Evaluation and Optimization of Emotion Detection Models for UAE Deployment

Introduction

Voice-based emotion recognition is being leveraged to analyze sales calls for live coaching and customer insight. In the United Arab Emirates (UAE), strict data privacy regulations require that sensitive audio data remain on local infrastructure. This report evaluates three emotion detection solutions for voice-only analysis, focusing on real-time performance, deployment feasibility in the UAE (local hosting and compliance), possible optimizations, and overall "UAE-readiness."

The models compared are:

- **MixedEmotions Toolbox** (EU project for multimodal emotion analysis)
- **Emolysis Toolkit** (open-source group emotion analysis, preferred for deployment)
- **Hume AI** (cloud API as a benchmark)

Key evaluation criteria include: real-time processing capability, deployment in the UAE (on-premises hosting vs. cloud), data privacy compliance, performance optimization options (model size reduction, latency tuning, hardware acceleration), and emotion granularity & accuracy. We also propose a UAE-compliant architecture and optimizations for Emolysis, as it is the preferred model for on-premise deployment. A comparison table is provided to summarize UAE-readiness across metrics.

MixedEmotions Toolbox

MixedEmotions is a comprehensive big-data toolbox for multilingual, multimodal emotion analysis developed under an EU H2020 project. It consists of modular Docker-based components orchestrated for scalable deployment (originally using Mesos). Relevant aspects for this voice-only use case include:

Real-Time Processing
MixedEmotions can analyze text, audio, and video inputs for emotions, but it was primarily designed for batch or scalable analysis pipelines rather than instantaneous streaming. The audio emotion module focuses on estimating emotional valence and arousal from audio. It uses signal processing (e.g. OpenSMILE features) and machine learning classifiers, outputting a continuous valence (positive/negative) and arousal (calm/excited) score for a given audio sample. This approach is computationally efficient – OpenSMILE feature extraction runs quickly on CPU – so short segments of calls can be processed fast. The latency per segment (e.g. a few seconds of audio) is low, enabling near-real-time analysis if audio is chunked appropriately. However, unlike a streaming API, one would need to repeatedly send audio frames to the service; some buffering might be needed, adding slight delay.

Deployment Feasibility (UAE)

The MixedEmotions platform was built with on-premises deployment in mind. All modules can be run as stand-alone Docker containers, meaning an organization can deploy only the needed components (e.g. the audio emotion recognizer) on local servers. There are open-source modules for emotion extraction, while some advanced features (like certain speech-to-text or emotion modules) were proprietary – for UAE deployment, one would use the open modules to avoid external dependencies. The audio emotion module (up_emotions_audio) is GPL-licensed open-source and does not require sending data to any cloud service. Thus, data stays local, satisfying compliance. Setting up the MixedEmotions environment is more complex than a single-model solution: it might involve deploying a REST service for the audio module and possibly the orchestration layer. But since the modules are decoupled, one can integrate the audio Docker container into a custom pipeline without the legacy Mesos orchestrator.

Integration with Voice Pipelines

MixedEmotions provides a RESTful web service interface for its audio emotion module. In practice, one could feed audio from sales calls (either in real-time chunks or as full call recordings) to this service and receive emotion outputs in JSON (valence and arousal). For real-time coaching, a possible integration is: the voice pipeline splits audio into short windows (e.g. 5 seconds), sends each to the MixedEmotions audio API, and interprets the valence/arousal stream to detect significant emotional events (e.g. spikes in arousal indicating stress). Since MixedEmotions also has fusion modules and data visualization components, it could be extended to combine audio with other signals (though here we focus on voice only). The integration effort is moderate – developers need to containerize the audio analysis and ensure the call streaming system can interact with it (perhaps via an asynchronous queue to batch audio frames). There may be some overhead in converting live audio to the required input format for the service.

Performance Optimization Strategies

The audio emotion recognition in MixedEmotions relies on computing a predefined set of acoustic features (e.g. pitch, energy, spectral features) and applying a predictive model. To optimize performance: one could reduce the feature set or processing frequency if latency is an issue (OpenSMILE allows customizing feature extraction profiles). However, given its lightweight nature, CPU processing already meets real-time in most cases. If needed, the service could be scaled horizontally by running multiple instances behind a load balancer to handle many simultaneous calls. Another strategy is to update the model – the default model was likely trained on acted datasets for valence/arousal. Training a new model on UAE-specific sales call data (to recognize frustration or engagement levels) would improve accuracy. This can be done offline and then deployed within the container (the framework allows plugging in new classifiers). Because the approach is not a large deep network, typical deep compression techniques (pruning, quantization) are less relevant here. Instead, ensure the Docker container has sufficient CPU and consider pinning threads to CPU cores for consistent low latency.

Emotion Granularity & Accuracy

MixedEmotions uses a 2-dimensional emotion representation for audio (or 3D if including dominance). Granularity is thus limited: it doesn't directly output discrete labels like "happy" or "angry," but rather continuous scores. This can capture subtle gradations (e.g. moderately positive vs highly positive emotion), which is useful for tracking customer sentiment trajectory over a call. However, it may not distinguish why valence is negative (frustration vs sadness are both negative valence). For sales call coaching, valence/arousal can identify moments of high stress or positivity, but additional interpretation is needed to map that to specific emotions. As an older research toolkit (circa 2018), its models might not match state-of-the-art deep learning performance. The valence/arousal predictions could be noisy for complex real-world audio. In quiet lab settings, such models achieve moderate correlation with human labels; on spontaneous call data, expect lower reliability. Improving this would require retraining on domain data or employing modern deep models. In summary, MixedEmotions provides a broad but shallow emotion analysis – good for overall sentiment trends, but less precise for distinct emotion categories compared to other solutions.

UAE-Readiness

MixedEmotions is strong on compliance and on-premise deployment (fully self-hostable). It can achieve real-time performance, but integration and maintenance are more involved due to its legacy architecture. Its valence/arousal output is useful but may need enhancement for fine-grained coaching insights. With tuning and possibly combining with other signals, it can serve UAE needs, though its out-of-the-box accuracy on sales calls may be limited relative to newer models.

Emolysis Toolkit (Preferred Model)

Emolysis is a modern open-source toolkit for multimodal group emotion analysis that we spotlight for UAE deployment. It was introduced as an ACII 2024 demo and offers real-time, synchronized processing of video and audio to assess group emotions. For our voice-only scenario (analyzing two-party sales calls), Emolysis provides a flexible foundation:

Overview and Capabilities

Emolysis ingests an input (originally designed for video with audio) and outputs the group's emotional state, along with valence and arousal metrics. In a call, the "group" is small (customer and agent), so Emolysis can effectively analyze the combined emotional atmosphere or be adapted to analyze each speaker separately. Notably, Emolysis supports multiple platforms (it's compatible with Android, iOS, Windows) and has an intuitive GUI for selecting modalities and target persons. This indicates a well-engineered, optimized system capable of running in diverse environments. The toolkit encourages developers to extend it to application-specific contexts – we take advantage of this by configuring Emolysis specifically for voice calls in a UAE-compliant setup.

Real-Time Processing

Emolysis is designed for near-real-time analysis of affect from multimedia input. It uses synchronized multimodal inference (e.g. facial cues, voice tone) to compute overall emotion. For voice-only use, the computational load is reduced (no face analysis), which improves speed. Emolysis likely employs deep learning models (the paper references deep affective computing) for tasks like speech emotion recognition. The provided Docker deployment uses NVIDIA GPU acceleration, implying that with a GPU, Emolysis can process video frames and audio on the fly. In a voice-only scenario, a modern GPU (or even a CPU with the video models disabled) should easily handle real-time streaming audio. The architecture processes input in streaming fashion (the GUI suggests it can continuously update group emotion), so it aligns with live call monitoring requirements. In summary, Emolysis can achieve low-latency inference for audio, especially when optimized as described below.

Deployment Feasibility (UAE)

Emolysis is fully self-hostable. It provides a Docker image that can be run with nvidea-docker for GPU support, exposing a local web service for the analysis app. This means we can deploy Emolysis on a secure UAE-based server (or edge device) and serve the emotion analysis internally. Data privacy is ensured: audio never leaves the local server, satisfying UAE data regulations. The toolkit is open-source (Apache 2.0 license), so there are no legal barriers to on-prem use. A recommended deployment architecture is to set up Emolysis as a microservice within the organization's network. The UAE-compatible architecture would involve the sales call audio being streamed (or segment-buffered) to the Emolysis service, which returns emotion inferences in real time, feeding into the coaching system. In this architecture, Emolysis's modality selection feature is utilized: only the audio modality analysis is enabled, and group-level outputs are interpreted for the two participants on the call. We might run two instances of Emolysis if we want separate analysis of customer and agent (by isolating each audio channel) and then combine results for a full picture. The system remains entirely within local infrastructure, addressing compliance and latency concerns. Additionally, Emolysis's cross-platform support means the same core engine could even be deployed at edge locations (for example, on an agent's device or branch office server if central data center latency is a concern).

Performance Optimization for Emolysis

To ensure Emolysis meets strict real-time and resource requirements on edge/private hardware, we propose several concrete optimizations:

- **Disable Unused Modalities:** For voice-only deployment, we turn off video processing. This avoids overhead from face detection or video analysis pipelines, freeing up CPU/GPU resources exclusively for audio. The Emolysis GUI/API allows selecting modalities, so we run it in audio-only mode.
- **Model Pruning and Compression:** We can apply pruning to Emolysis's audio model (if it uses a neural network for speech emotion). Removing redundant neurons/filters can shrink model size and speed up inference. Given Emolysis's research-grade models, there may be room to prune 20-30% of parameters with minimal accuracy loss. Similarly, using weight quantization (e.g. 16-bit or 8-bit precision) will accelerate compute, especially if running on CPU or utilizing Tensor Cores on NVIDIA GPUs. Quantized models use less memory and can leverage optimized instructions, reducing latency.
- **Batching and Pipeline Parallelism:** In a single call scenario, batching isn't applicable to live audio (since inference must be sequential per call). However, if the system analyzes multiple calls concurrently (e.g. multiple agents' calls at once), we can batch-process audio frames from different calls on the GPU. For example, processing 4 audio segments

in parallel on a GPU can improve throughput efficiency. We must keep batch sizes small to maintain low per-call latency. Additionally, we can pipeline audio capture and inference – while one chunk is being processed, the next chunk is simultaneously being recorded – to hide latency.

- **Hardware Acceleration:** Deploy Emolysis on a machine with a suitable GPU (or an edge AI accelerator). The Docker deployment already uses NVIDIA CUDA. For example, an NVIDIA Tesla T4 or similar can handle multiple streams of real-time audio emotion inference. If GPU is not available, we can use optimized inference engines like ONNX Runtime or OpenVINO on CPU with AVX instructions. Also consider using a Jetson Xavier (NVIDIA edge device) for smaller-scale edge deployment; Emolysis's mobile compatibility hints that its models can run on such hardware (with quantization and perhaps lower model complexity).
- **Pre-loading and Caching Models:** Ensure that all required model files are downloaded and loaded into memory at server startup (Emolysis allows mounting a cache of model checkpoints). This avoids any runtime delays or internet fetches and keeps inference deterministic. Also, any repetitive computations (like feature normalization constants) should be computed once and reused.
- **Stream Processing Optimization:** Emolysis could process audio in fixed time windows (e.g. 1-second or 2-second frames) to provide a rolling emotional assessment. We should tune the window length to balance resolution and stability. A smaller window gives faster response to emotion changes but might be noisy; a larger window smooths output but adds lag. We may choose an optimal middle ground (perhaps ~2 seconds) for coaching feedback. We also ensure overlapping windows (for smoother output) are handled efficiently – reusing computations between overlapping segments if possible.
- **Scalability:** For higher call volumes, Emolysis containers can be replicated. Using a container orchestration (Kubernetes or Docker Swarm), we can horizontally scale the emotion service. In a UAE data center, multiple Emolysis instances behind a load balancer could handle dozens of concurrent calls, with each GPU server running several instances if resources allow. This ensures the system can scale without impacting real-time performance as the business grows.

With these optimizations, Emolysis can run with low latency (well under 1 second per analysis window) on local hardware, making it suitable for live call coaching.

Emotion Granularity & Accuracy

Emolysis provides three primary outputs for a group: a categorical group emotion, and continuous valence and arousal values. The "group emotion" is an overall label (the paper doesn't list categories, but likely something like overall positive, neutral, or negative mood, or possibly a basic emotion label that best describes the group). For sales calls, this group-level label could translate to the general tone of the conversation (e.g. "customer satisfied" vs "customer frustrated"). The valence-arousal scores add nuance, allowing detection of intensity (arousal) and positivity/negativity (valence) beyond the label. This granularity is moderate – not as many discrete emotion types as Hume AI, but more informative than a single binary sentiment. It aligns well with coaching use: for instance, high arousal + negative valence could signify frustration or anger, cueing the system to advise the agent to adopt a calming strategy. Emolysis's accuracy in emotion recognition is state-of-the-art for group emotion in research settings. However, on voice-only data, accuracy will depend on the underlying speech emotion model. If Emolysis uses a pre-trained audio model (potentially trained on standard datasets), it should reliably detect broad emotional cues (happy vs angry vs neutral) in speech. There may be some mismatch for call center data (which can have more subtle expressions). We can improve accuracy by fine-tuning the audio model on annotated UAE call data if available, or by calibrating valence/arousal outputs to known successful interactions. Emolysis's advantage is that it is extensible – we can swap in a better speech emotion model or retrain its audio pipeline, given it's open-source and intended for extension. Overall, Emolysis is expected to deliver robust performance comparable to contemporary academic models. With the optimizations above, it will meet the accuracy and responsiveness needs of a production coaching system in UAE.

UAE-Readiness

Emolysis stands out as a top choice for UAE deployment. It is fully compliant (self-hosted, no external data transfer) and designed with cross-platform real-time operation in mind. Deployment is straightforward via Docker, and it can integrate into the voice pipeline as a microservice. After applying pruning, quantization, and focusing the pipeline on audio, Emolysis will achieve high throughput on local hardware. Its emotion analysis (group mood + valence/arousal) provides actionable insight for call coaching, though possibly with less granularity than Hume's cloud model. Given these strengths and the ability to iterate on the model locally, Emolysis is well-suited for an edge or private data center solution in the UAE.

Hume AI (Benchmark)

Hume AI offers a cloud-based emotion recognition API, and we use it as a benchmark for performance and capabilities. Hume's Expression Measurement platform can analyze voice (and other modalities) via a simple API, returning a rich set of emotional descriptors.

Real-Time Processing

Hume AI's service is built for low-latency streaming. Developers can send live audio to Hume's API (via WebSocket or REST streaming) and receive nearly instant analysis. The system is designed for applications like real-time voice interactions and call center analytics. In fact, Hume highlights call center use cases such as detecting caller frustration or distress in real time. This indicates that under ideal conditions, Hume's cloud can process audio with only milliseconds of inference delay, plus network transit time. In practice, if the API is hosted in a distant region (e.g. US or Europe), UAE users would incur network latency (perhaps 100–200+ ms). Hume's streaming optimization and possibly regional servers (if available) mitigate this, but it's not as immediate as an on-prem solution. Still, for benchmarking, Hume provides excellent real-time performance given a good internet connection.

Deployment & Compliance in UAE

This is where Hume falters for our needs. Hume AI is a cloud SaaS – meaning audio data from calls would have to be sent over the internet to Hume's servers for processing. Given UAE's data protection law and industry regulations, sending sensitive customer call audio abroad can violate compliance (especially for sectors like finance or telecom where data localization is mandated). Unless Hume offers an on-premise model package (as of now they do not publicly), it cannot be hosted in a private UAE environment. Thus, deployment in the UAE is not feasible from a regulatory standpoint. For a strict local-hosting requirement, Hume scores very low – it is essentially not UAE-ready due to the external cloud dependency.

Integration with Voice Pipelines

Technically, Hume AI is integration-friendly. It provides straightforward APIs and even SDKs (Python, Node) to send audio and receive JSON results. The documentation shows integration guides for platforms like Twilio (telephony) and LiveKit (WebRTC), which align exactly with voice call scenarios. For example, a sales call on Twilio could be streamed to Hume's API in real-time to analyze the emotions. This ease of integration is a major benefit – developers do not need to manage models or infrastructure, only route the audio to the API and handle responses. Hume's API returns a rich set of emotion signals; integrating that into the coaching logic (e.g. triggering alerts when "customer frustration" confidence goes high) is straightforward. If it weren't for data leaving the premises, Hume would score top marks in integration simplicity.

Emotion Granularity

Hume's strength is the granularity and richness of emotion data it provides. Its models, based on 10+ years of research by emotion scientists like Alan Cowen, can detect "hundreds of dimensions of human expression" across voice, face, and text. Specifically for voice, Hume offers multiple model outputs: Speech Prosody (tone, pitch dynamics), Vocal Expression (nuanced emotional attributes like amused, annoyed, polite, etc.), and Vocal Call Types (nonverbal vocal sounds like laughter, sighs, crying). Together, these give a very detailed profile of the speaker's emotional state. For instance, Hume could simultaneously provide: voice pitch contours, energy (prosody); estimated intensity of emotions like joy, anger, confusion (vocal expression); and detection of events like laughter or sobbing (call types). No other model in this comparison has such breadth. This high granularity is ideal for deep analysis – it can reveal subtle cues (e.g. customer's tone shows "frustration" rising while also "confusion" is present). In a sales coaching context, these detailed signals can power sophisticated insights (for example, correlating specific emotional nuances with conversion outcomes). Hume essentially sets the benchmark for what is possible in emotion analytics.

Accuracy

Hume AI's models are considered state-of-the-art in accuracy and nuance. Trained on large datasets and validated in research, they likely outperform open-source models in correctly identifying emotional states, especially subtle or compound emotions. Real-world performance is hard to measure without proprietary tests, but given that many companies trust Hume, we infer that Hume provides reliable results in many scenarios. In a direct comparison, one can expect Hume to catch emotional signals that simpler models might miss (e.g. low-level frustration that doesn't manifest as a clear change in valence). As a benchmark, Hume's output on some sample sales calls would serve to evaluate how close the open-source alternatives come in terms of identifying key emotional moments.

UAE-Readiness
Hume AI as a service is not UAE-ready for deployment due to data residency and privacy constraints – it scores low on that front. However, as a benchmark, it sets a high bar in terms of latency, emotion granularity, and accuracy. Its one-API solution is appealing technically, but from a strategic perspective, the lack of local hosting makes it unsuitable for our UAE deployment requirement. If Hume in the future provided an on-premise appliance or a UAE cloud region, it could be reconsidered. Until then, we use Hume’s capabilities as a yardstick to ensure our chosen solution (Emolysis) is optimized to approach similar performance within a compliant, self-hosted setup.

Comparative Analysis and UAE-Readiness Summary

The four models each offer distinct advantages. The table below summarizes their comparison across key metrics related to UAE deployment readiness: Local Deployment (ability to host on-prem and ease of integration), Compliance (data localization/privacy), Real-Time Latency, Emotion Granularity, and Accuracy/Reliability. Scores are rated as High, Medium, or Low for readiness in each aspect:

Model	**Local Deployment (On-Prem Integration)**	**Compliance (UAE Data Privacy)**	**Real-Time Latency (Inference Speed)**	**Emotion Granularity (Outputs)**	**Accuracy & Reliability (Voice Emotion)**
MixedEmotions	**Medium:** On-prem via Docker modules. Integration requires connecting REST modules (valence/arousal API). More setup effort (legacy orchestration).	**High:** Entire analysis done locally. Open-source (no cloud). No PII leaves UAE.	**High:** Fast processing of valence/arousal on CPU (openSMILE-based). Real-time feasible with short windowing.	**Low:** Outputs 2D valence & arousal only (continuous values). No direct emotion labels (requires interpretation for specific emotions).	**Low-Moderate:** Reliable for overall sentiment trends (positive/negative energy). May miss specific emotions (e.g. confusion). Older models less

					accurate on complex, noisy call data.
Emolysis (Preferred)	**High:** Easy on-prem deployment (Docker + GPU). Integrates as a local service with API/GUI. Mobile/edge capable. Flexible to customize for voice-only.	**High:** 100% local processing. Apache-2.0 open-source; no external data transfer. Meets UAE localization mandates.	**High:** Near-real-time multimedia processing. Audio-only mode is lightweight – sub-second inference on GPU. Scales via additional containers.	**Moderate:** Provides group emotion label + valence & arousal. Captures intensity and positivity, but fewer discrete categories than Hume. Sufficient for detecting satisfaction vs frustration.	**Moderate:** State-of-art research model for affect; accurate on clear group emotions. For calls, expected solid performance on broad states (happy vs unhappy). Fine-tuning can further improve it.
Hume AI (Cloud)	**Low:** No on-prem option (cloud API only). Simple API integration (SDKs, guides for Twilio etc.), but requires internet connectivit	**Low:** Voice data leaves country to third-party servers – not compliant with UAE privacy requirements. Not usable for sensitive call audio	**Medium:** Optimized for real-time (streaming API). Inference is fast, but network latency (~100ms+) adds to response time.	**High:** Extremely granular – hundreds of emotion dimensions (tone, nuanced expressions, vocal events). Far more detail than others (e.g. can	**High:** Proprietary models with very high accuracy on a wide range of expressions. Well-validated on diverse data; captures

	y.	in-region.	Dependent on internet reliability.	detect subtle emotions, laughter).	subtleties that open models might miss (e.g. mild frustration).

Both MixedEmotions and Emolysis all satisfy the core UAE requirement of local hosting and data privacy, whereas Hume AI – while powerful – fails that criterion. Emolysis stands out for real-time performance and easier deployment; MixedEmotions is deployable but with less out-of-the-box ease and a less rich output. Hume offers the richest analysis but at the cost of compliance and control.

In terms of real-time performance, all open-source options can be tuned to operate in or near real-time. Emolysis, using modern frameworks and possibly GPUs, achieve high throughput. MixedEmotions, using CPU-based feature extraction, is efficient and scalable for real-time with enough CPU resources. Hume's cloud is engineered for low latency but introduces network dependency.

Regarding emotion detail and accuracy, Hume is the gold standard (fine-grained and highly trained). Emolysis provides a balanced approach (some detail with valence/arousal and group mood), sufficient for distinguishing positive vs negative experiences and intensity. MixedEmotions gives a coarse but valuable measure of sentiment intensity. Accuracy-wise, open models may lag Hume, but Emolysis is expected to perform well on obvious emotional cues and can be improved with domain data. MixedEmotions might need augmentation (perhaps combining its output with speech-to-text sentiment analysis to compensate for lack of discrete emotion labels).

Conclusion and Recommendation

After a comprehensive evaluation, **Emolysis emerges as the preferred model for UAE deployment** of voice-based emotion recognition in sales calls. Emolysis offers an excellent balance of real-time performance, on-premise deployability, and sufficient emotional insight (group emotion and valence/arousal) to drive coaching feedback. By deploying Emolysis within a UAE data center and applying optimizations like modality focusing, model compression, and GPU acceleration, we ensure the solution is scalable, low-latency, and compliant with local data laws. Emolysis's open framework also allows continuous improvement (e.g. retraining on UAE-specific data, or extending to multi-speaker analysis), giving the organization full control and customization ability.

MixedEmotions provides a proven, though somewhat dated, approach. Its valence-arousal output can be useful for overall sentiment monitoring. It might serve well in combination with other signals (for instance, using valence trends plus keyword analysis from transcripts to identify moments of customer dissatisfaction). However, given the effort to integrate its older architecture, one might prefer to extract the core concept (valence/arousal analysis) and implement it with a more modern model or library within the Emolysis framework.

Hume AI, while not deployable due to compliance, has set a benchmark. In moving forward with Emolysis, it would be wise to periodically benchmark Emolysis's results against Hume on test calls (with non-sensitive data) to identify any gaps. For example, if Hume consistently detects a certain subtle emotion (like "hesitation") that correlates with outcomes, we can aim to incorporate similar detection in our local model pipeline. This benchmarking ensures our on-prem solution remains competitive with state-of-the-art cloud AI.

In conclusion, Emolysis stands out as the most suitable choice for UAE-based deployment, offering a powerful balance of real-time performance, on-premise flexibility, and actionable emotional insights. We recommend deploying Emolysis within a UAE-hosted private environment, leveraging its open-source architecture to ensure full compliance with local data laws. By applying key optimizations-such as pruning, quantization, GPU acceleration, and restricting to audio-only mode-the system can achieve low-latency performance at scale. This approach ensures a scalable, secure, and future-ready solution for real-time emotion analysis in sales environments.