MediSpeech

D2.1. STATE OF THE ART ANALYSIS

Work package	WP2: Use cases, requirements and specifications
Type	Document
Contractual date of deliverable	30 June 2026
Actual date of deliverable	08 September 2026
Dissemination level	Public
Partner responsible	University of Porto
Status	Finalized

Document history			
Version	**Date**	**Author**	**Description**
1.0		Mariia Zamyrova	State-of-the-art ASR
1.0		Henk van den Heuvel	State-of-the-art SD
1.0		Khiet Truong	State-of-the-art paralinguistic event detection
1.0		Dragos Balan	Text review
1.0		Maxim Popov	State-of-the-art Dutch NER eval.
1.1	09 July 2026	Filipe Bernardi	Updated and harmonised D2.1 State-of-the-art with new implementation-oriented sections and Portuguese ASR resources
1.2	24 August 2026	Erica Yang	Added two sections under Section 3.5, focusing on information extraction with and without LLM
1.3	25 August 2026	Tim Dong	Added Clinical Decision Support Systems State-of-the-art section
1.4	26 August 2026	Zeynep Gürler	Added Turkish contributions to ASR, speaker diarisation, medical NLP/NER, clinical decision support, and interoperability sections
1.5	27 August 2026	Filipe Bernardi	Consolidated partner contributions, harmonised content and structure, resolved comments and tracked changes, and performed final editorial and reference review
1.6	05 September 2026	Filipe Bernardi	Final consolidation after partner review; outstanding comments, scope issues, references and formatting resolved for submission

Document review		
Date	**Reviewer**	**Description**
07/09/2026	Whole consortium	Reviewed and approved

Partners involved (main)		
Company	**Name and surname**	**Email address**
University of Porto	Filipe Bernardi, Rute Almeida	fbernardi@med.up.pt rutealmeida@med.up.pt
Radboud University	Henk van den Heuvel	henk.vandenheuvel@ru.nl
University of Twente	Khiet Truong	k.p.truong@utwente.nl
Amsterdam UMC	Maxim Popov	m.popov@amsterdamumc.nl
Chilton Computing Ltd	Erica Yang, Andrew Tsui	erica.yang@chiltoncomputing.co.uk andrew.tsui@chiltoncomputing.co.uk
Bristol Heart Institute	Tim Dong, Gianni Angelini	qd18830@bristol.ac.uk g.d.angelini@bristol.ac.uk
MEDrecord	Joanna Morozowska	joanna@medrecord.io
Anadolu Sigorta	Zeynep Gürler	zgurler@anadolusigorta.com.tr

Table of Contents

1. EXECUTIVE SUMMARY

1.1. Purpose, scope, and key outcomes

This deliverable reviews the state of the art in automatic speech recognition, paralinguistic event detection, speaker diarisation, clinical information extraction, and clinical decision support. The review considers evidence from both academia and industry in the clinical contexts relevant to the project. Thus, this report provides input for the functional and technical requirements to be defined in T2.2 and will serve as a reference for MediSpeech system development, piloting, and evaluation. Beyond this component-level review, automated medical reporting is a clinical documentation pipeline rather than a single speech-to-text task. The evidence reviewed in this deliverable, therefore, also covers clinical documentation burden, ambient clinical documentation, clinical NLP and LLM-based report drafting, interoperability with electronic health records, terminology mapping, evaluation methods, privacy, safety, and human oversight. These aspects are included to ensure that the conclusions of D2.1 can be translated into the high-level use case requirements in D2.2 and the clinical requirements and baseline systems in D3.1.

1.2. Abbreviations

ABBREVIATION	MEANING
ASR	Automatic Speech Recognition
SD	Speaker Diarisation
NER	Named Entity Recognition
DNN	Deep Neural Network
HMM	Hidden Markov Model
E2E	End-to-end
WER	Word Error Rate
CER	Character Error Rate
RTFx	Inverse Real Time Factor
LSc	LibriSpeech clean
LSo	LibriSpeech other
MLS	Multilingual LibriSpeech
MMS	Massively Multilingual Speech
STT	Speech-to-text
NLP	Natural Language Processing
EHR	Electronic Health Record
LLM	Large Language Model
CDSS	Clinical Decision Support System
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
ICPC-2	International Classification of Primary Care, 2nd edition
LOINC	Logical Observation Identifiers Names and Codes
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms

2. INTRODUCTION

2.1. Background and context of deliverable

As a starting point for the work on MediSpeech, it is essential to have an overview of the SoTa of technologies relevant to the project and its later requirements. The aim of the MediSpeech project is to reduce administrative time waste in healthcare by creating an open digital healthcare ecosystem for automated medical reporting of patient-doctor conversations. In this context, the clinical problem addressed by MediSpeech should be framed not only as a speech-recognition challenge but also as a documentation, workflow, and patient-safety challenge. Administrative workload may reduce the time available for direct patient care, fragment clinicians' attention during consultations, and increase after-hours documentation. Automated medical reporting is therefore relevant insofar as it can support safer, more efficient, and more complete clinical documentation, rather than merely producing verbatim transcripts. Accordingly, documentation quality, correction burden, auditability, and patient safety should be considered important dimensions when reviewing the SoTa and deriving implications for subsequent requirements and evaluation activities. This overview helps project partners and those interested in the topic gain insights into relevant technologies and the SoTa, serving as a starting point for the project. The technologies reported on here are: Automatic Speech Recognition, Paralinguistic Event Detection, Speaker Diarisation, and Medical Term Recognition.

2.2. Relation to the project objectives and work package

This deliverable will provide input to the functional and technical requirements to be defined in T2.2 and will serve as a reference for MediSpeech system development, piloting, and evaluation. Further, the deliverable will provide guidance on the tasks in WP3 (AI-powered automated speech processing components), specifically T3.1 (Clinical requirements and baseline systems), T3.3 and T3.4 (Adoption of novel techniques for Speech Recognition), and T3.6 (NLP algorithms). D2.1 therefore serves as an upstream evidence base for subsequent WP2 and WP3 activities rather than as a detailed requirements or architecture specification. The state-of-the-art analysis informs D2.2 by identifying relevant technological possibilities, limitations, and evidence gaps that can be translated into use-case innovations and high-level functional and non-functional requirements. It also supports D3.1 by providing a reference framework for defining clinical requirements, baseline systems, and evaluation priorities for the automated speech-processing components. This linkage is important to ensure that later requirements are grounded not only in project objectives, but also in the current maturity, limitations, and risks of the underlying technologies.

3. STATE-OF-THE-ART OF THE TECHNOLOGY

3.1. Automatic Speech Recognition

3.1.1. Introduction

Automatic speech recognition (ASR) is the task of generating a transcript of speech from provided audio data. In the past several decades, with the introduction of deep neural network (DNN) architectures and with the research community amassing increasing amounts of diverse speech data, the field of ASR has seen a lot of progress, shifting from speaker-dependent single-language statistical Hidden Markov Models (HMMs) to the more recent speaker-independent multi-lingual Transformer-based DNN models.

It is important to note that, in addition to the model's architecture and training schedule, the data used to train it is a key factor in its accuracy. Models perform best on data that shares qualities with training data. Speech is highly nuanced and can vary across a wide range of factors. Speech differs among speakers based on factors such as age, gender, and accent. There can also be within-speaker differences, e.g., due to variations in pronunciation. Furthermore, there is a distinction between spontaneous and scripted speech scenarios, with spontaneous speech more likely to contain disfluencies (e.g., stuttering, pauses, filler words such as "uh" and "hmm") and self-corrections. It is also crucial to keep in mind the distinction between clean and noisy speech, and between speech from a single speaker and from multiple (overlapping) speakers. Lastly, models need to be adapted to the required languages and domain-specific jargon (e.g., medicine, law). In clinical settings, ASR performance should also be interpreted in relation to real-world healthcare conditions. Doctor-patient dialogue is often spontaneous, interrupted, acoustically variable, and rich in domain-specific vocabulary. Errors involving medication names, doses, numerical measurements, negation cues, temporal expressions, accents, background noise, or overlapping speech may directly affect documentation quality and patient safety.

The variety in speech makes it important not only to train models on diverse data but also to evaluate them on a representative set of datasets to obtain a clear overview of their abilities. Moving forward, as a reference for model performance error rates, we use the Open ASR leaderboard on Hugging Face, along with error rates from the original model reports. The Open ASR leaderboard is an initiative started by Gandhi et al. (2022) to establish a uniform protocol for evaluating end-to-end (E2E) ASR models. The authors selected 8 (English) datasets of various speaking styles (spontaneous vs scripted), domains (meetings, audiobooks, TED talks, etc.), and transcription formats (normalized vs punctuated/capitalized). The benchmark reports the average word error rate (WER) and the inverse real-time factor (RTFx) across all datasets. The word error rate is the ratio between the total number of words that were inserted, deleted or substituted in the predicted transcription and the total number of words in the correct reference transcription. The lower the WER, the fewer mistakes in the predicted transcription. Inverse RTFx is computed as the ratio between the length of the input audio and the compute time of the ASR model (in seconds). The higher the RTFx the faster the model is with respect to the length of its input. For MediSpeech, WER remains useful as a technical benchmark, but it is insufficient as the sole evaluation target for clinical ASR. A system may achieve an acceptable aggregate WER while still producing clinically significant errors in diagnoses, medication instructions, doses, measurements, allergies, referrals

or follow-up plans. Clinical ASR evaluation should therefore consider complementary dimensions such as medical term error rates, number/date/dose errors, semantic accuracy, clinically significant error review and correction time.

It is important to note that the benchmark retains any punctuation and/or capitalization in the target transcriptions before computing the WER in order to showcase the full capabilities of ASR models that produce punctuated transcriptions. As such, for models that do not support punctuation or capitalization, like wav2vec 2.0, the error rates can be higher than in the original paper. Out of the values reported in the Open ASR leaderboard, we focus on the models' average WER and RTFx and their individual WERs on the AMI (spontaneous, long-form, punctuated and cased), LibriSpeech (read, short-form, normalized) and TED-LIUM (oratory, long-form, normalized) datasets. Note, that to accommodate most models' input sizes, all of the datasets used in the Open ASR leaderboard are segmented into short sequences. The LibriSpeech testing set is additionally split into "clean" (LSc) and "other" (LSo) sets based on the quality of the audio recordings. The selected datasets are most frequently encountered in the papers that we surveyed and provide a sufficiently diverse overview of the models' performance.

Another important remark is that the Open ASR leaderboard uses HuggingFace implementations of models, such that the results in the leaderboard can deviate from those obtained for alternative model implementations. Lastly, the leaderboard is continuously updated, such that the values referenced in this report are timestamped.

For the MediSpeech project, we aim to work with real-time spontaneous (potentially noisy) Dutch speech. As such, in our overview of ASR models we put special emphasis on the languages they support, their accuracy on spontaneous and noisy speech data, and their processing speed/real-time streaming functionality. For European Portuguese, and particularly for spontaneous primary care consultations, the availability of robust clinical ASR evidence and benchmark datasets should be treated as a potential evidence gap. This is relevant for MediSpeech because performance observed in generic or non-clinical Portuguese speech may not necessarily transfer to primary care documentation scenarios.

3.1.2. Wav2vec 2.0 and Whisper

Before diving into the review of the most recent state-of-the-art ASR models, we provide some background information on wav2vec 2.0 and OpenAI's Whisper - models that have become a staple in the field. **Wav2vec 2.0** was introduced in 2020 by Baevski et al. It is a Transformer-based encoder-only model that learns robust acoustic representations from audio input. Because the model is encoder-only, it cannot transcribe speech on its own, and a decoding layer needs to be added to it. One of wav2vec2's important properties is that it is pre-trained in a self-supervised way, which allows it to use unlabeled training data (the original paper used ~53k hours of unlabeled read English speech from LibriVox). It creates training labels by discretizing intermediate audio features (via product quantization). After pre-training and learning to generate acoustic labels for audio, wav2vec 2.0 is fine-tuned for speech recognition by adding a (Connectionist Temporal Classification-CTC-based) decoder layer and training on labeled speech-text pairs. Because wav2vec 2.0 learns frame-wise labels, it can capture nuanced acoustic information and can be used not only for word-level but also phoneme-level recognition. However, the model's sensitivity to audio acoustics means it is not as robust to noise and is prone to misspelling words depending on the quality of the input audio and the speaker's pronunciation.

The authors of wav2vec2 report the best results of 1.8% and 3.3% on LSc and LSo, where the model checkpoint was pre-trained on LibriVox and fine-tuned on 960 hours of LibriSpeech with a Transformer-based language model for decoding. According to the Open ASR benchmark, the best performing wav2vec2 checkpoint on HuggingFace is “speechbrain/asr-wav2vec2-librispeech”. It uses the (pre-trained + self-trained) “facebook/wav2vec2-large-960h-lv60-self” checkpoint (Xu et al., 2020) with two additional DNN layers and fine-tunes it on LibriSpeech. As we previously mentioned, wav2vec 2.0 outputs normalized transcriptions (lowercase, without punctuation). Therefore, its average WER (14.35%) is much higher than its score on, for example, just the LibriSpeech dataset (1.77% and 3.83% WER on LSc and LSo). Its error rate on spontaneous speech, e.g., the AMI dataset, is reported to be 32.05%. However, the AMI dataset has cased and punctuated transcriptions, so we cannot attribute the high error rate purely to wav2vec’s inability to recognize acoustic artifacts unique to spontaneous speech. A more detailed analysis of model performance in the original ESB paper confirms that by normalizing the transcriptions the WER can be reduced by up to 10%. Notably, the average WER of the other wav2vec2 checkpoints, including “facebook/wav2vec2-large-960h-lv60-self”, which use the original single-layer decoder rises above 21% (with LibriSpeech WER above 11%). This potentially shows the importance of carefully selecting and fine-tuning the decoder when working with wav2vec2-based models.

The **Whisper** model was introduced in 2022 by Radford et al. It is also Transformer-based, but unlike wav2vec 2.0, it has both an encoder and a decoder. What distinguishes Whisper is the amount of data it was trained on. The model was trained in a weakly supervised manner with ~680.000 hours of both manually and automatically annotated English and multilingual speech from various sources on the Internet. Because it is trained on multiple data sources, Whisper has been shown to be robust to noisy speech. There are several sizes of Whisper available, out of which the “large” models have over 1500M parameters. In comparison, the wav2vec 2.0 “large” model has ~317M parameters. Another difference between wav2vec 2.0 and Whisper is that Whisper was trained to produce capitalized, punctuated transcriptions. In the Open ASR benchmark, we see that Whisper (large-v3) has an average WER of 7.44% and outperforms wav2vec 2.0 on almost all datasets, including AMI (15.95% vs 32.05% WER) and TED-LIUM (3.86% vs 7.58% WER).

One of Whisper’s downsides is that it can be challenging to obtain verbatim transcriptions from it as by default it standardizes the output text. The downside for both wav2vec 2.0 and Whisper, and for most Transformer-based models in general, is that they can process a limited amount of audio frames at once. The computational complexity of the attention module in the Transformer architecture grows rapidly with the size of the feature matrix. Therefore, the duration of input audio is often limited (for Whisper, it is 30 seconds), and longer audio is processed by chunking it into smaller sequences. Regarding the models’ performance speed, those with fewer parameters usually operate faster, and from the average RTFx measures we can indeed see that wav2vec 2.0 “large” model has triple the RTFx of Whisper (451.18 vs 145.51).

3.1.3. SoTa ASR systems: Academic solutions

Wav2vec2-based models

Following the success of wav2vec 2.0 and OpenAI’s Whisper, both academic and commercial research are investing in creating models that surpass their performance. A number of works focus on improving models by fine-tuning and adjusting the architectures of wav2vec 2.0 and Whisper. The wav2vec2 family of models is most prominently represented by **HuBERT** (Hsu et al., 2021)

and **XLSR** (Conneau et al., 2020). Compared to wav2vec 2.0, HuBERT uses the k-means clustering algorithm to generate discrete target "acoustic" frame labels for unsupervised pre-training. HuBERT was pre-trained and fine-tuned on similar proportions of data from LibriVox as wav2vec 2.0. Its "large" (300M) and "x-large" (1B) checkpoints have shown to outperform most wav2vec 2.0 checkpoints on the LibriSpeech testing sets by at most 2% WER, except for the wav2vec 2.0 models that use the self-training approach. In the Open ASR leaderboard, the HuBERT checkpoints also outperformed the majority of the wav2vec 2.0 checkpoints, except for the self-training ones, and obtained an average WER of 22.55%. Due to the similarity of HuBERT's and wav2vec2's architectures, their RTFx scores are also close (within the 450-520 value range), except for HuBERT's "x-large" model which is slower (RTFx of 361.32).

XLSR preserves the original architecture of wav2vec 2.0 but is fine-tuned on multilingual data via unsupervised cross-lingual representation learning, unlike HuBERT that used only English data. The biggest XLSR model, pre-trained on 56k hours of speech from 53 languages and fine-tuned on various amounts of data from the evaluation dataset, is not included in the Open ASR leaderboard, but in the original paper achieved an average WER below 11% on 8 languages (including English and Dutch) in the Multilingual LibriSpeech (MLS) dataset (Pratap et al., 2020). The RTFx of XLSR can be estimated to be the same as for other wav2vec2 models, because it preserves the original architecture.

Since the introduction of XLSR, there have been several efforts to further scale up the number of supported languages. For example, **XLS-R** (Babu et al., 2021) uses a wav2vec2-based architecture like XLSR but increases the model size up to 2B parameters and expands the training data to 128 languages (436K hours of pre-training data). For every model size, including 0.3B which is the same as XLSR-53, XLS-R outperforms XLSR-53 by an average of 1-3% WER, under both noisy spontaneous and clean read speech conditions. The **Massively Multilingual Speech** (**MMS**) project (Pratap et al., 2023) increases the number of languages to 1107 (491K hours of pre-training and 45K hours of fine-tuning data). The authors use the same architectures as XLS-R 0.3B and 1B and measure character and word error rates on the multilingual FLEURS dataset, which show that MMS outperforms both XLS-R (~13.3% vs ~14% CER) and Whisper "large-v2" (28.8% vs 44.3% WER). The MMS (1B) model is also benchmarked in the Open ASR leaderboard ("facebook/mms-1b-all"). Its average WER is reported to be 22.54, just above HuBERT "x-large", however, because this checkpoint uses additional adapter modules to optimize performance per language its RTFx is lower (230.79).

Whisper-based models

Similarly for Whisper, there are many models, e.g., **CrisperWhisper** (Wagner et al., 2024) and **WhisperX** (Bain et al., 2023), that are a result of optimizing its pipeline. WhisperX addresses Whisper's long-form transcription abilities by segmenting input audio based on voice activity detection (VAD) assignments, such that only segments containing speech are passed to the ASR model. Longer VAD segments are further split at the points of minimal speech activity to ensure that speech does not end abruptly. With this the authors make an assumption that segments are independent and can be processed in parallel without conditioning on previous output, as is suggested in the original Whisper paper. WhisperX additionally aims to correct Whisper's word-level timestamps by force-aligning them with wav2vec 2.0 phoneme-level timestamps. While WhisperX is not part of the Open ASR leaderboard, the authors report their own results comparing Whisper and WhisperX WER on the long-form TED-LIUM dataset (10.5% and 9.7% WER, respectively). They also found the precision and recall of WhisperX's word segmentation on the

AMI and Switchboard datasets to be higher than that of Whisper. By processing audio segments in parallel, WhisperX is further able to perform inference almost 12 times faster than Whisper. However, the additional force-alignment module might bring WhisperX's final performance speed down.

CrisperWhisper also aims to improve the time alignment of the output transcription, as well as the model's ability to recognize speech disfluencies, which Whisper usually avoids in its output. Firstly, it finetunes Whisper on decoder cross-attention cost to penalize misalignment between encoder and decoder output. Secondly, it changes the model tokenizer to more accurately assign the space token and applies pause heuristics to distinguish natural short pauses between speech segments and genuine (long) pauses. In the Open ASR leaderboard, CrisperWhisper has the lowest average WER out of the models we covered thus far, at 6.67%. It also beats Whisper's WER on the AMI dataset with 8.71% WER, showing its advantage for spontaneous speech transcription. Additionally, the authors of CrisperWhisper report word segmentation F1-scores on the AMI dataset that surpass WhisperX. A possible downside of CrisperWhisper is its performance speed, which may degrade because of the additions to the Whisper pipeline. Indeed, Open ASR estimates its RTFx to be 84.05 compared to Whisper's (large-v3) 145.51. Furthermore, CrisperWhisper currently only supports English because of the data it was fine-tuned on, compared to WhisperX which did not require fine-tuning and therefore is compatible with many Whisper's supported languages.

It is also worth mentioning models such as **Distil-Whisper** (Gandhi et al., 2023) and **faster-whisper** ("GitHub - SYSTRAN/faster-whisper: Faster Whisper transcription with CTranslate2", n.d.), that aim at speeding up Whisper's performance, rather than decreasing its error rates. Faster-whisper is a reimplementation of Whisper using a more efficient C2Translate toolkit. It claims to be up to 4 times faster than the original implementation while preserving its accuracy. Distil-Whisper reduces the number of model parameters by freezing Whisper's encoder and replacing the decoder with a compressed version, which is trained via knowledge distillation on 21k hours of English speech. The Open ASR leaderboard reports Distil-Whisper (distil-large-v3.5) to perform similarly to Whisper's large-v3 model with an average WER of 7.21%, and a difference of up to ±2% WER on individual datasets. The RTFx of Distil-Whisper is indeed higher than that of Whisper (202.03 vs 145.51).

Another aspect of Whisper's architecture that needs to be addressed is its compatibility with real-time computing. A simplistic approach for real-time ASR is buffering audio in small chunks and processing them consecutively. The authors of **Whisper-Streaming** (Macháček et al., 2023) attempt to improve upon that approach by providing the output of previous buffer states as context for the following ones and using agreement heuristics between buffer states to make transcriptions more robust. As the base Whisper model, they use faster-whisper (large-v2). Comparing offline and online performance of Whisper-Streaming on ESIC, a dataset of English European Parliament speeches (with interpretations available in German and Czech) (Macháček et al., 2021), the authors found that online WER differs from the offline WER of 7.9% by less than 2%. The model was also tested on the German and Czech interpretations with similar results. In addition to measuring the WER, the authors estimate the model's latency. They factored in computational latency and the delay between when a word was emitted by the ASR and its onset timestamp in the golden transcription. For English audio samples, the latency was measured to be between 3 and 6 seconds, and 4-7 seconds for the other two languages.

WhisperFlow by Wang et al. (2024) also addresses Whisper's real-time processing abilities by building upon Whisper-Streaming's strategy. Instead of padding the buffered audio segments to

Whisper's 30 second input limit, as is usually done with inputs of inconsistent lengths, the authors append only a 0.5 second "hush word". The "hush word" is a trainable parameter that is appended to raw input audio with the main purpose of reducing latency by using less padding and preventing hallucinations that can occur when input is shorter than 30 seconds. The authors also apply beam pruning to reduce the number of word candidates considered during decoding. Lastly, parallelizable operations within the pipeline (e.g., encoding) are allocated to the GPU while sequential operations – decoding, are allocated to the CPU to optimize compute power usage. The authors chose to use the smaller Whisper models (base, small and medium) for an optimal trade-off between accuracy and latency. Like for Whisper-Streaming, the latency is computed as the time elapsed between when a word is uttered in the input audio and when the model outputs the corresponding transcription. WhisperFlow has shown to reduce latency by 1.6-4.7 times while keeping the WER (on the long-form TED-LIUM 3 dataset) below 6%. In comparison, the WER of Whisper-Streaming and original Whisper on TED-LIUM 3 is slightly lower at around 4%.

NVIDIA NeMo

Aside from the wav2vec2- and Whisper-based models, new cohorts of ASR models are emerging. Such are NVIDIA's NeMo **Parakeet** and **Canary** models, based on the Fast Conformer architecture (Rekesh et al., 2023). Parakeet, an encoder-only model, supports CTC-based and RNNT-based decoding. Note, that there is a trade-off between speed and accuracy when using either of the decoders, where CTC is faster, but Recurrent Neural Network Transducer (RNNT) is expected to be more accurate (Majumdar & Koluguri, 2024). The "parakeet-tdt-0.6b-v2" (600M) checkpoint is currently in second place on the Open ASR leaderboard with an average WER of 6.05% (AMI – 11.16% and TED-LIUM – 3.38% WER). All of Parakeet's checkpoints have the highest RTFx in the range of 2-5.3k, with the smaller CTC-based checkpoints having RTFx above 4k.

NVIDIA's Canary is a multilingual encoder-decoder model (encoder – Fast Conformer, decoder - Transformer), capable of both transcription and translation, similar to Whisper. Canary utilizes a novel concatenated tokenizer (Dhawan et al., 2023) which allows it to extract token-level language information and handle input audio with code-switching (multiple language in a single recording), a functionality not present in any of the previously discussed models (Rastorgueva & Koluguri, 2024). The best performing Canary checkpoint (canary-1b-flash (883M)) has an average WER of 6.35% (AMI – 13.11% and TED-LIUM – 3.12% WER), with average RTFx above 1k. While neither Parakeet nor Canary come with real-time processing capabilities, NVIDIA provides sample code for enabling buffered streaming for its ASR models ("Automatic Speech Recognition (ASR) — NVIDIA NEMO Framework User Guide", n.d.). In terms of language support, Canary is currently available for 4 languages (English, German, French and Spanish) while the versions of Parakeet that are openly available only support English. Both Canary and Parakeet are trained on either fully public or a mixture of public and in-house data.

Multimodal LLMs

A special group of models is made up of multimodal LLMs trained for speech recognition. As more and more such models are being released, some of the prominent ones openly available for users are IBM's **Granite** (Saon et al., 2025), **Qwen-Audio** (Chu et al., 2023), Microsoft's **Phi-4-Multimodal** (Abouelenin et al., 2025) and **SpeechT5** (Ao et al., 2022), **SALMONN** (Tang et al., 2023), and, as of recently, Google's **Gemma 3n** (Sanseviero & Ballantyne, 2025). The general structure of a multimodal LLM combines an encoder that maps each of the input modalities into a shared representational space and a core LLM that performs reasoning on the transformed input

data. Qwen-Audio (8.4B) uses Whisper's (large-v2) encoder and a transformer-based Qwen-7B LLM and reports 2.0% and 4.2% WER on the LSc and LSo test sets. SpeechT5 uses a Transformer encoder-decoder architecture as its reasoning model and the feature extraction layer from wav2vec 2.0 as the audio encoder instead. The model is both pre-trained and fine-tuned on LibriSpeech and achieves 1.9% and 4.4% WER on the LSc and LSo test sets.

SALMONN uses a combination of Whisper and BEATs (Chen et al., 2023) encoders to encode both speech and non-speech audio events. The encoded features are passed to a Q-former (Li et al., 2023) layer to adapt the features for Vicuna LLM (13B) (Chiang et al., 2023). The authors report 2.1% and 4.9% WER on LSc and LSo. Granite (8.7B), inspired by SALMONN, also uses two layers (a Conformer (Gulati et al., 2020) and a Q-former) to encode the audio and adapt the encoding to LLM's representational space, and the "granite-3.3-instruct" LLM to decode the encoded audio. Like SpeechT5, it is trained on publicly available English datasets. The Open ASR leaderboard reports Granite's average WER as 5.85%, with 9.12% WER on the AMI dataset and 1.42% and 2.99% on LSc and LSo. Lastly, Phi-4-Multimodal (5.6B), similarly to SALMONN and Granite, encodes audio in two stages – first, with a convolutional neural network block and then with a multilayer perceptron space projector. The encoded audio features are passed to the Phi-4-Mini (3.8B) LLM. Open ASR estimates Phi-4-Multimodal's average WER at 6.14%, with 11.45% on the AMI dataset and 1.67% and 3.82% on LSc and LSo.

The range of language support for LLM-based ASR models is generally limited to English, like for SpeechT5 and SALMONN, or a few other languages (e.g., French, Spanish, Portuguese, and German) as for Granite, Phi-4-Multimodal and Qwen-Audio. Regarding RTFx, although the measure is not reported for every model **(**only Granite - 31.33 and Phi-4-Multimodal - 62.12**)**, it can be expected to be generally low due to the large number of parameters typically used for LLMs. Google's new multimodal LLM Gemma 3n claims to address both the limited language support and high compute demand. The model uses architectures like the MatFormer, that consists of a larger model with nested smaller models, such that only the core small models can be used during inference. With such techniques the number of effective parameters can be reduced to 2-4B. Gemma 3n also optimizes long-form and streaming input processing with the KV Cache Sharing approach that alters how the attention keys and values are stored in memory and shared with the top model layers. Additionally, Gemma 3n uses an audio encoder based on Google's Universal Speech Model (Zhang et al., 2023) and should be able to support ~35 languages.

3.1.4. SoTa ASR systems: Commercial solutions

Parallel to the progress made by academic research, many commercial ASR systems are becoming available. We surveyed a selection of commercial models that are widely used and regarded as state-of-the-art: **Google Cloud's Speech-to-Text (STT)** (Google cloud - speech-to-text transcription, 2025), **Amazon Transcribe** (Amazon transcribe - speech to text – AWS, 2025), **Microsoft Azure AI's STT** (Microsoft Azure AI speech, 2025), **Speechmatics** (Speechmatics - speech-to-text API, 2025), **AssemblyAI** (Assembly ai, speech-to-text api, 2025), **Deepgram** (Deepgram, speech to text API: Next-gen AI speech recognition, 2025) and **Rev AI** (Rev AI: Speech to Text API, 2025). All commercial models provide text punctuation and capitalization. The majority of commercial models are also multilingual, compared to academic models that are more often adapted to a small set of languages. Like Whisper, all of the mentioned services offer language detection, which avoids the need to predetermine the languages that will be in the audio recordings. Moreover, an increasing number of commercial models (e.g., Speechmatics and AssemblyAI) are extending their capabilities to support code-switching.

A common trait of commercial models is that in addition to the base speech recognition functionality, they include other functionalities like speaker diarisation and streaming speech

recognition. The ability to transcribe speech in real-time is what sets commercial models apart, as this functionality is rarely provided alongside academic models and needs to be implemented ad-hoc by users. All mentioned models support both diarisation and streaming. Additionally, AssemblyAI, Speechmatics, Deepgram and Rev AI explicitly offer filler word detection/omission, which is beneficial for verbatim transcriptions.

The biggest downside of commercial models is the low level of transparency. Companies' in-house resources allow them to gather million-hour private datasets which help them build SoTa models but at the same time limit our understanding of the models' capabilities and the ethics behind their development. The models' architectures are also often kept private, and most companies provide their ASR-services via an API. Out of the surveyed models, we found AssemblyAI and Google to provide most insight into the models they use, e.g., research from Ramirez et al. (2024) and Zhang et al. (2023). Despite the model architectures being hidden, all companies provide the option to fine-tune their models with custom vocabularies, which is useful for adapting to a particular domain.

Processing data externally via APIs also raises privacy and security risks when working with sensitive data, especially in the medical domain. In their study, Zeeshan et al. (2025) compared academic and commercial models on the subject of their suitability for being used as scribes during psychiatric appointments. To gage the level of data security of commercial models the authors contacted the sales teams. They concluded that Google's STT, Amazon Transcribe, and Azure STT were not suitable to be used as scribes in the field of psychiatry based on their external data processing regulations and prices. However, they considered API-services from AssemblyAI, Speechmatics and Deepgram good candidates, as they provide measures such as on-demand removal of user data and the option to opt out of data storage.

From our own inspection, we found that all services claim to store no customer data and store the data processing results (e.g., transcriptions) for a limited period, for example, no longer than 5 (Speechmatics and Google) or 7 days (AssemblyAI). Google, Amazon and Deepgram services can use customer data to improve their models but provide the option to opt out of data logging. For processing speech real-time, Azure and Google, for example, claim to process input audio only in the memory without temporarily storing it. Out of the services we consider in this review, Google, Azure and Amazon provide not just the speech processing software as their main product, but also cloud storage. They make use of the shared-responsibility approach, where the user is responsible for the sensitivity of the data that they choose to process and how they store it in the cloud, but the companies are responsible for providing data privacy and security for the cloud service as a whole. Other services like AssemblyAI, use a combination of in-house storage facilities and cloud storage from, e.g., Google ("Assemblyai case study page", n.d.). Companies like Speechmatics and Deepgram are also starting to offer their services on servers hosted by the customers locally ("Self-Hosted Voice AI", n.d.; Phipps, 2025).

Overall, we share the reasoning of Zeeshan et al. that certain models appear more secure than others, but deeper enquiries need to be made with the sales teams to make sure the security is sufficient for particular use cases. For the MediSpeech use case, it is important to point out that as of recently many of the companies offer services particularly for medical transcribing (e.g., AssemblyAI, Speechmatics, Deepgram Nova-3 Medical, Google STT medical and Amazon Transcribe Medical) and all of the surveyed services claim to be HIPAA and GDPR compliant/eligible, which gives some assurance of the level of data privacy and security those companies are ready to offer.

We considered several benchmarks that compare commercial models on publicly available data. Some of the commercial models were evaluated in the Open ASR leaderboard, where Speechmatics, AssemblyAI and Rev AI are reported to have 6.91%, 7.03% and 7.12% average

WER, respectively. Their RTFx measures were not reported. Kuhn et al. (2023) compared models based on publicly available English and German YouTube videos from the higher education domain and a subset of LSo. Their results showed that Whisper (large-v2) outperformed commercial models with 2.9%, 3.3% and 5% WER on the English YouTube, LibriSpeech and German YouTube datasets. Next best was Speechmatics with 3.3%, 3.6% and 8%, respectively. AssemblyAI, Azure, Rev AI and Amazon scored similarly with 4.4-4.5% on the English YouTube data but had more divergence on the other two datasets: 4.2-7% on Librispeech and 10.1-19.2% on German recordings. Of the four, Rev AI performed worst overall and Azure – the best. All of Deepgram's error rates fell between 8% and 19% while Google's exceeded 18%. The authors additionally observed that using custom vocabulary (for models that have this functionality) decreased the average WER but not significantly.

From AssemblyAI's internal research (Ramirez et al., 2024), they found that their Universal-1 model had the lowest average WER (7.6%) on English compared to Whisper large-v3, NVIDIA's Canary-1B, Azure, Deepgram Nova-2, Amazon Transcribe, and Google's STT. They compared the models based on 7 public datasets (including LibriSpeech and TED-LIUM) and 4 in-house datasets (from the podcast, broadcast, telephony and noisy data domains). Unversal-1 outperformed all models on the in-house data and 2 of the public datasets. It was closely followed by Canary-1B (8.1% WER), that performed best on 3 public datasets, and Whisper large-v3 (8.4% WER). The models were also compared on Spanish, German and French data. Universal-1 performed best on Spanish (average WER of 4.8%) and came in second for the other two languages, outperformed by Canary-1B for German (7.9% WER) and Amazon Transcribe for French (8.8% WER).

3.1.5. SoTa Dutch ASR systems

Academic solutions

Table 1: Performance of MMS, XLS-R, XLSR and Whisper (large-v2) as reported in Pratap et al. (2023) and Babu et al. (2021).

	Dataset	English (% WER)	Dutch (% WER)
MMS	FLEURS	10.7	13.7
XLS-R	MLS	12.9	11.6
XLSR	MLS	14.6	12.8
Whisper (large-v2)	FLEURS	4.2	6.7

Out of the models discussed in the previous subsection, multilingual models like XLSR, XLS-R, MMS, Parakeet and Whisper(X) support Dutch (Table 1). One observation to be made is that the authors of multilingual models often provide model results on only some of the languages and use different test datasets, which can make it difficult to gauge their relative performance. For example,

Parakeet is omitted in Table 1 as its performance on Dutch speech was not yet benchmarked at the time this report was compiled. With the Dutch Open Speech Recognition Benchmark initiative ("Dutch Open Speech Recognition Benchmark", n.d.), a space is being created to compare ASR model performance on various types of Dutch speech and share the results. Currently, the leading model in terms of accuracy and speed is faster-whisper (v2), scoring ~11% and ~22% WER on read and conversational Dutch speech, respectively.

Due to Dutch being underrepresented in the training data compared to, e.g. English, the models can have suboptimal WER. For this reason, fine-tuning and model adaptation are often used to improve performance. For example, NVIDIA released a Fast-Conformer model specifically fine-tuned for Dutch with approximately 621 hours of public training data, achieving 12.09% WER on MLS testing data (Żelasko & Chen, 2024). A popular toolkit for Dutch ASR is Kaldi_NL ("Kaldi_NL", n.d.). It hosts a collection of DNN-HMM models trained with the Kaldi toolkit (Povey et al., 2011) specifically for Dutch. Bălan et al. (2024) evaluated the performance of Whisper, MMS and Kaldi_NL models on two collections of Dutch and Flemish speech: N-Best 2008 and JASMIN-CGN. N-Best contains samples of telephone conversations and broadcast news speech. JASMIN-CGN contains samples of Dutch read and human-machine interaction (HMI) recordings. The recordings were collected from different participant groups based on age (children and adults) and mother tongue (native and non-native). Whisper large-v2 performed best on all datasets, with Kaldi_NL having second-best performance (Table 2).

Table 2: Performance of Whisper, MMS and Kaldi_NL on the Dutch partition of N-Best 2008 and JASMIN-CGN corpora (Bălan et al., 2024).

	N-Best 2008 (% WER)		JASMIN-CGN (% WER)			
	Broadcast	Conversational	Read (non-native)	Read (adult)	Conversational (non-native)	Conversational (adult)
Whisper (large-v2)	10.0	23.9	30.0	12.8	51.4	26.8
MMS	18.5	42.7	54.0	28.3	83.3	59.9
Kaldi_NL	12.6	38.6	45.3	20.9	60.0	44.0

Commercial solutions

While all commercial systems we previously mentioned support Dutch, few give information on the level of accuracy they can achieve for Dutch. Therefore, it is worth adding another service to the list – Soniox. Like other commercial STT services, it supports multiple languages and real-time transcription as well as medical transcription. What sets it apart is that its medical scribe is claimed to support multiple languages whereas for other services, e.g., Google STT, Amazon and Deepgram, medical transcription is limited to English. Soniox also provides performance

benchmarks for 60 languages (including Dutch) with 45-70 minutes of (unspecified) YouTube audio per language ("Speech-to-text benchmarks 2025", n.d.). Another ASR service that provides WER estimates for Dutch is ElevenLabs. They report 3.4% WER on the publicly available FLEURS dataset, which they compare to Deepgram's 11.3% WER. A new feature that ElevenLabs support is audio event tagging (e.g., laughter), although they do not yet offer real-time ASR ("Free Dutch Speech to Text Transcription", n.d.).

Table 3: Soniox WER benchmarks for English and Dutch partitions of the in-house YouTube audio dataset.

	English (% WER)	Dutch (% WER)
Soniox	6.5	9.4
AssemblyAI	9.3	10.6
ElevenLabs	11.1	11.2
Speechmatics	9.3	11.6
Amazon	11.4	12.0
Deepgram	9.9	12.1
Azure	9.4	14.4
Google	18.2	23.2

In addition to Soniox and ElevenLabs benchmarking, there has been independent research looking into the performance of commercial ASR models on Dutch speech. For example, Kasdorp et al. (2024) evaluated Azure and Google's STT model Chirp on regional dialects for Dutch and Flemish from the JASMIN-CGN dataset. In the dataset, Dutch speech is categorized into 4 regional dialects: West-Dutch, Transitional region, Peripheral region and Southern peripheral region. The results showed that both Azure and Google STT models performed best on read West-Dutch speech (Google: read – 24.8% and conversational – 25.2% WER; Azure: read – 15.7% and conversational – 20.5% WER).

Dutch ASR in healthcare

There is a growing number of companies specializing in speech transcription services specifically for Dutch medical conversations. **HealthTalk** ("HealthTalk", n.d.) provides a web-/mobile-application for transcribing doctor-patient conversations and generating medical reports in domains like mental and acute care, and child care. The system is able to transcribe and diarize hour-long

(real-time) recordings containing more than 2 speakers, with reports generated within minutes of the consultation ending. Person identification during consultations is reported to reach 99% accuracy, which HealthTalk uses to correctly assign transcribed data to the client, family members, or healthcare providers present. To improve terminology accuracy, the platform applies the MeSH and UMLS clinical vocabularies to standardise medical terms and diagnoses across transcripts and reports. The underlying language and speech models run on HealthTalk's own NVIDIA-based private-cloud infrastructure hosted in Amsterdam, which the company states improves accuracy and data security while remaining fully auditable by the healthcare provider. The platform also gives clinicians access to a client's full consultation history within minutes, covering prior treatments, diagnoses, and referrals. Another company is **Attendi**, providing their API services in the domains of elderly care, disability care and mental healthcare ("Intuitive Reporting – Attendi", n.d.). Their system creates offline punctuated transcriptions of recorded conversations. One more example is **SpraakLab** and their product **Tell James** created for ASR in the medical domain. SpraakLab's ASR provides real-time speech recognition with automatic punctuation and grammar-based rules that help it detect terminology, and can be hosted both on-premises and via an API ("Spraakherkenning van SpraakLab", n.d.). Tell James's accuracy on Dutch speech is claimed to be below 8% WER ("Domeinspecifieke taalmodellen - Tell James", n.d.). A fourth service worth mentioning is **Juvoly**. They provide real-time speech recognition with NER and transcription summarization functionalities ("Juvoly V2 WebSocket API", n.d.). As data for training models Juvoly use public datasets and, to better adapt the models to the medical domain and make them more inclusive, record fictitious conversations between medical students and between students learning Dutch as a second or third language ("FAQ | Juvoly", n.d.). Another service is **Medendo**. Medendo offers a web interface between Google and Azure to transcribe medical conversations in real time and process transcriptions into reports ("Medendo - Digital Assistant for Healthcare professionals"). HealthTalk is at the forefront of healthcare technology, specializing in speech transcription services tailored for the medical domain. Utilizing our proprietary MedQwen model — a medically fine-tuned Qwen architecture optimized with GRPO, alongside a fine-tuned Whisper-large-v3-turbo for Automatic Speech Recognition and Titanet-large for precise speaker diarisation (99% accuracy) — we deliver high-fidelity, context-aware transcriptions. Advanced techniques such as GPTQ/AWQ quantization, speculative decoding, and FlashInfer-optimized inference enable real-time, low-latency performance even on resource-constrained clinical hardware.

Besides commercial medical scribes there are multiple academic projects. One such project was started by Leiden University Medical Centre's **CAIRELab**. The initiative has since turned into a commercial product, software that generates templated medical reports in real time, under the name **Autoscriber** ("Autoscriber", n.d.). Another example is **Care2Report** (Nivel, Utrecht University, Radboudumc). They set out to create a multimodal medical reporting system that uses audio, video and sensor input. The proposed pipeline uses graphs to store and combine the information extracted from each modality and transforms those graphs into SOEP (SOAP) reports. At the prototype stage of the project, Google's STT is proposed as the audio input processor. Triples are then extracted from the transcription and matched against a medical ontology to find information relevant for the report. The matched triplets are stored in a graph, from which the report is generated with the NaturalOWL system (Maas et al., 2020; Molenaar et al., 2020). A third project example is the **HoMed** project (Nivel, Radboud, Utrecht and Twente universities). It focuses on developing methodology for applying ASR to sensitive (doctor-patient) conversations (Tejedor-García et al., 2024). As part of the research, publicly available ASR systems (wav2vec2, Whisper and Kaldi_NL) were evaluated on several sets of Dutch medical conversation recordings (Medicijnjournaal, medical video recordings and Nivel test and training sets). The results showed that Whisper (large-v2) outperformed the other systems, achieving the lowest WER of 10.9% on

the medical video recordings data, compared to 24.2% for wav2vec 2.0 and 28.4% for Kaldi_NL. In addition to evaluating SoTa models, Kaldi_NL was fine-tuned on some of the medical conversations and tested on the Nivel test set. While it did not outperform Whisper (57.1% vs 68.0% WER) it did improve compared to its baseline version (71.2% WER) ("Dutch Open Speech Recognition Benchmark", n.d.).

3.1.6. SoTa Portuguese ASR systems: resources, benchmarks, and healthcare applicability

Recent work on Portuguese ASR shows that European Portuguese is increasingly treated as a distinct language variety rather than as interchangeable with Brazilian Portuguese. This is relevant for MediSpeech because general multilingual ASR performance may vary substantially across Portuguese varieties, accents, speaking styles, and domains.

The CAMÕES benchmark was recently introduced as an open framework for European Portuguese and other Portuguese varieties. It includes 46 hours of European Portuguese test data across multiple domains and evaluates foundation models in zero-shot and fine-tuned settings, as well as E-Branchformer models trained from scratch. The authors also curated 425 hours of European Portuguese speech for training and fine-tuning, reporting relative WER improvements exceeding 35% over the strongest zero-shot foundation model. This suggests that European Portuguese-specific adaptation can substantially improve ASR performance, but also that generic multilingual models should not be assumed to be sufficient without local evaluation.

Additional resources are also emerging for European Portuguese speech. FalAR, a large-scale speaker-annotated corpus of European Portuguese parliamentary speech, contains approximately 5,800 hours of speech data, including around 4,850 hours with speaker identity annotations. Although parliamentary speech differs from clinical consultations, the corpus is relevant as evidence of the growing body of large-scale European Portuguese ASR resources and may support further model adaptation. In Brazilian Portuguese, recent work, such as the NURC-SP Audio Corpus, has also highlighted the importance of spontaneous speech datasets for ASR evaluation, reporting that spontaneous, accented speech remains challenging even with fine-tuned wav2vec 2-XLSR and Distill-Whisper models.

For the Portuguese MediSpeech use case, these developments indicate that Portuguese ASR should be evaluated using local and domain-relevant data whenever possible. Evidence from European Portuguese benchmarks and spontaneous Portuguese speech corpora supports the need to test performance not only on generic WER, but also on clinically relevant errors, medical terminology, numbers, medication names, speaker turns, and downstream documentation quality. At the current stage, published evidence on ASR for spontaneous European Portuguese primary care consultations appears limited, so any available partner-specific resources, EHR-linked workflows or domain adaptation strategies should be validated with the Portuguese technical and clinical partners before being presented as mature state-of-the-art evidence.

Beyond ASR-specific resources, recent Portuguese NLP and LLM initiatives also reinforce the importance of treating European Portuguese as a distinct target variety. AMALIA has been introduced as a fully open large language model prioritizing European Portuguese through targeted training and native pt-PT evaluation, while ALBA provides a linguistically grounded benchmark for assessing generative LLMs across European Portuguese language and cultural dimensions. Although these resources are not ASR systems, they are relevant to downstream MediSpeech

components such as transcript correction, summarization, clinical note drafting and Portuguese-language evaluation of LLM-supported documentation.

3.1.7. SoTa Turkish ASR systems: resources, benchmarks, and healthcare applicability

State-of-the-art automatic speech recognition for spontaneous conversational Turkish, including doctor–patient and doctor–caregiver clinical interactions, is expected to build on end-to-end recurrent neural transducer (RNN-T) architectures (Graves, 2012), which map the acoustic signal directly to text through three deep neural network modules optimised end-to-end and support both streaming and non-streaming operation from the same model. Two adaptation techniques from the recent ASR literature are particularly relevant to the medical-terminology and noisy-consultation conditions of the Turkish use case. External language-model integration for RNN-T (Zhou et al., 2022) allows subword recognition hypotheses to be re-scored against a domain-specific language model — typically implemented as an n-gram language model converted into a finite-state transducer — so that pronunciation-driven errors on domain terminology can be reduced without retraining the acoustic component. On-the-fly noise augmentation during training (Nguyen et al., 2020), in which noise samples collected from diverse environments are randomly mixed into training recordings, has been reported to improve robustness to unseen acoustic conditions without requiring in-situ noisy-domain recordings for every deployment setting. By analogy with the recent Turkish clinical NLP evidence discussed in Section 3.4.5, where domain-specific pre-training and fine-tuning on Turkish biomedical or clinical text consistently outperform generic and multilingual baselines, similar domain adaptation is expected to be an important factor for Turkish clinical ASR.

For the Turkish use case, published evidence on ASR performance in Turkish clinical consultations, particularly for paediatric neurodevelopmental and neuropsychiatric settings such as ASD and ADHD assessment, appears limited at the current stage. The absence of publicly available Turkish clinical ASR benchmarks means that domain- and use-case-specific validation is required.

3.2. Paralinguistic event detection

3.2.1. Paralinguistics

Paralinguistics has been defined and described throughout the years in many different ways. We start with one of the earlier descriptions of paralanguage by Trager (1958). Trager talks about speech, the process of talking, and how three kinds of events employing the vocal apparatus are happening during talking: a) language, b) variegated other noises, not having the structure of language – vocalizations, and c) modifications of all the language and other noises – voice qualities. The vocalizations and voice qualities together are being called paralanguage. Rephrased in a different and concise way, paralinguistics can be defined as the study into the *vocal but nonverbal elements (nonverbal means not using words) of communication by speech*[1]. We will start with describing these vocal nonverbal elements (in terms of features) and will then move to describing what kind of information these elements can convey. In the last 20 years, a field coined *computational paralinguistics* (Schuller & Batliner, 2013) has emerged that aims to automate processes in paralinguistics research and equip machine intelligence with paralinguistic analysis abilities.

3.2.2. Features

Paralinguistic features refer to a range of non-linguistic (non-verbal) elements that accompany spoken language and that convey information about how something is said (e.g., affect), rather than what is said. These elements include acoustic features, non-verbal vocalisations, and turn-taking behavior. Typically, **acoustic** features in paralinguistics research include (Lee et al., 2015):

Prosody-related measures

- Fundamental frequency (f0)
- Energy/loudness
- Speaking rate

Spectral-related measures

- Mel-frequency cepstral coefficients (MFCCs)
- Mel-filter bank Energy coefficients (MFBs)

Voice quality-related measures

- Jitter
- Shimmer
- Harmonic-to-noise ratio

Since it remains largely unknown which features perform best for which task, researchers have proposed a standard, minimal set of features, the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), that includes the features mentioned above (Eyben et al., 2016) as well as others. This featureset is widely adopted by the research community and is often used as a baseline to improve replicability.

Non-verbal vocalisations (also known as non-lexical sounds, see Ward, 2006) are vocal sounds made without words. Examples of non-verbal vocalisations include laughter, crying, moaning, sighing, as well as mm-hmm, mmm, uh, uh-huh, ah (Ward, 2006; Trouvain and Truong, 2012; Kamiloglu & Sauter, 2024), and can serve several functions. In particular, functions of interest here include emotion regulation (managing and shaping one's emotional experiences), social bonding (the formation and strengthening of interpersonal relationships), and group cohesion (the promotion of unity and cooperation within a social group).

Turn-taking: The way people take turns in conversations can say something about the speaker's state or how the conversation is going. Information about frequency and duration of turns, of overlaps, and quality of overlaps (competitive or non-competitive, see Kurtic et al., 2013; Truong, 2013) can all be indicative of a speaker's status (e.g., Beattie, 1981) or level of engagement (Jokinen et al., 2013). Silence is a specific phenomenon in turn-taking that, especially in patient-clinician encounters, can be meaningful in a conversation, as silences can be perceived differently in conversation, e.g., as awkward, invitational, or compassionate (Back et al., 2009).

Instead of using hand-engineered features, end-to-end learning has also been used to recognize emotions in speech for example (e.g., Trigeorgis et al., 2016). In end-to-end learning, the model learns directly from the raw waveform. However, there are some disadvantages, especially when applied to a paralinguistics-related task. End-to-end learning usually requires a large amount of data that is usually scarce for paralinguistics-related tasks. For example, spontaneous emotional speech data or speech data from people with dementia are not largely available. Furthermore, while post-processing methods exist to make the output of a machine learning model more interpretable,

generally end-to-end learning models work as „black box" models which is not always preferred. Especially if the model is used for diagnostics in the health domain, one would like it to be somewhat transparent.

These paralinguistic features not only have linguistic (e.g., marking of lexical stress, question vs. statement) or pragmatic functions (e.g., sarcasm, dialog acts); they also serve non-linguistic purposes and can reveal personal characteristics of the speaker. Laver & Trudgill (1979) describe three classes of markers of identity in speech:

a) Those that mark social characteristics, such as regional affiliation, social status, educational status, occupation and social role;
b) Those that mark physical characteristics, such as age, sex, physique and state of health;
c) Those that mark psychological characteristics of personality and affective state.

The variety of types of non-linguistic information that speech can convey is also reflected in the tasks of the yearly Computational Paralinguistics Challenges, see Table 4.

Table 4: A selection of computational paralinguistics tasks[2]

Emotion	Interest	Gender	Age
Sleepiness	Intoxication	Intelligibility	Likability
Personality	Autism	Conflict	Social signals
Physical load	Cognitive load	Eating condition	Parkinson's
Nativeness	Native language	Sincerity	Deception
Snoring	Cold	Addressee	Heart beats
Infant crying	Orca activity	Baby sounds	Styrian dialects
Masks	Elderly emotion	Breathing	Escalation
Covid speech	Covid cough	Stuttering	Vocalisations

Latif et al. (2021) identified several opportunities and challenges of speech processing healthcare. In addition to paralinguistic and emotional state processing, these opportunities include speech processing for psychological disorders and speech processing for diseases diagnostic and monitoring.

3.2.3. Emotion

Speech emotion recognition (SER) is the task of recognizing a person's emotional state from their speech. While sentiment analysis is a related task (i.e., estimating valence from text), SER refers to recognizing emotions based on acoustic features only. Typically, developing SER starts with a ***database*** containing labelled emotional speech samples from which acoustic features are extracted (***feature extraction***) that are subsequently used in a ***machine learning model*** to learn the mapping between features and label (Madanian et al., 2023). The development of SER has followed the rise of end-to-end techniques for automatic speech recognition (ASR), making some parts of the pipeline less explicit. However, there are some specific challenges that the field copes with and that are very different from the ones in ASR. In the following paragraphs, we will give a concise overview of current SER methods based on several recent surveys (Hashem et al., 2023; Singh & Goel., 2022; Madanian et al., 2023).

Databases Compared to ASR, the amount of available speech data for SER is considerably smaller, and the characteristics relevant for the database are very different. For emotional speech databases, it is important for example to distinguish among different "modes" of emotional expression (acted vs spontaneous vs elicited), types and number of emotions, and different emotion annotation methods (categories vs dimensions, categories vs continuous, self vs observed). While there is a need for spontaneous, naturalistic emotional speech spoken in a variety of different languages, most of emotional speech databases are acted and in English or Chinese (Singh & Goel, 2022).

Table 5: Most often used emotional speech databases (Hashem et al., 2023) with some added recent ones

Database	Speakers	Emotion annotation	Style	Amount
EMODB (Burkhardt et al., 2005)	5m, 5f	Neutral, anger, fear, joy, sadness, disgust, boredom	acted	535 sentences (25min)
SAVEE (Haq et al., 2008)	4m	Anger, disgust, fear, happiness, sadness, surprise	Acted	480 sentences (30min)
IEMOCAP (Busso et al., 2008)	5m, 5f	Anger, happiness, surprise, excited, disgust, sadness, frustration, fear, neutral + valence, activation, dominance	Scripted + improvised	Conversations, (12h)
RAVDESS (Livingstone & Russo, 2018)	12m, 12f	Happy,sad, angry, fearful, surprise, disgust, calm, neutral, acted	acted	1440 sentences (1,5h)
CREMA-D (Cao et al., 2014)	48m, 43f	Anger, disgust, fear, happy, netural, sad	Acted	7442 sentences (5,3h)
MSP-IMPROV (Busso et al., 2017)	6m, 6f	Anger, sadness, happiness, neutral + valence, activation, dominance	Improvised, read, natural	3249 recordings (2,3h)
MSP-Podcast (Lotfian & Busso, 2019)	?	Anger, sadness, happiness, surprise, fear, disgust, contempt, neutral + valence, arousal, dominance	Podcasts	84125 turns (27h)
LSSED (Fan et al., 2021)	335m, 485f	Anger, happiness, sadness, disappointment, boredom, disgust, excitement, fear, surprise, normal	Lab setting	147025 sentences (206h)

Feature extraction A second important step in the development of SER is feature extraction. Here, usually a choice is made between the use of handcrafted or (automatically) learned acoustic features. Handcrafted features typically consist of for example prosodic features, the eGeMAPS feature set or spectrograms (see previous subsection) as previously explained. Learned features

consist for example of wav2vec 2.0 embeddings, representations of speech that are automatically learned from the raw audio (Baevski et al., 2020).

Machine learning A third important step in the development of SER is the choice for a machine learning method. Given the relatively small amount of data, traditional machine learning models such as Support Vector Machines and Random Forest are often still used. Deep learning models, such as CNNs that use spectrograms, have also been used (e.g., Huang et al., 2014). Recently, there is a growing interest in using end-to-end models for SER (e.g., Wagner et al., 2023, Pepino et al., 2021, Chen et al., 2024) and in developing pre-trained emotional speech models (e.g., Chen et al., 2024). End-to-end models replace handcrafted feature extraction and the choice of a separate classifier; instead, the model learns both features and emotion classification jointly. According to a study by Wagner et al. (2023), transformer-based models show very good cross-corpus generational robustness and appear invariant to domain, speaker, and gender characteristics compared to a baseline CNN.

The performance of SER is heavily influenced by the database (and thereby its characteristics) used. To give an idea of the current SER performance on a frequently used database, we can have a closer look at the IEMOCAP database. In Wagner et al. (2023), some general observations were drawn based on recent works that have used the IEMOCAP database and end-to-end models (wav2vec 2.0, HuBERT): they found that fine-tuning pre-trained weights yields a 10% boost, additional ASR fine-tuning did not help with SER, larger architectures were typically better than the base ones, and HuBERT outperformed wav2vec 2.0. More particularly, they reported unweighted average recall rates between 60.0%-74.3% for end-to-end models performing a 4-class emotion recognition task on IEMOCAP. One of their wav2vec2-based models fine-tuned on the MSP-Podcast dataset to recognize dimensions of arousal, valence, and dominance was made publicly available[3]. However, in a survey carried out by Singh & Goel (2022), higher accuracies were found for the same database when using handcrafted features (55%-83%) and spectrograms (83%-89%) compared to automatically learned features (67%-73%). A wav2vec2 model[4] fine-tuned on IEMOCAP, developed by SpeechBrain, yielded an average accuracy of 75.3%. For the IEMOCAP database, it seems that using end-to-end modelling does not always yield the best SER performance. Another wav2vec2 model[5] was fine-tuned on TESS, CREMA-D, SAVEE and RAVDESS and yielded ~80% accuracy on the same datasets for 7 emotion classes. Finally, more recently a foundation WavLM-based model called Vesper[6] was released that is pre-trained on the LSSED dataset and can be fine-tuned for specific speech emotion recognition tasks and datasets.

3.2.4. Health state

As voice production depends on several physiological processes and systems in mind and body, diseases can affect normal voice production. In the context of voice, a **vocal biomarker** is a feature (or a combination of features) in the voice that has been identified and validated as associated with a clinical outcome (Fagherazzi et al., 2021; Kraus, 2018). The use of vocal biomarkers in combination with machine learning has been increasingly investigated for automatic diagnosis, risk prediction, and monitoring of **neurological** and **mental health** disorders (including **psychiatric** and **mood disorders**) (Low et al., 2020; Fagherazzi et al. 2021; Idrisoglu et al. 2023). Low et al. (2020) report in their survey that 50% of all studies on automated assessment of psychiatric disorders using speech target major depressive disorder, followed by 18% targeting schizophrenia, 16.5% targeting bipolar disorder and 8% post-traumatic stress disorder. Paralinguistic features such as f0 and speech rate seem to correlate well with depression and schizophrenia respectively (Low et al., 2020). A decrease in f0 and f0 range has been observed in depressed individuals while several studies found that speech rate is lower in people with schizophrenia (Low et al., 2020). In

Idrisoglu et al. (2023), a survey is presented on voice- and machine learning-based diagnosis of systematic conditions affective voice, nonlaryngeal aerodigestive disorders affecting voice, and neurological disorders affecting voice. They list that Parkinson's disease (60%) was the most targeted disease, followed by dementia (12%), cognitive impairment, ALS, depression and autism. With respect to machine learning models, 35% of all studies used support vector machines while 27% used neural networks. While the research is evolving, various surveys have identified similar challenges for the field. These include the use of unbalanced, closed datasets, as well as the under-researched task of monitoring instead of diagnostic testing (Fagherazzi et al., 2021; Idrisoglu et al., 2023). Finally, while in general the results are reasonable, many of the studies are relatively small feasibility studies that still need a validation against gold standards in order to enable large-scale clinical development (Fagherazzi et al., 2021).

3.2.5. Opportunities for paralinguistics in electronic healthcare records

Paralinguistic analysis has a long history in healthcare, as seen from the studies above; however, for other purposes than in the context of EHR. Typically, digital scribes who transcribe physician-patient encounters for medical note-taking or EHR purposes discard paralinguistic phenomena. In the following section, we will present potential benefits of incorporating paralinguistic aspects into the speech-to-text process, as speech contains more than words. In clinical encounters, paralinguistic signals such as hesitation, silence, distress, crying, laughter, backchannels and changes in speaking rhythm may provide contextual information about patient communication and clinician-patient interaction. However, their use in clinical documentation should remain cautious. Such signals may support communication analysis or flag segments for review, but they should not be converted into diagnostic claims without specific validation, transparency and governance.

Emotions are important in physician-patient encounters

For various reasons, emotions are considered to play important roles in physician-patient encounters. In medical note-taking for example, patients emphasized including the patients' social and family history, as well as patients' emotional reactions to diagnoses and health conditions (Hanson et al., 2012). The SOAP (Subjective, Objective, Assessment, Plan) structure is often adopted by healthcare practitioners to document information provided by patients (Podder et al., 2023). The first heading is Subjective and includes experiences, personal views or feelings of a patient (Podder et al., 2023) while in the second heading Objective, the same categories as in Subjective but instead measured or observed by the physician are described. The Objective heading generally begins with a statement regarding the patient's overall identification and impression, for example "62-year-old White female, alert, oriented, relaxed, and in no apparent distress" (Pearce et al., 2016). Hence, while not very explicitly and prominently present, information related to the patient's emotional state or feelings could be considered part of the medical note-taking which is relevant for the physician's overall impression of the patient, and for maintaining a successful physician-patient relationship.

It is widely acknowledged that physician empathy is important in the physician-patient relationship and has significant impact on a patient's recovery (e.g., Larson & Yao, 2005; Lelorain et al., 2012; Zachariae et al., 2003). Empathy involves understanding other people's feelings and perspectives. In a survey on associations between empathy measures and patient outcomes in cancer care, Lelorain et al. (2012) concluded that a clinician's empathy is associated with higher patient satisfaction, better psychosocial adjustment, lesser psychological distress and need for information. Likewise, in another survey, Zachariae et al. (2003) concluded that higher physician

attentiveness and empathy were associated with greater patient satisfaction, increased self-efficacy, and reduced emotional distress.

Hence, emotions are important in physician-patient encounters and SER technology could potentially enrich and support EHR with information about the patient's feelings. Since empathy is part of the physician's training, SER technology could also play a role in supporting the physician in their effective and empathetic communication by providing feedback and training the physician to become a better communicator (e.g., Chen et al., 2020). If paralinguistic information is explored in MediSpeech, the documentation workflow should distinguish observed communication cues from inferred emotional or health states. This distinction is important to avoid over-interpretation, bias and inappropriate inclusion of sensitive or weakly validated inferences in the EHR.

Potential for disease monitoring

As mentioned previously, speech can function as a vocal biomarker for several diseases and has great potential for diagnostic testing. Disease monitoring however has received less attention. While the main function of the physician-patient encounters is to monitor the patient's disease through the patient's responses, the encounters could potentially also be used for passive monitoring of the patient's disease by analysing the patient's speech (Fagherazzi et al., 2021; Idrisoglu et al., 2023).

Improving transcription accuracy: Nonverbal vocalisations are challenging for ASR

While nonverbal vocalisations can be indicative of a person's emotional and interpersonal state or even a person's health state and can convey important information in clinical conversations where a "mhm" could be a yes-answer in response to a physician's question (Tran et al., 2023), current open-source ASR models have trouble recognizing these nonverbal vocalisations correctly (5,6,8, Tran et al., 2023). Moreover, sometimes these nonverbal vocalisations are discarded all together which is the case for laughter. Tran et al. (2023) argue that not only the transcription of nonverbal vocalisations such as "uh-huh" and "mm-hm" is important in the context of clinical conversations, their intention (or dialog act) is just as important. For example, a "uh-huh" or "mhm" can express an acknowledgement/backchannel, or positive or negative affirmation. Although the proportion of clinically relevant nonverbal vocalisations is very small, Tran et al. (2023) still urge researchers and developers to improve the transcription accuracy of these nonverbal vocalisations that are also potentially relevant for the success of downstream tasks. CrisperWhisper is one of the few ASR models that is able to deal with disfluencies (Zusag et al., 2024). For non-verbal vocalizations such as laughter, there have been quite some work on automatic laughter detection (e.g., Gillick et al., 2021), but not many of these have integrated laughter detection in the ASR pipeline (Ho et al., 2025).

Downstream tasks can benefit from paralinguistic aspects

Several studies have pointed out how the extraction of paralinguistic information in physician-patient encounters can be relevant for downstream tasks. Since many downstream tasks such as summarization or other clinical documentation tasks that facilitate patient care rely on ASR transcriptions, the quality of these transcripts is critical for the success of these tasks (Tran et al., 2023; Quiroz et al., 2019). Hence, improving the transcription accuracy of nonverbal vocalisations that are usually poorly transcribed in ASR could support the success of several specific downstream tasks (Tran et al., 2023; Quiroz et al., 2019). ASR transcriptions can also help the development of AI-based decision-making systems (Ovesdotter Alm, 2016). For example,

paralinguistic aspects in speech, such as filled pauses, can reveal how certain physicians are in their diagnosis, which is critical information for decision-making systems.

[1] https://dictionary.apa.org/paralanguage
[2] http://www.compare.openaudio.eu/tasks/
[3] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
[4] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP
[5] https://huggingface.co/Dpngtm/wav2vec2-emotion-recognition
[6] https://github.com/HappyColor/Vesper

3.3. Speaker diarisation

3.3.1. Introduction

Speaker diarisation (SD) is the process of dividing an audio stream containing human speech into distinct segments based on speaker identity. This technique improves the clarity of ASR output by organizing speech into individual speaker turns.

SD involves two key steps—**speaker segmentation**, which detects transitions between speakers in an audio stream, and **speaker clustering**, which groups speech segments based on speaker characteristics. Note: SD does not identify speakers, it only allocates speech fragments to the various speakers in a recording. Combined with ASR, it states which speaker turns where uttered by speaker 1, by speaker 2, etc. If speaker identities are known, the labels, speaker 1, speaker 2, etc., can be replaced by these identities, e.g. Doctor X, Patient Y, in the MediSpeech context.

It is relevant to realise that the baseline for SD performance depends on the number of speakers that should be distinguished. In the context of doctor patient conversations, the number of speakers is typically two, giving a chance level of 50% for correct performance (and if the patient is accompanied by a third person, this gives a 33% chance level). In the MediSpeech clinical context, speaker diarisation has a direct documentation and safety role because the system must distinguish clinician speech, patient speech and, where relevant, speech from caregivers, guardians or other accompanying persons. This is especially important in consultations where clinically relevant information may be provided by someone other than the patient. Wrong speaker attribution can change the interpretation of a note, for example by presenting a patient-reported symptom as a clinician assessment, attributing caregiver information to the patient, or treating a clinician plan as a patient statement. Evaluation should therefore include word-level diarisation errors, role attribution errors and their clinical significance, not only aggregate DER.

Other factors affecting the performance of SD systems are background noise, speaker overlaps, and length of speaker turns (short speaker turns being harder to detect). In standard types of interviews the interviewee has the largest share of the conversation whereas the contribution of the interviewer is limited to short questions, be it sometimes with a longer introduction. Medical conversations of doctors and patients may deviate from this pattern with equal or larger shares of doctor contributions. Also, speaker overlaps are expected to be less frequent and shorter in duration.

Whereas performance of ASR systems heavily depends on the language involved (largely favouring English), this is less the case for SD where speaker properties are more relevant than language. However, dependencies still exist, especially since, in recent encoder-decoder architectures, all speech properties are encoded in an intertwined manner. Further, some SD methods are based on speech transcriptions rather than on speech acoustics (see below) which also increases their language dependency.

The performance of a SD system is expressed as Diarisation Error Rate (DER), which is the proportion of time that speech is allocated to a wrong speaker. A lower DER therefore points to a better performance. Another measure used in literature is Word-level DER (WDER), which is the proportion of words attributed to the wrong speaker. WDER is used to evaluate diarisation systems that incorporate ASR. Typically, the WDER of a system is lower than its DER. Alternatively, F scores (or F1 measure) are reported which is an accuracy score for the identification of speakers on each segment in an audio recording, which is calculated considering the duration of each segment.

Most systems work directly on the acoustic audio signal. The introduction and elaboration of Deep Learning models gave a boost in SD performance. Earlier systems based on other technologies typically have not been updated after 2020 due to lack of competitive potential compared to such DL approaches. The most popular SD methods these days are Pyannote, NVIDIA Nemo, and EEND. Pyannote is integrated into some Whisper implementations.

There are also Audio-Visual Speaker Diarisation methods and models (Mingote, et al., 2024) but we consider these less relevant for the MediSpeech context.

Recent reviews of SD systems can be found in Al-Hadithy et al. (2022), Nagavi et al. (2024), Park et al. (2022), O'Shaughnessy (2025), Rahim et al. (2025).

A dynamically updated list of ongoing SD research can be found here: https://wq2012.github.io/awesome-diarisation

Basically there are three approaches to speaker diarisation (SD):

- SD at equipment level
- Acoustic analysis with SD DL models
- Analysis of transcriptions with LLMs

which we will briefly discuss in the next subsections.

3.3.2. SD at equipment level (Microphone-Based Approaches)

These equipment-based methods rely on hardware and recording conditions to separate speakers. By employing specialized microphone arrays, beamforming, and spatial filtering techniques, systems can largely isolate speakers before any acoustic or linguistic processing. Such approaches are particularly beneficial in controlled environments (e.g., conference rooms, interrogation rooms or broadcast studios) where the positions of the speakers are known or can be reliably estimated.

Other applications of this method are found in online meetings and interviews where platforms such as Zoom or Teams are used. By using the transcription facilities of these platforms speaker separation is realised at the source being the microphone of the participants. A variant of this

approach in face-to-face meetings is to equip each participant with a (close-talk) microphone. Still, in such conditions speaker overlap still occurs especially for speakers positioned in close vicinity.

Performance:

Under controlled conditions with well-calibrated hardware, equipment-level diarisation systems have been reported to yield diarisation error rates (DER) in the range of 8–10% or lower. It is important to note, however, that performance is highly sensitive to room acoustics, background noise, and microphone placement. In less controlled or far-field settings, these systems can suffer from degradation if not paired with robust acoustic models.

By way of illustration, in the Dutch context a microphone array was used in an experiment in interrogation rooms of the Dutch police. DER performance was 12% (with five speakers in the room: the suspect, two interrogators and two lawyers).

For conversations recorded and transcribed via online meeting platforms such as Zoom and Teams, perfect speaker diarisation is possible. However for this optimal result, the transcription function of the platform must be used. Channel separation is not contained in downloads of the audio recordings of these platforms.

In general, this microphone-based approach is of less relevance for the context of the MediSpeech project where speaker separation by microphone is not practical. However, for online patient doctor meetings, the SoTa of current meeting platforms is more than sufficient.

3.3.3. Acoustic analysis with SD Deep Learning models

This is currently the most active and successful domain of speaker diarisation research. In these systems, raw audio is first transformed into discriminative acoustic feature representations (for example, using mel-frequency cepstral coefficients, i-vectors, or more recently x-vectors/d-vectors). Deep neural networks—often built with architectures such as convolutional or recurrent neural networks and more recently end-to-end (E2E) models—are used to learn speaker embeddings that characterize each speaker's voice. Two main strategies are commonly employed:

- **Two-Stage Approaches:** First, a segmentation step detects speaker change points, and then clustering (typically based on agglomerative hierarchical clustering) groups the segments using the learned embeddings.
- **End-to-End Neural Diarisation (EEND):** These systems jointly optimize segmentation and clustering in a single network, streamlining training and often yielding significant performance gains.

Another distinction that can be made is between unsupervised and supervised speaker diarisation. In the latter case the identity of one of the speakers is known and focused on during the diarisation process. This can be of benefit in the medical domain where the identity of a doctor is known beforehand. "This allows for focusing on a doctor's speech processing, and thus end up with more reliable and accurate results for different use cases, such as the detection and analysis of diagnosis, prescriptions, and procedures." (Khoma et al., 2023).

Performance:

In standard benchmarks—such as the CALLHOME dataset or meeting recordings—state-of-the-art deep learning systems have reported DER scores as low as 5–10% in controlled or synthetic conditions and around 15–20% in more challenging real-world scenarios. For example, reviews by Park et al. and others highlight that systems based on robust embedding techniques (e.g., using the ECAPA-TDNN architecture or the pyannote.audio toolkit) can push DER values into this lower range. In their review Al-Hadithy et al. (2022) point to the problem of domain mismatch: “When a model is built on data from one domain, it usually performs poorly when applied to data from another”. Khoma et al. (2023) expanded pyannote for tasks of supervised diarisation which led to a DER of 15.52% and an F value of 90.79% compared to a DER of about 25% for the unsupervised baseline system.

Medaramitta (2021) studied SD performance in simulated patient doctor interviews by medical students using standard Google and Amazon SD services and concluded that “the current speaker diarisation capabilities alone are not able to provide sufficient performance for our particular use case when integrating them into ReadMI for operating in in-person role-play training environments”. More recently, Tran et al. (2023) found more acceptable Word-level DER ranging between 1.3% for Amazon General ASR and 6.3% for Google General ASR. An interesting study by Sanni et al. (2025) on a benchmark dataset of 50 simulated medical and non-medical African-accented English conversations, using pyannote and NVIDIA Titanet, reports much higher DER scores for the medical part of the dataset (35% vs 12%) which is attributed to presence of speaker overlaps.

3.3.4. Analysis of transcriptions with LLMs

A new and promising avenue involves leveraging the power of large language models to refine speaker diarisation. In these systems, automatic speech recognition (ASR) converts speech segments to text, and LLMs subsequently analyze the transcribed content to determine speaker turns. Here the diarisation is based on text (the transcriptions) rather than on voice characteristics. The linguistic context—such as conversational cues, discourse markers, and even topic shifts—can help resolve ambiguities that pure acoustic analysis might miss. This approach is often combined with the acoustic SD methods described above.

Performance

While still an emerging area, early experiments indicate that integrating transcription analysis via LLMs with acoustic cues can produce a relative improvement compared to acoustic-only diarisation systems. For instance, if a baseline deep acoustic model yields a DER of approximately 10%, a combined system incorporating LLM insights might reduce the error rate by an additional margin—bringing it down into the high 8% region in some scenarios. It should be noted, however, that standardized benchmarks remain sparse given this method’s novelty, and performance gains are often domain-dependent (particularly benefiting conversational or meeting scenarios where language context is rich).

In the medical domain of doctor patient conversations, Nehrboss (2024) presents the CASCADE system with three specialist LLMs, one of which has to identify speaker roles. For medical dialogues he reports a decrease of WDER from 16.1% for the baseline system to 1.6% for the CASCADE system. As limitations of the approach, he mentions the confinement to only two speakers (and expects severe degradation for more speakers) and the heavy computational load of LLM processing, which permits offline processing only.

Consequently, diarisation should be evaluated at least at three levels: segmentation of turns, attribution of words to the correct speaker, and attribution of clinically salient entities or statements to the correct speaker role. Two-speaker consultations are the simplest case, but primary care, pediatrics, pregnancy care, and multidisciplinary interactions may involve additional participants. The system should therefore make uncertainty visible and should allow clinician correction of speaker labels before any information is inserted into the EHR.

Adedeji et al. (2024), investigating the PriMock57 dataset, found that LLMs, particularly through CoT (Chain-of-Thought) prompting of six commercial LLMs, improved the diarisation accuracy of existing ASR systems. Here, diarisation was combined with ASR to calculate doctor- and patient-specific WER scores. It was concluded “that the best-performing LLM and ASR system combinations, particularly GPT-4, Gemini Pro, and Gemini Ultra paired with Whisper 1, demonstrated superior diarisation accuracy compared to each and every ASR benchmark. These pairings yielded a lower Doctor-Specific Word Error Rate (D-WER)[1] and exhibited reduced variability in results, indicating more reliable and consistent diarisation performance“(p.11).

3.4. Named Entity Recognition of Medical Terms

3.4.1. Introduction

Named Entity Recognition (NER) is a fundamental task in the field of Natural Language Processing. Its primary function is to identify and classify predefined categories of objects, known as named entities, within unstructured text. These entities are rigid designators, referring to specific real-world objects such as people, organizations, locations, and dates. The process involves taking a string of text, such as a sentence, a paragraph, or an entire document, and programmatically locating and categorizing these entities.

When applied to the medical and biomedical domains, this task is specialized in Biomedical Named Entity Recognition (BioNER) or Biomedical Entity Linking (BEL). Here, the entities of interest shift from general categories to highly specific medical concepts, including diseases, symptoms, medications, treatments, procedures, anatomical parts, genes, and proteins. The overarching goal of BioNER is to transform the immense and ever-growing volume of unstructured text from sources such as Electronic Health Records (EHRs), clinical trial documents, biomedical research literature, and discharge summaries into structured, computable data.

This process is vital for the normalization and aggregation of entity mentions, which has applications in categorizing and improving search in medical and scientific literature, as well as extracting information from clinical notes and patient forums. For instance, a BEL model can link a patient's mention of "trouble with sleeping" to the medical concept "insomnia" in an ontology, enabling the aggregation of similar mentions across patients.

In MediSpeech, medical NER should be understood as part of the post-ASR clinical structuring layer. The task is not only to identify isolated terms, but also to extract clinically relevant problems, symptoms, diagnoses, medications, doses, measurements, laboratory values, risk factors, follow-up plans, and referrals from speech-derived text or draft notes.

3.4.2. Key Challenges

- **High intra-class variance:** Identical biomedical terms can appear in various forms, such as "MI" and "hartaanval" (heart attack) both referring to "myocard infarct".
- **Low inter-class variability:** Different biomedical terms can be very similar, like "candida" (yeast) and "cardia" (heart), despite having distinct meanings, only have a Levenstein distance of 2.
- **Noisy Free Text:** Texts generated by patients and medical professionals often contain spelling errors and personal abbreviations, making entity identification complex.
- **Vast Number of Entities:** The biomedical domain is enormous, with ontologies like the Unified Medical Language System (UMLS) containing over 3.3 million unique concepts.
- **Limited Labeled Datasets:** There is a scarcity of labeled biomedical entity linking datasets, especially in languages other than English.
- **Gap between lay language and clinical language**: patients don't use professional biomedical terms to describe things. For example, they could say "his heart stopped", instead of "cardiac arrest". This makes it difficult to accurately map patient-doctor conversations or text into clinical terms.

3.4.3. Evaluation Metrics

In biomedical entity linking, evaluating a model's performance requires a multifaceted approach using several key metrics. These metrics assess the model's ability to accurately map a mentioned entity in a text (e.g., "heart attack") to a specific concept in a standardized knowledge base like the Unified Medical Language System (UMLS).

Candidate Ranking and Accuracy

A primary metric for this task is **Recall@k**. This evaluates the two-stage process of entity linking: candidate generation (retrieving a list of potential concepts) and candidate ranking (ordering that list by likelihood). **Recall@k** measures whether the correct, or "gold," entity is present within the top 'k' ranked candidates proposed by the model. For example, Recall@5 checks if the correct answer is among the top five suggestions. A special case, **Recall@1**, is equivalent to **accuracy**, as it only considers the single highest-ranked candidate, which must be the correct one to be counted. This metric directly reflects the model's top-choice performance.

Balancing Errors in Classification

Standard classification metrics such as precision, recall, and **F1-score** are also fundamental.

- **Precision measures the proportion of a model's links that are correct, indicating its reliability (Precision=TP/(TP+FP)).**
- **Recall measures the proportion of all possible correct links that the model successfully identifies, indicating its comprehensiveness (Recall=TP/(TP+FN)).**

In medical contexts, both false positives (linking to a wrong concept) and false negatives (failing to link an existing concept) can have severe consequences. Therefore, the F1-score, which is the harmonic mean of precision and recall ($F1=2\cdot(Precision\cdot Recall)/(Precision+Recall)$), is particularly crucial. It provides a single, balanced score that reflects the model's ability to minimize both types of errors.

Semantic Correctness

Finally, to account for the complex and hierarchical nature of medical terminologies, **1-distance accuracy** is often used. This is a more lenient metric where a prediction is considered correct even if it's not the exact gold entity, as long as it is directly related to it in the knowledge base hierarchy. For instance, if the model links a mention of "Barrett's esophagus" to the broader concept "Esophageal Diseases" in UMLS, and these two concepts are connected by a single relational link (e.g., is_a), 1-distance accuracy would count this as a correct prediction. This metric offers a more nuanced evaluation of a model's semantic understanding, rewarding predictions that are contextually and clinically very close, even if not identical.

3.4.4. Resources for Dutch Medical NER

Murphy et al. (2025) have published a manually annotated corpus of Adverse Drug Events (ADE), specifically tailored for NER task. It focuses on annotating **Drug entities**, **Disorder entities** (encompassing diseases, syndromes, signs, symptoms, and findings under the UMLS Disorder semantic group), and **Qualitative Concept entities** (specifically signal words like 'toxic', 'nephrotoxic', and 'iatrogenic' relevant to ADEs).

The corpus comprises **102 anonymized clinical progress notes** from Intensive Care Unit (ICU) patients admitted to Amsterdam University Medical Centre in the Netherlands between November 2015 and January 2020. These notes were specifically selected because they were suspected of containing drug-related acute kidney injury (AKI) mentions, although the corpus annotates *any* ADE found

Personal attributes such as names, addresses, dates, and referring hospitals were removed and replaced with placeholders (e.g., <NAME>, <ADDRESS>). Rare diseases were replaced with <RARE_DISEASE>, and all dates were randomly shifted to a future range (2100-2200) while preserving timelines. This rigorous de-identification ensures patient privacy and makes the dataset anonymous and sharable upon request.

Content Volume: The corpus contains **16,470 labels**, including 14,355 entity labels (8,914 Disorder, 5,307 Drug, 134 Qualitative Concept) and 2,115 relation labels (614 ADE, 1,501 Indication). While primarily focused on drug-related AKI, the study identified **261 distinct ADE mentions** in total, with **158 (60.8%) pertaining to other types of ADEs**, underscoring the richness of clinical notes as a source of ADE information.

Mantra Gold-Standard Corpus (GSC) is a multilingual, manually annotated collection of texts designed for training and evaluating biomedical concept recognition systems. Its goal was to create the first gold-standard corpus for this task in languages other than English. The methodology developed to create the corpus is a key contribution, designed to ensure high-quality and consistent annotations across all languages. To reduce the manual workload, the process started by providing human annotators with automatically generated "preannotations" from five different concept recognition systems. For each language, three annotators then independently corrected these preannotations and added any missing concepts. After this, the separate annotations were automatically merged into a single "harmonized" set using the e-centroid method. This unified set was then reviewed by expert curators who adjudicated any discrepancies to create the final annotations. A major strength of the process was the use of parallel texts, which allowed the curated English annotations to serve as a high-quality reference and enabled final cross-language consistency checks to resolve any remaining differences.

Dutch Medical Concepts Repository repository by Utrecht UMC provides **instructions and code to create concept tables of Dutch medical names** (primary names, synonyms,

abbreviations, misspellings) based on UMLS, SNOMED CT, and Human Phenotype Ontology (HPO), which can then be used in named entity recognition and linking methods like **MedCAT**. This provides a foundational resource for building comprehensive concept vocabularies for Dutch medical language.

Another repository, released by Erasmus Medical Center (Nijmegen), contains clinical corpora in the Dutch language for validating and training biomedical natural language processing models. Both MedMentions and Mantra have been translated from English to Dutch using the Google Cloud Translate API and OpenAI GPT-4 API.

Hartendorp et al. introduced **WALVIS, a weakly labeled Dutch biomedical entity linking dataset** automatically generated using Wikidata and Wikipedia, addressing the lack of labeled datasets. This dataset is used to fine-tune BEL models like **MedRoBERTa.nl**, which is pretrained on anonymized Dutch hospital notes and further improved through self-alignment pretraining on a cleaned Dutch sample of UMLS and SNOMED. This demonstrates an approach to overcome data scarcity by leveraging existing knowledge bases and Wikipedia.

3.4.5. Turkish medical NLP, Named Entity Recognition and terminology resources

Turkish biomedical and clinical Natural Language Processing (NLP) can be considered a relatively low-resource setting compared with English. The limited availability of large-scale domain-specific corpora and annotated clinical datasets restricts the development and reproducible evaluation of Turkish medical NLP systems. In addition, Turkish is a morphologically rich and agglutinative language, where productive suffixation increases lexical variation and presents additional challenges for tokenisation and representation in downstream NLP tasks (Türkmen et al., 2023). Domain-specific terminology, abbreviations, spelling variations and non-standard expressions further increase the complexity of processing real-world clinical narratives.

Recent work demonstrates that domain adaptation can improve the representation of Turkish biomedical and clinical text. Türkmen et al. (2023) introduced **BioBERTurk**, a family of Turkish biomedical pretrained language models, and evaluated different pre-training strategies using Turkish biomedical corpora and radiology-related resources. The models were evaluated on a labelled dataset of Turkish head CT radiology reports. Continual pre-training of BERTurk with biomedical data produced better results than general-domain BERTurk, multilingual BERT and the evaluated baseline models, indicating the value of domain-specific adaptation for Turkish clinical NLP.

Resources directly targeting entity recognition in Turkish clinical text remain limited. Dilmaç and Alpkocak (2026) introduced a manually annotated benchmark dataset for de-identification of Turkish Electronic Health Records (EHRs) and evaluated traditional sequence-based approaches together with fine-tuned transformer models for detecting protected health information. The best fine-tuned transformer model achieved a macro F1 score of **92.20%**, compared with **83.00%** for the BiLSTM-CRF approach. Although de-identification focuses on protected health information rather than clinical entities such as symptoms or diagnoses, the study provides evidence that task-specific transformer fine-tuning can be effective for token-level entity extraction from Turkish clinical narratives (Dilmaç & Alpkocak, 2026).

Other recent Turkish medical NLP resources also highlight the importance of domain-specific adaptation and terminology normalisation. **TurkMedNLI** provides a Turkish medical natural language inference dataset derived through an LLM-supported translation and normalisation pipeline that explicitly addresses medical abbreviations and measurement expressions. Experiments with BERTurk showed differences between original and normalised medical text and demonstrated that performance on medical-domain data is substantially stronger when models are trained with relevant medical examples than when relying only on general-domain NLI resources (Oğul et al., 2025).

The relevance of domain adaptation is also demonstrated in real-world Turkish clinical narratives. Akbasli et al. (2025) evaluated domain-specific fine-tuning of a large language model on Turkish clinical notes from paediatric emergency admissions. The dataset included unstructured records containing typographical variation, and the fine-tuned model substantially outperformed the pretrained model in the evaluated classification task. Although the study addresses respiratory tract infection classification rather than general-purpose medical NER, it provides additional evidence that domain-specific adaptation can be important when processing noisy Turkish clinical text (Akbasli et al., 2025).

For the Turkish MediSpeech use case, medical information extraction should therefore extend beyond the recognition of isolated entity mentions. Relevant information may include complaints and symptoms, diagnoses, examinations and procedures, medications, tests and clinical observations, together with contextual attributes such as negation, temporality and uncertainty. Where appropriate, extracted mentions should also be normalised and linked to standard clinical concepts while preserving the original Turkish expression and its context. Evaluation should consequently consider entity-level precision, recall and F1 together with semantic correctness and terminology-mapping accuracy. The limited availability of representative Turkish clinical datasets, particularly for conversational and specialised clinical settings, remains an important gap for domain-specific validation within MediSpeech.

3.4.6. Model Evaluation

Liu et al. (2024) have performed an extensive evaluation of various approaches to NER in English. The study evaluated seven distinct NER models across three challenging medical datasets: Revised JNLPBA, BioCreative V CDR (BC5CDR), and Anatomical Entity Mention (AnatEM). The models were grouped into three categories: statistical machine learning, BERT-based deep learning, and LLMs.

The statistical models, **Hidden Markov Models (HMM)** and **Conditional Random Fields (CRF)**, delivered a moderate performance. While CRF generally outperformed HMM across all datasets, with higher accuracy scores, both fell short of the more modern architectures. For instance, on the Revised JNLPBA dataset, CRF achieved a microaverage F1 -score of 0.8476, whereas HMM scored 0.7265.

In contrast, the **BERT-based deep learning models**—BioBERT, RoBERTa, BigBird, and DeBERTa-consistently demonstrated strong performance and superior accuracy. Among them, **BioBERT** emerged as a top performer, achieving the highest accuracy on the Revised JNLPBA (0.932 microaverage) and AnatEM (0.8494 microaverage) datasets. Its success is largely attributed to its specialized pretraining on extensive medical literature, which allows it to capture domain-specific nuances effectively.

The fine-tuned LLM, **Gemma**, showed exceptional but variable performance. It achieved the highest accuracy on the BC5CDR dataset with a near-perfect microaverage score of 0.9962. This dataset, which contains only two entity types (disease and chemical), appears well-suited to Gemma's architecture. However, its performance was less pronounced on the more complex Revised JNLPBA and AnatEM datasets compared to BioBERT, indicating a need for further dataset-specific optimization.

To evaluate the performance of deep learning models on WALVIS, Hartendorp et al. use **MedRoBERTa.nl**, a language model pretrained on Dutch hospital notes, as their base model. They first perform self-alignment pretraining (SapBERT) using the enhanced Dutch ontology to teach the model to recognize synonyms. Following this, the model is fine-tuned on the newly created WALVIS dataset. At inference time, all terms from the enhanced ontology are converted into embeddings and stored in a computationally efficient FAISS index. When a new, unseen mention is provided, its embedding is generated and compared against the indexed embeddings to find the nearest neighbour, thereby linking it to the most similar concept in the ontology.

The quantitative results of the final model, evaluated on the Dutch portion of the Mantra GSC corpus, are as follows: the model achieved a **classification accuracy of 54.7% and** a **1-distance accuracy of 69.8%**. This performance represents a 10.1 percentage point improvement in classification accuracy and a 13.1 percentage point improvement in 1-distance accuracy compared to the base model without the study's specific pre-training and fine-tuning.

In a separate case study on patient-support forum data, a manual evaluation of a small sample found that for correctly recognized entities, the model linked approximately **65%** to a correct or closely related concept.

Entity-level evaluation should include precision, recall and F1 for clinically relevant categories, but also semantic correctness, assertion status, negation, temporality, uncertainty and mapping accuracy. These metrics should be linked to clinically significant error review and human adjudication

3.5. Ambient clinical documentation, medical scribes, and LLM-based reporting

Ambient clinical documentation systems build on ASR but extend beyond transcription. They capture the clinical encounter, identify speaker turns, extract clinically salient content and generate draft documentation such as consultation summaries, SOAP-style notes, referral drafts or patient instructions. Recent studies and reviews of AI scribes suggest potential reductions in documentation burden and improvements in perceived clinician experience, while also showing that note quality, editing burden, integration and safety remain active concerns (Duggan et al., 2025; Olson et al., 2025; Seth et al., 2024).

The main methodological implication is that an ambient scribe should not be evaluated as if it were only a transcription engine. A draft clinical note can be fluent and well-structured while still omitting clinically important information, introducing unsupported content, misrepresenting uncertainty, or over compressing the patient narrative. Human review is therefore a baseline requirement for documentation workflows unless and until a specific, validated and approved workflow justifies a different approach.

Recent evidence also suggests that the value of AI scribes should not be interpreted as full automation of clinical documentation. Reported benefits are mainly associated with reductions in perceived documentation burden, cognitive load and after-hours work, but these gains coexist with the need for clinician review, terminology correction and control of omissions or unsupported generated content. Studies evaluating LLM-generated clinical notes indicate that output quality may approach that of human documentation in selected settings, while still requiring structured review using instruments such as the PDQI-9 and clinically focused error assessment. Large-scale analyses of clinician-edited draft notes also suggest that post-editing often shifts generated text from patient-facing or conversational language towards more standardised clinical terminology. This reinforces the interpretation that the clinician's role is moving from sole author to accountable reviewer-editor, rather than disappearing from the documentation workflow (Morse et al., 2025; Palm et al., 2025; Cho et al., 2026; Accurx, 2026).

LLMs are relevant to this layer because they can support summarisation, information extraction, spell correction, report drafting, coding suggestions and transformation of transcript fragments into structured outputs. Domain-specific clinical language models have shown that large models can improve the extraction and interpretation of information from EHR narratives, but healthcare LLM reviews and evaluation frameworks continue to emphasize task-specific validation, human assessment and safety controls rather than generic deployment claims (Yang et al., 2022; Tam et al., 2024; Wang et al., 2025).

A useful design distinction is between documentation support, structured information extraction and clinical decision support. Documentation support and structured extraction may be appropriate early targets if outputs remain editable and reviewable. Active decision support, by contrast, requires stricter evidence boundaries, governance, explainability and responsibility allocation, because the system is no longer only helping to record care but may influence clinical judgement.

3.5.1. SoTa Clinical Information Extraction without LLMs

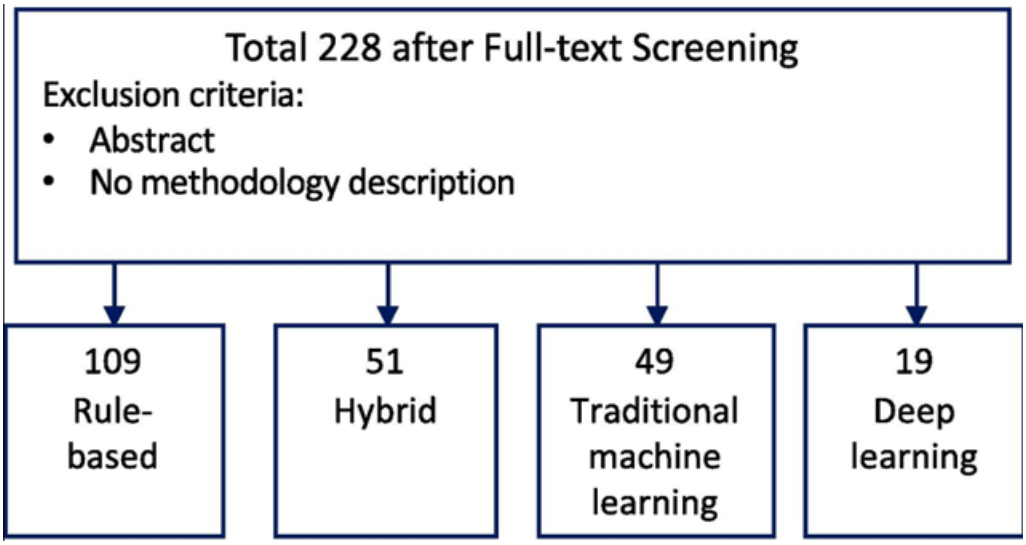
Automated clinical documentations created using ASR and annotated with name entity recognition are valuable sources of base information for clinicians, reducing their paperwork burdens and cognitive overheads of reporting from scratch. In a clinical environment, these documents are often used in conjunction with a wide range of lab test reports and follow-up visit notes (together, *clinical notes)* to form a complete picture of a patient for clinical decision making. This is an important dimension of MediSpeech clinical decision support systems. It enables our technology to go beyond voice-to-text transcription, SOAP-style report generation, name-entity recognition to cover relationship extraction, medication dose/quantity/volume extraction, lab test result extraction, post-operative complication and indication extraction for clinical decision support. The range of extended information extraction process provides a rich context for patient-doctor interactions, not just at one point in time, but throughout the clinical journey a patient goes through.

In practice, the most common information extraction strategy for clinical notes is still rule-based Natural Language Processing (NLP) and in some cases, together with sophisticated regular expression (for example, rules to deal with negations, and abbreviations) strategies, machine learning [Wang & Akella 2015]. The authors noted several challenges and advanced with the learning-based approaches which include

- **Generalisation:** Difficulty of generalisation (i.e. using a trained model for a new type of clinical dataset/notes renders the model inadequate)
- **Feature variations**: Loss of model accuracy when entities have many surface forms, for example, when a clinical term/phrase spells differently or briefly – causing ML features to collapse
- **Semantic features**: the "significantly boosted" (increased F1 by 34%) performance enhancement brought by incorporating semantic features, such as semantically related words of a concept or semantic type of a concept in SNOMED CT, to expand the "span" of training annotations or to simply use the type as a feature.

[Fu et. al 2020] classifies information extraction for clinical notes into four categories by studying 228 papers selected for the analysis, and they are: Rule-based, machine learning, deep learning, and hybrid (Rule-based + machine/deep learning). The literature screening process is summarised in Figure 1.

Figure 1: Literature screening flow for clinical information extraction studies



(Source: [Fu et. al 2020])

Although rule-based methods require significant manual interventions, it remains to be the most popular due to its transparency, low cost (both labour, and computation), and speed. The authors also noted that deep learning methods, whilst only covered by 8% of the papers, are gaining significant traction fast in adoption. Moreover, in the state-of-the-art benchmark results for the selected eight clinical tasks, such as clinical workflow optimisation, drug studies, deep learning occupies 6/8 of the top ranked benchmarks, showing remarkable achievements brought by deep learning models, including BioBERT, BERT-large, BERT-base (P+M).

3.5.2. SoTa Clinical Information Extraction and Reporting with LLMs and Knowledge Graphs

Recent advances in LLM have highlighted promising results for clinical text extraction with superior performance. It has been shown that LLM, specifically, instruction-tuned (prompt-engineered) LLaMA-2 and LLaMA-3-70B models, consistently outperformed the bidirectional encoder representations from transformers (BERT) model, a state-of-the-art deep learning model for information extraction, in both name entity and relation extraction tasks, across four types of clinical notes [Hu et. al. 2026]. The authors have shown that, with the 70B models, the F1 scores have increased by over 7% for NER and 4% for RE. An LLM based approach has been shown to offer better generalisability and adaptability to different NLP tasks, such as unseen clinical texts from different institutions. On the other hand, the BERT models tend to perform in high-volume and time-constrained tasks whilst perform poorly in unseen texts.

The evaluation was performed against a comprehensively annotated corpus of 1588 clinical notes from 4 data sources—UT Physicians (UTP) (1342 notes), Transcribed Medical Transcription Sample Reports and Examples (MTSamples) (146), Medical Information Mart for Intensive Care (MIMIC)-III (50), and Informatics for Integrating Biology and the Bedside (i2b2) (50), capturing 4 clinical entities (problems, tests, medications, other treatments) and 16 modifiers (e.g., negation, certainty). Whilst the LLM model performance is encouraging, the costs of achieving so were high. Infrastructure wise, to run the largest, i.e. 70B, LLM, 2 NVIDIA A100 80GB GPUs were used in the inference tasks. Computationally, the authors noted that the inference task runs 28 times slower when the LLaMA models are used, compared with BERT.

Whilst LLM demonstrates significant advancements in information extraction, the biggest challenge facing its full adoption in the clinical context is perhaps hallucination. In [Zhang et. al. 2023], the authors have shown that LLMs can occasionally hallucinate, i.e. fabricating irrelevant facts that were not relevant to the user input, generating contents that are self-contradictory, inconsistent and misaligned with established world knowledge. This has direct implications in clinical information extraction in that the capabilities exhibited in LLMs can vary depending on the type of text they deal with [**AB Das**, et. al 2025]. Further, the authors have shown that open weight LLMs, specifically Qwen2.5 and DeepSeek-v2, older versions released back in 2025, can perform reasonably well in admission notes but generally less consistent when it comes to follow-up notes.

Knowledge Graphs (KGs) have been investigated as one approach to grounding LLM outputs and reducing inconsistencies. It has been shown that KGs can enhance LLM performance across various phases at inference, model training, and output verification stages [Agrawal et. al. 2024, Lavrinovics et. al. 2024]. However, the authors in [Agrawal et. al. 2024] also noted careful adaption of KG usage is required to benefit LLM based reasoning in various use cases, including use of KGs in small LLMs, stepwise reasoning of KGs in larger models, for example.

In the context of clinical information extraction, incorporating KGs such as SNOMED, to augment LLM-based methods are gaining popularity [Škrlj et. al. 2025, Lavrinovics et. al. 2024, Agrawal et. al. 2024, Chang et. al. 2024, Alarcon-Serrano et. al. 2026]. However, due to the natural language nature of how people typically interact with LLM, i.e. using language-based prompt engineering and individualised chain of interactions involved in the reasoning and inference processes, the quality of extraction can vary significantly between different protocols [Alarcon-Serrano et. al. 2026]. This has opened up further research avenues in, for example, using a more structured reasoning

scaffolding to interact with LLMs to explicitly encode semantic or contextual success criteria in information retrieval.

3.6. Clinical Decision Support Systems

Early non-AI implementations involved clinicians and pharmacists in the United States who integrated data into Cerner, a primary clinical information system, to provide automated alert functions by comparing different rule-based configurations.(Welch and Kawamoto 2013) In oncology applications, deep learning models have been employed in clinical decision support systems to automatically extract features and categorize data depending on impact. They also suggest that these systems could be incorporated to forecast treatment outcomes and combine with longitudinal data to track the course of diseases. Utilizing algorithm recommendations derived from clinical guidelines, one study designed a CDSS for respective double-blind randomized controlled trial evaluation of effectiveness in managing psychiatric medications.(Zastrozhin et al. 2018) Four input steps were entered including data variant type, hospital and patient information, and language to generate a report recommendation on which drugs to have increased dose, decreased dose, standard dose, or stop using.(Zastrozhin et al. 2018) Using rule-based ontology, another study developed a CDSS named ClinGenWeb designed with the capability to report relevant clinical findings combined with factors (e.g. family history, environmental, behavioral, and clinical data) for assessment of individual increase in their prostate cancer risk.(Beyan and Son 2014a) The distributed software platform BioXM along with Zoho Reports business intelligence tool was used to rapidly prototype the user interface as well as incorporating existing prediction models prostate cancer risk.(Beyan and Son 2014a, 2014b)

3.6.1. ADHD Decision Support

Recent machine learning research on ADHD demonstrates that structured psychometric and demographic features — including WISC-IV subscales, continuous performance test scores, and behavioral rating scales — yield high discriminative performance when modelled with gradient-boosted ensembles. Navarro-Soria et al. (2025) showed that Random Forest and XGBoost classifiers, augmented with SHAP-derived feature attributions, successfully recover known neuropsychological correlates of ADHD such as working memory and inhibitory control deficits, providing clinically interpretable explanations alongside predictions. A critical methodological advance established by Mikolas et al. (2022) is the importance of training and validating models against real psychiatric referral cohorts rather than healthy controls, substantially improving clinical generalizability.

3.6.2. ASD Decision Support

Multimodal AI architectures currently represent the leading paradigm for ASD diagnostic support. Megerian et al. (2022) validated a Gradient Boosted Decision Tree system integrating caregiver questionnaires, clinician assessments, and video-derived behavioral features, demonstrating performance comparable to specialist evaluation. Crucially, the system introduced an indeterminate output class for low-confidence predictions — a design principle now widely recommended for trustworthy clinical AI. Wang et al. (2025) further demonstrated that machine learning errors in ASD classification are interpretable and aligned with clinically known sources of diagnostic ambiguity, advancing model transparency.

3.6.3. Explainability and Emerging Trends

SHAP (Lundberg & Lee, 2017) is widely used for post-hoc explainability in tabular clinical AI, enabling both global model auditing and patient-level feature attribution. Beyond established methods, the 2024–2026 period has seen growing interest in Large Language Models for clinical feature extraction from unstructured notes and Retrieval-Augmented Generation for evidence grounding, though regulatory guidance consistently discourages generative systems as standalone diagnostic tools. Trustworthy AI principles — encompassing calibration, fairness, uncertainty quantification, and human-in-the-loop governance — have emerged as essential design requirements rather than optional enhancements.

3.6.4. Research gaps and Turkish approach

Despite rapid progress, several gaps persist. The majority of published models rely on healthy-control designs that overestimate real-world performance. Formal clinical validation of model explanations — assessing whether SHAP outputs are perceived as interpretable and actionable by clinicians — remains limited. External multi-institutional validation is seldom reported, restricting confidence in generalizability. Finally, longitudinal clinical record integration and fusion of structured assessments with free-text clinical notes represent underexplored but clinically consequential directions.

The Turkish consortium's approach directly addresses the identified gaps through two independent binary classifiers (ASD and ADHD), trained on structured clinical and psychometric features from realistic psychiatric referral cohorts. Explainable ML models — Logistic Regression, Random Forest, XGBoost, and LightGBM — are combined with SHAP-based dual-scope explanations (global auditing and patient-level attribution), satisfying current regulatory requirements and clinical transparency standards. The system outputs probabilistic advisory evidence rather than deterministic diagnoses, preserving clinician authority and aligning with EU AI Act mandates (European Parliament and Council of the European Union, 2024). Independent classifier design further accommodates comorbid presentations without enforcing artificial mutual exclusivity.

3.7. Interoperability, clinical terminologies and EHR integration

The value of automated reporting depends on whether the output can be safely reused inside clinical information systems. Free-text transcripts are useful for review, but continuity of care, analytics, coding, quality monitoring and decision support usually require structured outputs, provenance information and terminology mappings. Interoperability should therefore be treated as a design requirement from the start rather than as a late-stage integration task.

HL7 FHIR provides a resource-based approach for exchanging healthcare information electronically, while terminologies such as SNOMED CT, LOINC, ICD and ICPC support different levels of semantic interoperability and classification (HL7 International, 2023a; Bodenreider et al., 2018). In practice, a MediSpeech output may need to preserve both the draft narrative and the structured concepts from which it was generated, so that clinicians can audit, correct and trace how each coded or structured field was produced.

Table 6: Standards and terminologies

Standard / terminology	Typical role	Relevance for automated reporting
HL7 FHIR	Exchange of structured health information through resources and APIs	Potential pattern for moving reviewed outputs, observations, notes or tasks into EHR workflows.
ICPC-2	Primary care classification of reasons for encounter, diagnoses and processes	Relevant where the use case involves primary care problem coding and longitudinal follow-up.
ICD-10 / ICD family	Diagnosis and morbidity classification	Relevant for reporting, analytics and alignment with established coding pathways.
SNOMED CT	Detailed clinical terminology	Potential reference terminology for concept normalisation and interoperability.
LOINC	Laboratory and clinical observations	Relevant to measurements, laboratory values and structured observations.

EHR integration should therefore be understood as semantic and workflow integration, not merely as text export. Recent ambient documentation products increasingly combine note generation with structured templates, letters, clinical observations and, in some cases, terminology coding such as SNOMED CT. This strengthens the relevance of HL7 FHIR resources and metadata such as Provenance for linking generated documentation to source transcripts, structured entities, clinician edits and final validated content. It also suggests that safer clinical NLP and coding pipelines should rely on explicit grounding in terminologies, guidelines and traceable mappings, rather than on free-text generation alone. In the European context, this direction is reinforced by the European Health Data Space, which places additional emphasis on interoperability, portability and governance of electronic health data (HL7 International, 2023a; HL7 International, 2023b; Gan et al., 2026; European Parliament and Council of the European Union, 2025).

For EHR integration, the system should support a review-and-commit workflow in which generated content is clearly marked as draft until approved by the clinician. Audit trails should record the source transcript, generated text, structured entities, human edits, final accepted content and any rejected suggestions. These requirements are important not only for usability, but also for accountability, incident review, data protection, and future model monitoring.

For the Turkish MediSpeech use case, interoperability is also relevant to the use of longitudinal patient information together with information extracted from the current doctor-patient conversation. The Smart Engine is expected to combine newly captured clinical information with relevant historical patient data, allowing previous diagnoses, examinations, tests, treatments and other available clinical records to provide additional context for downstream processing. This requires historical information to be retrieved and represented in a structured and traceable manner rather than being treated only as unstructured background text.

Semantic normalisation is therefore an important part of the Turkish workflow. Clinical expressions extracted from Turkish conversations and historical records may need to be linked to standard clinical concepts and classifications so that equivalent expressions can be interpreted consistently across different data sources and reused by downstream reporting and decision-support components. Where applicable, both the original Turkish expression and its normalised representation should be preserved, together with provenance information linking the structured concept to its source. Standards and terminologies such as HL7 FHIR, ICD, SNOMED CT and LOINC provide relevant reference points for this process, while the exact terminology set and level

of integration should be selected according to the available clinical systems, licensing conditions and requirements of the Turkish use case.

3.8. Evaluation, safety and governance considerations

Evaluation should be layered across the pipeline. WER remains useful for benchmarking ASR components, but it is insufficient as a clinical safety metric because it gives the same weight to harmless wording differences and errors that change diagnosis, medication, dose, numerical values, temporality or negation. Clinical evaluation should therefore include clinically significant error rates, entity-level precision and recall, semantic accuracy, speaker-attribution errors, correction time, documentation quality and workflow outcomes (Zhou et al., 2018; Afonja et al., 2024; Wang et al., 2025).

Table 7: Evaluation layer

Evaluation layer	Candidate metrics	Purpose
ASR layer	WER, CER, medical term error rate, dose/number/date error rate, latency	Measures transcription quality but does not by itself establish clinical safety.
Diarisation layer	DER, WDER, speaker-role accuracy for clinically salient statements	Checks whether the right content is attributed to the right participant.
Clinical NLP layer	Precision, recall and F1 for entities, negation, temporality, assertion status and coding	Assesses whether structured concepts are complete and clinically faithful.
Documentation layer	Completeness, correctness, concision, organisation, hallucination and omission rate, PDQI-9 or similar review	Assesses whether the generated note is usable and safe for clinician review.
Workflow layer	Editing time, after-hours documentation, user satisfaction, adoption barriers, perceived burden	Assesses whether the system improves actual work rather than only technical scores.
Governance layer	Audit trail, human override, incident reporting, privacy compliance, subgroup analysis	Captures risks not visible in purely technical benchmarks.

For LLM-supported documentation, evaluation also needs to capture risks that are not visible in ASR or NER metrics alone. Regulatory and policy discussions on generative AI in healthcare emphasise output variability, limited transparency of foundation models, the need for clear user information, human oversight, logging, cybersecurity and post-deployment monitoring. Recent work on AI scribes has also identified patient-safety signals in areas such as medication, treatment plans and clinically important omissions. For MediSpeech, this means that evaluation should combine technical metrics with clinician adjudication of clinically significant errors, structured review of generated notes, auditability of edits and explicit mechanisms for incident reporting and model monitoring (U.S. Food and Drug Administration, 2024; Dai et al., 2025; Palm et al., 2025; European Parliament and Council of the European Union, 2024; HL7 International, 2023b).

The safety framework should also include a taxonomy of clinically significant errors. High-priority categories include negation errors, medication and dose errors, numerical value errors, diagnosis or problem omissions, hallucinated findings, wrong speaker attribution and incorrect follow-up timing. Severity thresholds, adjudication rules, inter-rater agreement procedures and acceptable residual-risk levels should be specified later in the clinical validation plan.

The governance context reinforces the same design direction. Health conversations and transcripts are sensitive personal data, and AI-enabled systems used in healthcare must be developed with data minimisation, transparency, access control, logging, security, human oversight and post-

deployment monitoring. WHO guidance on AI for health and European frameworks such as the AI Act and the European Health Data Space support treating these aspects as design requirements rather than administrative add-ons (World Health Organization, 2021; European Parliament and Council of the European Union, 2024; European Parliament and Council of the European Union, 2025).

For the Turkish MediSpeech use case, evaluation should cover the complete processing chain from extracted clinical information to its use in downstream reporting and recommendation components. In addition to entity-level precision, recall and F1, evaluation should consider semantic correctness, negation, temporality, uncertainty and terminology-mapping accuracy. For LLM-supported extraction and documentation, clinically relevant omissions, unsupported information, factual inconsistencies and the amount of clinician correction required should also be monitored. Since structured information may subsequently contribute to recommendation and decision-support functions, errors should be assessed not only at the component level but also according to their potential downstream clinical impact. Human review, traceability to the source conversation and historical records, audit logging and the ability to override generated outputs should therefore remain part of the validation and governance process.

4. CONCLUSIONS

This deliverable reviewed the main technological areas relevant to MediSpeech: automatic speech recognition, paralinguistic event detection, speaker diarisation, clinical information extraction, ambient clinical documentation, clinical decision support, interoperability, evaluation, safety and governance. Overall, the evidence shows rapid progress across AI-driven speech and language technologies, but also confirms that clinical deployment requires more than high-performing speech-to-text models.

For ASR, recent academic and open-source systems show strong progress in multilingual coverage, robustness and processing speed. wav2vec-based approaches such as XLSR, XLS-R and MMS have expanded language coverage, while Whisper-based systems such as WhisperX, CrisperWhisper, Distil-Whisper, Whisper-Streaming and WhisperFlow address long-form transcription, disfluency recognition, timestamping, latency and real-time or near-real-time processing. NVIDIA NeMo models such as Parakeet and Canary also show strong benchmark performance, and multimodal LLMs such as IBM Granite and Google Gemma 3n illustrate the growing convergence between speech recognition and broader language understanding. However, these systems still vary substantially in language support, clinical robustness, transparency, computational cost and suitability for real-world healthcare workflows.

Commercial ASR systems, including Google Cloud Speech-to-Text, Amazon Transcribe, Microsoft Azure AI Speech, Speechmatics, AssemblyAI, Deepgram and Rev AI, provide mature functionalities such as multilingual transcription, punctuation, speaker diarisation, streaming and filler-word handling. Some benchmarks suggest competitive WER values, including results reported for AssemblyAI Universal-1 and other commercial systems. Nevertheless, commercial solutions also raise important issues around transparency, private training data, API-based processing, customisation, privacy, cost and deployment conditions, which are particularly relevant when sensitive clinical conversations are involved.

The Dutch ASR and healthcare-scribe landscape shows that both generic multilingual models and language-specific systems are relevant to MediSpeech. Whisper, XLSR, XLS-R, MMS, Kaldi_NL, Dutch Fast-Conformer models, Soniox and ElevenLabs illustrate the range of academic and commercial options for Dutch speech. In Dutch healthcare, commercial and academic solutions such as HealthTalk, Attendi, SpraakLab/Tell James, Juvoly, Medendo, Autoscriber, Care2Report and HoMed show that medical transcription and draft reporting are already active areas of development. At the same time, performance remains dependent on domain, recording setting, accent, conversational style and clinical vocabulary. For the Portuguese use case, the emerging evidence on European Portuguese ASR resources and benchmarks reinforces the need for local, language-specific and domain-relevant validation rather than assuming transfer from generic multilingual systems.

For the Turkish use case, the SoTa similarly highlights the importance of combining language- and domain-adapted clinical NLP with structured information extraction, terminology normalisation and access to relevant historical patient information. The limited availability of representative Turkish clinical datasets remains an important challenge, particularly for conversational and specialised clinical settings. MediSpeech should therefore validate Turkish clinical NLP components using domain-relevant data and assess not only entity recognition performance, but also semantic correctness, contextual interpretation, terminology mapping and the downstream impact of extracted information on reporting and recommendation workflows.

For paralinguistic event detection, the reviewed literature shows potential value for identifying non-verbal vocalisations, disfluencies, emotion-related cues and other communication signals that may affect clinical interpretation or downstream documentation. End-to-end models based on wav2vec2 or WavLM, as well as systems such as CrisperWhisper, indicate progress in recognising conversational phenomena that are often poorly captured by standard ASR. However, paralinguistic information should be handled cautiously in MediSpeech: it may support communication analysis or flag segments for review, but should not be converted into diagnostic or emotional inferences without specific validation, transparency and governance.

Speaker diarisation is central to the clinical usefulness of automated documentation. Microphone-based approaches can perform well in controlled settings, deep learning approaches such as pyannote, NVIDIA NeMo and EEND represent the current mainstream, and emerging LLM-based approaches can use transcript context to improve speaker-turn attribution. Reported performance varies across settings, with lower DER in controlled conditions and higher error rates in more complex real-world or medical conversations. For MediSpeech, diarisation should therefore be evaluated not only through aggregate DER, but also through word-level diarisation, speaker-role attribution and the clinical significance of assigning statements, symptoms, plans or decisions to the wrong participant.

Clinical information extraction and entity linking remain essential for transforming transcripts or draft notes into clinically useful and structured documentation. BERT-based models, especially domain-pretrained models such as BioBERT and Dutch clinical models such as MedRoBERTa.nl, remain strong baselines, while LLMs such as Gemma show promising but variable performance depending on task complexity and dataset characteristics. The field continues to face challenges related to terminology variation, noisy text, limited labelled datasets, non-English language coverage and the complexity of biomedical knowledge bases. For MediSpeech, NER evaluation should therefore go beyond simple accuracy and include precision, recall, F1, Recall@k, semantic correctness, assertion status, negation, temporality, uncertainty and terminology-mapping accuracy.

The review also confirms that ambient clinical documentation systems should be understood as documentation pipelines rather than isolated ASR tools. These systems combine transcription, diarisation, summarisation, information extraction, terminology mapping and draft note generation. Recent evidence suggests potential reductions in documentation burden, cognitive load and after-hours work, but note quality, omissions, hallucinations, editing burden, workflow integration and clinician trust remain active concerns. Human review remains essential, especially where generated outputs may influence clinical documentation, coding, follow-up or downstream care. For CDSS, current evidence similarly supports an advisory, clinician-mediated role with task-specific validation, explainability and explicit handling of uncertainty.

Interoperability, terminology mapping and EHR integration are therefore central to implementation readiness. Automated reporting should not be treated as simple text export: its value depends on whether outputs can be reviewed, corrected, traced and reused safely inside clinical information systems. Standards and terminologies such as HL7 FHIR, ICPC-2, ICD, SNOMED CT and LOINC provide relevant reference points for structured and interoperable outputs.

Finally, evaluation, safety and governance should be treated as design requirements rather than late-stage compliance tasks. WER remains useful for ASR benchmarking, but it is insufficient as a clinical safety metric. MediSpeech evaluation should include clinically significant errors, medical terminology errors, numbers, doses, dates, negation, speaker-role attribution, entity-level accuracy, note completeness, hallucination and omission rates, correction time, workflow outcomes and user burden. Privacy, auditability, provenance, human oversight, incident reporting and post-deployment monitoring should be embedded into the system design.

Taken together, the SoTa identifies both technical maturity and persistent evidence gaps across the end-to-end clinical documentation pipeline. The relevant comparison is not only between individual ASR or NLP models, but between clinical documentation pipelines with different levels of validation, integration, transparency and deployment readiness. These findings provide the evidence base for D2.2 use-case requirements and for D3.1 clinical requirements and baseline systems.

5. REFERENCES

Ambient clinical documentation, medical scribes and LLM-based reporting

Cho, H. N., Guo, Y., Sutari, S., Chow, E., Tam, S., Perret, D., Pandita, D., & Zheng, K. (2026, March 18). Consumer-to-clinical language shifts in ambient AI draft notes and clinician-finalized documentation: A multi-level analysis. arXiv:2603.18327. https://doi.org/10.48550/arXiv.2603.18327

Morse, J., Gilbert, K., Shin, K., Cooke, R., Rose, P., Sullivan, J., & Sisante, A. (2025). A custom-built ambient scribe reduces cognitive load and documentation burden for telehealth clinicians. arXiv:2507.17754. https://doi.org/10.48550/arXiv.2507.17754

Palm, E., Manikantan, A., Pepin, M. E., Mahal, H., & Belwadi, S. S. (2025, May 15). Assessing the quality of AI-generated clinical notes: A validated evaluation of a large language model scribe. arXiv:2505.17047. https://doi.org/10.48550/arXiv.2505.17047

Vaid, S., Weldon, M., Dunn, J., Davis, S., Lonergan, K., Li, H., Franc, J., Abdalla, M., Baumgart, D. C., Hayward, J., & Mitchell, J. R. (2026, March 5). Berta: An open-source, modular tool for AI-enabled clinical documentation. arXiv:2603.23513. https://doi.org/10.48550/arXiv.2603.23513

Duggan, M. J., Gervase, J., Schoenbaum, A., Hanson, W., Howell, J. T. III, Sheinberg, M., & Johnson, K. B. (2025). Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. JAMA Network Open, 8(2), e2460637. https://doi.org/10.1001/jamanetworkopen.2024.60637

Olson, K. D., Meeker, D., Troup, M., Barker, T. D., Nguyen, T., Manders, J., Stults, C. D., Jones, B. D., Shah, S. D., Shah, T., & Schwamm, L. H. (2025). Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. JAMA Network Open, 8(10), e2534976. https://doi.org/10.1001/jamanetworkopen.2025.34976

Seth, P., Carretas, R., & Rudzicz, F. (2024). The Utility and Implications of Ambient Scribes in Primary Care. JMIR AI, 3, e57673. https://doi.org/10.2196/57673

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine, 7, 258. https://doi.org/10.1038/s41746-024-01258-7

ASR - Automatic Speech Recognition

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., Chen, D., Chen, D., Chen, J., Chen, W., Chen, Y., Chen, Y., Dai, Q., Dai, X., Fan, R., . . . Zhou, X. (2025, March 3). PHI-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LORAS. *arXiv:2503.01743v2*. https://arxiv.org/abs/2503.01743

Amazon transcribe - speech to text – AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/pm/transcribe/

Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., & Wei, F. (2022). SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. In Proceedings of the 60th Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), 5723–5738. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.393

Assemblyai case study page. (n.d.). Google Cloud. https://cloud.google.com/customers/assemblyai?hl=en

Assembly ai, speech-to-text api. (n.d.). AssemblyAI. https://www.assemblyai.com/products/speech-to-text

Automatic Speech Recognition (ASR) — NVIDIA NEMO Framework User Guide. (n.d.). NVIDIA. https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/asr/intro.html

Autoscriber. (n.d.). Autoscriber. https://nl.autoscriber.com/

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Patrick, V. P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2021, November 17). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv: 2111.09296*. https://arxiv.org/abs/2111.09296

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, June 20). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv: 2006.11477*. https://arxiv.org/abs/2006.11477

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023, March 1). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *arXiv: 2303.00747*. https://arxiv.org/abs/2303.00747

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., & Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: A systematic review. Journal of the American Medical Informatics Association, 26(4), 324–338. https://doi.org/10.1093/jamia/ocy179

Carvalho, C., Teixeira, F., Botelho, C., Pompili, A., Solera-Ureña, R., Paulo, S., Julião, M., Rolland, T., Mendonça, J., Pereira, D., Trancoso, I., & Abad, A. (2025, August 27). CAMÕES: A comprehensive automatic speech recognition benchmark for European Portuguese. arXiv:2508.19721. https://doi.org/10.48550/arXiv.2508.19721

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X. & Wei, F. (2023). BEATs: Audio Pre-Training with Acoustic Tokenizers. *Proceedings of the 40th International Conference on Machine Learning*. PMLR 202:5178-5193. https://proceedings.mlr.press/v202/chen23ag.html

Chen, H., Wang, X., Lan, J., Ding, H., Jiang, Y., Yang, M., Xu, D., Luo, J., Ng, N.-C., Cheng, G. W. Y., Mao, Y., & Yoo, J. S. (2025, September 24). MMedFD: A real-world healthcare benchmark for multi-turn full-duplex automatic speech recognition. arXiv:2509.19817. https://doi.org/10.48550/arXiv.2509.19817

Chiang, W., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023, March 30). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., & Zhou, J. (2023, November 14). Qwen-Audio: Advancing universal audio understanding via unified Large-Scale Audio-Language models. *arXiv: 2311.07919v2*. https://arxiv.org/abs/2311.07919

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020, June 24). Unsupervised cross-lingual representation learning for speech recognition. *arXiv:2006.13979v2*. https://arxiv.org/abs/2006.13979

Deepgram, speech to text API: Next-gen AI speech recognition. (n.d.). Deepgram. https://deepgram.com/product/speech-to-text

Dhawan, K., Rekesh, D., Ginsburg, B. (2023). Unified Model for Code-Switching Speech recognition and Language identification based on concatenated tokenizer. *In Proceedings of the 6th Workshop on*

Computational Approaches to Linguistic Code-Switching (pp. 74–82). Association for Computational Linguistics. https://aclanthology.org/2023.calcs-1.7.pdf

Domeinspecifieke taalmodellen - Tell James. (n.d.). Tell James. https://telljames.nl/waarom-tell-james/domeinspecifieke-taalmodellen/

Dutch Open Speech Recognition Benchmark. (n.d.). GitHub - Open Spraaktechnologie. https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/RU/wer.html

FAQ | Juvoly. (n.d.). Juvoly. https://www.juvoly.nl/faq

Free Dutch Speech to Text Transcription. (n.d.). ElevenLabs. https://elevenlabs.io/speech-to-text/dutch

Gandhi, S., Patrick, V. P., & Rush, A. M. (2022, October 24). ESB: a benchmark for Multi-Domain End-to-End speech recognition. *arXiv: 2210.13352*. https://arxiv.org/abs/2210.13352

Gandhi, S., Patrick, V. P., & Rush, A. M. (2023, November 1). Distil-Whisper: Robust knowledge distillation via Large-Scale Pseudo Labelling. *arXiv: 2311.00430*. https://arxiv.org/abs/2311.00430

GitHub - SYSTRAN/faster-whisper: Faster Whisper transcription with CTranslate2. (n.d.). GitHub. https://github.com/SYSTRAN/faster-whisper

Google cloud - speech-to-text transcription. Google Cloud. (2025). https://cloud.google.com/speech-to-text

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020, October). Conformer: Convolution-augmented transformer for speech recognition. *In Proceedings of Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China*. pp. 5036–5040. https://doi.org/10.21437/Interspeech.2020-3015

HealthTalk. (n.d.). HealthTalk. https://healthtalk.ai/

Hsu, W., Bolte, B., Tsai, Y. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021, June 14). HuBERT: Self-Supervised Speech Representation Learning by masked prediction of hidden units. *arXiv: 2106.07447*. https://arxiv.org/abs/2106.07447

Intuitive Reporting - Attendi. (n.d.). Attendi. https://attendi.nl/en/attendi-english/

Juvoly V2 WebSocket API. (n.d.). Juvoly *API Documentation*. https://documentation.juvoly.nl/v2/websocket.html

Kasdorp, S. A., Zhang, Y., Scharenborg, O. E. (2024). *State-of-the-art automatic speech recognition systems on Dutch regional dialects*. https://repository.tudelft.nl/record/uuid:1b5b5a3c-288f-4301-935c-96f7163512f9

GitHub - opensource-spraakherkenning-nl/Kaldi_NL: Code related to the Dutch instance and user groups of the KALDI speech recognition toolkit. (n.d.) GitHub. https://github.com/opensource-spraakherkenning-nl/Kaldi_NL

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2023). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, *16*(4), 1–23. https://doi.org/10.1145/3636513

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, January 30). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv: 2301.12597*. https://arxiv.org/abs/2301.12597

Lima, R., Leal, S. E., Candido Junior, A., & Aluísio, S. M. (2024, September 10). A large dataset of spontaneous speech with the accent spoken in São Paulo for automatic speech recognition evaluation. arXiv:2409.15350. https://doi.org/10.48550/arXiv.2409.15350

Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., van de Water, R., Dulmen, A.M., Dalpiaz, F., van Deemter, K. & Brinkkemper, S. (2020). The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare. http://hdl.handle.net/10125/64184

Macháček, D., Dabre, R., & Ondřej Bojar. (2023). Turning Whisper into Real-Time Transcription System. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 17–24). Association for Computational Linguistics. https://aclanthology.org/2023.ijcnlp-demo.3.pdf

Macháček, D., Žilinec, M., Bojar, O. (2021) Lost in Interpreting: Speech Translation from Source or Interpreter? *Proc. Interspeech 2021, 2376-2380*. doi: 10.21437/Interspeech.2021-2232. https://www.isca-archive.org/interspeech_2021/machacek21_interspeech.pdf

Majumdar, S., & Koluguri, N. R. (2024, April 18). Pushing the Boundaries of Speech Recognition with NVIDIA NeMo Parakeet ASR Models. *NVIDIA Technical Blog*. https://developer.nvidia.com/blog/pushing-the-boundaries-of-speech-recognition-with-nemo-parakeet-asr-models/

Medendo - Digital Assistant for Healthcare professionals. (n.d.). Medendo. https://www.medendo.com/

Microsoft Azure AI speech. (n.d.). Microsoft. https://azure.microsoft.com/en-us/products/ai-services/ai-speech

Molenaar, S., Maas, L., Burriel, V., Dalpiaz, F. & Brinkkemper, S. (2020). Medical Dialogue Summarization for Automated Reporting in Healthcare. In: Dupuy-Chessa, S., Proper, H. (eds) Advanced Information Systems Engineering Workshops. CAiSE 2020. Lecture Notes in Business Information Processing, vol 382. Springer, Cham. https://doi.org/10.1007/978-3-030-49165-9_7

Ng, J. J. W., Wang, E., Zhou, X., Zhou, K. X., Goh, C. X. L., Sim, G. Z. N., Tan, H. K., Goh, S. S. N., & Ng, Q. X. (2025). Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: A systematic review. BMC Medical Informatics and Decision Making, 25, Article 236. https://doi.org/10.1186/s12911-025-03061-0

Nix, A., James, R., Borgholt, L., Ekner, A. B., Krumm, L., Severin, J., Engel, D., Maaløe, L., & Havtorn, J. (2026, May 15). Symphony for speech-to-text: Supporting real-time medical voice interfaces. arXiv:2605.16545. https://doi.org/10.48550/arXiv.2605.16545

Phipps, B. (2025, May 20). The return of on-premise: Why enterprise AI's head is no longer in the cloud. *Speechmatics*. https://www.speechmatics.com/company/articles-and-news/the-return-of-on-prem-why-enterprise-is-no-longer-in-the-cloud

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: a Large-Scale Multilingual Dataset for Speech research. *Interspeech 2022*. https://doi.org/10.21437/interspeech.2020-2826

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W., Conneau, A., & Auli, M. (2023, May 22). Scaling speech technology to 1,000+ languages. *arXiv: 2305.13516*. https://arxiv.org/abs/2305.13516

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December 6). Robust speech recognition via Large-Scale Weak Supervision. *arXiv: 2212.04356*. https://arxiv.org/abs/2212.04356

Ramirez, F. M., Chkhetiani, L., Ehrenberg, A., McHardy, R., Botros, R., Khare, Y., Vanzo, A., Peyash, T., Oexle, G., Liang, M., Sklyar, I., Fakhan, E., Etefy, A., McCrystal, D., Flamini, S., Donato, D., &

Yoshioka, T. (2024, April 15). Anatomy of Industrial Scale Multilingual ASR. *arXiv: 2404.09841*. https://arxiv.org/abs/2404.09841

Rastorgueva, E., & Koluguri, N. R. (2024, April 18). New Standard for Speech Recognition and Translation from the NVIDIA NeMo Canary Model. *NVIDIA Technical Blog*. https://developer.nvidia.com/blog/new-standard-for-speech-recognition-and-translation-from-the-nvidia-nemo-canary-model/

Rekesh, D., Koluguri, N. R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., & Ginsburg, B. (2023, May 8). Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. *arXiv: 2305.05084*. https://arxiv.org/abs/2305.05084

Rev AI: Speech to Text API. (n.d.). Rev AI. https://www.rev.ai/

Sanni, M., Abdullahi, T., Kayande, D. D., Ayodele, E., Etori, N. A., Mollel, M. S., Yekini, M., Okocha, C., Ismaila, L. E., Omofoye, F., Adewale, B. A., & Olatunji, T. (2025, February 6). Afrispeech-Dialog: A benchmark dataset for spontaneous English conversations in healthcare and beyond. arXiv:2502.03945. https://doi.org/10.48550/arXiv.2502.03945

Sanseviero, O., & Ballantyne, I. (2025, June 26). Introducing Gemma 3n: The developer guide. https://developers.googleblog.com/en/introducing-gemma-3n-developer-guide/

Saon, G., Dekel, A., Brooks, A., Nagano, T., Daniels, A., Satt, A., Mittal, A., Kingsbury, B., Haws, D., Morais, E., Kurata, G., Aronowitz, H., Ibrahim, I., Kuo, J., Soule, K., Lastras, L., Suzuki, M., Hoory, R., Thomas, S., . . . Kons, Z. (2025, May 13). Granite-speech: open-source speech-aware LLMs with strong English ASR capabilities. *arXiv: 2505.08699v2*. https://arxiv.org/abs/2505.08699v2

Self-Hosted Voice AI. (n.d.). DeepGram. https://deepgram.com/self-hosted

Speechmatics - speech-to-text API. (n.d.). Speechmatics. https://www.speechmatics.com

Speech-to-text benchmarks 2025. (n.d.). Soniox. https://soniox.com/benchmarks

Spraakherkenning van SpraakLab. (n.d.). SpraakLab. https://www.spraaklab.nl/technologie/spraakherkenning

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., & Zhang, C. (2023, October 20). SALMONN: towards generic hearing abilities for large language models. *arXiv: 2310.13289v2*. https://arxiv.org/abs/2310.13289

Teixeira, F., Carvalho, C., Julião, M., Botelho, C., Solera-Ureña, R., Paulo, S., Rolland, T., Peters, B., Trancoso, I., & Abad, A. (2026, May 26). FalAR: A large-scale speaker-annotated European Portuguese speech corpus of parliamentary sessions. arXiv:2605.27062. https://doi.org/10.48550/arXiv.2605.27062

Tejedor-García, C., van den Heuvel, H., van Hessen, A., van Dulmen, S. & Pieters, T. (2024, April 26). An Innovative Methodology Utilizing AI-based Automatic Speech Recognition for Transcribing Dutch Patient-Provider Consultation Recordings. Zenodo. https://doi.org/10.5281/zenodo.11071085

Wagner, L., Thallinger, B., & Zusag, M. (2024, August 29). CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions. *arXiv: 2408.16589*. https://arxiv.org/abs/2408.16589

Wang, R., Xu, Z., & Lin, F. X. (2024, December 15). WhisperFlow: speech foundation models in real time. *arXiv: 2412.11272*. https://arxiv.org/abs/2412.11272

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., & Auli, M. (2020, October 22). Self-training and Pre-training are Complementary for Speech Recognition. *arXiv: 2010.11430*. https://arxiv.org/abs/2010.11430

Zeeshan, R., Bogue, J., and Naveed Asghar, M. (2025, July 2). Relative Applicability of Diverse Automatic Speech Recognition Platforms for Transcription of Psychiatric Treatment Sessions. *IEEE Access*, vol. 13, pp. 117343-117354, 2025. https://ieeexplore.ieee.org/document/11063333

Żelasko, P., & Chen, Z. (2024, January 16). New support for Dutch and Persian released by NVIDIA NEMO ASR. *NVIDIA Technical Blog*. https://developer.nvidia.com/blog/new-support-for-dutch-and-persian-released-by-nemo-asr/

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., . . . Wu, Y. (2023, March 2). Google USM: Scaling Automatic Speech Recognition beyond 100 Languages. *arXiv: 2303.01037*. https://arxiv.org/abs/2303.01037

Clinical documentation burden and EHR workload

Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., & Gilchrist, V. J. (2017). Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. Annals of Family Medicine, 15(5), 419–426. https://doi.org/10.1370/afm.2121

Speaker diarisation

Adedeji, A., Joshi, S., & Doohan, B. (2024). The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*. https://arxiv.org/abs/2402.07658

Al-Hadithy, T. M., Frikha, M., & Maseer, Z. K. (2022, October). Speaker Diarization based on Deep Learning Techniques: A Review. In *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 856-871). https://ieeexplore.ieee.org/abstract/document/9932710

Ghosh, A., Fuhs, M., Kim, B., Chowdhury, A., & Woszczyna, M. (2025). ASR-synchronized speaker-role diarization. arXiv.

Khoma, Volodymyr, Yuriy Khoma, Vitalii Brydinskyi, and Alexander Konovalov. "Development of supervised speaker diarization system based on the pyannote audio processing library." *Sensors* 23, no. 4 (2023): 2082. https://www.mdpi.com/1424-8220/23/4/2082

O'Shaughnessy, D. (2025). Speaker Diarization: A Review of Objectives and Methods. *Applied Sciences*, *15*(4), 2002. https://www.mdpi.com/2076-3417/15/4/2002

Medaramitta, R., 2021. Evaluating the Performance of Using Speaker Diarization for Speech Separation of In-Person Role-Play Dialogues. https://corescholar.libraries.wright.edu/etd_all/2517/

Mingote, Victoria, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. "Audio-Visual Speaker Diarization: Current Databases, Approaches and Challenges." *arXiv preprint arXiv:2409.05659* (2024). https://arxiv.org/abs/2409.05659

Nagavi, T. C., Samanvitha, S., Sudhanva, S., Shivakumar, S., & Hullur, V. (2024). Comprehensive Analysis of State-of-the-Art Approaches for Speaker Diarization. *Automatic Speech Recognition and Translation for Low Resource Languages*, 427-444. https://doi.org/10.1002/9781394214624.ch19

Nehrboss, W. K. (2024). CASCA: Leveraging Role-Based Lexical Cues for Robust Multimodal Speaker Diarization via Large Language Models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)* (pp. 214-224). https://aclanthology.org/2024.icnlsp-1.24.pdf

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317. https://www.sciencedirect.com/science/article/pii/S0885230821001121

Rahim, M. Z., Juan, S. S., & Junaini, S. N. (2025). Systematic Literature Review of Speaker Diarization Techniques: Toward Bridging Gaps in Low-resourced Languages Using Machine Learning. *Applications of Modelling and Simulation*, *9*, 22-36. https://arqiipubl.com/ojs/index.php/AMS_Journal/article/view/801

Sanni, M., Abdullahi, T., Kayande, D. D., Ayodele, E., Etori, N. A., Mollel, M. S., ... & Olatunji, T. (2025). Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. *arXiv preprint arXiv:2502.03945*. https://arxiv.org/abs/2502.03945

Tran, B. D., Mangu, R., Tai-Seale, M., Lafata, J. E., & Zheng, K. (2023, April). Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. In *AMIA Annual Symposium Proceedings* (Vol. 2022, p. 1072). https://pmc.ncbi.nlm.nih.gov/articles/PMC10148344/

Paralinguistic event detection

Alm, C. O. (2016). Language as Sensor in Human-Centered Computing: Clinical Contexts as Use Cases. Language and Linguistics Compass, 10(3), 105–119. https://doi.org/10.1111/lnc3.12171

Back, A. L., Bauer-Wu, S. M., Rushton, C. H., & Halifax, J. (2009). Compassionate Silence in the Patient–Clinician Encounter: A Contemplative Approach. Journal of Palliative Medicine, 12(12), 1113–1117. https://doi.org/10.1089/jpm.2009.0175

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Beattie, G. W. (1981). Interruption in conversational interaction, and its relation to the sex and status of the interactants*. Linguistics, 19(1–2). https://doi.org/10.1515/ling.1981.19.1-2.15

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. Interspeech 2005, 1517–1520. https://doi.org/10.21437/Interspeech.2005-446

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., & Provost, E. M. (2017). MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. IEEE Transactions on Affective Computing, 8(1), 67–80. https://doi.org/10.1109/TAFFC.2016.2515617

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

Chen, W., Xing, X., Chen, P., & Xu, X. (2024). Vesper: A Compact and Effective Pretrained Model for Speech Emotion Recognition. IEEE Transactions on Affective Computing, 15(3), 1711–1724. https://doi.org/10.1109/TAFFC.2024.3369726

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, 7(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Fagherazzi, G., Fischer, A., Ismael, M., & Despotovic, V. (2021). Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. Digital Biomarkers, 5(1), 78–88. https://doi.org/10.1159/000515346

Fan, W., Xu, X., Xing, X., Chen, W., & Huang, D. (2021). LSSED: A Large-Scale Dataset and Benchmark for Speech Emotion Recognition. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 641–645. https://doi.org/10.1109/ICASSP39728.2021.9414542

Gillick, J., Deng, W., Ryokai, K., & Bamman, D. (2021). Robust Laughter Detection in Noisy Environments. Interspeech 2021, 2481–2485. https://doi.org/10.21437/Interspeech.2021-353

Hanson, J. L., Stephens, M. B., Pangaro, L. N., & Gimbel, R. W. (2012). Quality of outpatient clinical notes: A stakeholder definition derived through qualitative research. BMC Health Services Research, 12(1), 407. https://doi.org/10.1186/1472-6963-12-407

Haq, S., Jackson, P. J. B., & Edge, J. (2008). Audio-visual feature selection and reduction for emotion classification. Proceedings of the International Conference on Auditory-Visual Speech Processing.

Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. Speech Communication, 154, 102974. https://doi.org/10.1016/j.specom.2023.102974

Ho, P. H., Bălan, D. A., Heylen, D. K. J., & Truong, K. P. (2025). Enhancing Transcripts of Open-Source Automatic Speech Recognition Models Through Fine-Tuning with Laughter and Speech-Laugh. Interspeech 2025, 4513–4517. https://doi.org/10.21437/Interspeech.2025-2193

Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech Emotion Recognition Using CNN. Proceedings of the 22nd ACM International Conference on Multimedia, 801–804. https://doi.org/10.1145/2647868.2654984

Idrisoglu, A., Dallora, A. L., Anderberg, P., & Berglund, J. S. (2023). Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review. Journal of Medical Internet Research, 25, e46105. https://doi.org/10.2196/46105

Jokinen, K., Furukawa, H., Nishida, M., & Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. ACM Transactions on Interactive Intelligent Systems, 3(2), 1–30. https://doi.org/10.1145/2499474.2499481

Kamiloğlu, R. G., & Sauter, D. A. (2024). Voices without words: The spectrum of nonverbal vocalisations. European Review of Social Psychology, 1–36. https://doi.org/10.1080/10463283.2024.2418714

Kraus, V. B. (2018). Biomarkers as drug development tools: Discovery, validation, qualification and use. Nature Reviews Rheumatology, 14(6), 354–362. https://doi.org/10.1038/s41584-018-0005-9

Kurtić, E., Brown, G. J., & Wells, B. (2013). Resources for turn competition in overlapping talk. Speech Communication, 55(5), 721–743. https://doi.org/10.1016/j.specom.2012.10.002

Larson, E. B., & Yao, X. (2005). Clinical Empathy as Emotional Labor in the Patient-Physician Relationship. JAMA, 293(9), 1100. https://doi.org/10.1001/jama.293.9.1100

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2021). Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. IEEE Reviews in Biomedical Engineering, 14, 342–356. https://doi.org/10.1109/RBME.2020.3006860

Layer, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer & H. Giles (Eds.), Social Markers in Speech. Cambridge University Press.

Lee, C.-C., Kim, J., Metallinou, A., Busso, C., Lee, S., & Narayanan, S. S. (2015). Speech in Affective Computing. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), The Oxford Handbook of Affective Computing. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199942237.013.021

Lelorain, S., Brédart, A., Dolbeault, S., & Sultan, S. (2012). A systematic review of the associations between empathy measures and patient outcomes in cancer care. Psycho-Oncology, 21(12), 1255–1264. https://doi.org/10.1002/pon.2115

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391

Lotfian, R., & Busso, C. (2019). Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. IEEE Transactions on Affective Computing, 10(4), 471–483. https://doi.org/10.1109/TAFFC.2017.2736999

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. https://doi.org/10.1002/lio2.354

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent Systems with Applications, 20, 200266. https://doi.org/10.1016/j.iswa.2023.200266

Pearce, P. F., Ferguson, L. A., George, G. S., & Langford, C. A. (2016). The essential SOAP note in an EHR age. The Nurse Practitioner, 41(2), 29–36.

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings (arXiv:2104.03502). arXiv. https://doi.org/10.48550/arXiv.2104.03502

Podder, V., Lew, V., & Ghassemzadeh, S. (2023). SOAP Notes. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK482263/

Quiroz, J. C., Laranjo, L., Kocaballi, A. B., Berkovsky, S., Rezazadegan, D., & Coiera, E. (2019). Challenges of developing a digital scribe to reduce clinical documentation burden. Npj Digital Medicine, 2(1), 114. https://doi.org/10.1038/s41746-019-0190-1

Schuller, B. W., & Batliner, A. M. (2013). Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing (1st ed.). Wiley. https://doi.org/10.1002/9781118706664

Singh, Y. B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, 245–263. https://doi.org/10.1016/j.neucom.2022.04.028

Trager, G. L. (1958). Paralanguage: A First Approximation. Stud. Linguist., 13, 1–12.

Tran, B. D., Latif, K., Reynolds, T. L., Park, J., Elston Lafata, J., Tai-Seale, M., & Zheng, K. (2023). "Mm-hm," "Uh-uh": Are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? Journal of the American Medical Informatics Association, 30(4), 703–711. https://doi.org/10.1093/jamia/ocad001

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669

Trouvain, J., & Truong, K. P. (2012). Comparing Non-Verbal Vocalisations in Conversational Speech Corpora. 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals.

Truong, K. P. (2013). Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. Interspeech 2013, 1404–1408. https://doi.org/10.21437/Interspeech.2013-368

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10745–10759. https://doi.org/10.1109/TPAMI.2023.3263585

Ward, N. (2006). Non-lexical conversational sounds in American English. Pragmatics & Cognition, 14(1), 129–182. https://doi.org/10.1075/pc.14.1.08war

Zachariae, R., Pedersen, C. G., Jensen, A. B., Ehrnrooth, E., Rossen, P. B., & Von Der Maase, H. (2003). Association of perceived physician communication style with patient satisfaction, distress, cancer-related self-efficacy, and perceived control over the disease. British Journal of Cancer, 88(5), 658–665. https://doi.org/10.1038/sj.bjc.6600798

Zusag, M., Wagner, L., & Thallinger, B. (2024). Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. Interspeech 2024, 1265-1269, https://doi.org/10.21437/Interspeech.2024-731.

Named Entity Recognition of Medical Terms

Gan, Y., Nguyen, D. D., Lin, Y., Zhong, P., Vu, T., Duong, L., & Li, Y.-F. (2026, April 9). RAG-Coding: Enhancing LLM medical coding with structured external knowledge. arXiv:2605.27377. https://doi.org/10.48550/arXiv.2605.27377

Nunes, M., Boné, J., Ferreira, J. C., Chaves, P., & Elvas, L. B. (2024). MediAlbertina: An European Portuguese medical language model. Computers in Biology and Medicine, 182, Article 109233. https://doi.org/10.1016/j.compbiomed.2024.109233

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. npj Digital Medicine, 5, Article 194. https://doi.org/10.1038/s41746-022-00742-2

Kartchner, D., Deng, J., Lohiya, S., Kopparthi, T., Bathala, P., Domingo-Fernández, D., & Mitchell, C. S. (2023). A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14720–14736). Association for Computational Linguistics.

Hartendorp, F., Seinen, T., van Mulligen, E., & Verberne, S. (2024). Biomedical entity linking for Dutch: Fine-tuning a self-alignment BERT model on an automatically generated Wikipedia corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1656–1665). ELRA and ICCL.

Liu, S., Wang, A., Xiu, X., Zhong, M., & Wu, S. (2024). Evaluating Medical Entity Recognition in Health Care: Entity Model Quantitative Study. JMIR Medical Informatics, 12, e59782. https://doi.org/10.2196/59782

Shaitarova, A., Zaghir, J., Lavelli, A., Krauthammer, M., & Rinaldi, F. (2023). Exploring the latest highlights in medical natural language processing across multiple languages: A survey. *Yearbook of Medical Informatics*, *32*(01), 203–213.

Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A. A., Roberts, A., Bendayan, R., Richardson, M. P., Stewart, R., Shah, A. D., Wong, W. K., Ibrahim, Z., Teo, J. T., & Dobson, R. J. B. (2021). Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artificial Intelligence in Medicine, 117, 102083. https://doi.org/10.1016/j.artmed.2021.102083

University Medical Center Utrecht. (2022). *Dutch-medical-concepts* (Version 1.11.0) [Computer software]. GitHub. https://github.com/umcu/dutch-medical-concepts

University Medical Center Utrecht. (n.d.). *Dutch-medical-wikipedia* [Computer software]. GitHub. Retrieved September 8, 2025, from https://github.com/umcu/dutch-medical-wikipedia

Mosteiro, P., Wang, R., Scheepers, F., & Spruit, M. (2025). Investigating De-Identification Methodologies in Dutch Medical Texts: A Replication Study of Deduce and Deidentify. Electronics, 14(8), 1636. https://doi.org/10.3390/electronics14081636

National Library of Medicine. (2025). *UMLS Metathesaurus Precomputed Subsets (2025AA)* [Data set]. U.S. Department of Health and Human Services, National Institutes of Health. https://www.nlm.nih.gov/research/umls/licensedcontent/umlske_download.html

Wever, L., Dalpiaz, F., Brinkkemper, S., Scheepers, F., & Vermond, R. (2023). *Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting* [Preprint]. arXiv. https://arxiv.org/abs/2311.13274

Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., & Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, *22*(5), 948–953.

University Medical Center Utrecht. (2024). *ClinIp: A Python library for performing NLP on clinical text written in Dutch* (Version 0.9.4) [Computer software]. GitHub. https://github.com/umcu/clinlp

Park, Y., Son, G., & Rho, M. (2024). Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances. *Applied Sciences*, *14*(20), 9302. https://doi.org/10.3390/app14209302

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, *19*, 59-99.

Landolsi, M. Y., Hlaoua, L., & Ben Romdhane, L. (2023). Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, *65*(2), 463-516.

Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, *177*, 105122.

Sandoval, A. M., Díaz, J., Llanos, L. C., & Redondo, T. (2019). Biomedical term extraction: NLP techniques in computational medicine. *IJIMAI*, *5*(4), 51-59.

Hahn, U., & Oleynik, M. (2020). Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, *29*(01), 208-220.

Gumiel, Y. B., Silva e Oliveira, L. E., Claveau, V., Grabar, N., Paraiso, E. C., Moro, C., & Carvalho, D. R. (2021). Temporal relation extraction in clinical texts: a systematic review. *ACM Computing Surveys (CSUR)*, *54*(7), 1-36.

Türkmen, H., Dikenelli, O., Eraslan, C., Çallı, M. C., & Özbek, S. S. (2023). BioBERTurk: Exploring Turkish Biomedical Language Model Development Strategies in Low-Resource Setting. Journal of Healthcare Informatics Research, 7(4), 433–446. https://doi.org/10.1007/s41666-023-00140-7.

Dilmaç, F., & Alpkocak, A. (2026). De-Identification of Electronic Health Records Using Deep Learning and Transformers. Applied Sciences, 16(4), 1692. https://doi.org/10.3390/app16041692.

Oğul, İ. Ü., Soygazi, F., & Bostanoğlu, B. E. (2025). TurkMedNLI: A Turkish medical natural language inference dataset through large language model based translation. PeerJ Computer Science, 11, e2662. https://doi.org/10.7717/peerj-cs.2662.

Akbasli, I. T., Birbilen, A. Z., & Teksam, O. (2025). Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-English languages. BMC Medical Informatics and Decision Making, 25, 154. https://doi.org/10.1186/s12911-025-02871-6.

Clinical Information Extraction without LLMs

[Wang & Akella 2015] Wang C, Akella R. A Hybrid Approach to Extracting Disorder Mentions from Clinical Notes. AMIA Jt Summits Transl Sci Proc. 2015 Mar 25;2015:183-7. PMID: 26306265; PMCID: PMC4525272.

[Fu et. al 2020] Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, Hongfang Liu. Clinical concept extraction: A methodology review, Journal of Biomedical Informatics, Volume 109, 2020, 103526, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103526.

Clinical Information Extraction and Reporting with LLMs and Knowledge Graphs

[Hu et. al. 2026] Yan Hu, et. al, "Information extraction from clinical notes: are we ready to switch to large language models?", Journal of the American Medical Informatics Association, Volume 33, Issue 3, March 2026, Pages 553–562, https://doi.org/10.1093/jamia/ocaf213.

[Zhang et al. 2023] Y. Zhang, et. al, "Siren's song in the ai ocean: a survey on hallucination in large language models", v1 arXiv preprint arXiv:2309.01219 (2023) and v3 https://arxiv.org/abs/2309.01219v3 (v3, updated in 2025).

[AB Das, et. al 2025] AB Das, et. al, "Hallucinations and Key Information Extraction in Medical Texts: A Comprehensive Assessment of Open-Source Large Language Models", arXiv:2504.19061v3, 2025.

[Agrawal et. al. 2024] Agrawal et. al., "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey", https://arxiv.org/html/2311.07914v2, 2024.

[Lavrinovics et. al. 2024] Lavrinovics et. al., "Knowledge Graphs, Large Language Models, and Hallucinations: An NLP Perspective", https://arxiv.org/html/2411.14258v1, 2024.

[Škrlj et. al. 2025] Škrlj et. al, "From Symbolic to Neural and Back: Exploring Knowledge Graph–Large Language Model Synergies", https://arxiv.org/html/2506.09566v1, 2025.

[Chang et. al. 2024] Chang et. al, "Use of SNOMED CT in Large Language Models: Scoping Review", https://pubmed.ncbi.nlm.nih.gov/39374057/, 2024.

[Alarcon-Serrano et. al. 2026] Alarcon-Serrano et. al, "Assessing open LLMs' ability to identify biomedical taxonomic relationships: A SNOMED CT-based experimental evaluation", https://www.sciencedirect.com/science/article/pii/S0950705126006088, 2026.

Clinical Decision Support Systems

Welch, B.M.; Kawamoto, K. Clinical Decision Support for Genetically Guided Personalized Medicine: A Systematic Review. *J Am Med Inform Assoc* **2013**, *20*, 388–400, doi:10.1136/amiajnl-2012-000892.

Zastrozhin, M.S.; Sorokin, A.S.; Agibalova, T.V.; Grishina, E.A.; Antonenko, A.P.; Rozochkin, I.N.; Duzhev, D.V.; Skryabin, V.Y.; Galaktionova, T.E.; Barna, I.V.; et al. Using a Personalized Clinical Decision Support System for Bromdihydrochlorphenylbenzodiazepine Dosing in Patients with Anxiety Disorders Based on the Pharmacogenomic Markers. *Human Psychopharmacology: Clinical and Experimental* **2018**, *33*, e2677, doi:10.1002/hup.2677.

Beyan, T.; Son, Y.A. Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 2: The Incorporation of SNP into the National Health Information System of Turkey. *JMIR Medical Informatics* **2014**, *2*, e3555, doi:10.2196/medinform.3555.

Beyan, T.; Son, Y.A. Incorporation of Personal Single Nucleotide Polymorphism (SNP) Data into a National Level Electronic Health Record for Disease Risk Assessment, Part 3: An Evaluation of SNP

Incorporated National Health Information System of Turkey for Prostate Cancer. *JMIR Medical Informatics* **2014**, *2*, e3560, doi:10.2196/medinform.3560.

Faraone, S. V., et al. (2021). The World Federation of ADHD International Consensus Statement. Neuroscience & Biobehavioral Reviews, 128, 789–818.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30, 4765–4774.

Megerian, J. T., et al. (2022). Evaluation of an AI-based medical device for diagnosis of ASD. npj Digital Medicine, 5(1), 57.

Mikolas, P., Vahid, A., Bernardoni, F., Süß, M., Martini, J., Beste, C., & Bluschke, A. (2022). Training a machine learning classifier to identify ADHD based on real-world clinical data from medical records. Scientific Reports, 12, 12934. https://doi.org/10.1038/s41598-022-17126-x

Navarro-Soria, I., Rico-Juan, J. R., Juárez-Ruiz de Mier, R., & Lavigne-Cervan, R. (2025). Prediction of attention deficit hyperactivity disorder based on explainable artificial intelligence. Applied Neuropsychology: Child, 14(4), 474–487. https://doi.org/10.1080/21622965.2024.2336019

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28, 31–38. https://doi.org/10.1038/s41591-021-01614-0

Wang, Y.-C., Cheng, C.-Y., Wu, C.-S., Lee, C.-C., & Gau, S. S.-F. (2025). Clinical correlates of errors in machine-learning diagnostic model of autism spectrum disorder: Impact of sample cohorts. Autism, 29(12), 3083–3099. https://doi.org/10.1177/13623613251360271

Interoperability, clinical terminologies and EHR integration

Bodenreider, O., Cornet, R., & Vreeman, D. J. (2018). Recent Developments in Clinical Terminologies—SNOMED CT, LOINC, and RxNorm. Yearbook of Medical Informatics, 27(1), 129–139. https://doi.org/10.1055/s-0038-1667077

European Parliament and Council of the European Union. (2025). Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union, L 2025/327. https://eur-lex.europa.eu/eli/reg/2025/327/oj

HL7 International. (2023a). FHIR v5.0.0: Overview. https://hl7.org/fhir/R5/

HL7 International. (2023b). FHIR v5.0.0: Provenance. https://hl7.org/fhir/R5/provenance.html

LOINC. (2026). The international standard for identifying health measurements, observations, and documents. Regenstrief Institute. https://loinc.org/

SNOMED International. (2026). The value of SNOMED CT. https://www.snomed.org/value-proposition

World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians. (1998). Classification of Primary Care: ICPC-2 (2nd ed.). Oxford University Press.

Evaluation, safety and governance considerations

Dai, J., Huang, A., Nasrallah, C., Croci, R., Soleimani, H., Pollet, S. J., Adler-Milstein, J., Murray, S. G., Yazdany, J., & Chen, I. Y. (2025, December 1). Patient safety risks from AI scribes: Signals from end-user feedback. arXiv:2512.04118. https://doi.org/10.48550/arXiv.2512.04118

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144

and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Simplício, A., Vinagre, G., Ramos, M. M., Tavares, D., Ferreira, R., Attanasio, G., Alves, D. M., Calvo, I., Vieira, I., Guerra, R., Furtado, J., Canaverde, B., Paulo, I., Ramos, V., Glória-Silva, D., Faria, M., Treviso, M., Gomes, D., Gomes, P., Semedo, D., Martins, A., & Magalhães, J. (2026, March 27). AMALIA technical report: A fully open source large language model for European Portuguese. arXiv:2603.26511. https://doi.org/10.48550/arXiv.2603.26511

U.S. Food and Drug Administration. (2024). Executive summary for the Digital Health Advisory Committee meeting: Total product lifecycle considerations for generative AI-enabled devices. https://www.fda.gov/media/182871/download

Vieira, I., Calvo, I., Paulo, I., Furtado, J., Ferreira, R., Tavares, D., Glória-Silva, D., Semedo, D., & Magalhães, J. (2026, March 27). ALBA: A European Portuguese benchmark for evaluating language and linguistic dimensions in generative LLMs. arXiv:2603.26516. https://doi.org/10.48550/arXiv.2603.26516

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. ISBN 9789240029200. https://www.who.int/publications/i/item/9789240029200

World Health Organization. (2025). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. World Health Organization. ISBN 978-92-4-008475-9. https://www.who.int/publications/i/item/9789240084759

Afonja, T., Olatunji, T., Ogun, S., Etori, N. A., Owodunni, A., & Yekini, M. (2024). Performant ASR Models for Medical Entities in Accented Speech. Interspeech 2024, 2315–2319. https://doi.org/10.21437/Interspeech.2024-2261

Zhou, L., Blackley, S. V., Kowalski, L., Doan, R., Acker, W. W., Landman, A. B., Kontrient, E., Mack, D., Meteer, M., Bates, D. W., & Goss, F. R. (2018). Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professioal Transcriptionists. JAMA Network Open, 1(3), e180530. https://doi.org/10.1001/jamanetworkopen.2018.0530

Commercial solutions and implementation readiness

Accurx. (2026). Accurx Scribe: An overview. https://support.accurx.com/

Microsoft. (2025). Microsoft Dragon Copilot provides the healthcare industry's first unified voice AI assistant that enables clinicians to streamline clinical documentation, surface information and automate tasks. https://news.microsoft.com/

Tandem Health. (2026). Security built for healthcare. https://www.tandemhealth.ai/data-security