will be compared with the State of the Art results which will guide us to improve. Coding issues or bugs will be detected with routine checks as well. The deployment stage can create time delays as the platform where the model works has to be discussed with the data provider in the further stages of the project.

roject has received funding from the European Union's Horizon 2020
ch and innovation programme under grant agreement No 951981