FOSTERLANG

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

Funded by the European Union

Automatic Speech Recognition for Minoritized Languages

By:

Dr. Arvind Kumar and Mr. Paul Trilsbeek
(Max Planck Institute for Psycholinguistics, The Netherlands)

Session Roadmap

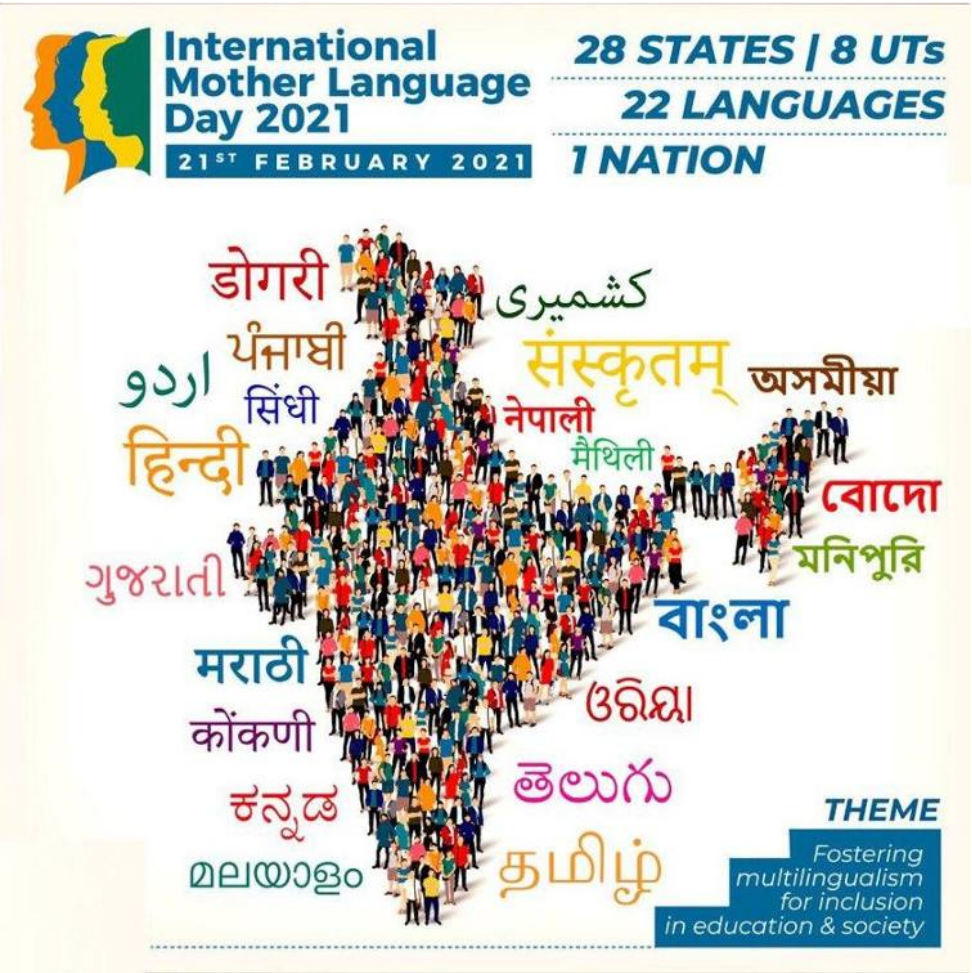
- ❑ Introduction & Motivation
- ❑ Mentimeter Interaction
- ❑ ASR Fundamentals
- ❑ Challenges in Low-Resource ASR
- ❑ Overview of Modern ASR Models
- ❑ Evaluation metrics
- ❑ Public Dataset
- ❑ Data Augmentation Techniques
- ❑ Steps to build ASR
- ❑ Hands on session
- ❑ Conclusion and direction for future work

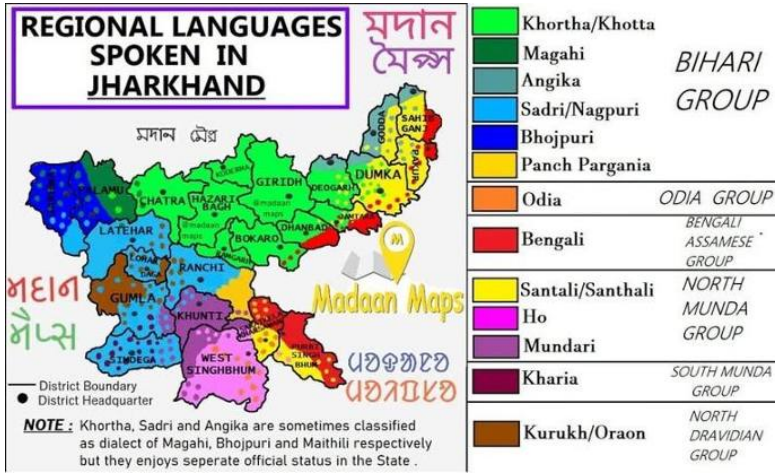
Section I: Introduction and Motivation

Our Team

Arvind Paul Caroline

My background





In India, there are close to 121 major languages and roughly **1,599** other languages and dialects

Mentimeter Poll

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

menti.com

8108 6214

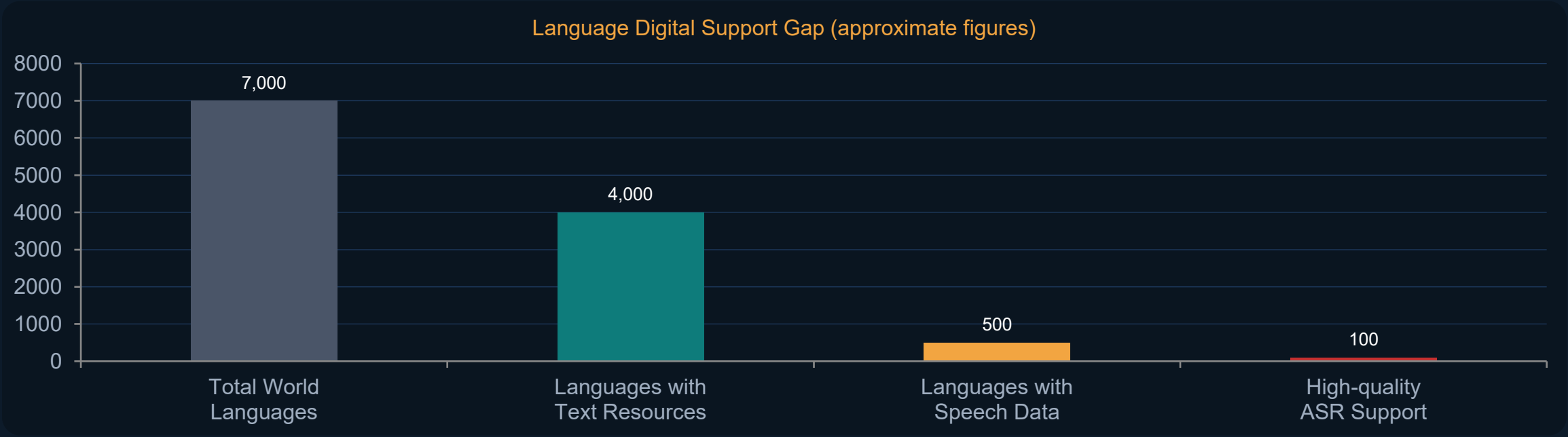
Icebreaker Poll

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

"How many languages currently have high-quality ASR systems?"

Less than 50	50 – 100	100 – 500	More than 1,000



MOTIVATION

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Why Speech Technology Matters

Speech is the Most Natural Human Interface

Voice Assistants

Accessibility

Auto Subtitles

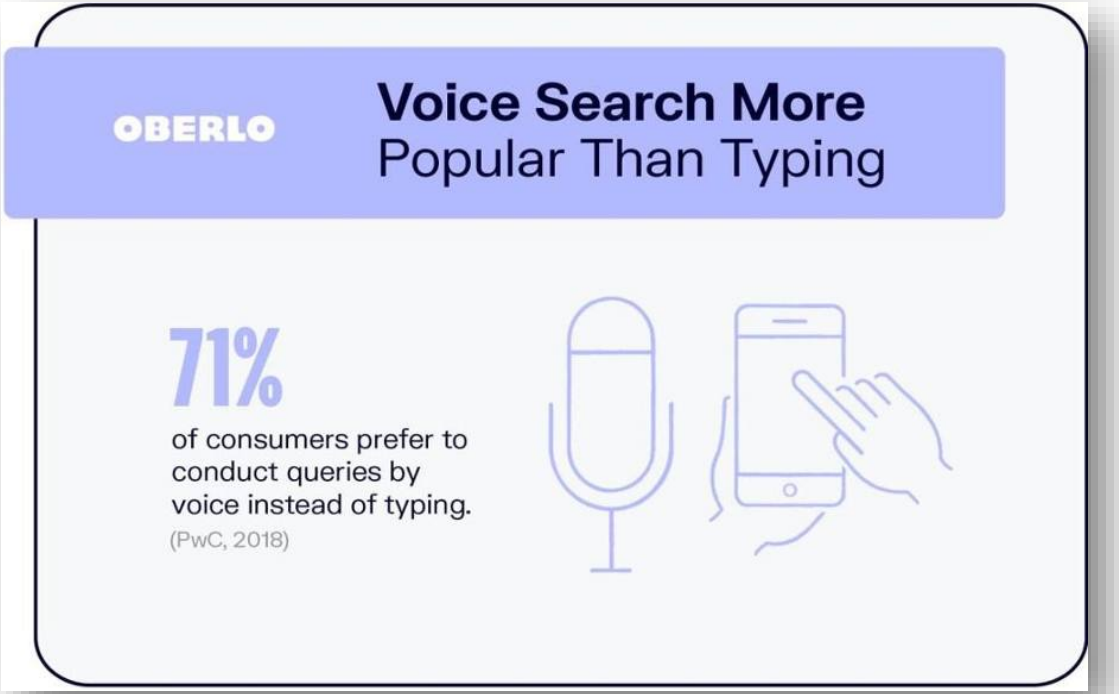
Healthcare Docs

Education

Call Centres

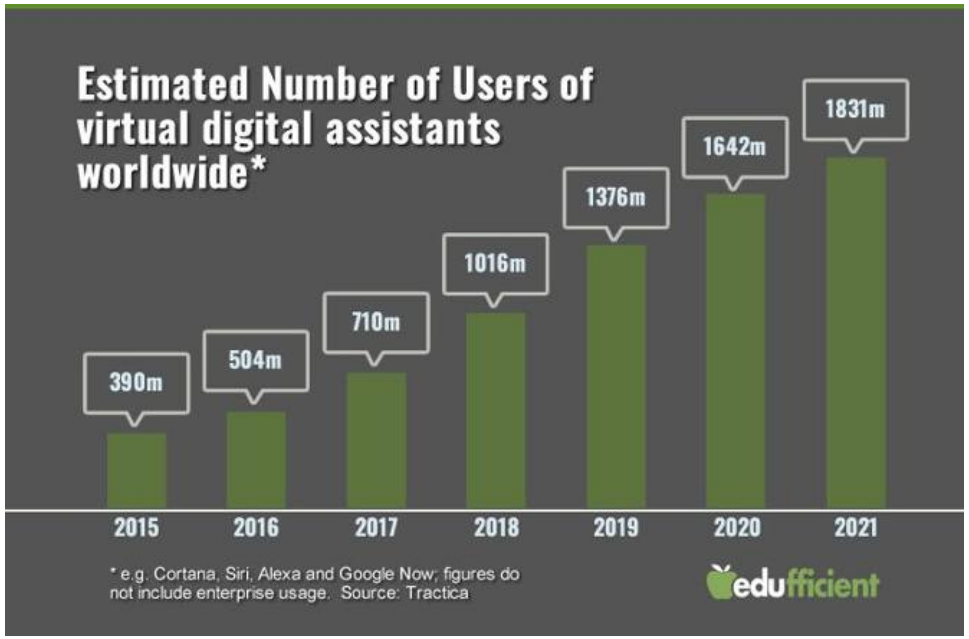
Smart Devices

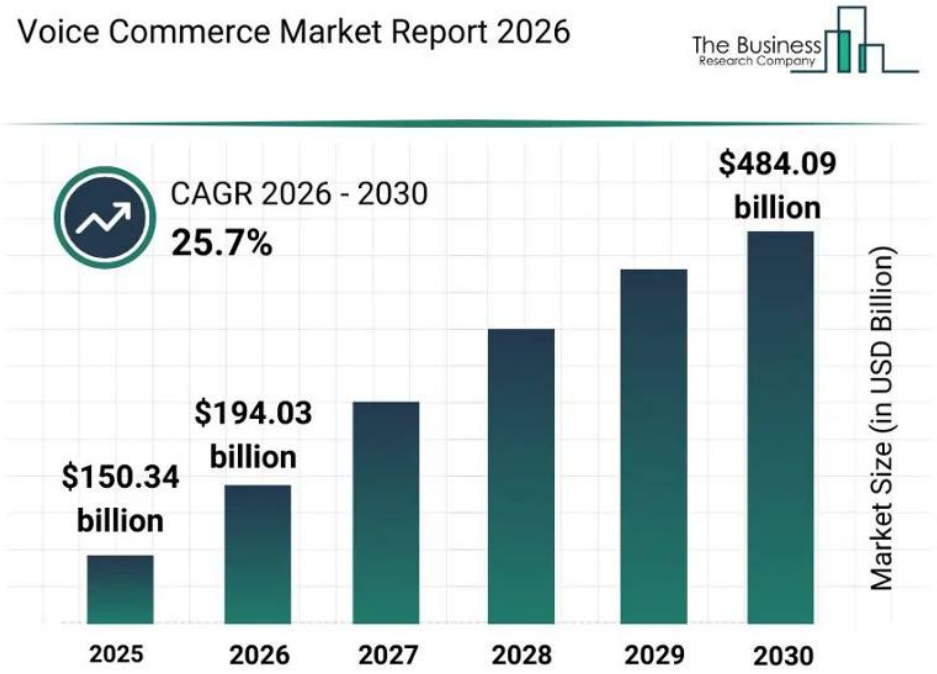
Why Speech Technology Matters



https://www.zionmarketresearch.com/report/speech-and-voice-recognition-technologies-market

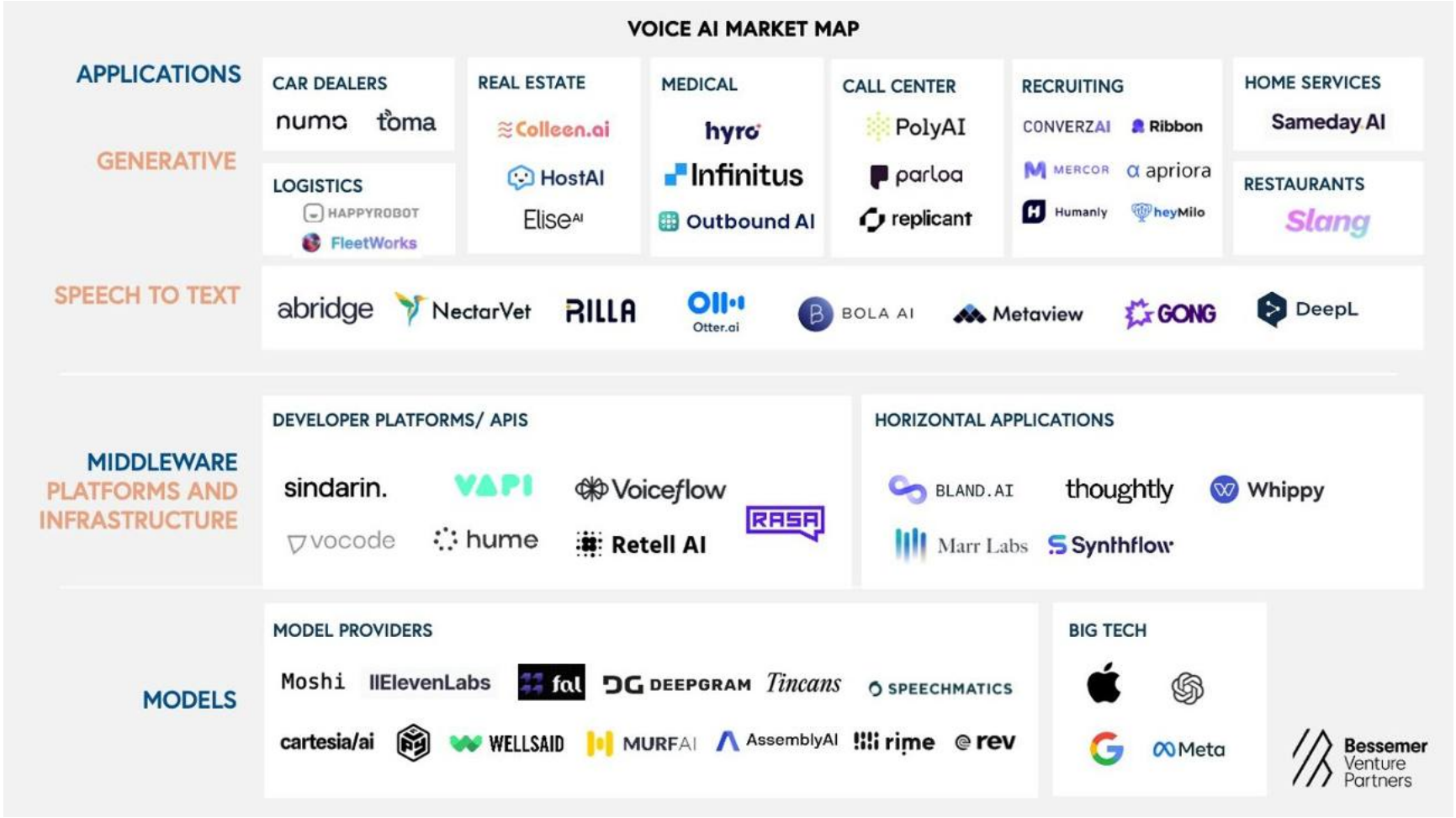
Why Speech Technology Matters





Source: https://voicebot.ai/2018/03/05/voice-shopping-reach-40-billion-u-s-5-billion-uk-2022/

Marketspace in Voice Tech



Open-Source Toolkits and Frameworks

1. NVIDIA NeMo
2. OpenAI Whisper
3. SpeechBrain
4. Kaldi
5. ESPnet
6. Vosk

Deep Learning Libraries

1. PyTorch
2. TensorFlow
3. Hugging Face Transformers

Cloud-Based and Managed Services

1. Amazon Transcribe
2. Google Cloud Speech-to-Text
3. Deepgram

Why Speech Technology is growing so fast?

1. Low-cost internet and smartphone penetration
2. Explosion of deep learning and transformer-based models
3. Availability of massive speech datasets
4. Growth in computational power
5. Open-source ecosystems such as Hugging Face
6. Illiterate populations(Developing Countries)
7. Faster interaction(Smart TV etc)
8. Hands-free control

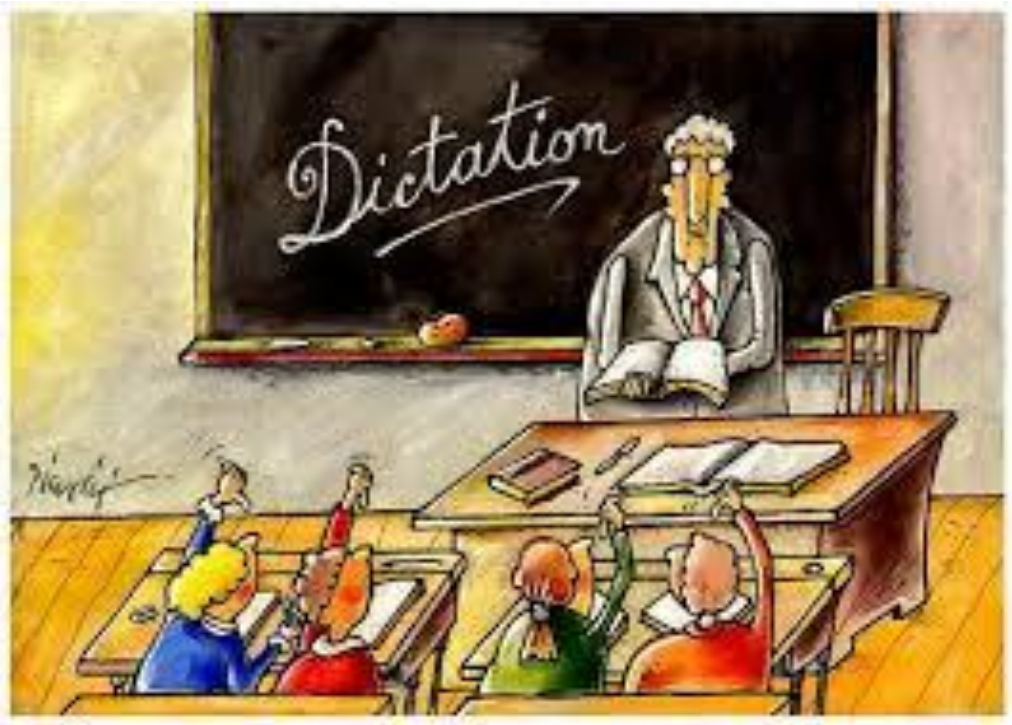
Speech Technology is the Future of HCI

How ASR Can Help Revitalize Minority Languages?

- ASR enables speakers to use their language in modern digital environments, increasing the language's relevance in daily life.
- Speech technology helps create digital presence for minority languages through subtitles, transcription, search, voice assistants, and educational tools.
- Automatic transcription allows oral traditions, folk stories, songs, and conversations to be documented and archived efficiently.
- ASR reduces the effort required to create written resources for historically oral languages.
- Educational applications can support language learning for younger generations through pronunciation feedback and interactive speech-based learning.
- Elder speakers' recordings can be preserved and transcribed before linguistic knowledge is lost.
- Voice technologies increase visibility and prestige of minority languages

Section II:
Fundamentals of ASR

What is Automatic Speech Recognition?



Convert Voice to Text

Applications:

- **Voice assistants**
- **Dictation systems**
- **Call center analytics**
- **Accessibility**
- **Any many more…**

Evolution of ASR: Rule-Based to Foundation Models

1950s–1980s	1990s	2010s	2018+	Today
Rule-Based & Template Matching (DTW)	Hidden Markov Models (HMMs)	Deep Neural Networks (DNNs)	Transformer-Based (End to End) Models	Self-Supervised Foundation Models

Key Milestones

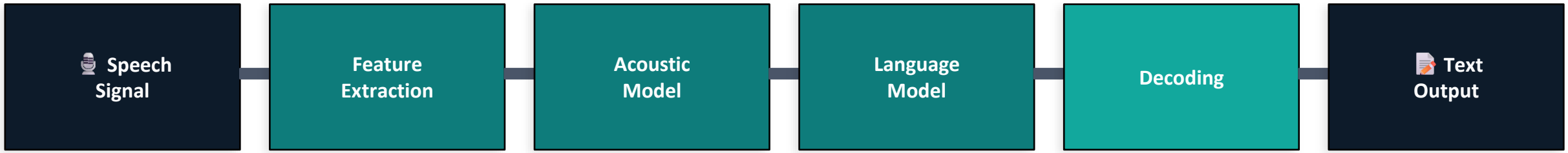
Kaldi | **DeepSpeech** | **wav2vec 2.0** | **Whisper** | **MMS (Meta)** | **SeamlessM4T**

How Automatic Speech Recognition Works?

FOUNDATIONS

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG



Example Audio: → **Output Text:** "Pedrok doitasunez susenduzuen luskadeko orkestra sinfonikoa”

Acoustic Modelling

Maps audio features (MFCCs, filterbanks) to phonemes or sub-words using neural networks.

Pronunciation Modelling

Handles the mapping between written words and their spoken phoneme sequences.

Language Modelling

Predicts likely word sequences using statistical or neural language models.

End-to-End ASR

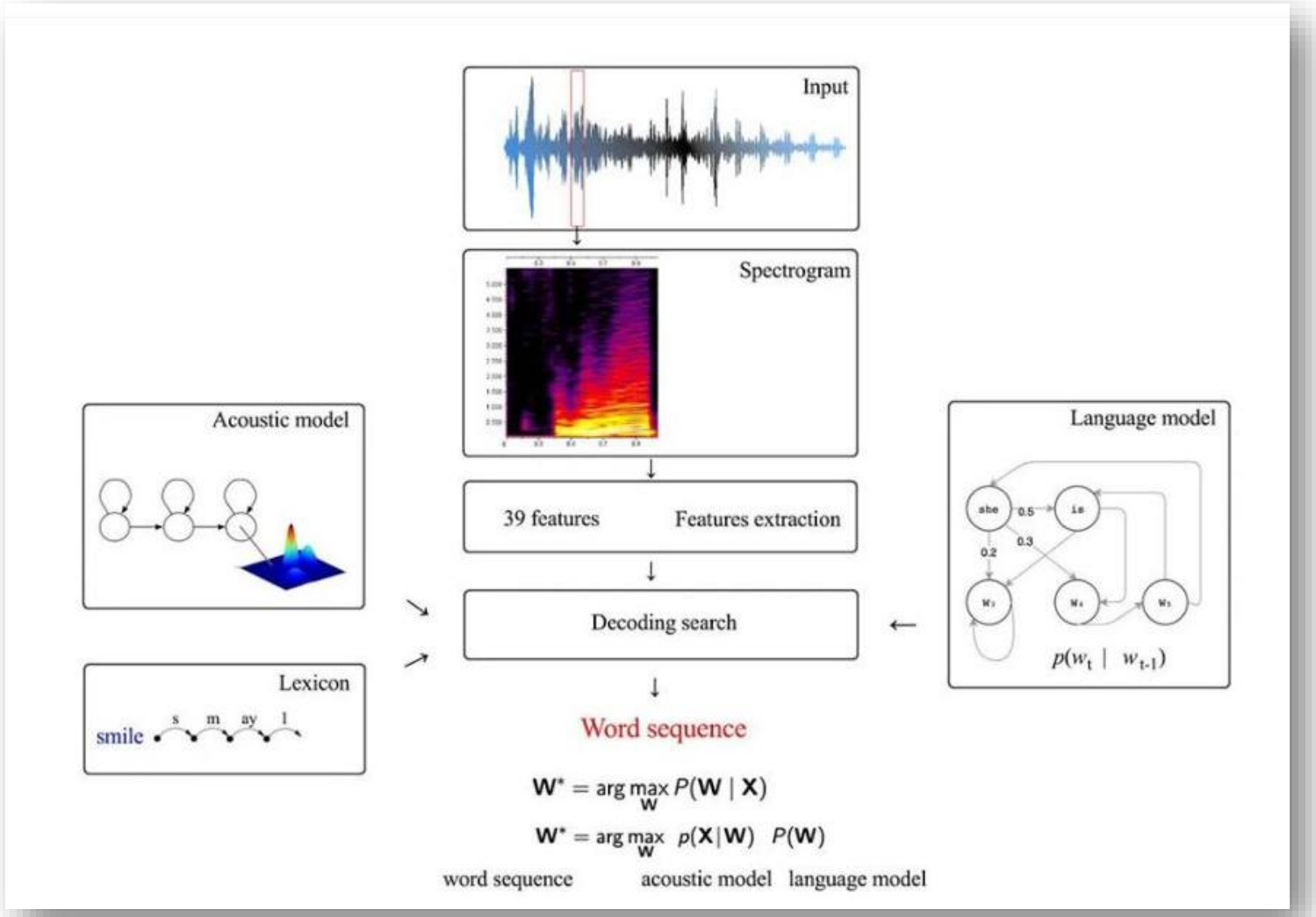
Modern DNNs unify acoustic + language models into a single trainable system.

How Automatic Speech Recognition Works?

FOUNDATIONS

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG



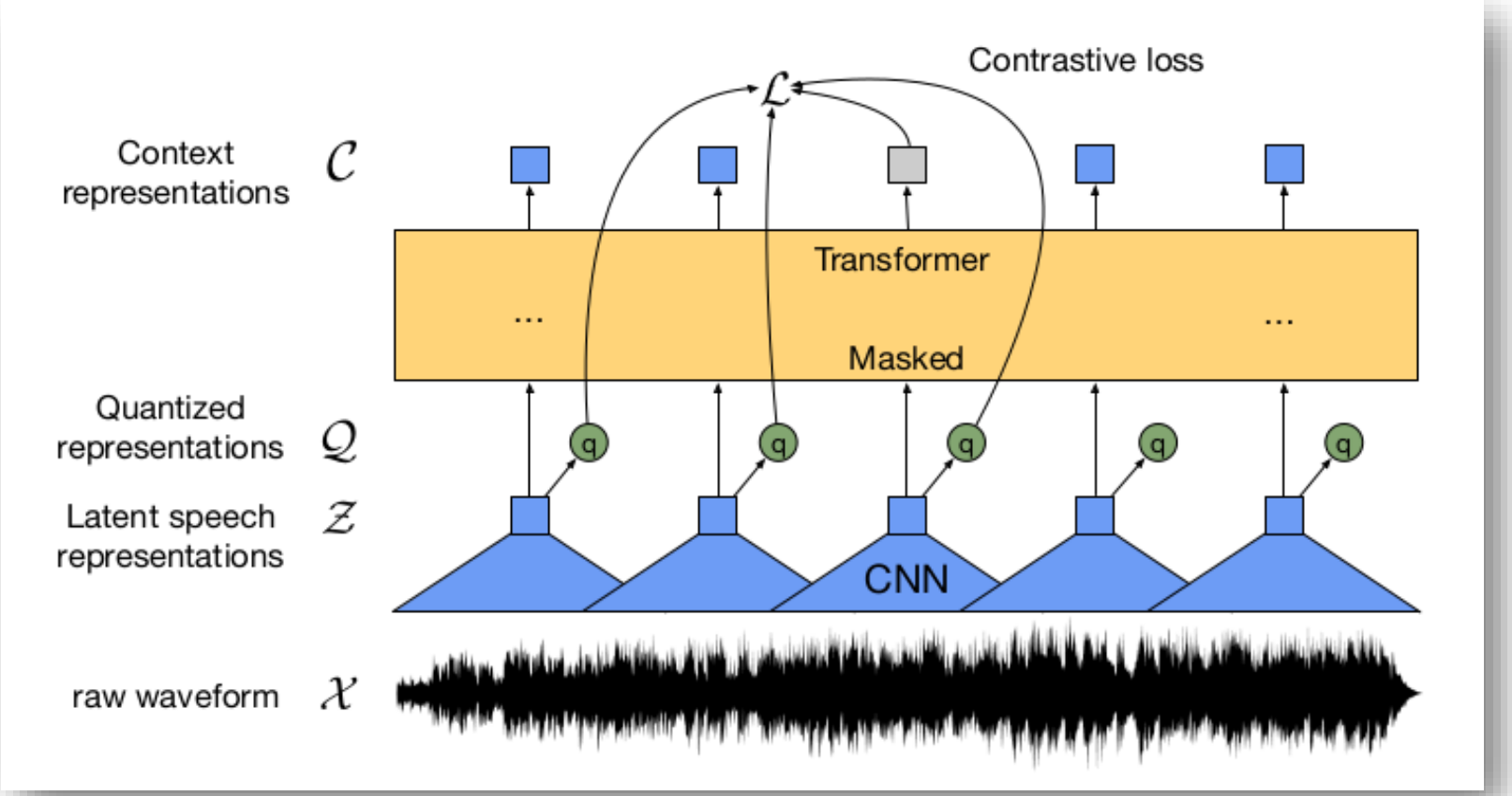
The primary objective of speech recognition is to build a statistical model to infer the text sequences ***W*** **(say "cat sits on a mat") from a sequence of feature vectors *X*.**

The distribution of features for a phone can be modeled with a **Gaussian Mixture Model (GMM).**

The transition between phones and the corresponding observable can be modeled with the **Hidden Markov Model (HMM).**

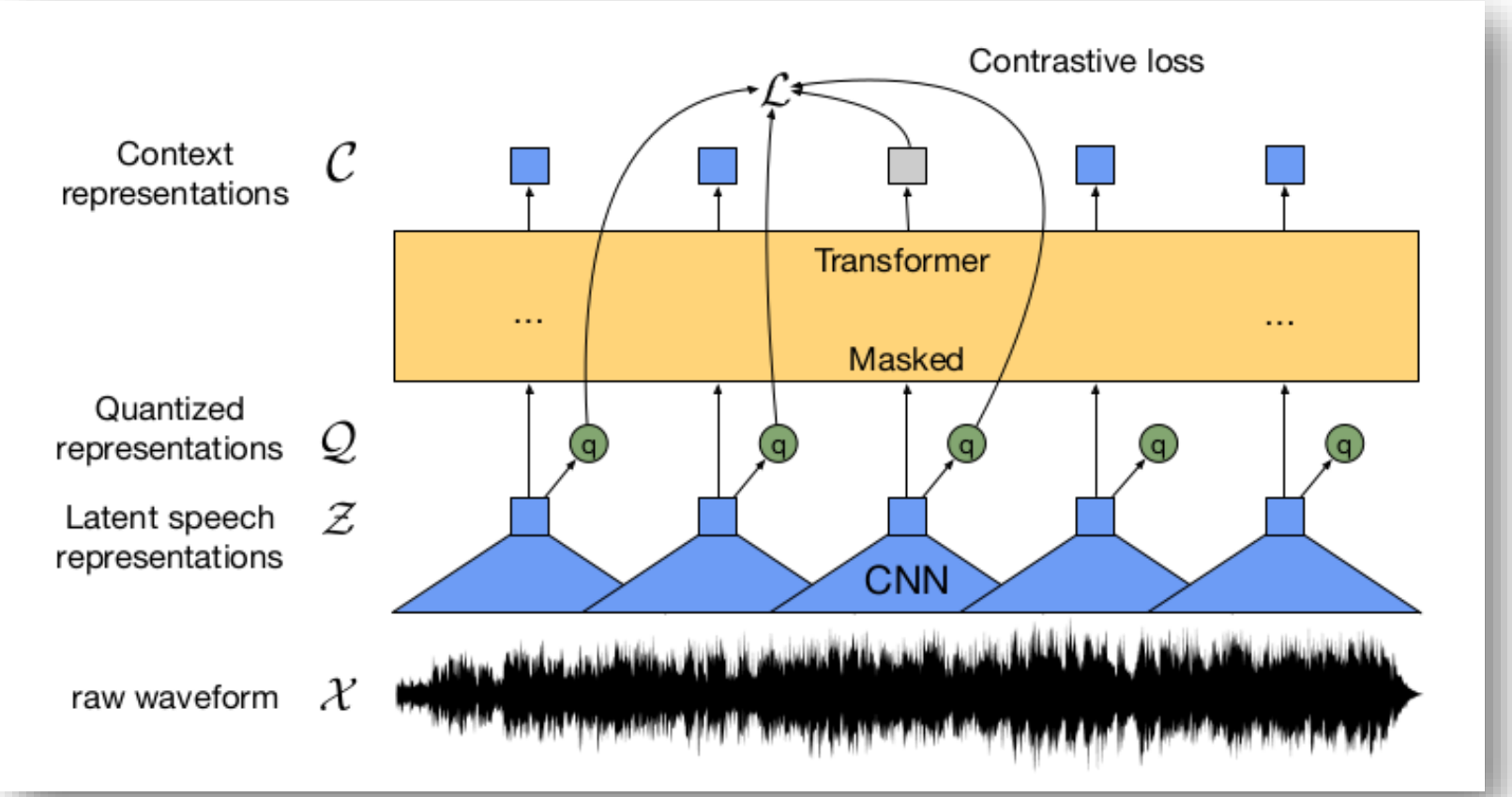
Combining information on the lexicon, the acoustic model and the language model, we can find the optimal phone sequence with the **Viterbi decoder.**

Working of End-to-End ASR



Deep learning models like
- Wav2Vec 2.0
- OpenAI Whisper

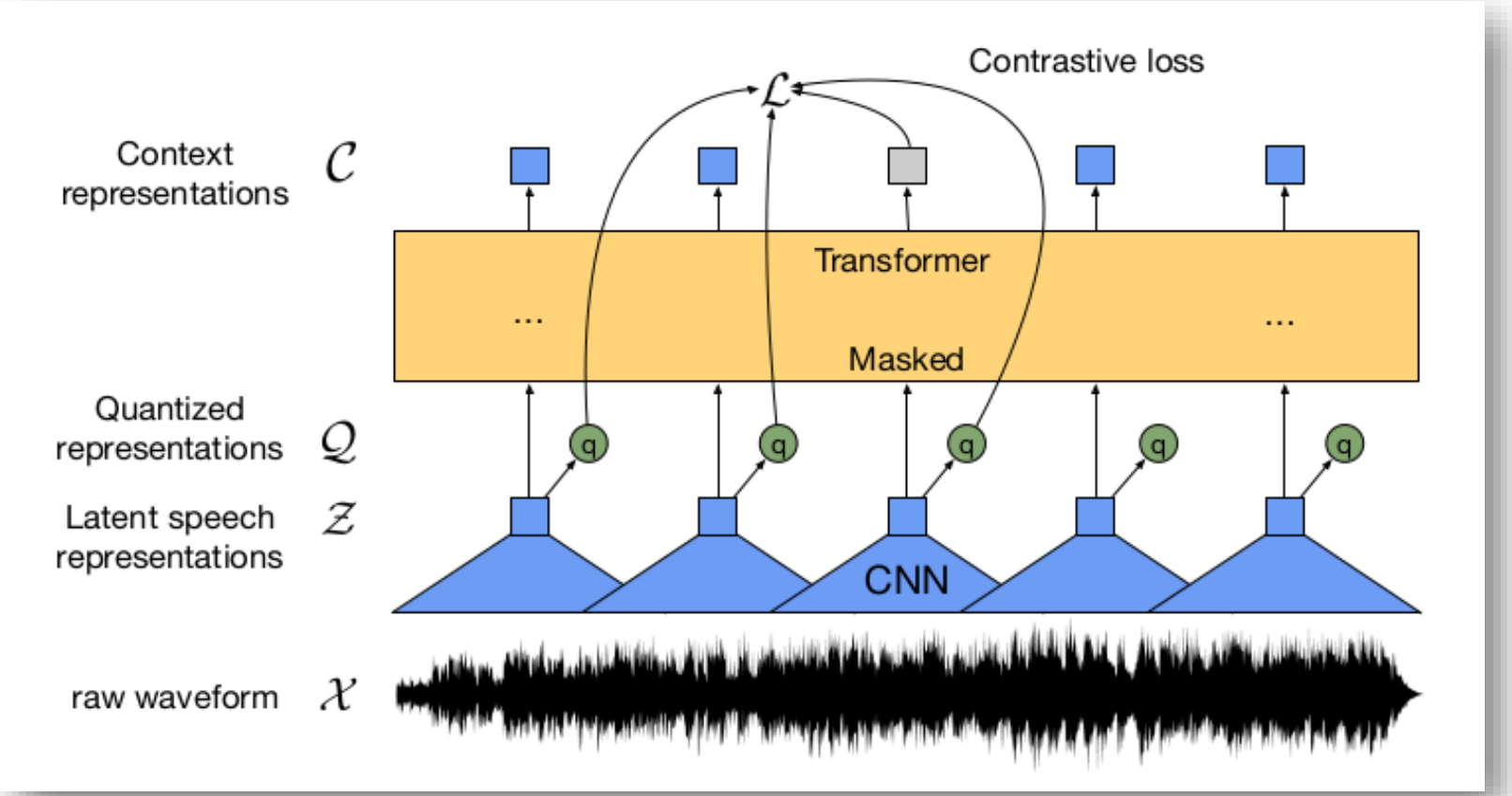
Step-wise Explanation



Step 1: Raw Speech Waveform
Continuous time signal sampled at **16 kHz**

Step 2: Feature Encoder
A stack of **1-D convolution layers** extracts low-level acoustic patterns and captures **local temporal patterns and** produce **latent speech representation**

Step-wise Explanation

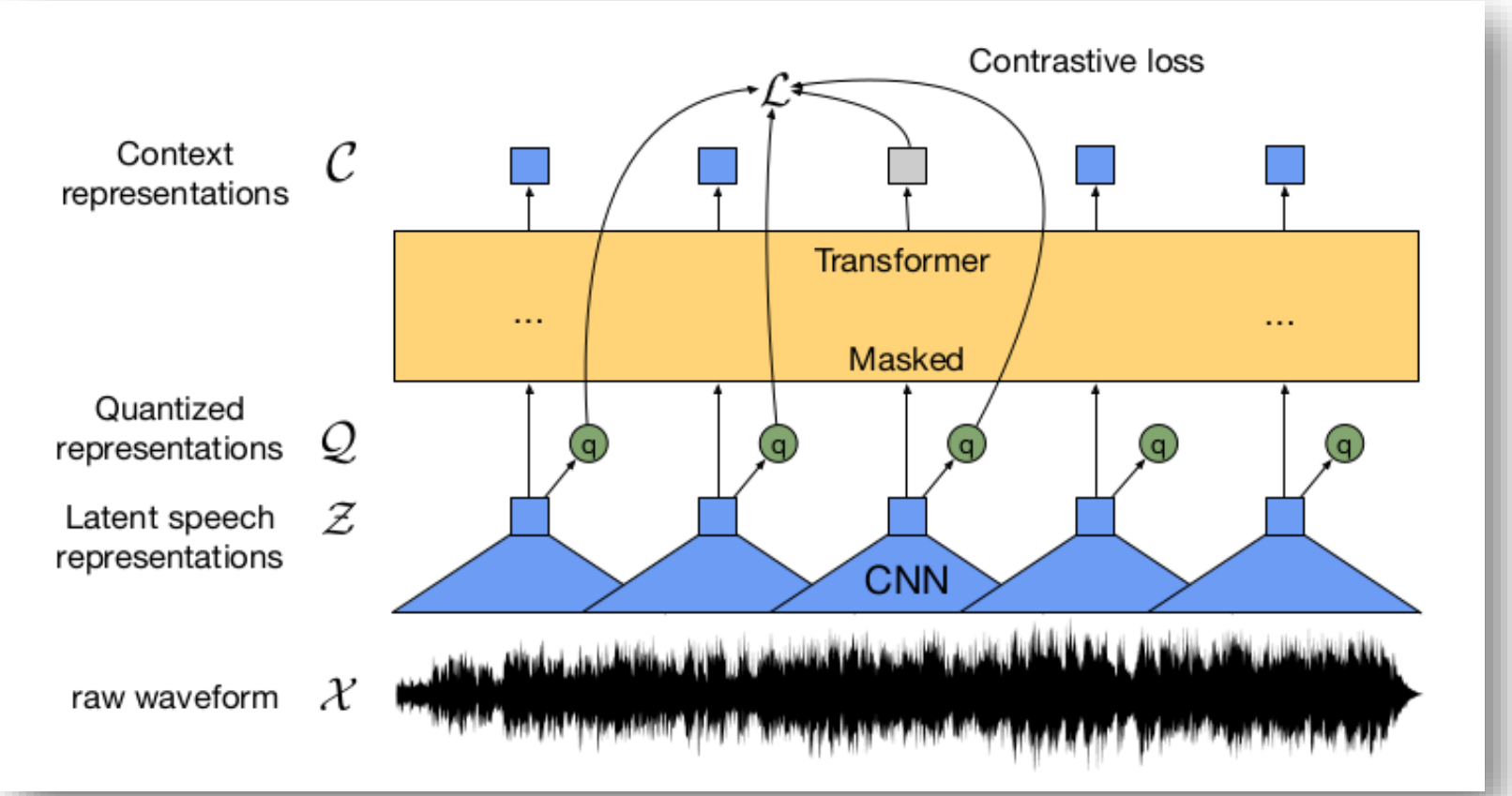


Step 3: **Latent Speech Representation**

- Dense continuous vectors representing speech frames.
- Encodes phonetic information

Step 4: **Vector Quantization**
Convert continuous speech embeddings into **discrete units**

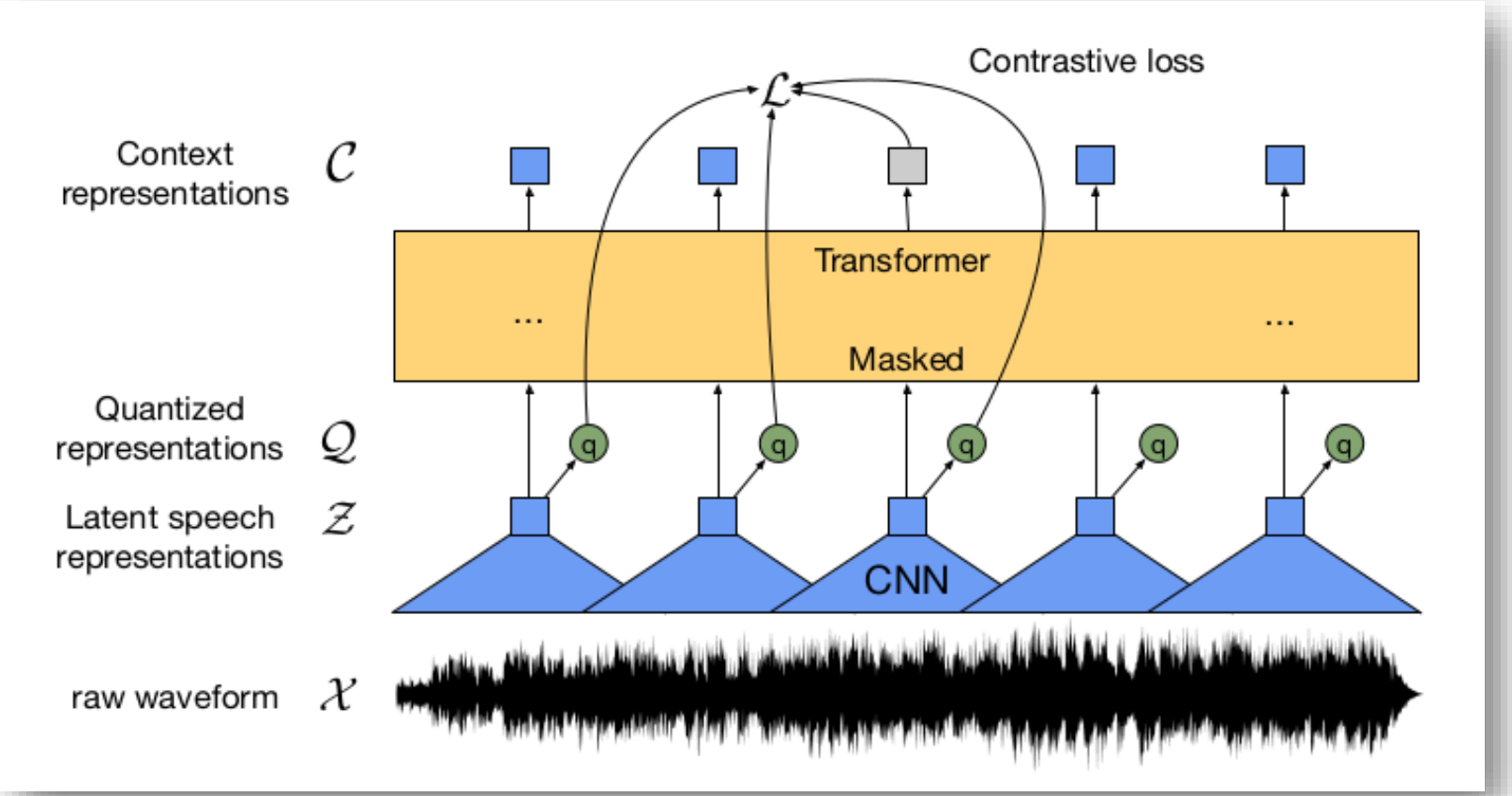
Step-wise Explanation



Step 5: **Quantized Speech Representation**
Discrete tokens representing speech units. These behave like **pseudo-phonemes** or **acoustic units**.

Step 6: **Context Network**
Usually a **Transformer encoder which** Capture **long-range dependencies**

Step-wise Explanation



Step 7: **Prediction / ASR Head**
Final module converts contextual embeddings to text.

Step 8: **Text Output**
Usually a **Transformer encoder which** Capture **long-range dependencies**

Why building ASR is complex? Interdisciplinary Approach

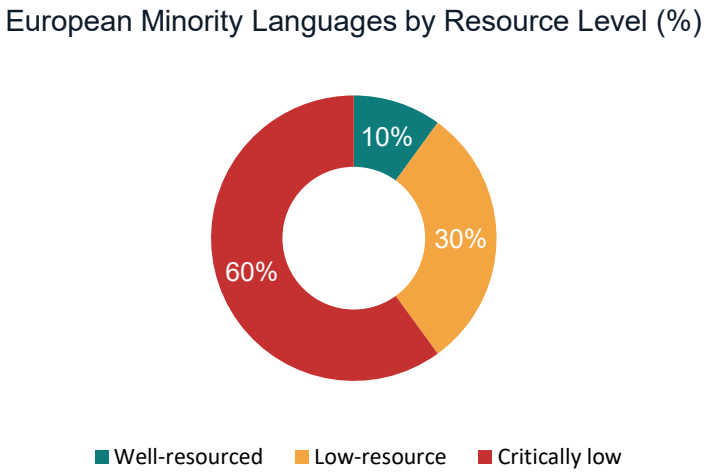
1. **Signal Processing**
2. **Physics (Acoustics)**
3. **Pattern Recognition**
4. **Communication and Information Theory**
5. **Linguistics**
6. **Physiology**
7. **Computer Science**
8. **Psychology**

Source: Fundamentals of Speech Recognition

What Makes a Language "Low-Resource"?

Characteristics

- Limited transcribed speech data
- Small or noisy text corpora
- Few or no NLP resources
- Limited funding & community support
- Lack of standard orthography



Examples

Basque

Breton

Galician

Occitan

Welsh

Frisian

Many indigenous and regional languages worldwide face the same challenges — from Māori to Sámi, Tibetan to Quechua. Digital exclusion amplifies existing inequalities.

Examples of Low Resource Language

Spain

- Amazigh (in Melilla)
- Aragonese
- Aranese / Occitan (in Catalonia)
- Asturian (in Asturias, parts of León, Zamora, Salamanca, Cantabria, Extremadura)
- Basque (in Euskadi and Navarre)
- Catalan (in Catalonia, Balearic Islands, Valencia (Valencian), and part of Aragon)
- Darija (in Ceuta)
- Extremaduran (in Extremadura)
- Fala (in Extremadura)
- Galician (in Galicia, parts of Asturias, León, Zamora)
- Leonese
- Portuguese (in Extremadura)

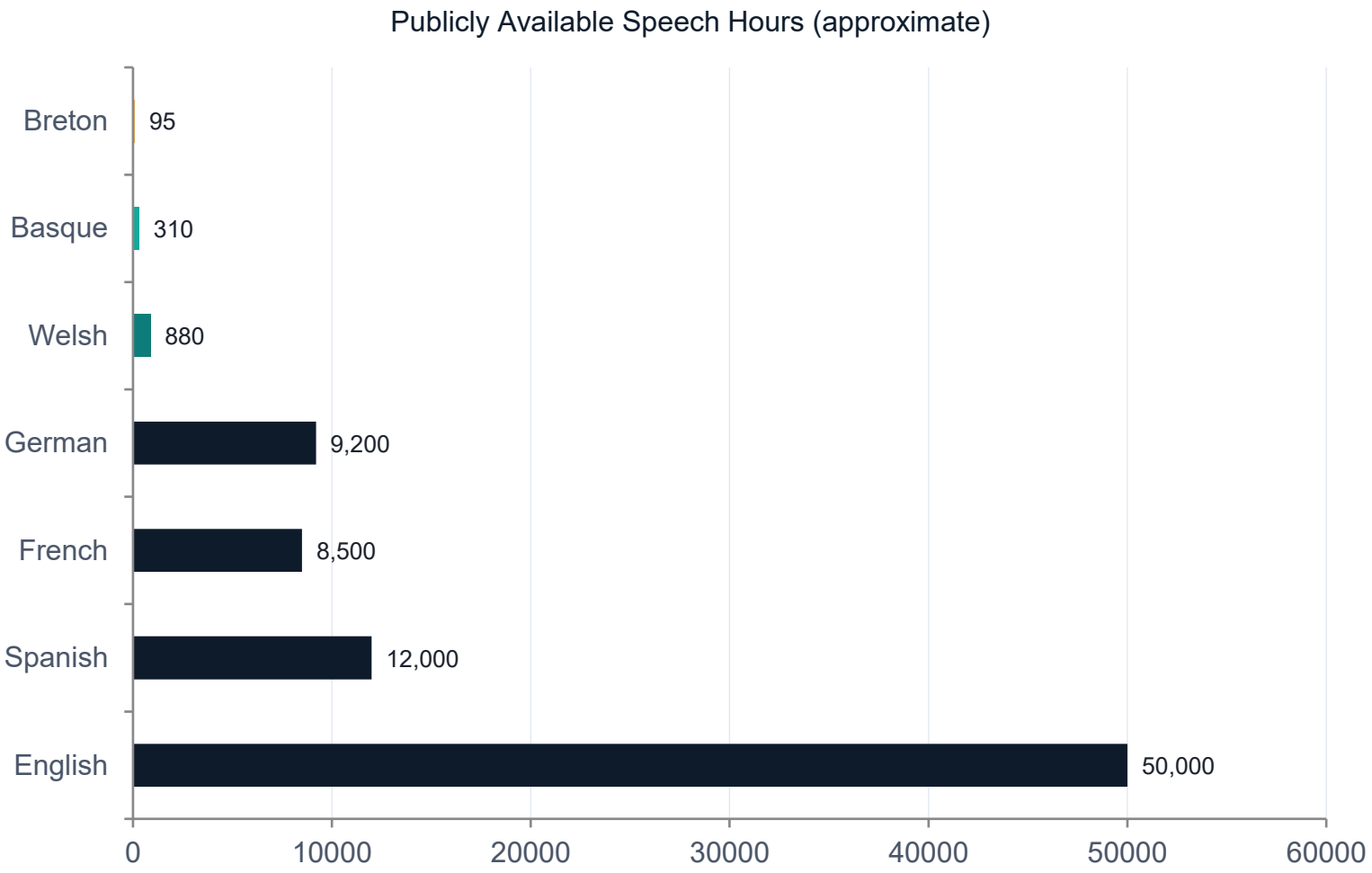
Source: https://fosterlang.al.uw.edu.pl/languages-map/language-statistics/

DATA IMBALANCE

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Most Languages Have Almost No Digital Speech Data



Key Message

Modern ASR systems are data-hungry. English has >500× more data than Breton.

Discussion

Can multilingual learning reduce the dependency on huge per-language datasets?

CHALLENGES

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Why is ASR Difficult for Minoritised Languages?

⚙ Technical Challenges

- Insufficient labelled data
- Code-switching
- Dialectal variation
- Background noise
- Pronunciation diversity
- Lack of text corpora
- Limited compute resources

🌍 Social & Ethical Challenges

- Language preservation ethics
- Community participation
- Ethical AI & accountability
- Data ownership & consent

⚠️ *Remember: Technology should empower communities, not replace them.*

CHALLENGES

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Code-Switching: Multilingual Reality in Daily Speech

1. **Today meeting ke baad we will submit the report (English-Hindi).**
2. Gaur klasean profeak dijo que el proyecto mañana entregatu behar dugu.(Basque-Spanish)

English Hindi

Mixed Vocabularies

Code-switched speech draws from two or more language lexicons simultaneously, creating out-of-vocabulary challenges.

Pronunciation Variation

Phoneme inventories differ across languages; switched words may be pronounced with L1 or L2 phonology.

Rapid Transitions

Speakers switch mid-sentence or even mid-word, making language boundary detection extremely difficult for models.

Common code-switching pairs: Basque–Spanish · Hindi–English · Catalan–Spanish · Welsh–English

Traditional vs End-to-End ASR

Aspect	Traditional ASR	End-to-End ASR
Architecture	Multiple separate components	Unified single architecture
Pronunciation	Dictionary required	Learned directly from data
Complexity	Complex training pipelines	Simpler end-to-end training
Multilingual	Hard to scale across languages	Better cross-lingual adaptation
Low-resource	Struggles without enough data	Transfer learning helps greatly
Interpretability	Easier to inspect each module	Black-box, harder to debug

Key End-to-End Architectures

CTC

RNN-Transducer

Transformer

Encoder-Decoder (Seq2Seq)

STATE OF THE ART

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Self-Supervised Learning: Learning from Unlabelled Audio

Key Idea: Train on massive amounts of raw audio without transcripts → dramatically reduces labelled data requirements.

wav2vec 2.0 (Meta AI)

Quantised contrastive self-supervised learning on raw audio.

HuBERT (Meta AI)

Offline clustering of hidden units as pseudo-labels for SSL training.

XLS-R (Meta AI)

53-language cross-lingual extension of wav2vec 2.0 (1B params).

Whisper (OpenAI)

Weakly supervised on 680K hrs; strong zero-shot multilingual ASR.

STATE OF THE ART

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Multilingual ASR: One Model, Many Languages

Whisper

99 languages

Zero-shot multilingual transcription & translation

XLS-R

128 languages

Best-in-class low-resource fine-tuning via transfer, 436,000 hours of speech data

SeamlessM4T

100+ languages

Speech-to-speech & speech-to-text in one model

MMS (Meta)

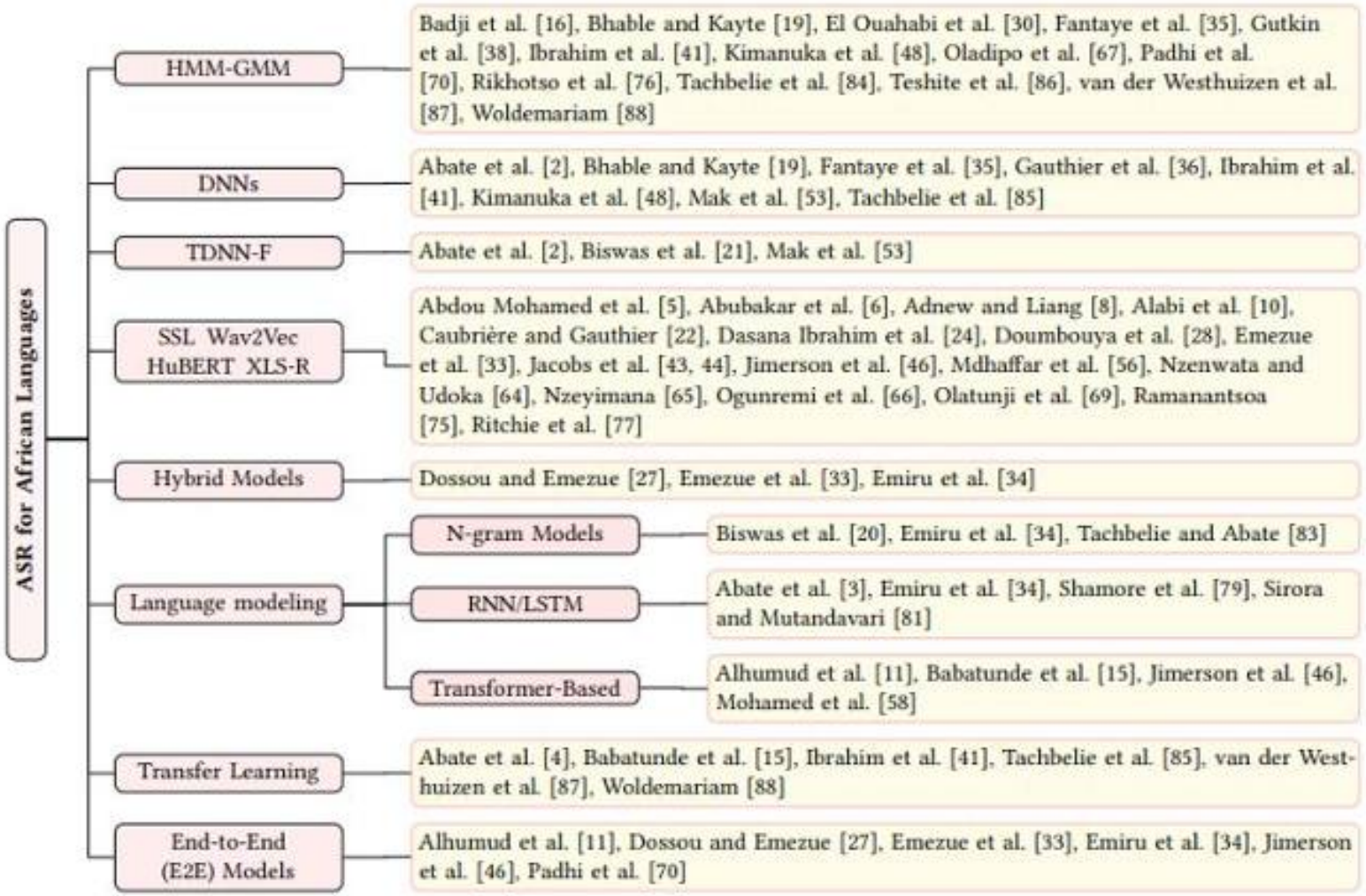
1,100+ languages

Breakthrough coverage of endangered languages

Advantages of Multilingual Approaches

- Cross-lingual transfer learning enables knowledge sharing across related languages
- Shared phonetic representations reduce per-language data requirements significantly
- A single model is cheaper to maintain than dozens of monolingual models

Model Architectures



Source: Automatic Speech Recognition (ASR) for African Low-Resource Languages: A Systematic Literature Review

STATE OF THE ART

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

OpenAI Whisper: Architecture & Capabilities

Large-Scale Training

Trained on 680,000 hours of multilingual and multilingual web-collected audio, covering 99 languages.

Noise Robustness

Robust to accents, background noise, technical vocabulary — trained on diverse in-the-wild audio.

Multitask

Performs transcription, language ID, and translation in a single model with task tokens.

Easy Deployment

Available in Tiny to Large-V3 variants; excellent zero-shot performance without fine-tuning.

⚠️ **Limitations:** Bias toward high-resource languages · Computationally heavy for edge deployment · Hallucinations on silent audio · Poor performance on truly endangered languages not well-represented in training data.

WER: How ASR Performance is Measured

Word Error Rate

$$WER = (S + D + I) \div N$$

S = Substitutions D = Deletions I = Insertions N = Total Reference Words

Example

Reference: "I love speech recognition"

Prediction: "I love speech **recognizer**"

WER = 1/4 = 25% (1 substitution)

2. Character Error Rate (CER)

Used especially for:

- morphologically rich languages
- Asian languages
- low-resource languages

Formula:

$$CER = \frac{S+D+I}{N_{char}}$$

Measures errors at character level instead of word level.

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Reported accuracies for few Low Resource Languages

Language	Best WER (%)	Model / Approach	Source
Catalan	≤5%	Fine-tuned Whisper + LLM (Whisper-LM)	de Zuazo et al., 2025
Basque	~5–10%	Whisper + LLM integration	de Zuazo et al., 2025
Galician	~5–10%	Whisper + LLM integration	de Zuazo et al., 2025
Welsh	~10–15%	Fine-tuned wav2vec 2.0 / XLS-R	Jones, 2022; community models
Maltese	~15–20%	Fine-tuned XLS-R 2B	Williams et al., SIGUL 2023
Irish (Gaelic)	**10.9%** (domain-matched) / **30.65%** (open benchmark)	ABAIR Fotheidil system (best); omniASR 7B (best open model)	Lonergan et al. 2025; BlasBench 2026
Scottish Gaelic	~26% (older); improved ~18% rel. with hybrid HMM+SSL	Hybrid HMM + self-supervised pretraining	Evans et al. 2022; Klejch et al. 2025

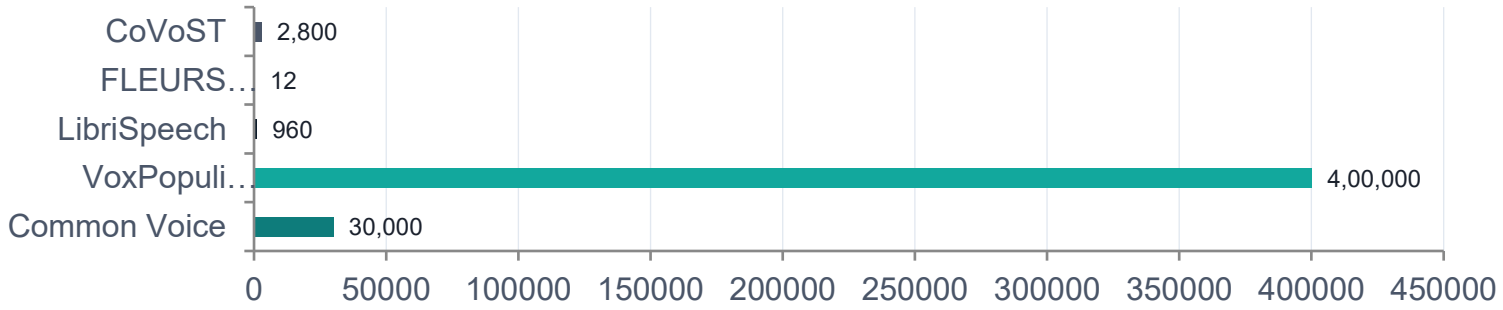
RESOURCES

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Public Datasets for ASR Research

Dataset	Languages	Size / Scale	Primary Use
Mozilla Common Voice	100+ incl. minority	30,000+ hrs total	Speech collection & training
FLEURS	102 languages	~12 hrs / language	Evaluation benchmark
VoxPopuli	23 European (European Parliament)	400K hrs unlabelled	Self-supervised pretraining
LibriSpeech	English	960 hrs	English benchmarking
CoVoST 2	21 languages	2880 hrs (Multilingual)	Speech translation



💡 Open datasets democratise speech research — enabling universities and community groups to build ASR without proprietary data.

METHODOLOGY

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Data Augmentation: Expanding Small Datasets

Noise Injection

Add background noise (street, crowd, music) to simulate real-world recording conditions.

Speed Perturbation

Stretch or compress audio by 0.9×–1.1× to simulate different speaking rates.

Pitch Shifting

Raise or lower pitch slightly without changing tempo to create speaker variety.

SpecAugment

Mask random frequency bands and time steps in the spectrogram during training.

Synthetic Speech (TTS)

Generate new training utterances from text using text-to-speech for the target language.

Cross-lingual Transfer

Borrow data from acoustically similar languages to supplement low-resource training.

Ethical & Social Dimensions of ASR for Minoritised Languages

? Who owns the speech data?

Native speakers and communities — not corporations. Data governance frameworks must be co-designed.

? Can AI help language preservation?

Yes — but as a tool, not a replacement. Technology should support revitalisation, not create digital artifacts of dying languages.

? How do we avoid bias?

Diverse speaker representation, dialectal coverage, and iterative community feedback loops are essential.

? What about consent and privacy?

Explicit informed consent, right to withdraw data, and community-level approval mechanisms must be standard practice.

"Technology should empower communities, not replace them."

RESEARCH TRENDS

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Where is ASR Research Going?

Large Multilingual Foundation Models

Models trained across hundreds of languages simultaneously, enabling universal speech understanding.

Speech Translation & Speech-to-Speech

Seamless cross-lingual communication without intermediate text — SeamlessM4T leads this direction.

Zero-Shot ASR

Adapting to entirely new languages using only textual or phonological descriptions.

On-Device / Edge ASR

Efficient quantised models that run entirely on mobile devices — privacy-preserving and offline-capable.

Federated Learning

Training across distributed devices without centralising sensitive community speech data.

Low-Power Deployment

ASR for embedded systems in low-connectivity regions where minority languages are often spoken.

BONUS

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Where is ASR Research Going(cont.)?

ASR for Dialects

Sub-regional variety modelling within a single language.

Code-Switching ASR

Joint multilingual decoding and language boundary detection.

Accent Adaptation

Online speaker adaptation to unseen accents.

Speech for Healthcare

Clinical ASR for documentation in minority language hospitals.

Multimodal Speech Systems

Audio + video lip reading for noisy environments.

Accessibility Applications

Real-time captions and voice interfaces for underserved communities.

Low-resource ASR remains an open, impactful, and under-explored research area — your contribution matters.

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Steps to build ASR for your minority language

- Collect speech recordings in the target language
- Prepare accurate text transcriptions
- Ensure audio quality and remove noisy/corrupted files
- Standardize text normalization (punctuation, spelling, script consistency)
- Split dataset into training, validation, and test sets (optional)
- Choose a pretrained multilingual ASR model (Example: Whisper, XLS-R)
- Select or build an appropriate tokenizer
- Convert audio into required sampling rate (commonly 16 kHz)
- Extract input features or use raw waveform input
- Map speech audio to tokenized text labels
- Configure fine-tuning parameters
- Train the model on minority-language speech data
- Monitor validation loss and evaluation metrics
- Evaluate using WER and/or CER

Low-resource ASR remains an open, impactful, and under-explored research area — your contribution matters.

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Steps to build ASR for your minority language(contd.)

- Analyse common recognition errors
- Perform data augmentation if dataset is small
- Fine-tune further using additional domain-specific speech
- Test robustness on unseen speakers and environments
- Deploy model for inference or real-world applications
- Continuously improve using community feedback and new speech data

Low-resource ASR remains an open, impactful, and under-explored research area — your contribution matters.

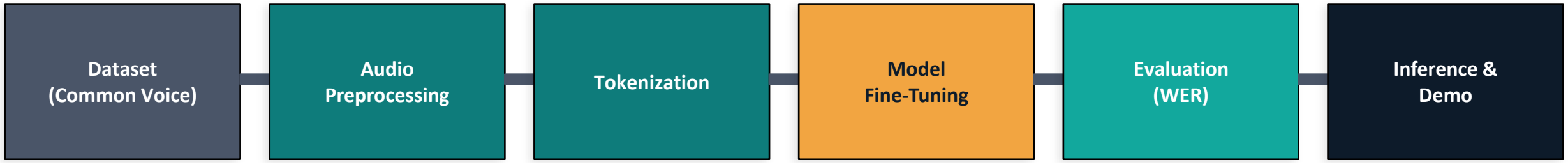
Section III:
Handson Session

HANDS-ON

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Fine-Tuning Workflow Overview



Rest of the sessions will be conducted on Google Colab !!

1. **ASR using Whisper and Wav2Vec2**
2. **Fine-Tune XLSR-Wav2Vec2 using Turkish Commonvoice Dataset**

Shared document for codes:
https://drive.google.com/drive/folders/1OdUf0gRG-AIjmqWCG1sCYDL75r3QQqb6?usp=sharing

Key Takeaways

01 Speech is the most natural human interface — languages without ASR face real digital exclusion and loss of cultural participation.

02 Minoritised languages face a severe data gap, but self-supervised models like wav2vec 2.0 dramatically reduce labelled data requirements.

03 Multilingual foundation models (Whisper, XLS-R, MMS) have transformed what is possible for low-resource languages.

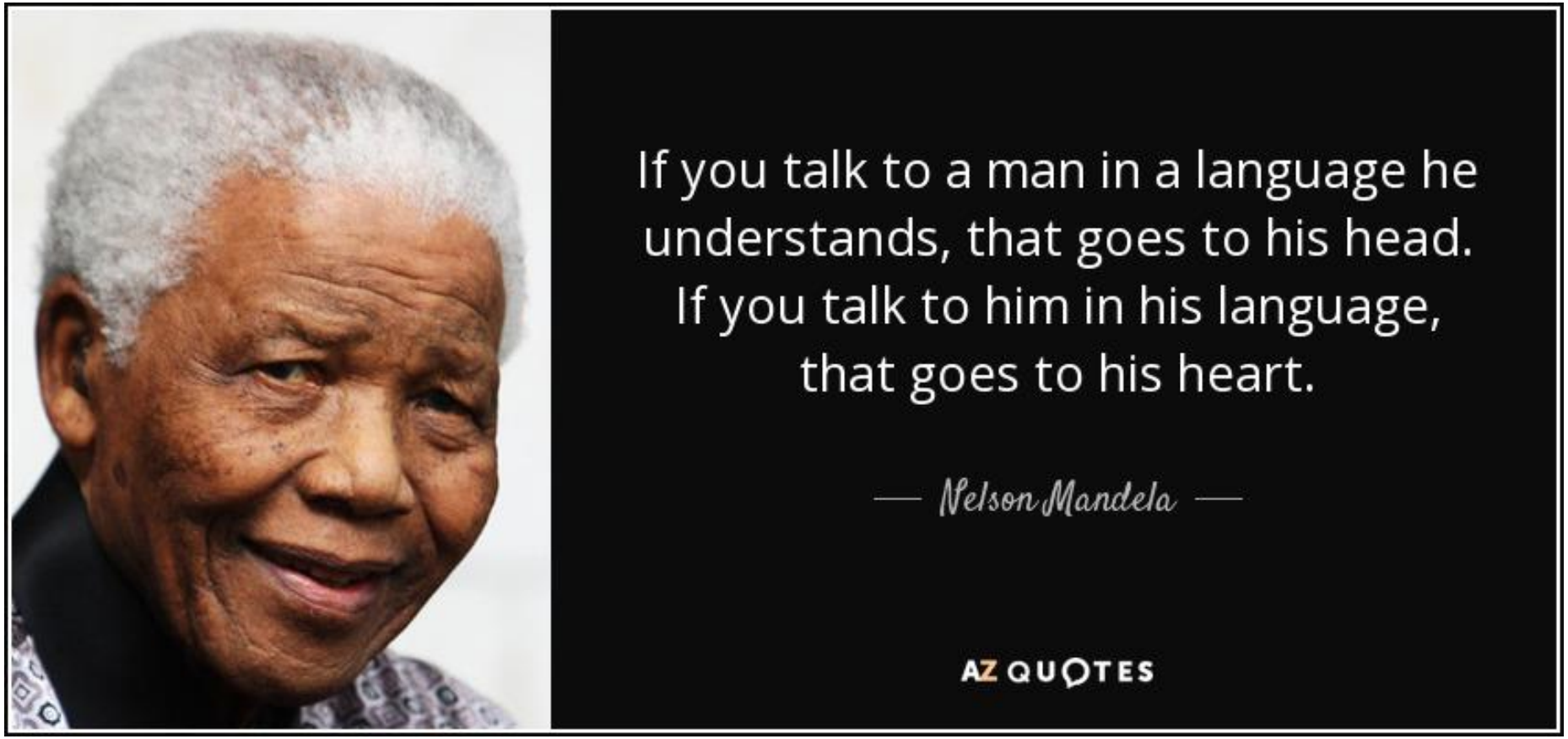
04 Fine-tuning on as little as 10–100 hours of transcribed data can yield a usable ASR system using free tools on Google Colab.

05 Ethical and community dimensions must be central — data ownership, consent, and language revitalisation goals matter as much as WER.

"Speech technology can play a major role in preserving linguistic diversity."

References

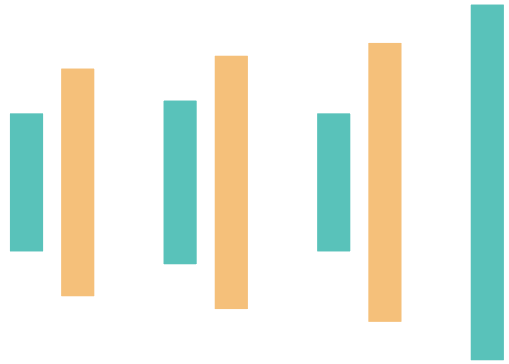
1. de Zuazo, X., Navas, E., Saratxaga, I., & Rioja, I. H. (2025). Whisper-lm: Improving asr models with language models for low-resource languages. *arXiv preprint arXiv:2503.23542*.
2. Klejch, O., Lamb, W., & Bell, P. (2025). A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic. *arXiv preprint arXiv:2506.04915*.
3. Jones, D. (2022, June). Development and evaluation of speech recognition for the welsh language. In *Proceedings of the 4th Celtic language technology workshop within LREC2022* (pp. 52-59).
4. Lonergan, L., Saratxaga, I., Sloan, J., Bravo, O. M., Qian, M., Chiaráin, N. N., ... & Chasaide, A. N. (2025, January). Fotheidil: an Automatic Transcription System for the Irish Language. In *Proceedings of the 5th Celtic Language Technology Workshop* (pp. 35-45).
5. Williams, A., Demarco, A., & Borg, C. (2023). The applicability of wav2vec2 and whisper for low-resource maltese asr.
6. Raj, J., & Conway, J. (2026). BlasBench: An Open Benchmark for Irish Speech Recognition. *arXiv preprint arXiv:2604.10736*.


If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart.
— Nelson Mandela —
AZ QUOTES

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

FOSTERLANG

Questions & Discussion



arvind.kumar@mpi.nl

"Every language deserves a digital voice."