# NVIDIA Riva Speech AI의 학술적 기반 및 고성능 추론 엔진에 대한 전문 보고서

## I. NVIDIA Riva 소개: 대화형 AI 추론을 위한 통합 플랫폼

### A. Riva 마이크로서비스 스택 및 배포 패러다임 정의

NVIDIA® Riva는 GPU 가속을 기반으로 하는 다국어 음성 및 번역 마이크로서비스의 집합체로 정의되며, 이는 완벽하게 맞춤 설정 가능하고 실시간 대화형 AI 파이프라인 구축을 목표로 한다.1 Riva는 업계 최고 수준의 자동 음성 인식(ASR), 텍스트 음성 변환(TTS), 그리고 신경망 기계 번역(NMT) 기능을 포함하며 2, 모든 클라우드, 데이터 센터, 엣지, 또는 임베디드 장치에 배포될 수 있는 유연성을 제공한다.1 Riva를 통해 기업들은 챗봇을 매력적이고 표현력이 풍부한 다국어 AI 음성 에이전트 또는 아바타로 변환하는 음성 및 번역 기능을 추가할 수 있다.1

Riva 플랫폼은 컨테이너화되어 있으며, 저지연 스트리밍 시나리오와 고처리량 오프라인 시나리오 모두에 최적화된 gRPC 기반 마이크로서비스 형태로 제공된다.3 이러한 서비스 구조는 Riva가 수백에서 수천 개의 병렬 스트림으로 쉽게 확장될 수 있도록 한다.3

### B. NVIDIA AI 스택 내 Riva의 맥락: NeMo, TensorRT, 및 Triton

Riva의 고성능 추론 능력은 NVIDIA의 더 넓은 AI 플랫폼 내 여러 계층의 기술적 통합에 기인한다.

#### 1. NeMo 훈련 기반

Riva에서 활용되는 모델들은 주로 NVIDIA NeMo(Neural Modules) 툴킷을 사용하여 개발 및 훈련된다. NeMo는 재사용성, 추상화 및 구성을 통해 AI 애플리케이션을 생성하기 위한 프레임워크 독립적인 파이썬 툴킷이다.4 이는 신경망의 개념적 블록인 뉴럴 모듈을 중심으로 구축되었으며, ASR 및 자연어 처리(NLP)를 위한 확장 가능한 사전 구축 모듈 컬렉션을 제공한다.4 NeMo는 최신 NVIDIA GPU에서 분산 훈련 및 혼합 정밀도에 대한 내장 지원을 제공하여 4, Riva에 배포되는 모델의 학술적 및 개발적 전조 역할을 수행한다. 결과적으로 Riva는 NeMo로 훈련된 모델을 위한 최적화된 추론 계층으로 기능한다.

#### 2. 추론 최적화 (TensorRT)

Riva의 근본적인 성능은 NVIDIA TensorRT™에 의해 구동되는 그래프 최적화 및 혼합 정밀도 기술의 직접적인 결과이다.3 TensorRT 최적화는 훈련된 모델을 단일 명령을 사용하여 온프레미스 또는 클라우드에서 쉽게 내보내고, 최적화하며, 음성 서비스로 배포할 수 있도록 매우 효율적인 배포 가능한 그래프 표현으로 변환한다.6

#### 3. 서빙 인프라 (Triton)

최적화된 모델의 배포 구조는 NVIDIA Triton™ Inference Server를 사용한다.6 Triton은 동시 스트림 스케줄링 및 모델 확장을 관리하여, Riva가 수백 또는 수천 개의 병렬 스트림을 원활하게 처리할 수 있도록 지원한다.3

#### 통합적 아키텍처 관찰

이러한 구조(NeMo $\rightarrow$ TensorRT $\rightarrow$ Triton $\rightarrow$ Riva 마이크로서비스)는 의도적인 추상화 계층을 나타낸다. NeMo는 학술적 및 개발 복잡성을 처리하고, TensorRT는 심층 최적화 복잡성을 처리하며, Triton은 서비스 인프라 복잡성을 관리한다. 최종적으로 Riva는 최종 서비스를 위한 단순하고 통일된 API(gRPC/Python 클라이언트 7)를 제공한다. 이러한 구조적 분리는 각 계층에 기여하는 엔지니어들이 다른 계층을 방해하지 않고 자신의 특정 구성 요소를 최적화할 수 있도록 하여, 지속적이고 시너지 효과를 내는 성능 향상으로 이어진다.

## II. 근본적인 ASR 아키텍처: Riva 정확도를 정의하는 학술 논문

Riva의 ASR 핵심은 NeMo 관련 학술 및 기술 논문에서 광범위하게 문서화된 아키텍처를 활용한다. 이들 아키텍처는 GPU에서 최대의 병렬 처리가 가능하도록 설계되었다.

### A. 컨볼루션 CTC 모델: Citrinet 및 MarbleNet

Riva에서 사용되는 초기 ASR 아키텍처는 주로 컨볼루션 기반의 Connectionist Temporal Classification (CTC) 모델을 중심으로 전개된다.

#### 1. Citrinet 아키텍처

Citrinet은 종단 간 컨볼루션 CTC 기반 ASR 모델로 상세히 설명된다.8 이 모델은 국지적 및 전역적 문맥 정보를 모두 포착하기 위해 QuartzNet 8의 1D 시간-채널 분리 가능 컨볼루션과 ContextNet에서 성공적으로 사용된 Squeeze-and-Excitation (SE) 메커니즘을 결합한다.8 Citrinet은 23개 블록과 235개 컨볼루션 레이어로 구성된 깊고 순수한 컨볼루션 구조를 특징으로 하며, 이는 GPU에서 높은 병렬 처리에 유리하게 작용한다.8

#### 2. VAD를 위한 MarbleNet

강력한 대화형 파이프라인을 위해 Riva는 Voice Activity Detection (VAD)에 MarbleNet 아키텍처 기반의 모델을 통합한다.9 VAD는 ASR 및 화자 분할(Speaker Diarization, SD)의 첫 번째 단계로 사용되는 중요한 전처리 단계이다.9 시스템이 실제 음성 세그먼트만 처리하여 컴퓨팅 리소스를 절약하도록 함으로써, 전체 RTFx(Real-Time Factor, 실시간 처리량)를 극대화하는 데 필수적이다.

### B. 하이브리드 및 트랜스포머 기반 모델: Conformer 및 변형

더 높은 정확도를 달성하기 위해 Riva는 트랜스포머 기반의 요소와 컨볼루션 네트워크를 통합한 하이브리드 모델을 지원한다.

#### 1. Conformer

Riva는 ASR 정확도를 높이기 위해 컨볼루션 스트림에 멀티 헤드 어텐션과 피드 포워드 네트워크를 통합한 Conformer와 같은 고급 하이브리드 모델을 활용할 수 있다.8 Citrinet의 경우, Attention-enhanced Citrinet은 더 적은 레이어와 블록으로 Conformer보다 더 빠른 수렴 속도와 낮은 문자 오류율을 달성할 수 있음을 보여주었다.8

#### 2. 추론 최적화를 위한 Squeezeformer

추가적인 학술 작업은 배포 효율성을 다룬다. 예를 들어, **Squeezeformer-CTC**는 Conformer의 간소화된 변형으로 설계되었으며, Temporal U-Net 스타일의 시간 축소 방식을 사용하여 메모리 소비와 FLOPs(부동 소수점 연산)를 효과적으로 줄인다.10 이는 고처리량 Riva 배포에 훨씬 더 적합하게 만든다.

#### 아키텍처 설계에 대한 분석

ASR 분야의 학술적 연구(Citrinet 및 Conformer와 같은)는 아키텍처 선택(순수 CNN 대 하이브리드)이 수렴 속도와 정확도 사이의 절충을 포함한다는 것을 보여준다.8 그러나 Squeezeformer와 같이 메모리/FLOPs를 줄이기 위해 명시적으로 고도로 최적화된 변형을 개발하는 것은 10, NVIDIA의 내부적인 학술 초점이 단순히 최고 정확도 달성(학술 벤치마크)에서 추론 그래프의 효율성 극대화(Riva 배포 메트릭)로 이동했음을 입증한다. 이는 모델이 TensorRT에 도달하기 전에 '추론 준비' 상태가 되도록 보장하는 것이다.

## III. 가속의 엔진: 실시간 ASR을 위한 GPU 가속 디코딩

이 섹션은 Riva의 저지연 ASR 성능에 직접적으로 기여한 주요 학술 논문, 즉 GPU 가속 가중 유한 상태 변환기(WFST) 빔 검색 디코더(ArXiv: 2311.04996) 11에 초점을 맞춘다.

### A. 알고리즘적 배경: CUDA WFST 디코더

이 연구는 Connectionist Temporal Classification (CTC) 모델을 위해 맞춤 제작된 기존 GPU 가속 빔 검색 디코더의 확장인 특수 CUDA WFST 디코더의 구현 및 평가를 상세히 설명한다.11 이 디코더는 Python 기반 머신러닝 프레임워크와의 사용 편의성을 위해 DLPack 기반 Python 바인딩과 함께 제공되며, 오프라인 및 온라인 시나리오 모두에서 명시적으로 벤치마킹되었다.11

### B. 오프라인 성능 검증: 처리량 (RTFx) 및 WER

연구 결과, CUDA WFST 디코더는 CPU 기반 디코더(Flashlight Decoder) 대비 압도적인 처리량 이점을 제공하는 것으로 나타났다.11

#### 1. 처리량 (RTFx) 이점

CUDA WFST 디코더는 종단 간 파이프라인 처리량 측면에서 Flashlight Decoder보다 월등히 우수했으며, 이는 RTFx(벽시계 시간당 전사된 오디오 시간)로 측정되었다.11

* CUDA Compact WFST 토폴로지의 처리량 가속 비율은 모델 크기와 데이터 세트에 따라 $4.63$배에서 $7.29$배 사이였다.11
* 가장 큰 처리량 가속은 Conformer CTC Medium 모델(test-other 데이터셋)에서 관찰되었으며, 비율은 **$7.3$배** (Flashlight: 536 RTFx; CUDA Compact: 3910 RTFx)에 달했다.11
* 가장 작은 가속은 $4.6$배(Conformer CTC Large 모델, test-clean)로, 이 역시 상당한 성능 향상이다.11

#### 2. 정확도 (WER) 향상

주목할 점은 처리량뿐만 아니라 정확도에서도 이득이 있었다는 것이다.11 **Compact WFST 토폴로지**는 동적 CPU 기반 디코더보다 일관되게 더 나은 WER을 달성했다.

* Conformer CTC Large 모델(Librispeech test-clean)의 경우, Flashlight WER은 $2.31%$였으나, CUDA Compact WER은 \*\*$2.21\%$\*\*로 개선되었다.
* 이러한 정확도 향상은 WFST 그래프에서 언어 모델 가중치가 "왼쪽으로 밀려" 검색을 미리 알릴 수 있기 때문이며, 이는 동적 디코더에서는 불가능한 최적화 기법이다.11 이는 단순한 하드웨어 가속을 넘어선 근본적인 알고리즘적 우위를 의미한다.

#### III.B. 섹션 표: 오프라인 처리량 및 WER 비교

오프라인 추론 성능 비교 (CUDA WFST 디코더 대 Flashlight 디코더)

| **모델 및 데이터셋** | **디코더 유형** | **WER (%)** | **처리량 (RTFx)** | **가속 비율** |
| --- | --- | --- | --- | --- |
| Conformer CTC Large (Test-Clean) | Flashlight (CPU) | $2.31$ | $488$ | $1.0\times$ (기준선) |
| Conformer CTC Large (Test-Clean) | CUDA Compact WFST (GPU) | **$2.21$** | $2261$ | $4.6\times$ |
| Conformer CTC Medium (Test-Other) | Flashlight (CPU) | $4.26$ (Large 모델 테스트) | $536$ | $1.0\times$ (기준선) |
| Conformer CTC Medium (Test-Other) | CUDA Compact WFST (GPU) | $4.10$ (Large 모델 테스트) | **$3910$** | **$7.3\times$** |

### C. 온라인 스트리밍 성능: 실시간 서비스 지연 시간 분석

Riva 마이크로서비스 패러다임에서 가장 중요한 지표는 실시간 지연 시간이다.3 CUDA WFST 디코더는 콜 센터와 같은 실시간 애플리케이션에 필수적인 정량화된 저지연 성능을 제공한다.2

#### 1. 총 지연 시간 감소

CUDA Compact Topology는 Flashlight Decoder(평균 총 지연 시간 $430.0$ ms)에 비해 평균 총 지연 시간을 거의 **$8$배** 낮춘 **$54.5$ ms**를 달성했다.11

#### 2. P99 지연 시간 및 큐 지연 시간 분석

로드 하에서의 견고한 신뢰성을 나타내는 P99 총 지연 시간은 Flashlight Decoder의 $541.5$ ms에서 $78$ ms로 대폭 감소했다.11 지연 시간 감소에 가장 크게 기여한 요소는 큐 지연 시간으로, $376.2$ ms에서 불과 $24.4$ ms로 감소하여 **$15.4$배**의 개선을 보였다. 이는 GPU 리소스 관리의 효율성이 크게 향상되었음을 의미한다.11

#### III.C. 섹션 표: 온라인 스트리밍 지연 시간 비교

온라인 스트리밍 지연 시간 비교 (CUDA Compact WFST 대 Flashlight 디코더)

| **지표 (ms)** | **기준선 (Flashlight Decoder)** | **CUDA Compact Topology** | **성능 개선 비율** |
| --- | --- | --- | --- |
| 평균 총 지연 시간 | $430.0$ | **$54.5$** | 약 $7.9\times$ |
| P99 총 지연 시간 | $541.5$ | **$78.0$** | 약 $6.9\times$ |
| 평균 큐 지연 시간 | $376.2$ | **$24.4$** | 약 $15.4\times$ |

### D. 고급 기능 분석: 단어 부스팅

이 연구는 Riva 사용자에게 매우 관련성이 높은 도메인 맞춤화 기능인 단어 부스팅(Word Boosting)의 효과에 대한 기술적 지침도 제공했다.11 단어 부스팅은 정확도에 해를 끼치지 않으면서 특정 단어에 가중치를 부여하는 데 사용된다. 논문은 부스팅 값이 최대 빔 폭의 **절반**을 넘지 않도록 설정하는 것이 경험적인 규칙으로 권장된다고 제안한다. 이는 정확도의 감소를 피하면서 실용적인 가이드라인을 제공한다.11

## IV. 근본적인 TTS 아키텍처: 비자기회귀적 합성

Riva는 높은 충실도와 고속 아키텍처를 기반으로 하는 업계 최고의 텍스트 음성 변환(TTS) 기능을 제공하며, 이는 NeMo를 통해 문서화되었다.2

### A. 종단 간, 비자기회귀적 생성 (FastPitch 및 HiFi-GAN)

고속, 고품질 합성을 위한 주요 모델은 **FastPitchHifiGanE2E**와 같은 종단 간 모델을 포함한다.12

#### 1. FastPitch 구성 요소

FastPitch 아키텍처는 트랜스포머 기반 인코더와 명시적인 피치(pitch)/지속 시간 예측기를 특징으로 하는 비자기회귀적 흐름을 채택한다.12 비자기회귀적 생성은 모든 음성 토큰을 병렬로 합성할 수 있게 하여, 순차적(자기회귀적) 모델에 내재된 지연 시간을 우회하는 데 중요하며, 이는 실시간 대화형 AI에 필수적이다.

#### 2. HiFi-GAN 보코더

HiFi-GAN 구성 요소는 고충실도 신경망 보코더 역할을 하며, 중간 표현으로부터 원시 오디오를 생성한다.12 이는 속도와 자연스러운 음성을 생성하는 능력으로 알려진 GAN(Generative Adversarial Network) 기반 모델이다.13 두 구성 요소의 공동 훈련은 중간 스펙트로그램 종속성을 제거하여 파이프라인을 더욱 간소화한다.12 모델 훈련 시 모든 손실은 HiFi-GAN에서 가져오며 피치 및 지속 시간 예측기에 대한 추가 손실이 포함된다.12

#### TTS의 지연 시간 및 충실도 분석

Riva의 가치 제안은 '실제와 같은 음성'과 실시간 응답성(low-latency)을 요구한다.2 FastPitch(고속, 비자기회귀적)와 HiFi-GAN(고충실도, 효율적인 GAN 보코더)을 선택함으로써, 기반이 되는 학술적 아키텍처는 진정한 대화 흐름에 필요한 서브-100ms 지연 시간 목표를 충족시키면서 MOS(Mean Opinion Score)를 최대화하는 데 중점을 둔다. 이는 인간-컴퓨터 상호 작용 요구사항이 아키텍처 선택을 주도하고 있음을 보여준다.

## V. 경험적 검증 및 지속적인 성능 최적화

이론적인 학술적 속도 향상을 넘어, 지속적인 엔지니어링 노력은 Riva가 성능 리더십을 유지하도록 보장한다.

### A. 버전 업데이트를 통한 지속적인 지연 시간 감소 정량화

업계 분석에 따르면 Riva 릴리스 간의 지속적인 최적화는 특히 고부하 조건에서 측정 가능한 지연 시간 감소를 가져온다.14 이러한 개선은 새로운 디코더의 통합과 TensorRT 및 Triton 내 커널 수준의 미세 조정을 통해 달성된다.

#### 1. Riva 2.1.0 대 1.10.0-beta 검증

초기 업그레이드에서 성능 향상이 확인되었다. 고처리량 모드에서 128개의 동시 스트림을 처리할 때, Riva 2.1.0은 1.10.0-beta 버전에 비해 Tesla V100 GPU에서 약 **$100$ ms** 더 빠르게 작동하는 것으로 나타났다.14 데이터 몬스터즈(Data Monsters)의 분석가들은 20개 이상의 동시 스트림을 가진 워크로드에서 눈에 띄는 성능 개선을 확인할 수 있다고 지적했다.14

#### 2. Riva 2.8.1 대 2.1.0 검증

이후의 최적화는 더욱 상당한 이득을 가져왔다. Tesla V100 GPU에서 Conformer-CTC ASR 모델을 스트리밍 모드로 테스트한 결과, Riva 2.8.1에서 384개의 동시 스트림은 Riva 2.1.0의 동일 구성 대비 거의 **$280$ ms** 더 적은 지연 시간을 달성했다.15 이러한 개선은 V100뿐만 아니라 T4 GPU에서도 관찰되었으며, 모든 가속기에서 성능이 증가했다.15

#### 지속적인 엔지니어링 동력 분석

CUDA WFST 디코더에서 얻은 초기 $7.3$배의 속도 향상은 단기적인 구조적 이득이다.11 그러나 후속 산업 보고서 14에서 입증된 지속적이고 대규모의 증분적 감소($100$ ms, $280$ ms)는 TensorRT와 Triton 내에서 커널을 조정하고 메모리 할당을 개선하는 등 엄격하고 지속적인 엔지니어링 최적화가 이루어지고 있음을 증명한다. 이러한 지속적인 엔지니어링 주기는 순수한 학술 논문에서는 간과될 수 있지만, 생산 배포 성공을 유지하는 데 결정적이다.

#### V. 섹션 표: Riva 버전 업데이트를 통한 ASR 스트리밍 지연 시간 감소 (V100 GPU)

| **버전 비교** | **사용 GPU** | **동시 스트림 수** | **관찰된 지연 시간 감소** |
| --- | --- | --- | --- |
| 2.1.0 대 1.10.0-beta | Tesla V100 | $128$ | 약 $100$ ms |
| 2.8.1 대 2.1.0 | Tesla V100 | $384$ | 약 $280$ ms |

## VI. 사용자 정의, 다국어 지원 및 운영 무결성

### A. 다국어 적응의 학술적 과제 (한국어 ASR 사례 연구)

Riva/NeMo는 광범위한 사용자 정의를 지원한다.2 비주류 언어를 위한 학술적 프로세스는 대규모 사전 훈련 모델을 미세 조정하는 것을 포함한다.

#### 1. 한국어 ASR 연구 사례

한국어 음성 인식을 위한 연구는 OpenAI의 Whisper 모델과 같은 다국어 ASR 시스템을 약 1,000시간의 한국어 음성 데이터로 미세 조정하는 것을 목표로 했다.17 실험 결과, 미세 조정을 통해 문자 오류율(CER) 측면에서 한국어 음성 인식 기능이 크게 향상되었으며, 모델 크기가 증가함에 따라 성능이 향상되는 경향이 포착되었다.17

#### 2. 성능 트레이드오프 분석

이러한 학술적 작업은 새로운 저자원 언어(예: 한국어)를 위해 미세 조정한 후, 원래의 주요 언어(예: 영어)의 성능이 저하되는 근본적인 문제를 강조했다.17 이는 강력하고 일반화된 다국어 모델을 동시에 개발하는 것이 여전히 활발한 연구 영역임을 확인시켜주며, 전이 학습 과정에서 모델 가중치의 신중한 균형 조정이 필요함을 시사한다.

### B. 운영 무결성: 고성능 배포의 보안 영향

최적화가 속도에 초점을 맞추는 반면, Riva의 프로덕션 준비 상태는 운영 보안 우려를 제기한다.18

#### 1. 취약성 발견

보안 연구에 따르면, 클라우드 배포에서 Riva API 엔드포인트의 잘못된 구성 패턴이 여러 조직에서 발견되었으며, 이는 심각한 위험을 초래한다.18

#### 2. 악용 시나리오

노출된 API에 대한 무단 액세스는 공격자가 두 가지 주요 방식으로 시스템을 악용할 수 있도록 한다. 첫째, 공격자는 GPU 리소스와 API 키를 무제한으로 남용하여 막대한 재정적 비용을 초래할 수 있다.18 둘째, 조직의 독점 AI 모델이나 추론 서비스를 노출시켜 지적 재산(IP) 도난의 위험을 초래할 수 있다.18

#### 가속화와 운영 위험의 상관관계 분석

Riva의 핵심 학술적 기여는 처리량($7.3\times$ RTFx)이다.11 이러한 가속은 기술적 성과이지만, 그만큼 심각한 운영 위험을 초래한다. 노출된 API가 악용될 경우, GPU 사용 비용이나 IP 도난의 규모가 처리량 증가에 비례하여 증가한다. 따라서 보안 취약점을 상세히 다룬 논문 18은 성능 논문의 중요한 보완 자료로 간주되어야 하며, 최적화가 엄격한 MLOps 및 DevSecOps 통제와 결합되어야 진정한 프로덕션 수준의 배포 성공을 이룰 수 있음을 강조한다.

## VII. 결론: 학술적 기여의 종합 및 미래 궤적

### A. 핵심 학술적 기반의 종합

NVIDIA Riva의 핵심은 전문화된 학술 및 엔지니어링 논문에 의해 검증된 세 가지 주요 기둥 위에 구축되었다.

1. **모델 효율성:** NeMo에서 파생된 아키텍처(Citrinet, Conformer, FastPitch)는 최대의 GPU 활용도를 위해 설계되었다.
2. **알고리즘적 디커플링:** CUDA WFST 디코더 논문은 하드웨어 인식 알고리즘 설계를 통해 ASR 디코딩의 전통적인 CPU/지연 시간 병목 현상을 해결함으로써, 실시간 역량을 근본적으로 변화시켰다.
3. **지속적인 최적화:** TensorRT/Triton 미세 조정은 지속적인 성능 보고서를 통해 검증되었으며, 신속한 기능 출시와 지속적인 운영 효율성 향상을 보장한다.

### B. 미래 연구 궤적

Riva는 ASR/TTS 기능을 대규모 언어 모델(LLMs) 및 검색 증강 생성(RAG)과 통합할 수 있는 역량을 기반으로 하며, 이는 앞으로의 학술적 초점이 생성형 AI와의 결합에 있음을 시사한다.2

구체적인 미래 연구 방향은 다음과 같다.

* **생성형 AI 통합:** Riva의 Canary-Qwen-2.5B ASR 모델 및 Magpie TTS 모델과 같이 2, LLM 및 RAG와의 심리스한 통합을 활용하여 정교한 가상 비서 및 디지털 아바타를 구현하는 데 중점을 둘 것이다.2
* **견고한 다국어 일반화:** 특정 언어 미세 조정에서 관찰된 성능 트레이드오프 문제를 해결하기 위해, 견고하고 일반화된 다국어 모델 개발에 대한 추가적인 학술 연구가 필요하다.17
* **운영 보안 프레임워크:** 고처리량 GPU 가속 마이크로서비스와 관련된 고유한 IP 및 리소스 위험을 완화하기 위해, 자동화된 보안 제어 및 모범 사례를 개발하는 배포 구조에 대한 지속적인 학술적 검토가 요구된다.18

#### Works cited

1. Riva Speech AI SDK - Get Started - NVIDIA Developer, accessed October 28, 2025, <https://developer.nvidia.com/riva>
2. Riva | Speech and Translation AI - NVIDIA, accessed October 28, 2025, <https://www.nvidia.com/en-us/ai-data-science/products/riva/>
3. Overview — NVIDIA Riva - NVIDIA Docs, accessed October 28, 2025, <https://docs.nvidia.com/deeplearning/riva/user-guide/docs/overview.html>
4. NeMo: a toolkit for building AI applications using Neural Modules, accessed October 28, 2025, <http://learningsys.org/neurips19/assets/papers/36_CameraReadySubmission_NeMo.pdf>
5. NVIDIA-NeMo/NeMo: A scalable generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (Automatic Speech Recognition and Text-to-Speech) - GitHub, accessed October 28, 2025, <https://github.com/NVIDIA-NeMo/NeMo>
6. NVIDIA Riva, accessed October 28, 2025, <https://docs.nvidia.com/riva/index.html>
7. NVIDIA Riva - GitHub, accessed October 28, 2025, <https://github.com/nvidia-riva>
8. arXiv:2209.00261v1 [cs.CL] 1 Sep 2022, accessed October 28, 2025, <https://arxiv.org/pdf/2209.00261>
9. Riva Marblenet Voice Activity Detection - NGC Catalog - NVIDIA, accessed October 28, 2025, <https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/voiceactivitydetection_marblenet>
10. Models — NVIDIA NeMo Framework User Guide, accessed October 28, 2025, <https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/asr/models.html>
11. GPU-Accelerated WFST Beam Search Decoder for CTC-based ..., accessed October 28, 2025, <https://arxiv.org/pdf/2311.04996>
12. TTS: Fastpitch-HifiGAN - NGC Catalog - NVIDIA, accessed October 28, 2025, <https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/tts_en_e2e_fastpitchhifigan>
13. FastPitch\_GermanTTS\_Training.ipynb - Colab - Google, accessed October 28, 2025, <https://colab.research.google.com/github/NVIDIA/NeMo/blob/stable/tutorials/tts/FastPitch_GermanTTS_Training.ipynb>
14. Boosting Performance: A Comparative Analysis of NVIDIA Riva 2.1.0 vs Riva 1.10.0, accessed October 28, 2025, <https://www.datamonsters.com/post/boosting-performance-a-comparative-analysis-of-nvidia-riva-2-1-0-vs-riva-1-10-0>
15. Comparison of NVIDIA Riva 2.8.1 vs Riva 2.1.0 - Data Monsters, accessed October 28, 2025, <https://www.datamonsters.com/post/comparison-of-nvidia-riva-2-8-1-vs-riva-2-1-0>
16. Performance comparison of NVIDIA Riva 2.1.0 vs Riva 1.10.0 | by Data Monsters | Product AI, accessed October 28, 2025, <https://medium.com/product-ai/performance-comparison-of-riva-2-1-0-vs-riva-1-10-0-3cf6e13cfc67>
17. [논문]대형 사전훈련 모델의 파인튜닝을 통한 강건한 한국어 음성인식 모델 구축, accessed October 28, 2025, <https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=JAKO202330043250854>
18. NVIDIA Riva Vulnerabilities Leave AI-Powered Speech and Translation Services at Risk, accessed October 28, 2025, <https://www.trendmicro.com/en_us/research/25/d/nvidia-riva-vulnerabilities.html>