Natural Language and Speech Processing Laboratory

Institute for Informatics and Automation Problems of NAS RA
Karen Avetisyan

Natural language and speech processing laboratory

- **Institute for Informatics and Automation Problems**
- **Research Directions**
  - **Speech Technologies**
  - **Data Pipelines**
  - **Natural Language Processing**

1-year period tasks

- **Speaker Diarization**
- **Speech Enhancement**
- **Emotion Recognition**
- **TTS for Armenian Dialects**
- **ASR for Armenian Dialects**

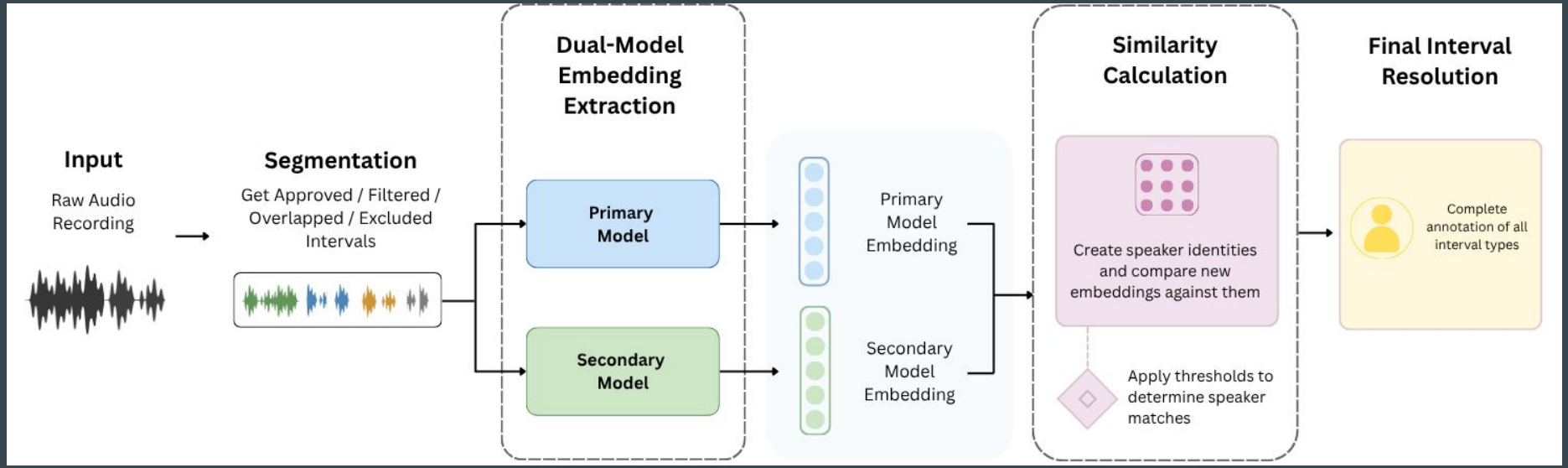
Speaker Diarization

What is Speaker Diarization?

Speaker diarization answers the question "who spoke when". It takes a raw audio recording and segments it in time, labelling each part with the identity of the speaker. It is a core component in meeting transcription, call analytics, and voice-enabled systems.

1. Segmentation
2. Embedding Extraction
3. Speaker Assignment
4. Overlap Handling

Speaker Diarization: System Overview



Speaker Diarization: Results

DER (lower is better)

System	AISHELL-4	AliMeeting	AMI-IHM	AMI-SDM	CallHome	MSDWild	MSP-Podcast	VoxConverse	Average
Pyannote	0.1173	0.1275	0.1755	0.2157	0.2164	0.0805	0.2909	0.1118	**0.167**
DiariZen	0.1011	0.2070	0.2519	0.1390	0.1274	0.2068	0.2745	0.0919	**0.175**
Ultra-SortFormer	0.1896	0.2325	0.1492	0.1836	0.2562	0.2983	0.3122	0.1520	**0.2217**
Proposed	0.1701	0.2245	0.2787	0.2029	0.2017	0.2194	0.2763	0.1103	**0.2105**

Processing Time (minutes)

System	AISHELL-4	AliMeeting	AMI-IHM	AMI-SDM	CallHome	MSDWild	MSP-Podcast	VoxConverse	Average
Pyannote	107	19.07	14.02	14.22	23.04	11.02	19.12	63.69	**33.9**
DiariZen	14.65	12.36	10.15	9.93	24.48	11.23	15.73	51.75	**18.79**
Ultra-SortFormer	1.49	1.19	0.97	0.96	2.34	1.41	1.58	5.07	**15.01**
Proposed	5.61	5.93	3.76	3.85	9.55	9.73	6.26	21.31	**8.25**

Speech Enhancement

What is Speech Enhancement?

Speech Enhancement is the process of improving the quality of a voice recording or live speech so it becomes clearer and easier to understand.

Noisy audio

Enhanced audio

Speech Enhancement

What is Speech Enhancement?

Speech Enhancement is the process of improving the quality of a voice recording or live speech so it becomes clearer and easier to understand.

Noisy audio

Enhanced audio

Speech Enhancement

What is Speech Enhancement?

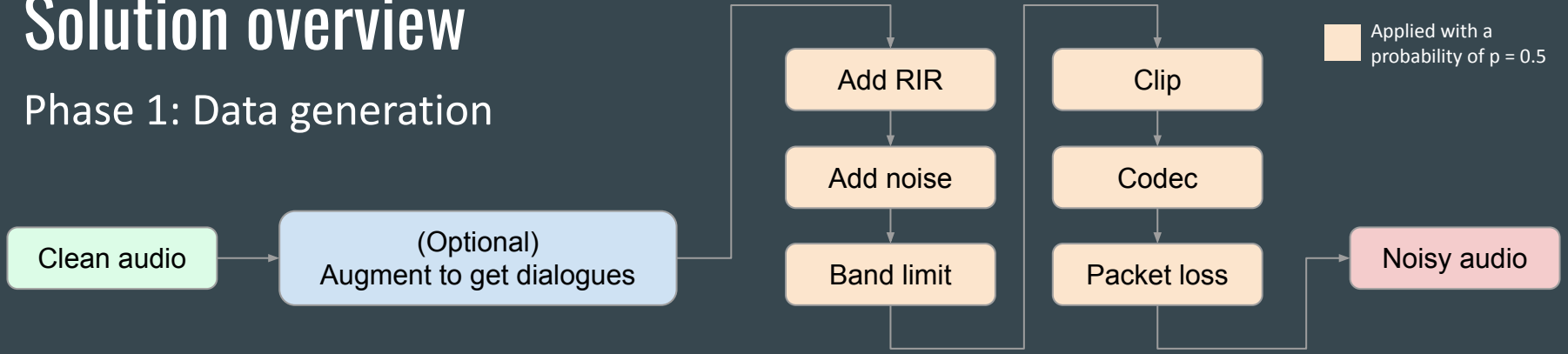
Speech Enhancement is the process of improving the quality of a voice recording or live speech so it becomes clearer and easier to understand.

Noisy audio

Enhanced audio

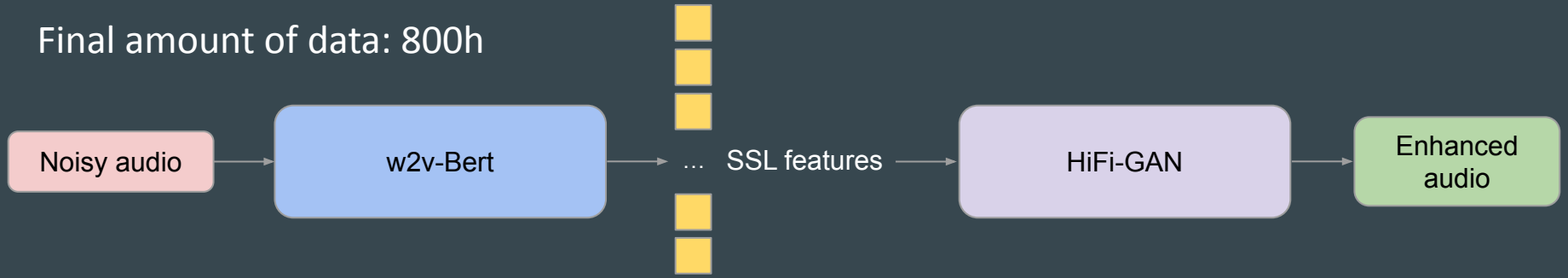
Speech Enhancement: Solution overview

Phase 1: Data generation



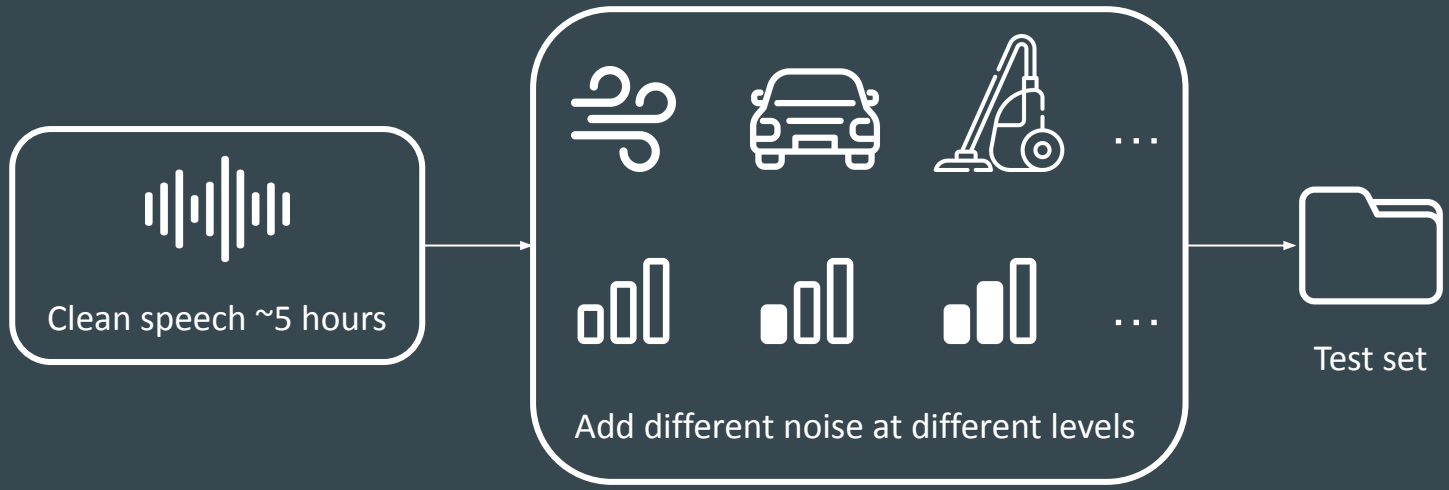
Phase 2: Model training

Final amount of data: 800h



Speech Enhancement: Test set

- Collected around 5 hours of clean Armenian speech.
- Added noise to the clean recordings using the same method described in the previous slide.
- Used different noise types and noise levels to make the test set more realistic.



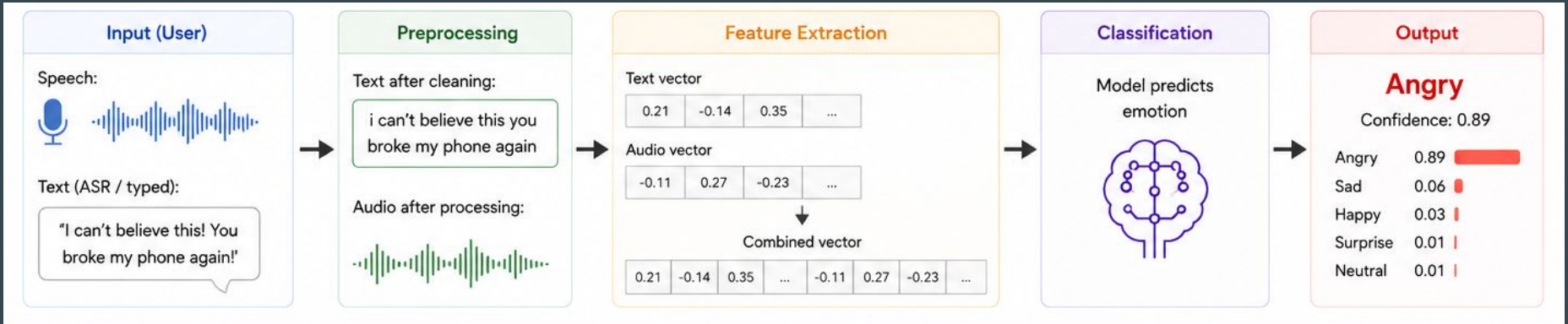
Speech Enhancement: Results

Model	PESQ	STOI	DNSMOS		
			SIG	BAK	OVL
FlowSE	**2.96**	0.91	3.32	3.81	2.97
SEMamba	2.43	0.83	3.16	3.56	2.85
MP-SENet	2.67	0.93	3.43	3.65	2.95
CMGAN	2.74	**0.94**	3.23	3.61	2.94
Ours	2.94	0.92	**3.96**	**4.12**	**3.46**

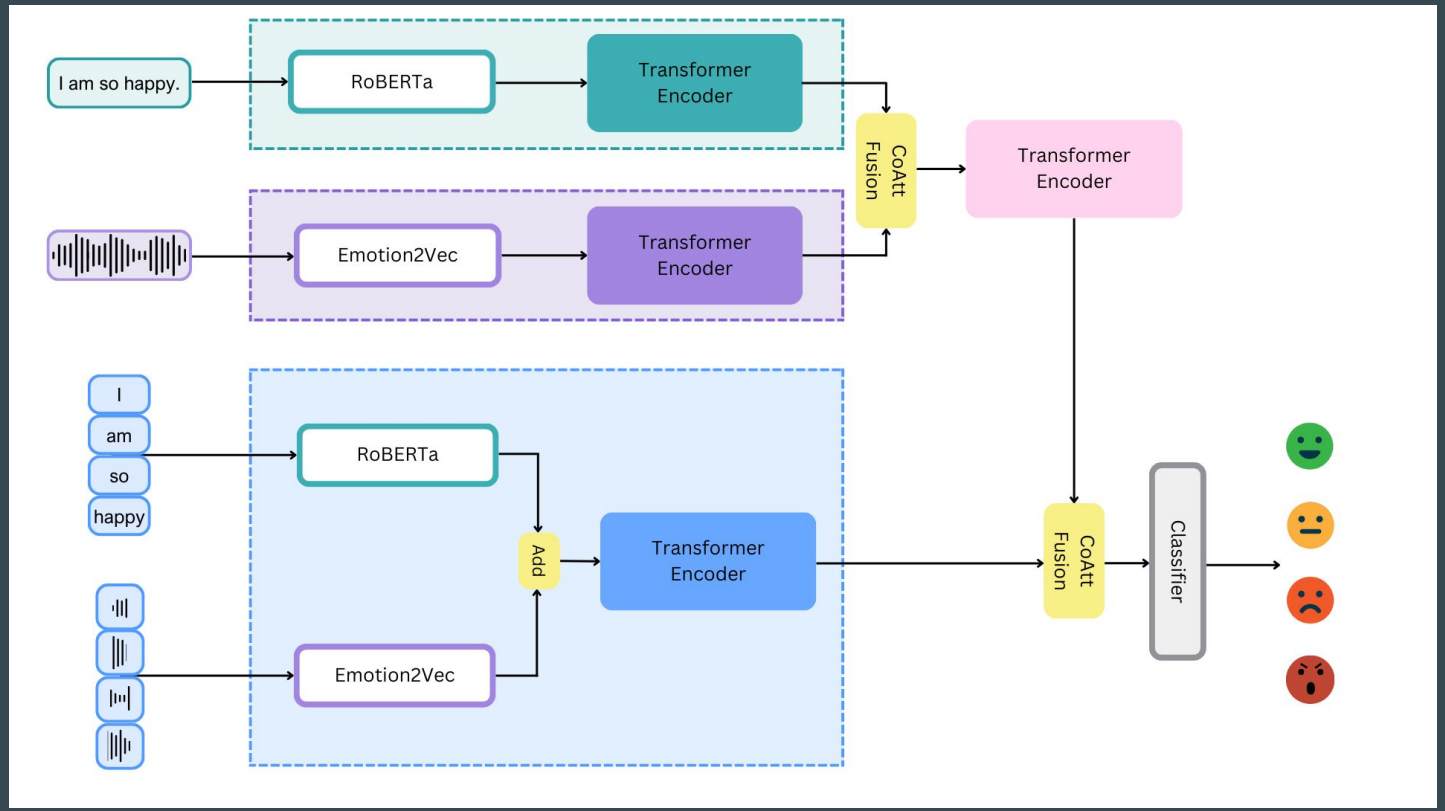
Emotion Recognition

What is Emotion Recognition?

Emotion Recognition is the process of identifying and analyzing human emotions from speech, facial expressions, text, or other signals. It helps detect feelings such as **happiness, sadness, anger, fear, or excitement.**



Emotion Recognition: Solution Overview



Emotion Recognition: Results

We tested proposed solution on two emotion recognition benchmarks: IEMOCAP and CMU-MOSI.

Approach	Year	IEMOCAP (%)
Sun et al.	2024	74.00
Lei et al.	2024	75.68
Wang et al.	2023	75.20
Mai et al.	2024	75.29
Zhao et al.	2024	75.70
Nguyen et al.	2024	77.22
ATENet		**78.35**

Approach	Year	CMU-MOSI (%)
Sun et al.	2020	-
Zhao et al.	2024	-
Wu et al.	2024	48.2
Lei et al.	2020	48.9
Yang et al.	2020	44.9
ATENet		**51.1**

Dialects TTS




Dialects TTS




Dialects TTS




Dialects TTS




Dialects TTS




ASR for Dialects: Results

- Western Armenian
- Lori
- Artsakh
- Mush

Model	WER	CER
Whisper Large v2	16.52±0.19	5.69±0.1
Whisper Large v3	**15.75±0.25**	**5.33±0.12**
SeamlessM4T v2	21.95±0.33	7.91±0.09
ArmSpeech	94.8	40.7
ASPRAM	79.3	27.2
NVIDIA-hy ASR	65.1	27.4
WAV	56.8	19.5
hispeech.ai	58.1	22.0
Chirp v2	63.3	26.9
Universal-1	86.5	42.1
Scribe v1	84.5	43.5

Thank you!