Hedy leva o NVIDIA Nemotron 3 Diarization ao iPhone e ao Mac

O Hedy 3.12 traz Nemotron 3 Diarization da NVIDIA para iPhone e Macs com Apple Silicon. O modelo roda no seu dispositivo para identificar quem disse o quê após uma sessão.

Publicado por Julian Pscheid · 23 de setembro de 2026
Ler este artigo online: https://www.hedy.ai/pt/post/nemotron-3-diarization-on-device/

Cinco colegas em uma conversa animada ao redor de uma mesa de reunião espaçosa, com uma linha luminosa de som para cada voz atravessando a mesa até um MacBook Pro

Resposta rápida Nemotron 3 Diarization da NVIDIA é um modelo que identifica quem falou e quando em uma conversa. Ele chega a iPhones compatíveis e Macs com Apple Silicon no Hedy 3.12, onde roda no seu dispositivo após o fim de uma sessão e atualiza qual falante disse o quê. As palavras e as marcas de tempo da sua transcrição permanecem como estão, e os rótulos de falantes ao vivo continuam funcionando como hoje.

A NVIDIA lançou hoje Nemotron 3 Diarization, seu modelo de diarização de falantes mais capaz até agora, e o Hedy oferece suporte ao modelo desde o dia 0 no Hedy 3.12 para iPhone e Mac. É o modelo de diarização mais preciso que já vimos e roda no seu próprio dispositivo, o que torna mais confiável o registro de quem disse o quê depois que uma reunião termina.

Os rótulos de falantes transformam uma transcrição em uma conversa. Quando deixam de corresponder às pessoas certas, uma fala é atribuída à pessoa errada e o resumo repete o erro. É esse o problema que o modelo foi criado para resolver.

O que Nemotron 3 Diarization faz?

A diarização responde a uma pergunta sobre uma gravação: quem falou e quando. Ela separa os falantes no áudio e marca os trechos em que cada um fala. Transcrever as palavras é outra tarefa, assim como saber o nome de cada pessoa.

Nemotron 3 Diarization é o modelo mais recente da NVIDIA para essa tarefa. Ele se baseia na pesquisa da NVIDIA sobre Streaming Sortformer (https://arxiv.org/abs/2507.18446) , que mantém estável o rótulo de cada falante ao longo de uma conversa. Segundo a documentação do modelo da NVIDIA:

• Ele foi treinado com áudio de conversas, incluindo reuniões, chamadas telefônicas e podcasts.

• Ele lida com falas sobrepostas, quando duas pessoas falam ao mesmo tempo.

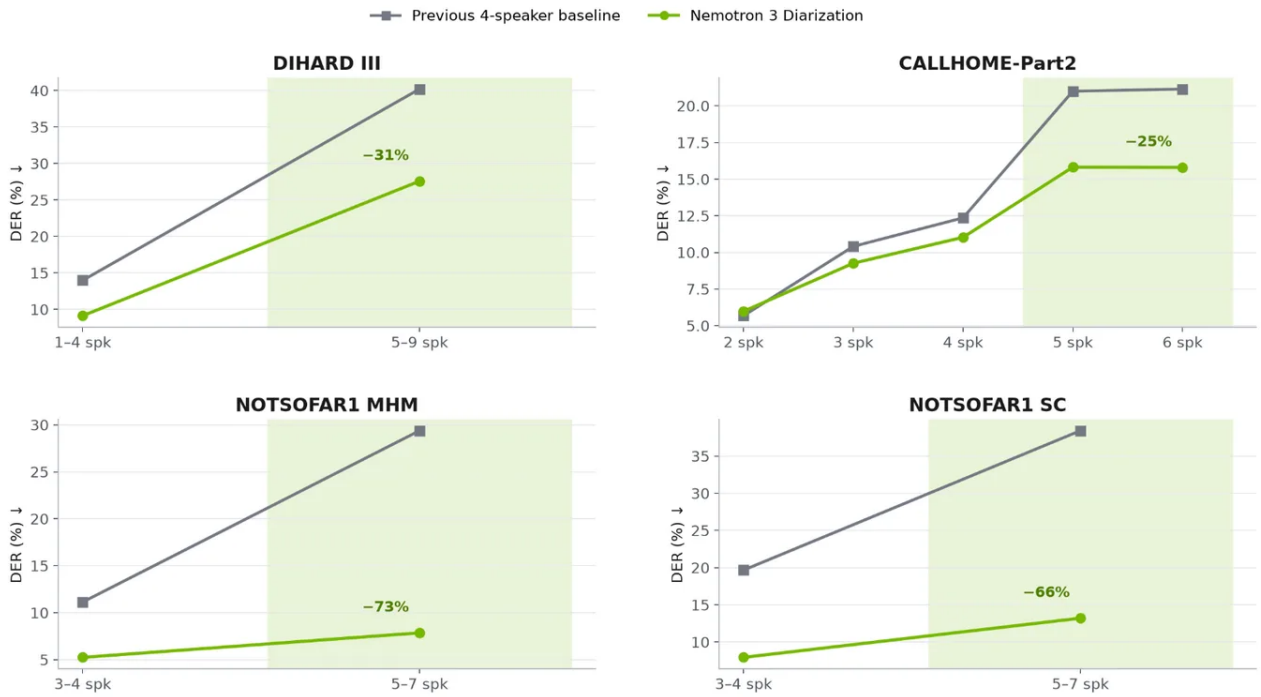
• Ele consegue acompanhar até oito falantes em uma gravação.

• Ele rotula os falantes na ordem em que falam pela primeira vez, então a primeira voz vira Speaker 1.

• É um modelo compacto, com cerca de 100 milhões de parâmetros, o que torna viável executá-lo em um celular.

Ele também supera serviços na nuvem em dois benchmarks publicados. Segundo o post de lançamento da NVIDIA (https://huggingface.co/blog/nvidia/nemotron-diarization) , o modelo ocupa o primeiro lugar no ranking de diarização da VoiceArena (https://voicearena.com/diarization-bench) com taxa de erro de diarização de 14,72% (a proporção do tempo de fala rotulado incorretamente), à frente do sistema seguinte, com 19,3%. No benchmark publicado pela Argmax (https://github.com/argmaxinc/OpenBench/blob/update-diarization-benchmarks/BENCHMARKS.md#diarization-error-rate-der) , teve a menor taxa de erro entre seis sistemas testados, três deles serviços de diarização na nuvem, enquanto rodava em um Mac.

A NVIDIA também relata ganhos em relação ao seu modelo de diarização anterior em todos os benchmarks que testou, e a diferença aumenta quando mais de quatro pessoas estão falando, situação em que os rótulos de falantes costumam falhar.



Gráficos de linhas da NVIDIA que comparam a taxa de erro de diarização por número de falantes em quatro benchmarks. Nemotron 3 Diarization fica bem abaixo do modelo anterior da NVIDIA para quatro falantes, e a diferença é maior com cinco ou mais falantes.

Gráfico: NVIDIA, do post de lançamento do Nemotron 3 Diarization (https://huggingface.co/blog/nvidia/nemotron-diarization) . Quanto menor, melhor; as áreas sombreadas representam mais de quatro falantes. A NVIDIA publica o modelo no Hugging Face (https://huggingface.co/nvidia/Nemotron-3-Diarization) junto com os resultados completos da avaliação.

O que Nemotron 3 Diarization muda no Hedy?

O Hedy já rotula os falantes enquanto você conversa, usando o motor de reconhecimento de fala Nemotron (/post/nemotron-on-device-speech-engine/) . Esses rótulos ao vivo permanecem exatamente como estão.

A novidade acontece depois que você encerra uma sessão. Os rótulos ao vivo precisam ser definidos no momento da fala, então o Hedy agora examina a conversa novamente após seu término. Ele executa Nemotron 3 Diarization na sessão e atualiza a qual falante pertence cada parte da transcrição.

Essa etapa só altera a atribuição dos falantes. Todas as palavras da sua transcrição e todas as marcas de tempo permanecem iguais. Depois, os chips de falantes acima da transcrição (/help/understanding-hedy-transcripts/) funcionam como hoje: toque em um deles para dar um nome ao falante ou una dois chips que representem a mesma pessoa.

Nemotron 3 Diarization roda no dispositivo?

O objetivo do Hedy é oferecer apoio em reuniões, ouvindo em tempo real e mantendo sua conversa privada. Por isso, o reconhecimento de fala roda no seu dispositivo por padrão, assim como a identificação dos falantes.

Nemotron 3 Diarization segue a mesma regra, e nos empolga saber que um modelo tão capaz é pequeno o suficiente para rodar em um celular. Ele roda no seu iPhone ou Mac por meio do Core ML da Apple, e nenhum áudio é enviado a um serviço na nuvem para identificar quem estava falando. Para executar essa etapa após uma sessão, o Hedy mantém o áudio da sessão em um cache temporário no dispositivo e o apaga em seguida. O guia de reconhecimento de fala (/help/speech-recognition-providers-in-hedy/) explica esse cache e a configuração que o controla.

A etapa de IA que escreve seu resumo e suas notas é uma escolha separada. Você também pode mantê-la no seu dispositivo com o Processamento de IA Local (/help/local-ai-processing/) .

Quais dispositivos são compatíveis com Nemotron 3 Diarization?

Nemotron 3 Diarization roda em iPhones compatíveis e Macs com Apple Silicon que executam o Hedy 3.12 ou posterior. O Hedy no Windows e no Android mantém seus rótulos de falantes atuais.

O Hedy 3.12 está sendo disponibilizado ao longo das próximas 24 horas, então procure a atualização no Hedy. Depois de recebê-la, abra Configurações !’ Fala e IA e instale a atualização do modelo de reconhecimento de fala que o Hedy oferece ali.

O modelo se aplica a sessões capturadas com o motor de reconhecimento de fala Nemotron, que separa os falantes no Hedy. Se você usa Whisper ou um provedor de reconhecimento de fala na nuvem hoje, mudar para Nemotron nessa mesma tela de configurações permite ter rótulos de falantes ao vivo e depois da sessão.

Perguntas frequentes

O que é Nemotron 3 Diarization?

É um modelo de diarização de falantes da NVIDIA. A diarização é a etapa que identifica quem falou e quando em uma gravação. Nemotron 3 Diarization foi criado para conversas como reuniões e chamadas, lida com pessoas falando ao mesmo tempo e consegue acompanhar até oito falantes.

Qual é a precisão do Nemotron 3 Diarization?

Segundo a NVIDIA, ele ocupa o primeiro lugar no ranking de diarização da VoiceArena, com taxa de erro de diarização de 14,72%, à frente do sistema seguinte, com 19,3%. No benchmark publicado pela Argmax, teve a menor taxa de erro entre seis sistemas testados, incluindo três serviços de diarização na nuvem. A taxa de erro de diarização representa a parcela do tempo de fala atribuída ao falante errado, não detectada ou marcada indevidamente como fala; quanto menor, melhor.

Os rótulos de falantes durante uma sessão ao vivo vão mudar?

Não. Os rótulos ao vivo que você vê durante uma sessão continuarão funcionando como hoje. Nemotron 3 Diarization faz seu trabalho depois que a sessão termina.

Ele roda no meu dispositivo?

Sim. O modelo roda no seu iPhone ou Mac por meio do Core ML da Apple. Seu áudio não é enviado a um serviço na nuvem para identificar quem estava falando.

A quais sessões ele se aplica?

Às sessões capturadas com o motor de reconhecimento de fala Nemotron. O modelo lê o áudio da sessão, então esse áudio precisa estar no seu dispositivo: a gravação de áudio deve estar ativada ou o cache temporário de áudio do Nemotron deve estar ativado. O cache vem ativado por padrão e é apagado após o processamento.

Ele sabe o nome das pessoas?

Não. O modelo diferencia as vozes e as rotula como Speaker 1, Speaker 2 e assim por diante. Os nomes ainda vêm do Hedy, que os identifica pelas apresentações, pelo convite do seu calendário, pelo contexto do Tópico e pelos nomes que você mesmo digita.

Quais dispositivos são compatíveis com ele?

iPhones compatíveis e Macs com Apple Silicon que executam o Hedy 3.12 ou posterior. O Hedy no Windows e no Android mantém seus rótulos de falantes atuais.

Hedy AI · Coaching de IA ao vivo para conversas importantes
Experimente o Hedy grátis: https://www.hedy.ai/pt/downloads/
https://www.hedy.ai/pt/post/nemotron-3-diarization-on-device/