

Hedy bringt NVIDIA Nemotron 3 Diarization auf iPhone und Mac

Hedy 3.12 bringt NVIDIAs Nemotron 3 Diarization auf iPhones und Apple Silicon Macs. Das Modell läuft auf Ihrem Gerät und ordnet nach einer Sitzung zu, wer was gesagt hat.

Veröffentlicht von Julian Pscheid · 23. September 2026

[Diesen Artikel online lesen: https://www.hedy.ai/de/post/nemotron-3-diarization-on-device/](https://www.hedy.ai/de/post/nemotron-3-diarization-on-device/)



Fünf Kolleginnen und Kollegen in einem lebhaften Gespräch an einem großen Besprechungstisch. Für jede Stimme fließt eine eigene leuchtende Schalllinie über den Tisch zu einem MacBook Pro.

Kurzantwort NVIDIAs Nemotron 3 Diarization ist ein Modell, das ermittelt, wer in einem Gespräch wann gesprochen hat. Es kommt mit Hedy 3.12 auf unterstützte iPhones und Apple Silicon Macs, läuft dort nach dem Ende einer Sitzung auf Ihrem Gerät und aktualisiert, welcher Sprecher was gesagt hat. Die Wörter und Zeitstempel Ihres Transkripts bleiben unverändert, und die Live-Sprecherlabels funktionieren weiter wie bisher.

NVIDIA hat heute Nemotron 3 Diarization veröffentlicht, sein bisher leistungsfähigstes Modell zur Sprecherzuordnung, und Hedy unterstützt es ab Tag 0 in Hedy 3.12 auf iPhone und Mac. Es ist das genaueste Modell zur Sprecherzuordnung, das wir bisher gesehen haben, und läuft auf Ihrem eigenen Gerät, sodass sich nach einem Meeting verlässlicher nachvollziehen lässt, wer was gesagt hat.

Sprecherlabels machen aus einem Transkript ein Gespräch. Wenn sie verrutschen, wird ein Zitat der falschen Person zugeordnet und die Zusammenfassung übernimmt den Fehler. Dieses Problem soll das Modell lösen.

Was macht Nemotron 3 Diarization?

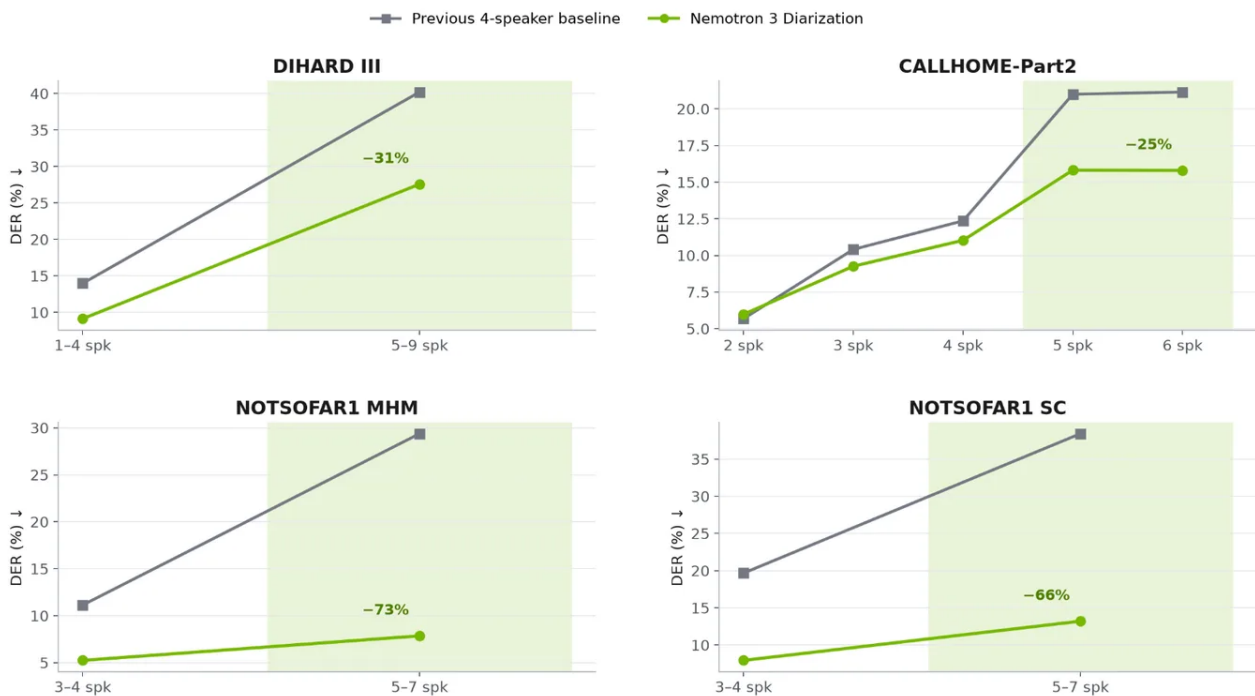
Bei der Sprecherzuordnung geht es um eine Frage: Wer hat wann gesprochen? Das Modell teilt das Audio nach Sprechern auf und markiert, wann jeder von ihnen spricht. Die gesprochenen Wörter zu transkribieren ist eine andere Aufgabe. Die Namen der Personen zu kennen ebenfalls.

Nemotron 3 Diarization ist NVIDIA's neuestes Modell für diese Aufgabe. Es baut auf NVIDIA's Forschung zu Streaming Sortformer (<https://arxiv.org/abs/2507.18446>) auf, die dafür sorgt, dass das Label eines Sprechers im Verlauf eines Gesprächs stabil bleibt. Laut NVIDIA's Modelldokumentation:

- Es wurde mit Audio aus Gesprächen trainiert, darunter Meetings, Telefonate und Podcasts.
- Es verarbeitet überlappende Sprache, also Momente, in denen zwei Personen gleichzeitig sprechen.
- Es kann in einer Aufnahme bis zu acht Sprecher verfolgen.
- Es kennzeichnet Sprecher in der Reihenfolge, in der sie erstmals sprechen. Die erste Stimme wird also zu Speaker 1.
- Es ist ein kompaktes Modell mit etwa 100 Millionen Parametern. Deshalb ist es realistisch, es auf einem Smartphone auszuführen.

Es schlägt auch Cloud-Dienste in zwei veröffentlichten Benchmarks. Laut NVIDIA's Beitrag zur Veröffentlichung (<https://huggingface.co/blog/nvidia/nemotron-diarization>) steht das Modell mit einer Fehlerrate bei der Sprecherzuordnung von 14,72 % auf Platz eins der VoiceArena-Bestenliste für Sprecherzuordnung (<https://voicearena.com/diarization-bench>). Das nächstplatzierte System liegt bei 19,3 %. Die Fehlerrate gibt an, welcher Anteil der Sprechzeit falsch zugeordnet wird. Im veröffentlichten Benchmark von Argmax (<https://github.com/argmaxinc/OpenBench/blob/update-diarization-benchmarks/BENCHMARKS.md#diarization-error-rate-der>) erzielte es auf einem Mac die niedrigste Fehlerrate unter sechs getesteten Systemen, darunter drei Cloud-Dienste zur Sprecherzuordnung.

NVIDIA berichtet außerdem, dass das Modell auf jedem getesteten Benchmark besser abschneidet als sein vorheriges Modell zur Sprecherzuordnung und der Abstand bei mehr als vier Sprechern größer wird, also dort, wo Sprecherlabels bisher oft durcheinandergerieten.



Liniendiagramme von NVIDIA vergleichen die Fehlerrate bei der Sprecherzuordnung nach Anzahl der Sprecher auf vier Benchmarks. Nemotron 3 Diarization liegt deutlich unter NVIDIAs früherem Modell für vier Sprecher; bei fünf oder mehr Sprechern ist der Abstand am größten.

Grafik: NVIDIA, aus dem Beitrag zur Veröffentlichung von Nemotron 3 Diarization (<https://huggingface.co/blog/nvidia/nemotron-diarization>) . Niedriger ist besser; die schattierten Bereiche zeigen mehr als vier Sprecher. NVIDIA veröffentlicht das Modell auf Hugging Face (<https://huggingface.co/nvidia/Nemotron-3-Diarization>) zusammen mit den vollständigen Ergebnissen seiner Evaluierung.

Was ändert Nemotron 3 Diarization in Hedy?

Hedy kennzeichnet Sprecher bereits während des Gesprächs mit der Nemotron-Sprach-Engine (/post/nemotron-on-device-speech-engine/) . Diese Live-Labels bleiben genau wie bisher.

Neu ist, was nach dem Ende einer Sitzung passiert. Live-Labels müssen sofort festgelegt werden. Deshalb prüft Hedy die Zuordnung noch einmal, wenn das Gespräch vorbei ist. Es verarbeitet die Sitzung mit Nemotron 3 Diarization und aktualisiert, welchem Sprecher die einzelnen Teile des Transkripts zugeordnet sind.

Dieser Durchlauf ändert nur die Sprecherzuordnung. Jedes Wort in Ihrem Transkript und jeder Zeitstempel bleibt unverändert. Danach funktionieren die Sprecher-Chips über Ihrem Transkript (/help/understanding-hedy-transcripts/) wie bisher: Tippen Sie auf einen Chip, um dem Sprecher einen Namen zu geben, oder führen Sie zwei Chips zusammen, die zur selben Person gehören.

Läuft Nemotron 3 Diarization auf dem Gerät?

Hedy soll Sie während eines Meetings in Echtzeit unterstützen und Ihr Gespräch privat halten. Deshalb läuft die Spracherkennung standardmäßig auf Ihrem Gerät. Das gilt auch für die Sprecherzuordnung.

Nemotron 3 Diarization folgt demselben Prinzip, und wir freuen uns, dass ein so leistungsfähiges Modell klein genug ist, um auf einem Smartphone zu laufen. Es läuft über Apples Core ML auf Ihrem iPhone oder Mac. Um zu ermitteln, wer gesprochen hat, wird kein Audio an einen Cloud-Dienst gesendet. Für

den Durchlauf nach einer Sitzung hält Hedy das Audio der Sitzung in einem temporären Cache auf Ihrem Gerät und löscht es danach. Der Leitfaden zur Spracherkennung (/help/speech-recognition-providers-in-hedy/) erklärt diesen Cache und die Einstellung dafür.

Der KI-Schritt, der Ihre Zusammenfassung und Notizen erstellt, ist eine separate Entscheidung. Mit Local AI Processing (/help/local-ai-processing/) können Sie auch ihn auf Ihrem Gerät ausführen.

Welche Geräte unterstützen Nemotron 3 Diarization?

Nemotron 3 Diarization läuft auf unterstützten iPhones und Apple Silicon Macs mit Hedy 3.12 oder neuer. Hedy auf Windows und Android behält seine bisherigen Sprecherlabels.

Hedy 3.12 wird in den nächsten 24 Stunden ausgerollt. Achten Sie deshalb in Hedy auf das Update. Sobald Sie es haben, öffnen Sie Einstellungen !' Sprache & KI und installieren Sie das Sprachmodell-Update, das Hedy dort anbietet.

Es gilt für Sitzungen, die mit der Nemotron-Sprach-Engine erfasst wurden. Diese Engine trennt die Sprecher in Hedy. Wenn Sie derzeit Whisper oder einen Cloud-Sprachanbieter verwenden, können Sie auf derselben Einstellungsseite zu Nemotron wechseln. Dann erhalten Sie Sprecherlabels live und nach der Sitzung.

Häufig gestellte Fragen

Was ist Nemotron 3 Diarization?

Es ist ein Modell von NVIDIA zur Sprecherzuordnung. Dabei wird ermittelt, wer in einer Aufnahme wann gesprochen hat. Nemotron 3 Diarization ist für Gespräche wie Meetings und Anrufe entwickelt, verarbeitet gleichzeitig sprechende Personen und kann bis zu acht Sprecher verfolgen.

Wie genau ist Nemotron 3 Diarization?

Laut NVIDIA steht es mit einer Fehlerrate bei der Sprecherzuordnung von 14,72 % auf Platz eins der VoiceArena-Bestenliste, vor dem nächstplatzierten System mit 19,3 %. Im veröffentlichten Benchmark von Argmax hatte es die niedrigste Fehlerrate unter sechs getesteten Systemen, darunter drei Cloud-Dienste zur Sprecherzuordnung. Die Fehlerrate erfasst Sprechzeit, die dem falschen Sprecher zugeordnet oder übersehen wird, sowie Abschnitte, die fälschlich als Sprache markiert werden. Je niedriger der Wert, desto besser.

Ändern sich die Sprecherlabels während einer laufenden Sitzung?

Nein. Die Live-Labels, die Sie während einer Sitzung sehen, funktionieren weiterhin wie bisher. Nemotron 3 Diarization arbeitet erst nach dem Ende der Sitzung.

Läuft es auf meinem Gerät?

Ja. Das Modell läuft über Apples Core ML auf Ihrem iPhone oder Mac. Ihr Audio wird nicht an einen Cloud-Dienst gesendet, um zu ermitteln, wer gesprochen hat.

Für welche Sitzungen gilt es?

Für Sitzungen, die mit der Nemotron-Sprach-Engine erfasst wurden. Das Modell liest das Audio der Sitzung und benötigt es deshalb auf Ihrem Gerät: Entweder ist die Audioaufnahme aktiviert oder der temporäre Audiocache für Nemotron. Der Cache ist standardmäßig aktiviert und wird nach der Verarbeitung gelöscht.

Kennt es die Namen der Personen?

Nein. Das Modell unterscheidet Stimmen und kennzeichnet sie als Speaker 1, Speaker 2 und so weiter. Die Namen kommen weiterhin von Hedy, das sie aus Vorstellungen, Ihrer Kalendereinladung und dem Themenkontext sowie aus Namen übernimmt, die Sie selbst eingeben.

Welche Geräte unterstützen es?

Unterstützte iPhones und Apple Silicon Macs mit Hedy 3.12 oder neuer. Hedy auf Windows und Android behält seine bisherigen Sprecherlabels.

Hedy AI · Live-KI-Coaching für wichtige Gespräche

Hedy kostenlos testen: <https://www.hedy.ai/de/downloads/>

<https://www.hedy.ai/de/post/nemotron-3-diarization-on-device/>