IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF TEXAS
MIDLAND-ODESSA DIVISION

DATASCRIPTION LLC Plaintiff, v. META PLATFORMS, INC., Defendant.	CIVIL ACTION NO. 7:26-cv-00183 **JURY TRIAL DEMANDED**

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff Datascription, LLC ("Datascription" or "Plaintiff"), for its Complaint against Defendant Meta Platforms, Inc. ("Meta" or "Defendant"), alleges as follows:

PARTIES

1. Plaintiff Datascription, LLC is a limited liability company organized under the laws of Delaware, with its principal place of business at 15511 Carmenita Road, Santa Fe Springs, California 90670.

2. Defendant Meta Platforms, Inc., is a corporation organized and existing under the laws of Delaware. Meta has a place of business located at 2230 Eberhardt Road, Temple, Texas. Meta has a registered agent at Corporation Service Company, 211 E. 7th Street Suite 620, Austin, Texas 78701.

JURISDICTION AND VENUE

3. This is an action for patent infringement arising under the Patent Laws of the United States, 35 U.S.C. § 1 et seq., including 35 U.S.C. § 271 for the infringement of United States Patent No. 9,230,547 ("the '547 Patent" or "Asserted Patent").

4. This Court has subject-matter jurisdiction pursuant to 28 U.S.C. §§ 1331 and 1338(a).

5. Meta is subject to this Court's personal jurisdiction because it conducts substantial, continuous, and systematic business in Texas and in this District; directs its services to users located in this District; and operates major corporate facilities in this District.

6. Defendant Meta has been using the Wav2vec algorithm for data extraction and transcription from non-transcribed source media such as Facebook live, an uploaded video, Instagram Live, and Reels (the "Accused Instrumentalities") in Meta datacenters, which activity infringes the Asserted Patent in violation of 35 U.S.C. § 271.

7. Plaintiff's cause of action arises, at least in part, from Meta's contacts with and activities in and/or directed at this District and the State of Texas.

8. This Court has personal jurisdiction over Meta pursuant to Tex. Civ. Prac. & Rem. Code § 17.041 *et seq*.

9. Meta has previously agreed that it is subject to personal jurisdiction in the Western District of Texas. *See, e.g.*, *Resonant Systems, Inc. v. Meta Platforms, Inc.*, No. 7:25-cv-35; *Eight KHz, LLC v. Meta Platforms, Inc.*, No. 6:22-cv-575; *Mullen Indus. LLC v. Meta Platforms, Inc.,* No. 1:24-cv-354.

10. Venue is proper in this District under 28 U.S.C. §§ 1391(b)-(d) and/or § 1400(b). Meta has transacted business in this District and has committed acts of infringement in this District by using the Accused Instrumentalities in Meta datacenters. Meta also has multiple regular and established places of business in this District.

11. For example, Meta's Austin, Texas office[1] employs almost 1,500 individuals. In addition, Meta's datacenter located in this District contains tens of thousands of Meta-owned servers running Meta's core caching, storage, social-graph, video, image, ranking, CDN, and machine-learning infrastructure.[2] This facility, located at 2230 Eberhardt Road, Temple, Texas, 76504, is a permanent, physical Meta campus that operates on a continuous basis with Meta staff and contractors on site.[3]

12. Meta deploys and operates the Accused Instrumentalities on servers located at its datacenter campus(es) in this District.[4]

THE PATENT-IN-SUIT

9. On January 5, 2016, the United States Patent and Trademark Office duly and legally issued the '547 Patent, entitled "Metadata extraction of non-transcribed audio and video streams." A true and correct copy of the '547 Patent is attached as Exhibit A. The priority date of the '547 Patent is no later than May 21, 2015.

10. Datascription is the owner by assignment of all right, title, and interest in and to the '547 Patent, including the right to bring suit for past, present, and future patent infringement, and to collect past, present, and future damages.

11. The Asserted Patent is valid and enforceable under the United States Patent Laws.

12. Inventors Kenny De Angelis and Jonathan Wilder met in 2007 and forged their relationship through a number of startup ventures. Building on the concept of sentiment extraction

[1] https://www.chron.com/culture/article/meta-austin-office-20221773.php
[2] https://www.facebook.com/TempleDataCenter/; https://www.govtech.com/products/new-data-center-coming-to-temple-texas-near-meta-campus; https://engineering.fb.com/2025/09/29/data-infrastructure/metas-infrastructure-evolution-and-the-advent-of-ai/
[3] https://datacenters.atmeta.com/wp-content/uploads/2025/02/Metas-Temple-Data-Center.pdf
[4] https://datacenters.atmeta.com/us-locations/

from text, Datascription evolved into a comprehensive multimedia search and analysis tool capable of indexing images, sound, and video to provide Google-like searchability within thousands of hours of footage.

13. Datascription's platform specifically addressed a critical gap in the business intelligence market: the industry's historical inability to analyze video content at the same scale as text. Recognizing that video would soon dominate global data traffic, Datascription developed technology to transform real-time or archived video streams into searchable text via speech recognition, while simultaneously performing frame-by-frame object detection and optical character recognition.

14. The '547 Patent is directed to a server-based system and methods for converting non-transcribed media (audio, video, imagery, and potentially printed/analog sources) into a time-aligned, machine-transcribed dataset and then extracting searchable metadata from it. At a high level, the server extracts audio/video streams, uses a speech recognition engine to produce a time-aligned textual transcription (e.g., word-level timestamps), and may store the transcription along with time-aligned video frames and other extracted information in a database/data warehouse accessible over a network by client devices.

15. An exemplary theme is normalizing audio loudness across different source files onto a single universal amplitude scale so that sound levels become comparable and queryable across a corpus. The patent describes examples of extracting audio amplitude over timed intervals/frames, such as generating an audio histogram, identifying the loudest and softest frames, assigning normalized min/max values (e.g., 0 and 100), and mapping each frame's dB value into that normalized scale (including by comparing frames against the histogram and/or using

Euclidean distance to select the closest normalized amplitude value). The normalized, time-aligned amplitude values may then be stored per frame in the database for later querying.

16. By way of further example, on top of transcription and amplitude normalization, the system may perform time-aligned analytics/metadata extraction on both text and video frames—e.g., sentiment scoring, NLP entity/topic/theme extraction, demographic estimation, and psychological profile estimation from the transcript; and OCR, facial recognition, and object recognition from time-aligned video frames—and results may be stored and searchable via a GUI/API. The patent describes exemplary combined, constraint-based searching across modalities (text + audio amplitude + visual detections), such as finding high-amplitude segments with many faces, or locating utterances of a word at peak loudness, as well as downstream uses like generating digital advertising based on extracted textual metadata.

COUNT I

Infringement of U.S. Patent No. 9,230,547

17. Datascription repeats and incorporates by reference each preceding allegation.

18. Meta has at least used the Accused Instrumentalities that incorporate one or more of the inventions claimed in the ’547 Patent in its owned and operated datacenters within the United States.

19. Meta has infringed and continues to infringe, either literally or under the doctrine of equivalents, one or more claims, including at least claim 1 of the ’547 Patent in violation of 35 U.S.C. § 271, *et seq*, by using software in its datacenters that practices each element of the claim, including the Wav2vec software.

20. Claim 1 of the ’547 patent is reproduced below with the addition of labels [a] - [j], corresponding to portions of the claim.

[Preamble] A computer based method for transcribing and extracting metadata from a non-transcribed source media, comprising the steps of:

[1a] extracting an audio stream from the non-transcribed source media by a processor-based server;

[1b] speech recognition processing of the audio stream to transcribe the audio stream into a time-aligned textual transcription by a speech recognition engine to provide a time-aligned machine transcribed media;

[1c] extracting time-aligned audio frames from the audio stream by an audio frame engine;

[1d] processing the time-aligned audio frames to extract audio amplitudes by a timed interval, to measure aural amplitudes of the extracted audio amplitudes and assign a numerical value to each extracted audio amplitude to provide time-aligned aural amplitudes by a server processor;

[1e] generating an audio histogram of the audio stream by the server processor;

[1f] normalizing the audio stream to a single, normalized, universal amplitude scale by determining a loudest frame with a loudest sound and a softest frame with a softest sound within the audio stream by the server processor;

[1g] assigning a normalized minimum amplitude value to the softest frame of the audio stream and a normalized maximum amplitude value to the loudest frame of the audio stream;

[1h] comparing each frame of the audio stream to the loudest frame and the softest frame by utilizing the audio histogram and assigning a normalized amplitude value between the normalized minimum amplitude value and the normalized maximum amplitude value to said each frame in accordance with a result of the comparison;

[1i] processing the time-aligned machine transcribed media by the server processor to extract time-aligned textual metadata associated with the source media; and

[1j] storing the time-aligned machine transcribed media, the time-aligned audio frames, the time-aligned aural amplitudes, time-aligned textual metadata and the normalized amplitude value of each frame of the audio stream in a database.

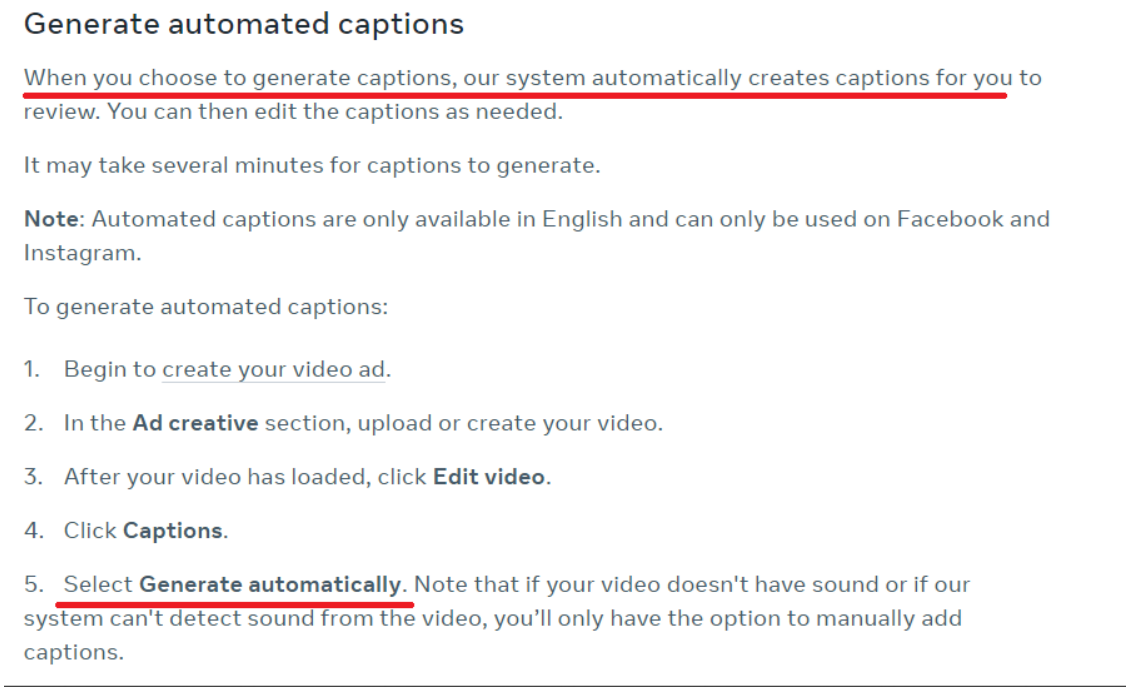
21. The ’547 Patent Accused Instrumentalities satisfy each and every limitation of at least claim 1 of the ’547 Patent, literally or under the doctrine of equivalents, as described in the non-limiting example set forth below. This non-limiting example is preliminary and is not intended to limit Plaintiff’s right to modify this non-limiting example or allege that other of Meta’s products

and services in Meta's datacenters infringe the identified claim, or any other claims, of the '547 Patent.

"[Preamble] A computer based method for transcribing and extracting metadata from a non-transcribed source media, comprising the steps of:"

22. The Accused Instrumentality discloses a computer-based method for transcribing and extracting metadata from a non-transcribed source media (e.g., a Facebook live, an uploaded video, Instagram live and Reels, etc.).

23. For example, as shown below, Meta datacenters provide automated transcriptions for live and uploaded media over the Facebook platform. Meta's own published information on the transcription service, as well as reporting on the same, emphasize that it is now used automatically for videos on Facebook. It utilizes the Wav2vec algorithm to analyze speech and generate transcriptions.



Generate automated captions

When you choose to generate captions, our system automatically creates captions for you to review. You can then edit the captions as needed.

It may take several minutes for captions to generate.

Note: Automated captions are only available in English and can only be used on Facebook and Instagram.

To generate automated captions:

1. Begin to create your video ad.
2. In the **Ad creative** section, upload or create your video.
3. After your video has loaded, click **Edit video**.
4. Click **Captions**.
5. Select **Generate automatically**. Note that if your video doesn't have sound or if our system can't detect sound from the video, you'll only have the option to manually add captions.

https://www.facebook.com/business/help/1675722002698686?id=603833089963720

When captions are added to your video, they may increase watch time, distribution and accessibility. To help you save time, captions are automatically generated when you:

- Upload a reel on a mobile device from your Facebook Page or profile in professional mode
- Upload a video on demand on your computer from your Facebook Page or Meta Business Suite

In the above cases, auto-generated captions are turned on by default. You can turn them off for all your videos or on individual videos when you upload them. Learn more about how to turn off auto-generated captions. You can also turn on a setting that will require your review of auto-generated captions before they publish.

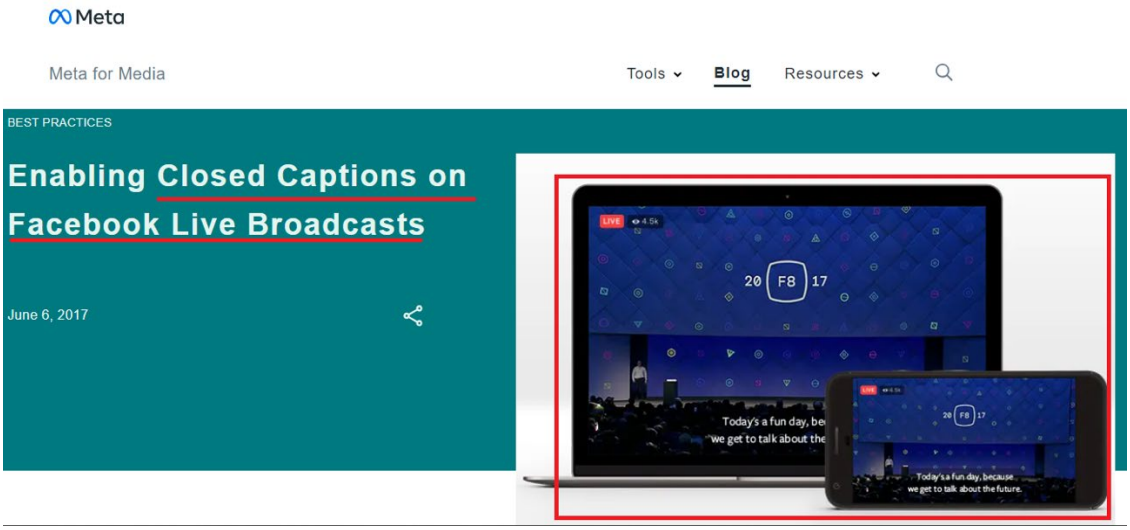
Auto-generated captions are given an accuracy rating to indicate how accurate we think they are based on things like sound quality, speech clarity and background noise.

https://www.facebook.com/business/help/385195769594602

New Automated Captions Powered by AI

September 15, 2020

https://about.fb.com/news/2020/09/new-automated-captions-powered-by-ai/



https://www.facebook.com/formedia/blog/enabling-closed-captions-on-facebook-live-broadcasts

Facebook AI researchers and engineers jumped into action and have now made live video content more accessible by enabling automatic closed captions for Facebook Live and Workplace Live. Already, six languages are supported: English, Spanish, Portuguese, Italian, German and French. Facebook Live automatic captions are helping governments disseminate crucial public health information, and ensuring that millions of viewers across the world – whether they have hearing loss, or are just watching where audio is not available – get the message. And, as workplace policies evolve, automatic captioning has become essential for employers to keep their staff and customers informed through safety updates.

The speed and scale of this AI-powered technology was only possible thanks to advances Facebook AI has made in automated speech recognition (ASR) over the past few years.

https://about.fb.com/news/2020/09/new-automated-captions-powered-by-ai/

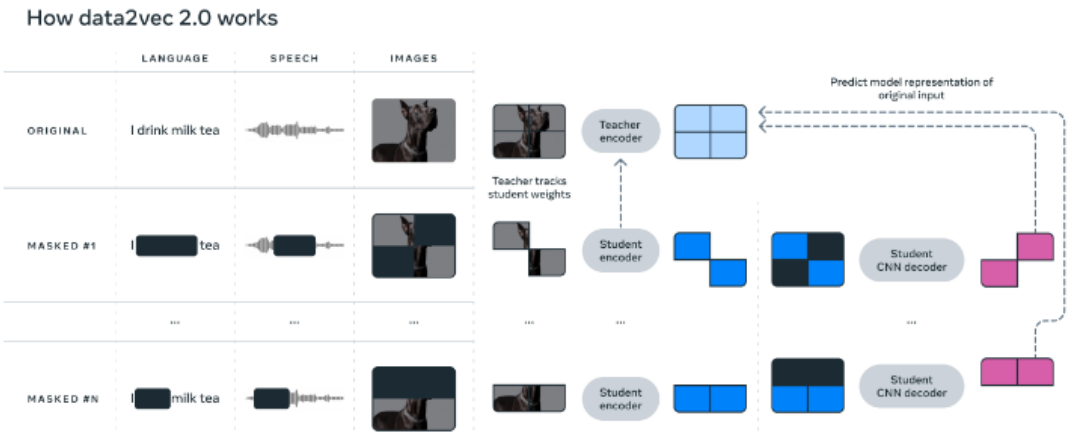
How it works:

We built Libri-light using more than 60,000 hours of unlabeled speech in English from LibriVox, a large repository of public domain audiobooks. In addition to filtering corrupted and duplicated data and adding speech activity, speaker, and genre metadata to make Libri-light useful in the context of ASR training, we built baselines systems and evaluation metrics on top of the popular LibriSpeech ASR benchmark.

https://ai.meta.com/blog/a-new-open-benchmark-for-speech-recognition-with-limited-or-no-supervision/

How data2vec 2.0 works

The general idea of self-supervised learning is for machines to learn the structure of images, speech, and text simply by observing the world. Advances in this area have led to many breakthroughs in speech (e.g., wav2vec 2.0), computer vision (e.g., masked autoencoders), and natural language processing (e.g., BERT). But modern systems can be computationally demanding, as training very large models requires many GPUs.



https://ai.meta.com/blog/ai-self-supervised-learning-data2vec/

What are Instagram Auto-Generated Captions?

Instagram's auto-generated captions are the text transcriptions of any speech found in an Instagram video or Reel. They appear on the screen in coordination with the speech in the video, meaning that they synchronize with the visual content for context. As the name suggests, these captions are automatically generated by the app without you having to manually input them.

https://influencermarketinghub.com/instagram-auto-generated-captions/

- **Instagram Live**
  - Navigate to your Instagram profile page by clicking your profile picture in the bottom right-hand corner of the app.
  - Tap the hamburger menu in the top right corner and scroll down to Your app and media.
  - Tap Accessibility > Captions, and toggle Instagram's auto-generated captions to On.
  - All Instagram Live broadcasts will now show with auto-generated captions.
- **Facebook Live**
  - Click your profile picture in the top right corner of the screen.
  - Select Settings and Privacy > Settings > Videos (located in the left column).
  - Toggle Always Show Captions to On.
  - All Facebook Live broadcasts will now show with auto-generated captions.

https://www.rev.com/blog/caption-blog/live-stream-captions

"[a] extracting an audio stream from the non-transcribed source media by a processor-based server"

24. The Accused Instrumentality discloses extracting an audio stream (e.g., audio from Facebook live, an uploaded video, Instagram live, reels etc.) by a processor-based server.

25. For example, as shown below, Meta datacenters provide automated transcriptions for live and uploaded media over Facebook and Instagram. For example, Facebook utilizes its automated speech recognition (ASR) algorithms to analyze speech and generate transcriptions. It extracts raw audio data from the video to predict a sequence of words from the extracted audio.

Facebook AI researchers and engineers jumped into action and have now made live video content more accessible by enabling automatic closed captions for Facebook Live and Workplace Live. Already, six languages are supported: English, Spanish, Portuguese, Italian, German and French. Facebook Live automatic captions are helping governments disseminate crucial public health information, and ensuring that millions of viewers across the world – whether they have hearing loss, or are just watching where audio is not available – get the message. And, as workplace policies evolve, automatic captioning has become essential for employers to keep their staff and customers informed through safety updates.

The speed and scale of this AI-powered technology was only possible thanks to advances Facebook AI has made in automated speech recognition (ASR) over the past few years.

https://about.fb.com/news/2020/09/new-automated-captions-powered-by-ai/

What are Instagram Auto-Generated Captions?

Instagram's auto-generated captions are the text transcriptions of any speech found in an Instagram video or Reel. They appear on the screen in coordination with the speech in the video, meaning that they synchronize with the visual content for context. As the name suggests, these captions are automatically generated by the app without you having to manually input them.

https://influencermarketinghub.com/instagram-auto-generated-captions/

MTIA has been deployed in our data centers and is now serving models in production. We are already seeing the positive results of this program as it's allowing us to dedicate and invest in more compute power for our more intensive AI workloads.

The results so far show that this MTIA chip can handle both low complexity and high complexity ranking and recommendation models which are key components of Meta's products. Because we control the whole stack, we can achieve greater efficiency compared to commercially available GPUs (graphics processing units).

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

> Facebook's services rely on fleets of servers in data centers all over the globe — all running applications and delivering the performance our services need. This is why we need to make sure our server hardware is reliable and that we can manage server hardware failures at our scale with as little disruption to our services as possible.

https://engineering.fb.com/2020/12/09/data-center-engineering/how-facebook-keeps-its-large-scale-infrastructure-hardware-up-and-running/

> Meta is the latest major **hyperscale cloud company** that has adopted AMD EPYC CPUs to power its data centers. Both companies worked together to define an open, cloud-scale, single-socket server designed for performance and power efficiency, based on the 3rd Gen EPYC processor.

https://techhq.com/2021/11/amd-strikes-chip-deal-to-power-metas-data-centers/

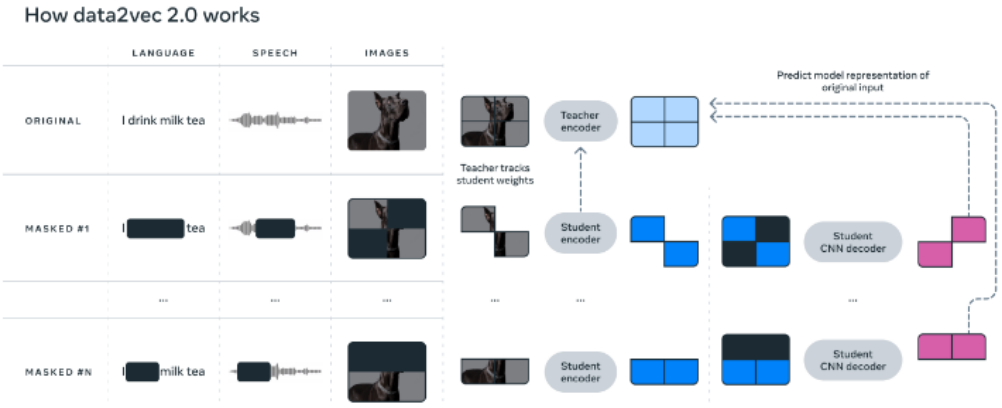
"[b] speech recognition processing of the audio stream to transcribe the audio stream into a time-aligned textual transcription by a speech recognition engine to provide a time-aligned machine transcribed media"

26. The Accused Instrumentality uses speech recognition processing (e.g., ASR processing, etc.) of the audio stream (e.g., audio from Facebook live, an uploaded video, Instagram live, reels, etc.) to transcribe the audio stream into a time-aligned textual transcription by a speech recognition engine (e.g., ASR using Wav2vec, etc.) to provide a time-aligned machine transcribed media (e.g., automatically transcribed audio, etc.).

27. For example, as shown below, Meta provides automated transcriptions for live and uploaded media on Facebook. It utilizes the Wav2vec algorithm to analyze speech and generate transcriptions. It extracts raw audio data from the video to predict a sequence of words from the extracted audio. Wav2vec is used for ASR and generate time-aligned transcripts from unlabeled audio data.

How data2vec 2.0 works

The general idea of self-supervised learning is for machines to learn the structure of images, speech, and text simply by observing the world. Advances in this area have led to many breakthroughs in speech (e.g., wav2vec 2.0), computer vision (e.g., masked autoencoders), and natural language processing (e.g., BERT). But modern systems can be computationally demanding, as training very large models requires many GPUs.



https://ai.meta.com/blog/ai-self-supervised-learning-data2vec/

- We're releasing our code for wav2vec, an algorithm that uses raw, unlabeled audio to train automatic speech recognition (ASR) models.
- This self-supervised approach beats traditional ASR systems that rely solely on transcribed audio, including a 22 percent accuracy improvement over Deep Speech 2, while using two orders of magnitude less labeled data.
- Wav2vec trains models to learn the difference between original speech examples and modified versions, often repeating this task hundreds of times for each second of audio, and predicting the correct audio milliseconds into the future.
- Reducing the need for manually annotated data is important for developing systems that understand non-English languages, particularly those with limited existing training sets of transcribed speech. Wav2vec is also part of our ongoing commitment to self-supervised training, which could accelerate the development of AI systems across the field.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

Although self-supervision has shown promise in natural language processing (NLP) tasks — including RoBERTa, Facebook AI's optimized pretraining method that recently topped the leaderboard for a major NLP benchmark — wav2vec applies the approach specifically to speech. Our algorithm does not require transcriptions, and our model learns from unlabeled audio data.

Most current ASR models train on the log-mel filter bank features of speech data, meaning audio that's been processed to make vocal features stand out. Our approach instead turns raw speech examples into a representation — specifically, a code — that can be fed into an existing ASR system. Using wav2vec's representations as inputs enables the algorithm to work with a wide variety of existing speech recognition models, making unlabeled audio data more widely useful for speech-related AI research.

One of the primary challenges in building wav2vec was dealing with the continuous nature of speech data, which makes it difficult to directly predict the data. We addressed this issue by using a pretraining regime inspired in part by the popular NLP algorithm word2vec. This algorithm learns representations by training a model to distinguish between the true data and a set of distractor samples.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

For wav2vec, we created an architecture consisting of two multilayer convolutional neural networks stacked on top of each other. The encoder network maps raw audio input to a representation, where each vector covers about 30 milliseconds (ms) of speech. The context network uses those vectors to generate its own representations, which cover a larger span of up to a second.

The model then uses these representations to solve a self-supervised prediction task. Within each 10-second audio clip that the model is trained on, wav2vec generates a number of distractor examples, which swap out 10 ms of the original audio with sections from elsewhere in the clip. The model must then determine which version is correct. And this selection process is repeated multiple times for each 10-second training clip, essentially quizzing the model to discern accurate speech sounds from distractor samples hundreds of times per second.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

We preprocessed the data to improve quality and to make it usable by our machine learning algorithms. To do so, we trained an alignment model on existing data in over 100 languages and used this model together with an efficient forced alignment algorithm that can process very long recordings of about 20 minutes or more. We applied multiple rounds of this process and performed a final cross-validation filtering step based on model accuracy to remove potentially misaligned data. To enable other researchers to create new speech datasets, we added the alignment algorithm to PyTorch and released the alignment model.

Thirty-two hours of data per language is not enough to train conventional supervised speech recognition models. This is why we built on wav2vec 2.0, our prior work on self-supervised speech representation learning, which greatly reduced the amount of labeled data needed to train good systems. Concretely, we trained self-supervised models on about 500,000 hours of speech data in over 1,400 languages — this is nearly five times more languages than any known prior work. The resulting models were then fine-tuned for a specific speech task, such as multilingual speech recognition or language identification.

https://ai.meta.com/blog/multilingual-model-speech-recognition/

FORCED ALIGNMENT WITH WAV2VEC2

Author: Moto Hira

This tutorial shows how to align transcript to speech with `torchaudio`, using CTC segmentation algorithm described in CTC-Segmentation of Large Corpora for German End-to-end Speech Recognition.

https://pytorch.org/audio/stable/tutorials/forced_alignment_tutorial.html

---

What are Instagram Auto-Generated Captions?

Instagram's auto-generated captions are the text transcriptions of any speech found in an Instagram video or Reel. They appear on the screen in coordination with the speech in the video, meaning that they synchronize with the visual content for context. As the name suggests, these captions are automatically generated by the app without you having to manually input them.

https://influencermarketinghub.com/instagram-auto-generated-captions/

"[c] extracting time-aligned audio frames from the audio stream by an audio frame engine"

28. The Accused Instrumentality extracts time-aligned audio frames (e.g., audio frames, etc.) from the audio stream (e.g., audio from Facebook live, an uploaded video, etc.) by an audio frame engine.

29. For example, as shown below, Meta extracts time-aligned audio frames from the audio stream to provide automated transcriptions for live and uploaded media on Facebook. It utilizes the Wav2vec algorithm to analyze speech and generate transcriptions that are time-aligned with the audio stream so that the two appear in synchrony.

"[d] processing the time-aligned audio frames to extract audio amplitudes by a timed interval, to measure aural amplitudes of the extracted audio amplitudes and assign a numerical value to each extracted audio amplitude to provide time-aligned aural amplitudes by a server processor"

30. Upon information and belief, the Accused Instrumentality processes the time-aligned audio frames (e.g., audio frames, etc.) to extract audio amplitudes by a timed interval, to measure aural amplitudes (e.g., loudness, etc.) of the extracted audio amplitudes and assign a numerical value (e.g., loudness level such as a LUFS (Loudness Units Full Scale) value, etc.) to each extracted audio amplitude to provide time-aligned aural amplitudes (e.g., time-aligned LUFS values for the audio, etc.) by a server processor (e.g., processor at Meta server).

31. For example, as shown below, Facebook and Instagram datacenters normalize the loudness levels (aural amplitude numerical value) of audio streams using an xHE-AAC coding scheme from the Fraunhofer Institute. As also indicated by the Audio Engineering Society's explanation of loudness and normalization concepts, this coding scheme calculates loudness levels (assigning a numerical value of LUFS) of the video/audio and normalizes the LUFS to a standard or target LUFS value. For example, the loudness levels are measured and normalized in a time-

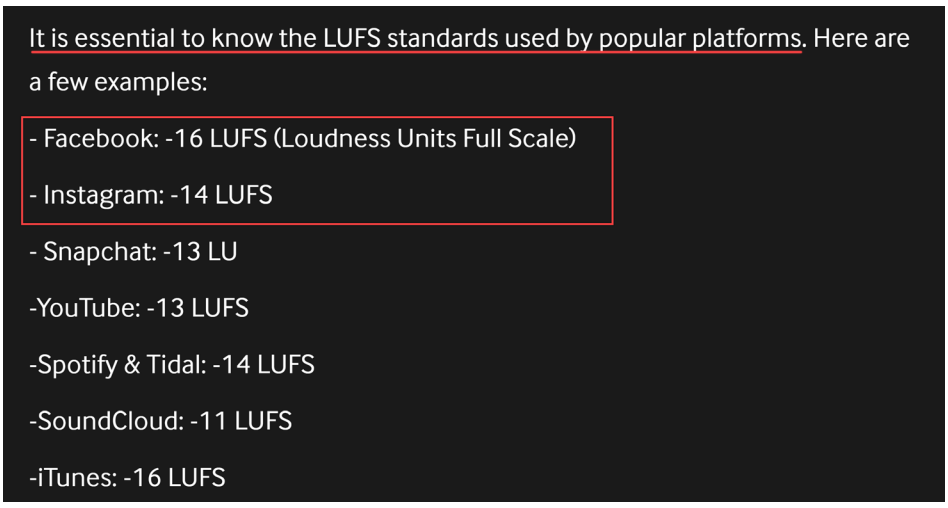
aligned manner. Further, on information and belief, Facebook and Instagram perform loudness normalization based on measurements consistent with ITU-R BS.1770.

Profiles

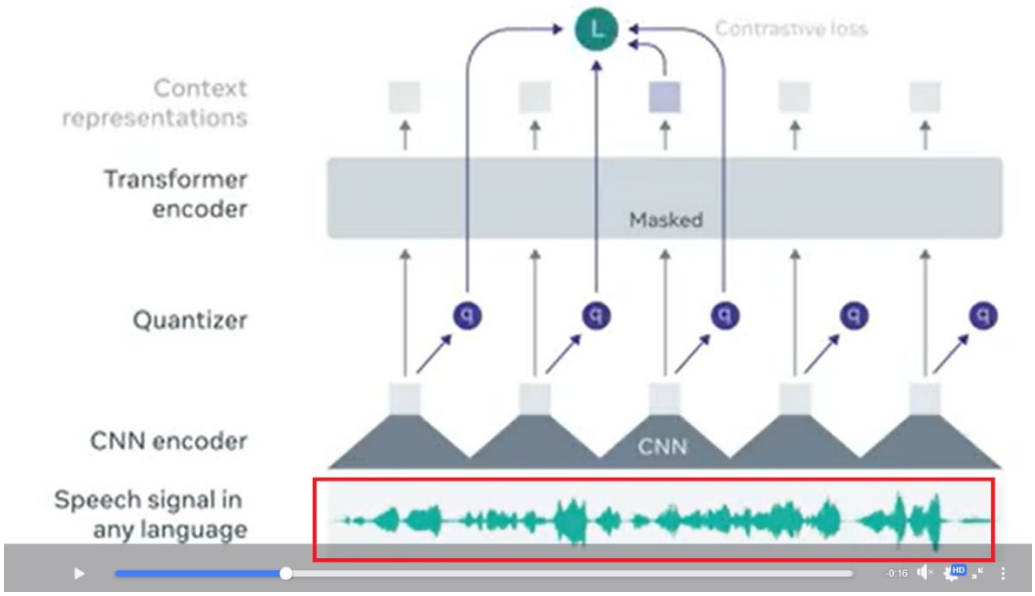
The following profiles are defined for identifying the constraints used in recommendations for different platforms and standards.

Loudness profile	Maximum loudness (LKFS)	Minimum loudness (LU)	Maximum True Peak (dBTP)
standard_a85	-22	-26	-2
standard_r128	-22.5	-23.5	-1
service_amazon	-13	-15	-1
service_apple	-15	-17	-1
service_facebook	-15	-17	-1

https://docs.dolby.io/media-apis/docs/loudness

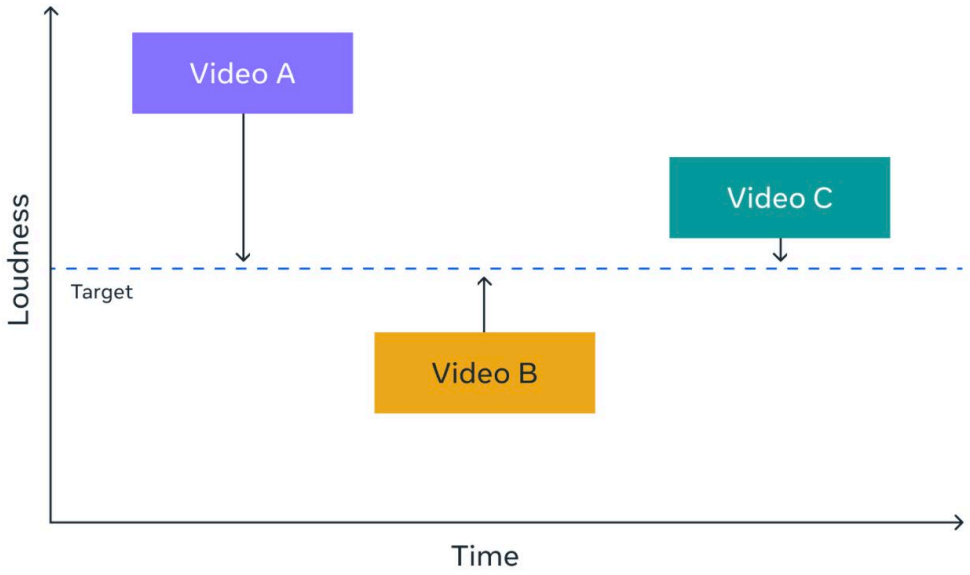


https://starsoundstudios.com/blog/lufs-social-media-platform-standards-mastering-music



https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

When people play these videos sequentially, they can perceive some audio as being too loud or too quiet. This creates listener fatigue from having to constantly adjust the volume.



xHE-AAC's integrated loudness management system solves for loudness inconsistency while meticulously preserving creator intent by bringing the average loudness of all sessions to the same target level and managing the dynamic range of each session to fit the playback environment.

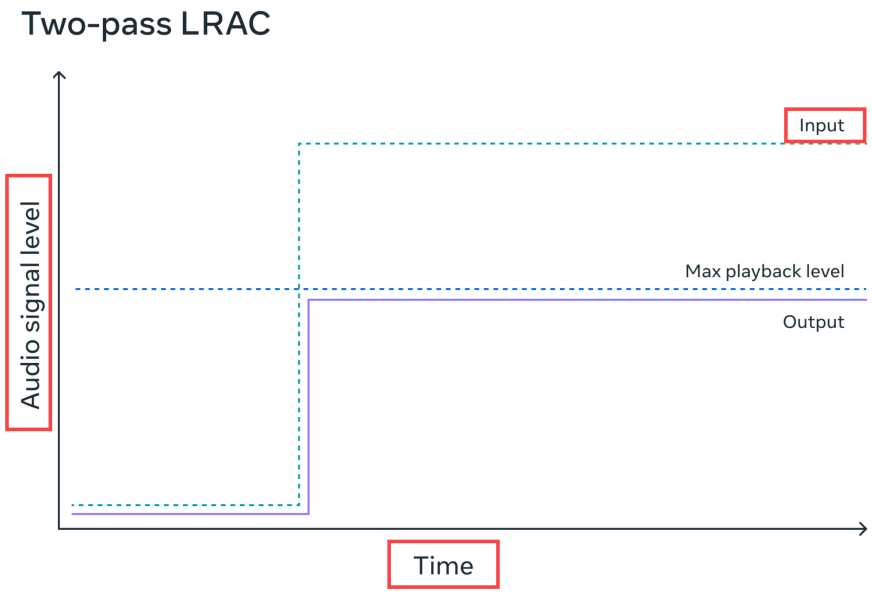
Instead of burning in a specific target level and dynamic range compression (DRC) profile during encoding, xHE-AAC allows us to leave the original audio characteristics untouched and delegate loudness management processing to the client via loudness metadata, for the optimal audio experience based on context.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

How we deployed xHE-AAC

We generate xHE-AAC bitstreams using an encoder SDK provided by the Fraunhofer Institute for Integrated Circuits IIS, and then prepare the resulting audio files for DASH streaming with shaka-packager. The xHE-AAC encoder's two-pass encoding mode is used to measure the input loudness envelope and average program loudness on the first pass and perform the actual audio data compression on the second pass. As an added benefit, two-pass encoding allows us to use loudness range control (LRAC) DRC, which mitigates pumping artifacts otherwise introduced by single-pass DRC algorithms.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/



https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

2 Match your audio levels

The next step is to match your audio levels to the platform's standards and expectations. Audio levels are measured in decibels (dB) and indicate how loud or quiet your sound is. Different platforms have different loudness targets and thresholds, which affect how your audio is perceived and processed. For example, YouTube uses a loudness target of -14 LUFS (Loudness Units Full Scale), while Facebook and Instagram use -16 LUFS. If your audio is too loud or too quiet, it can be distorted, compressed, or normalized by the platform, which can reduce its quality and impact. You can use audio editing software like Audacity or Adobe Audition to measure and adjust your audio levels.

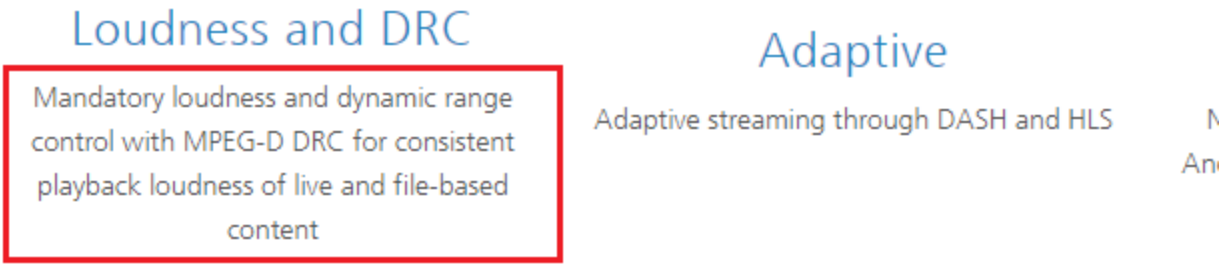
https://www.linkedin.com/advice/3/how-can-you-follow-audio-standards-different-platforms

Key Facts

Experience superior audio and video streaming with xHE-AAC

xHE-AAC, the latest generation of the AAC codec family, is the ideal solution for today's audio and video streaming services – be it movies, music, audiobooks or podcasts. With adaptive DASH/HLS streaming from 12 to more than 320 kbit/s for stereo, as well as improved speech quality and stereo imaging, xHE-AAC noticeably improves reception and sound quality. Its mandatory MPEG-D DRC loudness and dynamic range control and seamless bitrate adjustment guarantee the best user experience in any playback situation. xHE-AAC decoding is natively supported in Android, Fire OS, iOS, and Windows. In addition, xHE-AAC is the mandatory audio codec for Digital Radio Mondiale (DRM).

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html



https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

Loudness and Dynamic Range Control

MPEG-D DRC – Loudness and Dynamic Range Control – provides mandatory loudness control for xHE-AAC to play back content at a consistent volume and offers dynamic range control processing to provide the best possible user experience for listening on any platform and in any environment.

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

5 Normalization Principles

A. Loudness and Normalization

ITU-R BS.1770 defines a method for measuring Integrated Loudness of audio: a frequency-weighted level-gated measurement of average power over an interval of time. BS.1770 Loudness is an electric signal measurement relative to digital full scale, not an acoustic measurement. Absolute Loudness is measured in LUFS:, loudness units relative to full scale. Relative loudness with respect to a reference loudness is measured in LU, loudness units. A unit of loudness difference (LU) is equivalent to a decibel (dB). The measurement is frequency-weighted to approximate the sensitivity of the ear to different frequencies, and is level-gated to emphasize the parts of the audio contributing most to the sensation of loudness. EBU - TECH 3341 and ITU-R BS.1771 provide additional information about loudness measurement.

Loudness Normalization adjusts the loudness of content to match a desired Distribution Loudness by applying uniform attenuation or gain. This reduces annoying loudness jumps and the incentive to produce loud content.

https://www.aes.org/technical/documentDownloads.cfm?docID=731%20

Despite the complexity of human perception to sound, combining the elements of RMS measurement and frequency weighting make it possible to define loudness measurements accurately and efficiently.To address the need for an objective measure of loudness for broadcasting use, the ITU-R formed a group to work to study this. The work resulted in ITU-R BS.1770, which specifies an algorithm for the objective measure of the loudness of an audio programme. It is based on an algorithm by Soulodre of the Communications Research Centre, Canada, see [X1 X2 X3] and Recommendation ITU-R BS.1770 for more details. It defines how to compute Integrated Loudness – a single number that represents the average perceived loudness over the entire audio content, which might be a song, a music album or a whole program. The algorithm is conceptually very simple:

- A specific frequency weighting scheme known as K-weighting is applied to the incoming audio signal.
- RMS measurements are computed for small blocks of the K-weighted audio. Some of these RMS measurements are intentionally discarded if they are too quiet. Learn More: How is RMS computed and how to interpret it?
- At the end, the integrated loudness is computed as the average of the remaining RMS measurements. Learn More: How the Loudness Meter Works

https://aes2.org/resources/audio-topics/loudness-project/loudness-basics/

Normalization Technique

Normalization is the active level matching of file-based audio content, such as a record album or radio program, to a defined "target loudness." Normalizing avoids listener annoyance by matching loudness from one content asset to the next but it does not affect the quality of the content. At its most basic, file-based audio content is normalized by these steps:

- The full length of the audio content is measured for its Integrated Loudness, in LUFS. See the Loudness Basics section for measurement
- The amplitude of the entire audio content is then corrected so that the Integrated Loudness matches the target loudness. For example, if the target loudness is -24 LUFS (which is often used for audio content creation), and the content measures -27 LUFS, a gain offset of +3 LU to the content produces the target loudness.

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

When applying normalization, the loudness of the original audio (shown on the left in the figure) might be above or below the desired Distribution Loudness. In this case its Integrated Loudness is -13 LUFS, which is greater than the "target" of -18 LUFS. (See *AES TD1008* for specific recommendations on target loudness.) Normalization attenuates the audio by 5 LU to the target value. This process is commonly called Downward Normalization.

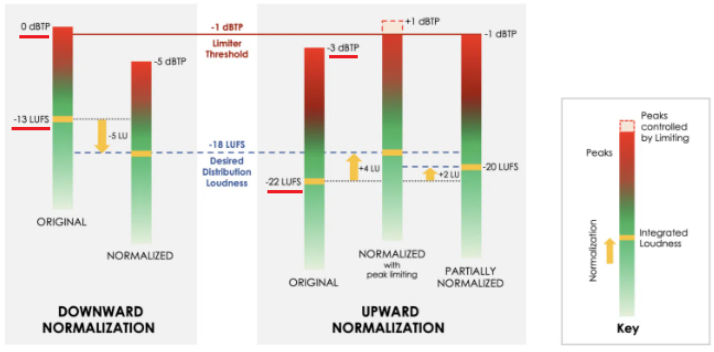


Figure 1 - Illustration of Downward Normalization and Upward Normalization to a Target Loudness of -18 LUFS

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

MTIA has been deployed in our data centers and is now serving models in production. We are already seeing the positive results of this program as it's allowing us to dedicate and invest in more compute power for our more intensive AI workloads.

The results so far show that this MTIA chip can handle both low complexity and high complexity ranking and recommendation models which are key components of Meta's products. Because we control the whole stack, we can achieve greater efficiency compared to commercially available GPUs (graphics processing units).

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

Facebook's services rely on fleets of servers in data centers all over the globe — all running applications and delivering the performance our services need. This is why we need to make sure our server hardware is reliable and that we can manage server hardware failures at our scale with as little disruption to our services as possible.

https://engineering.fb.com/2020/12/09/data-center-engineering/how-facebook-keeps-its-large-scale-infrastructure-hardware-up-and-running/

Meta is the latest major **hyperscale cloud company** that has adopted AMD EPYC CPUs to power its data centers. Both companies worked together to define an open, cloud-scale, single-socket server designed for performance and power efficiency, based on the 3rd Gen EPYC processor.

https://techhq.com/2021/11/amd-strikes-chip-deal-to-power-metas-data-centers/

"[e] generating an audio histogram of the audio stream by the server processor"

32. The Accused Instrumentality generates an audio histogram of the audio stream (e.g., audio from Facebook live, an uploaded video, Instagram Live, Reels, etc.) by the Facebook server processor.

33. For example, upon information and belief, and as shown above and below, the amplitude of the entire audio file is normalized to match the integrated loudness value to the target loudness value. The audio stream is analyzed to calculate the integrated loudness of the audio stream for normalization. The integrated loudness is calculated by taking the average of the amplitudes for an audio stream. The calculation of the average requires knowledge of the frequency of each amplitude, i.e., audio histogram. Thus, for every amplitude, the frequency is determined.

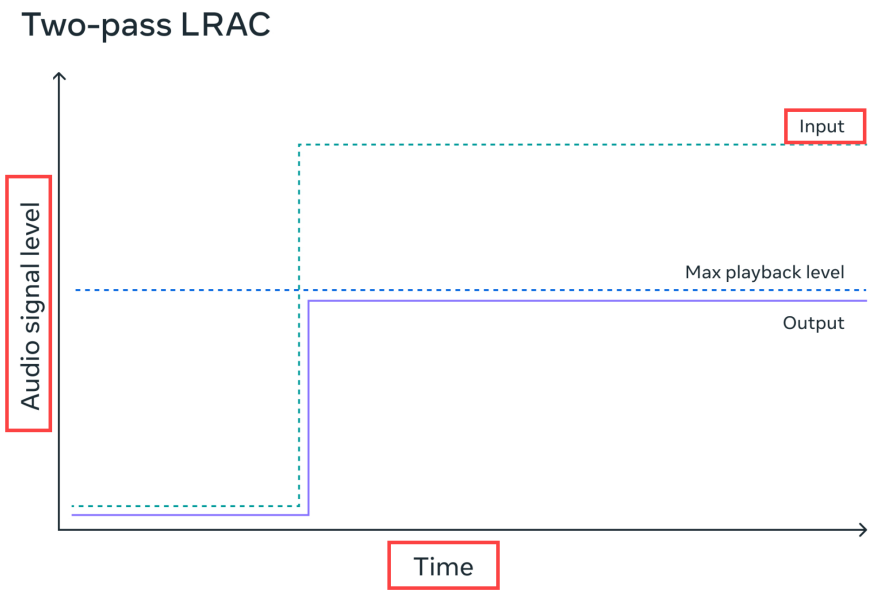
xHE-AAC's integrated loudness management system solves for loudness inconsistency while meticulously preserving creator intent by bringing the average loudness of all sessions to the same target level and managing the dynamic range of each session to fit the playback environment.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

How we deployed xHE-AAC

We generate xHE-AAC bitstreams using an encoder SDK provided by the Fraunhofer Institute for Integrated Circuits IIS, and then prepare the resulting audio files for DASH streaming with shaka-packager. The xHE-AAC encoder's two-pass encoding mode is used to measure the input loudness envelope and average program loudness on the first pass and perform the actual audio data compression on the second pass. As an added benefit, two-pass encoding allows us to use loudness range control (LRAC) DRC, which mitigates pumping artifacts otherwise introduced by single-pass DRC algorithms.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/



https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

2 Match your audio levels

The next step is to match your audio levels to the platform's standards and expectations. Audio levels are measured in decibels (dB) and indicate how loud or quiet your sound is. Different platforms have different loudness targets and thresholds, which affect how your audio is perceived and processed. For example, YouTube uses a loudness target of -14 LUFS (Loudness Units Full Scale), while Facebook and Instagram use -16 LUFS. If your audio is too loud or too quiet, it can be distorted, compressed, or normalized by the platform, which can reduce its quality and impact. You can use audio editing software like Audacity or Adobe Audition to measure and adjust your audio levels.

https://www.linkedin.com/advice/3/how-can-you-follow-audio-standards-different-platforms

Key Facts

Experience superior audio and video streaming with xHE-AAC

xHE-AAC, the latest generation of the AAC codec family, is the ideal solution for today's audio and video streaming services – be it movies, music, audiobooks or podcasts. With adaptive DASH/HLS streaming from 12 to more than 320 kbit/s for stereo, as well as improved speech quality and stereo imaging, xHE-AAC noticeably improves reception and sound quality. Its mandatory MPEG-D DRC loudness and dynamic range control and seamless bitrate adjustment guarantee the best user experience in any playback situation. xHE-AAC decoding is natively supported in Android, Fire OS, iOS, and Windows. In addition, xHE-AAC is the mandatory audio codec for Digital Radio Mondiale (DRM).

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

Loudness and DRC

Mandatory loudness and dynamic range control with MPEG-D DRC for consistent playback loudness of live and file-based content

Adaptive

Adaptive streaming through DASH and HLS

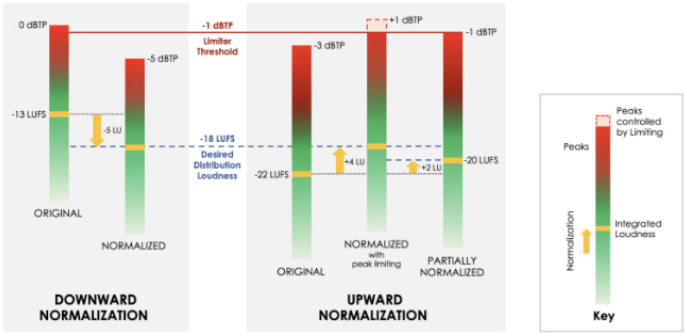
https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

Loudness and Dynamic Range Control

MPEG-D DRC – Loudness and Dynamic Range Control – provides mandatory loudness control for xHE-AAC to play back content at a consistent volume and offers dynamic range control processing to provide the best possible user experience for listening on any platform and in any environment.

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

When applying normalization, the loudness of the original audio (shown on the left in the figure) might be above or below the desired Distribution Loudness. In this case its Integrated Loudness is -13 LUFS, which is greater than the "target" of -18 LUFS. (See *AES TD1008* for specific recommendations on target loudness.) Normalization attenuates the audio by 5 LU to the target value. This process is commonly called Downward Normalization.



https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

Normalization Technique

Normalization is the active level matching of file-based audio content, such as a record album or radio program, to a defined "target loudness." Normalizing avoids listener annoyance by matching loudness from one content asset to the next but it does not affect the quality of the content. At its most basic, file-based audio content is normalized by these steps:

- The full length of the audio content is measured for its Integrated Loudness, in LUFS. See the Loudness Basics section for measurement
- The amplitude of the entire audio content is then corrected so that the Integrated Loudness matches the target loudness. For example, if the target loudness is -24 LUFS (which is often used for audio content creation), and the content measures -27 LUFS, a gain offset of +3 LU to the content produces the target loudness.

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

Despite the complexity of human perception to sound, combining the elements of RMS measurement and frequency weighting make it possible to define loudness measurements accurately and efficiently.To address the need for an objective measure of loudness for broadcasting use, the ITU-R formed a group to work to study this. The work resulted in ITU-R BS.1770, which specifies an algorithm for the objective measure of the loudness of an audio programme. It is based on an algorithm by Soulodre of the Communications Research Centre, Canada, see [X1 X2 X3] and Recommendation ITU-R BS.1770 for more details. It defines how to compute Integrated Loudness – a single number that represents the average perceived loudness over the entire audio content, which might be a song, a music album or a whole program. The algorithm is conceptually very simple:

- A specific frequency weighting scheme known as K-weighting is applied to the incoming audio signal.
- RMS measurements are computed for small blocks of the K-weighted audio. Some of these RMS measurements are intentionally discarded if they are too quiet. Learn More: How is RMS computed and how to interpret it?
- At the end, the integrated loudness is computed as the average of the remaining RMS measurements. Learn More: How the Loudness Meter Works

https://aes2.org/resources/audio-topics/loudness-project/loudness-basics/

Normalization, is the active level matching of audio content to a defined or "target" loudness. Normalizing avoids listener annoyance and their need to reach for the volume control when the content changes. To normalize audio, the amplitude of the audio signal is scaled uniformly over the length of the content, such as a full record album or radio program. For example, if the content originally measured -27 LUFS, +3 LU of gain would result in a new loudness of -24 LUFS. See the Loudness Normalization section on the Home Page for a full description.

https://aes2.org/resources/audio-topics/loudness-project/loudness-basics/

MTIA has been deployed in our data centers and is now serving models in production. We are already seeing the positive results of this program as it's allowing us to dedicate and invest in more compute power for our more intensive AI workloads.

The results so far show that this MTIA chip can handle both low complexity and high complexity ranking and recommendation models which are key components of Meta's products. Because we control the whole stack, we can achieve greater efficiency compared to commercially available GPUs (graphics processing units).

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

Facebook's services rely on fleets of servers in data centers all over the globe — all running applications and delivering the performance our services need. This is why we need to make sure our server hardware is reliable and that we can manage server hardware failures at our scale with as little disruption to our services as possible.

https://engineering.fb.com/2020/12/09/data-center-engineering/how-facebook-keeps-its-large-scale-infrastructure-hardware-up-and-running/

Meta is the latest major **hyperscale cloud company** that has adopted AMD EPYC CPUs to power its data centers. Both companies worked together to define an open, cloud-scale, single-socket server designed for performance and power efficiency, based on the 3rd Gen EPYC processor.

https://techhq.com/2021/11/amd-strikes-chip-deal-to-power-metas-data-centers/

"[f] normalizing the audio stream to a single, normalized, universal amplitude scale by determining a loudest frame with a loudest sound and a softest frame with a softest sound within the audio stream by the server processor"

34. Upon information and belief, and as shown above and below, the Accused Instrumentality normalizes the audio stream (e.g., audio from Facebook live, an uploaded video, etc.) to a single, normalized, universal amplitude scale (e.g., a normalized LUFS scale, etc.) by determining a loudest frame with a loudest sound (e.g., loudest frame of the audio, etc.) and a softest frame with a softest sound (e.g., softest frame of the audio, etc.) within the audio stream (e.g., audio from Facebook live, an uploaded video, Instagram Live, Reels, etc.) by the Meta server processor.

35. For example, as shown below, Facebook and Instagram running on Meta datacenters normalize the loudness levels (aural amplitude numerical value). On information and belief, the Accused Instrumentality calculates loudness levels of the video/audio in LUFS and normalizes the calculated LUFS to a standard or target LUFS value, based on ITU-R BS.1770. It calculates an integrated loudness value of the audio stream and determines its difference with a target loudness value. For example, this difference is later used to normalize all the audio frames within the program including the loudest and softest frames.

Profiles

The following profiles are defined for identifying the constraints used in recommendations for different platforms and standards.

Loudness profile	Maximum loudness (LKFS)	Minimum loudness (LU)	Maximum True Peak (dBTP)
standard_a85	-22	-26	-2
standard_r128	-22.5	-23.5	-1
service_amazon	-13	-15	-1
service_apple	-15	-17	-1
service_facebook	-15	-17	-1

https://docs.dolby.io/media-apis/docs/loudness

It is essential to know the LUFS standards used by popular platforms. Here are a few examples:

- Facebook: -16 LUFS (Loudness Units Full Scale)
- Instagram: -14 LUFS
- Snapchat: -13 LU
-YouTube: -13 LUFS
-Spotify & Tidal: -14 LUFS
-SoundCloud: -11 LUFS
-iTunes: -16 LUFS

https://starsoundstudios.com/blog/lufs-social-media-platform-standards-mastering-music

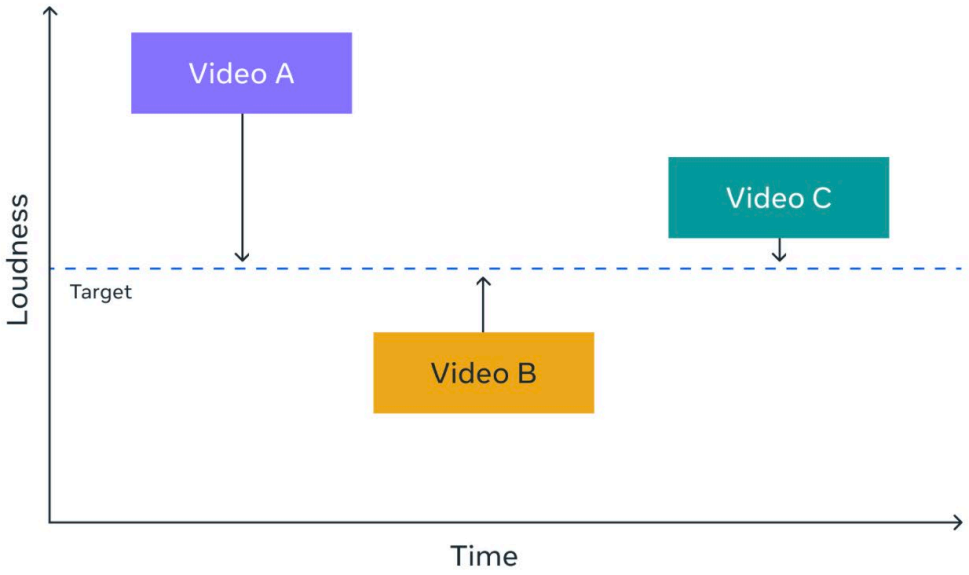
Why xHE-AAC is being embraced at Meta

- We're sharing how Meta delivers high-quality audio at scale with the xHE-AAC audio codec.
- xHE-AAC has already been deployed on Facebook and Instagram to provide enhanced audio for features like Reels and Stories.

At Meta, we serve every media use case imaginable for billions of people across the world — from short-form, user-generated content, such as Reels, to premium video on demand (VOD) and live broadcasts. Given this, we need a next-generation audio codec that supports a range of operating points with excellent compression efficiency and modern, system-level audio features.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

When people play these videos sequentially, they can perceive some audio as being too loud or too quiet. This creates listener fatigue from having to constantly adjust the volume.



xHE-AAC's integrated loudness management system solves for loudness inconsistency while meticulously preserving creator intent by bringing the average loudness of all sessions to the same target level and managing the dynamic range of each session to fit the playback environment.

Instead of burning in a specific target level and dynamic range compression (DRC) profile during encoding, xHE-AAC allows us to leave the original audio characteristics untouched and delegate loudness management processing to the client via loudness metadata, for the optimal audio experience based on context.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

How we deployed xHE-AAC

We generate xHE-AAC bitstreams using an encoder SDK provided by the Fraunhofer Institute for Integrated Circuits IIS, and then prepare the resulting audio files for DASH streaming with shaka-packager. The xHE-AAC encoder's two-pass encoding mode is used to measure the input loudness envelope and average program loudness on the first pass and perform the actual audio data compression on the second pass. As an added benefit, two-pass encoding allows us to use loudness range control (LRAC) DRC, which mitigates pumping artifacts otherwise introduced by single-pass DRC algorithms.

https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

Key Facts

Experience superior audio and video streaming with xHE-AAC

xHE-AAC, the latest generation of the AAC codec family, is the ideal solution for today's audio and video streaming services – be it movies, music, audiobooks or podcasts. With adaptive DASH/HLS streaming from 12 to more than 320 kbit/s for stereo, as well as improved speech quality and stereo imaging, xHE-AAC noticeably improves reception and sound quality. Its mandatory MPEG-D DRC loudness and dynamic range control and seamless bitrate adjustment guarantee the best user experience in any playback situation. xHE-AAC decoding is natively supported in Android, Fire OS, iOS, and Windows. In addition, xHE-AAC is the mandatory audio codec for Digital Radio Mondiale (DRM).

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

Loudness and DRC

Mandatory loudness and dynamic range control with MPEG-D DRC for consistent playback loudness of live and file-based content

Adaptive

Adaptive streaming through DASH and HLS

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

Loudness and Dynamic Range Control

MPEG-D DRC – Loudness and Dynamic Range Control – provides mandatory loudness control for xHE-AAC to play back content at a consistent volume and offers dynamic range control processing to provide the best possible user experience for listening on any platform and in any environment.

https://www.iis.fraunhofer.de/en/ff/amm/broadcast-streaming/xheaac.html

5 Normalization Principles

A. Loudness and Normalization

ITU-R BS.1770 defines a method for measuring Integrated Loudness of audio: a frequency-weighted level-gated measurement of average power over an interval of time. BS.1770 Loudness is an electric signal measurement relative to digital full scale, not an acoustic measurement. Absolute Loudness is measured in LUFS:, loudness units relative to full scale. Relative loudness with respect to a reference loudness is measured in LU, loudness units. A unit of loudness difference (LU) is equivalent to a decibel (dB). The measurement is frequency-weighted to approximate the sensitivity of the ear to different frequencies, and is level-gated to emphasize the parts of the audio contributing most to the sensation of loudness. EBU - TECH 3341 and ITU-R BS.1771 provide additional information about loudness measurement.

Loudness Normalization adjusts the loudness of content to match a desired Distribution Loudness by applying uniform attenuation or gain. This reduces annoying loudness jumps and the incentive to produce loud content.

https://www.aes.org/technical/documentDownloads.cfm?docID=731%20

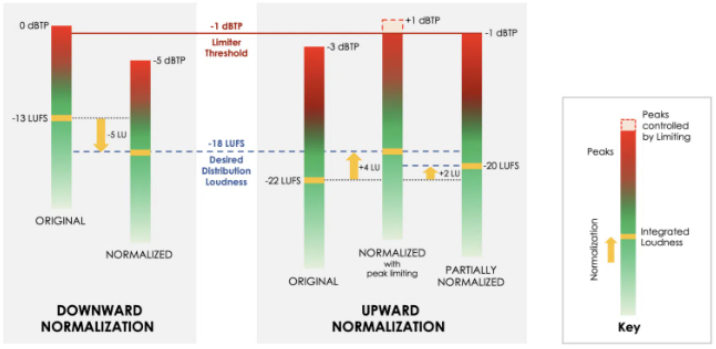
Normalization Technique

Normalization is the active level matching of file-based audio content, such as a record album or radio program, to a defined "target loudness." Normalizing avoids listener annoyance by matching loudness from one content asset to the next but it does not affect the quality of the content. At its most basic, file-based audio content is normalized by these steps:

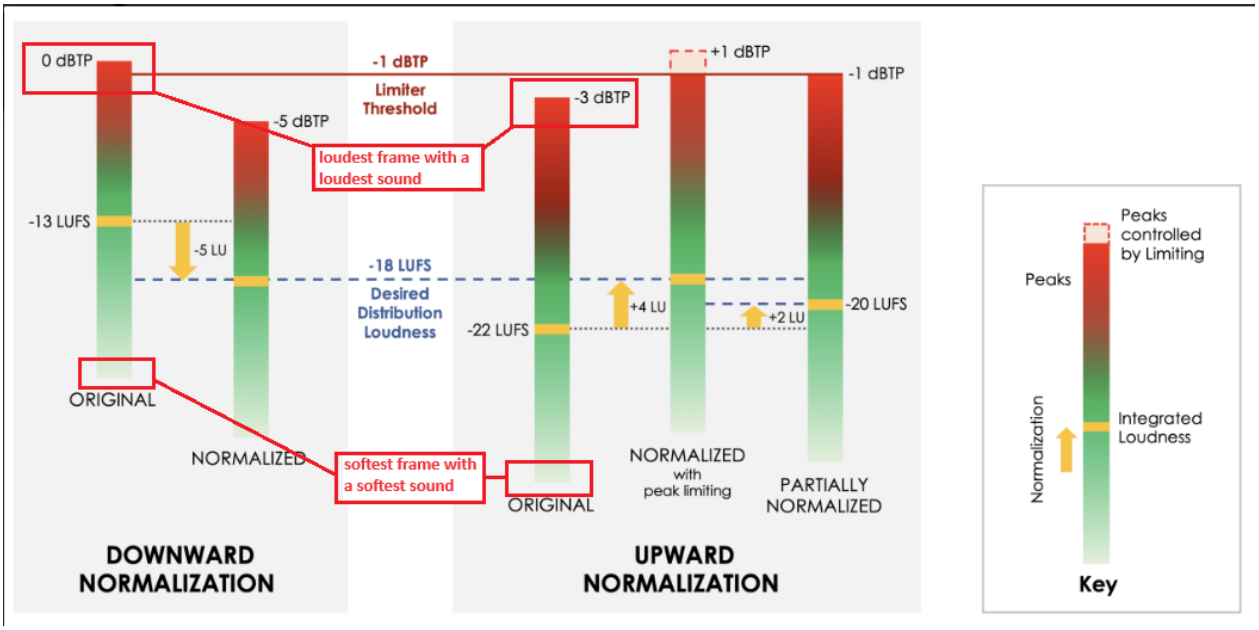
- The full length of the audio content is measured for its Integrated Loudness, in LUFS. See the Loudness Basics section for measurement
- The amplitude of the entire audio content is then corrected so that the Integrated Loudness matches the target loudness. For example, if the target loudness is -24 LUFS (which is often used for audio content creation), and the content measures -27 LUFS, a gain offset of +3 LU to the content produces the target loudness.

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

When applying normalization, the loudness of the original audio (shown on the left in the figure) might be above or below the desired Distribution Loudness. In this case its Integrated Loudness is -13 LUFS, which is greater than the "target" of -18 LUFS. (See *AES TD1008* for specific recommendations on target loudness.) Normalization attenuates the audio by 5 LU to the target value. This process is commonly called Downward Normalization.



https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/



https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

MTIA has been deployed in our data centers and is now serving models in production. We are already seeing the positive results of this program as it's allowing us to dedicate and invest in more compute power for our more intensive AI workloads.

The results so far show that this MTIA chip can handle both low complexity and high complexity ranking and recommendation models which are key components of Meta's products. Because we control the whole stack, we can achieve greater efficiency compared to commercially available GPUs (graphics processing units).

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

Facebook's services rely on fleets of servers in data centers all over the globe — all running applications and delivering the performance our services need. This is why we need to make sure our server hardware is reliable and that we can manage server hardware failures at our scale with as little disruption to our services as possible.

https://engineering.fb.com/2020/12/09/data-center-engineering/how-facebook-keeps-its-large-scale-infrastructure-hardware-up-and-running/

Meta is the latest major **hyperscale cloud company** that has adopted AMD EPYC CPUs to power its data centers. Both companies worked together to define an open, cloud-scale, single-socket server designed for performance and power efficiency, based on the 3rd Gen EPYC processor.

https://techhq.com/2021/11/amd-strikes-chip-deal-to-power-metas-data-centers/

"[g] assigning a normalized minimum amplitude value to the softest frame of the audio stream and a normalized maximum amplitude value to the loudest frame of the audio stream"

36. Upon information and belief, and as explained above and below, the Accused Instrumentality assigns a normalized minimum amplitude value to the softest frame of the audio stream (e.g., softest frame in the audio + difference between the integrated loudness value and the

target loudness value, etc.) and a normalized maximum amplitude value to the loudest frame of the audio stream (e.g., peak limit threshold, loudest frame in the audio + difference between the integrated loudness value and the target loudness value, etc.).

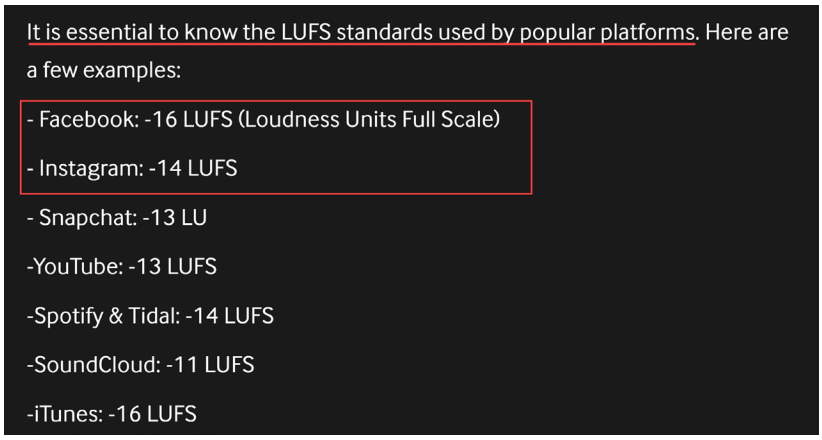
37. For example, as shown below, Facebook and Instagram running on Meta datacenters normalize the loudness levels (aural amplitude numerical value) based on measurements consistent with ITU-R BS.1770. It calculates loudness levels of the video and normalizes the LUFS to a standard LUFS value. It calculates an integrated loudness value of the audio stream and determines its difference with a target loudness value. For example, this difference is used to normalize all the audio frames including the loudest and softest frames.

Profiles

The following profiles are defined for identifying the constraints used in recommendations for different platforms and standards.

Loudness profile	Maximum loudness (LKFS)	Minimum loudness (LU)	Maximum True Peak (dBTP)
standard_a85	-22	-26	-2
standard_r128	-22.5	-23.5	-1
service_amazon	-13	-15	-1
service_apple	-15	-17	-1
service_facebook	-15	-17	-1

https://docs.dolby.io/media-apis/docs/loudness



https://starsoundstudios.com/blog/lufs-social-media-platform-standards-mastering-music

xHE-AAC's integrated loudness management system solves for loudness inconsistency while meticulously preserving creator intent by bringing the average loudness of all sessions to the same target level and managing the dynamic range of each session to fit the playback environment.

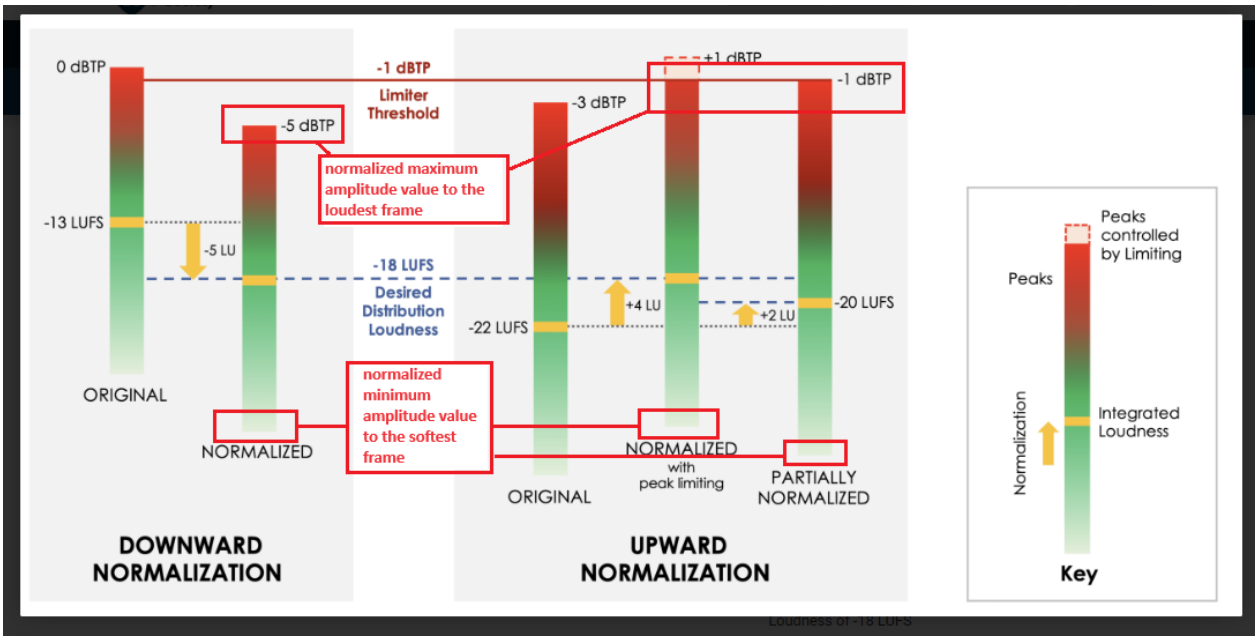
https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

Normalization Technique

Normalization is the active level matching of file-based audio content, such as a record album or radio program, to a defined "target loudness." Normalizing avoids listener annoyance by matching loudness from one content asset to the next but it does not affect the quality of the content. At its most basic, file-based audio content is normalized by these steps:

- The full length of the audio content is measured for its Integrated Loudness, in LUFS. See the Loudness Basics section for measurement
- The amplitude of the entire audio content is then corrected so that the Integrated Loudness matches the target loudness. For example, if the target loudness is -24 LUFS (which is often used for audio content creation), and the content measures -27 LUFS, a gain offset of +3 LU to the content produces the target loudness.

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/



https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

"[h] comparing each frame of the audio stream to the loudest frame and the softest frame by utilizing the audio histogram and assigning a normalized amplitude value between the normalized minimum amplitude value and the normalized maximum amplitude value to said each frame in accordance with a result of the comparison"

38. The Accused Instrumentality compares each frame of the audio stream (e.g., audio from Facebook live, an uploaded video, Instagram Live, Reels, etc.) to the loudest frame and the softest frame by utilizing the audio histogram and assigning a normalized amplitude value (e.g., a normalized LUFS value) between the normalized minimum amplitude value (e.g., normalized lowest loudness, etc.) and the normalized maximum amplitude value (e.g., normalized highest loudness, etc.) to said each frame in accordance with a result of the comparison (e.g., matching integrated loudness to target loudness value, etc.).

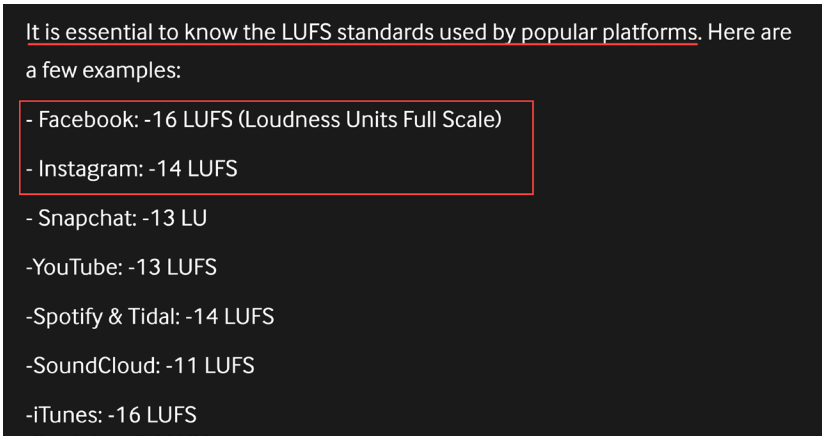
39. For example, as shown below, the amplitude of the entire audio file is normalized to match the integrated loudness value to the target loudness value. It calculates an integrated loudness value of the audio stream and determines its difference with the target value. For example, this difference is used to normalize all the audio frames including the loudest and softest frames.

Profiles

The following profiles are defined for identifying the constraints used in recommendations for different platforms and standards.

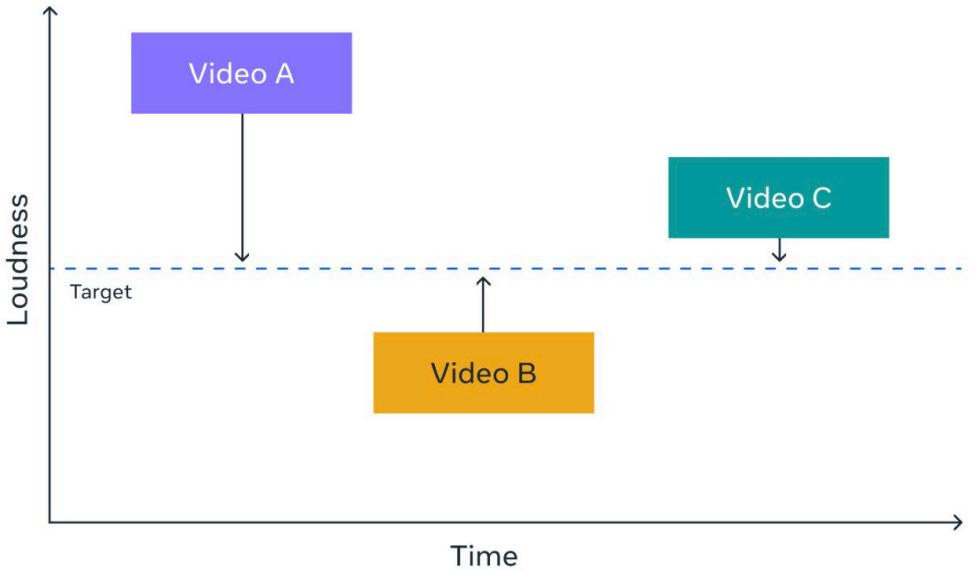
Loudness profile	Maximum loudness (LKFS)	Minimum loudness (LU)	Maximum True Peak (dBTP)
standard_a85	-22	-26	-2
standard_r128	-22.5	-23.5	-1
service_amazon	-13	-15	-1
service_apple	-15	-17	-1
service_facebook	-15	-17	-1

https://docs.dolby.io/media-apis/docs/loudness



https://starsoundstudios.com/blog/lufs-social-media-platform-standards-mastering-music

When people play these videos sequentially, they can perceive some audio as being too loud or too quiet. This creates listener fatigue from having to constantly adjust the volume.



xHE-AAC's integrated loudness management system solves for loudness inconsistency while meticulously preserving creator intent by bringing the average loudness of all sessions to the same target level and managing the dynamic range of each session to fit the playback environment.

Instead of burning in a specific target level and dynamic range compression (DRC) profile during encoding, xHE-AAC allows us to leave the original audio characteristics untouched and delegate loudness management processing to the client via loudness metadata, for the optimal audio experience based on context.

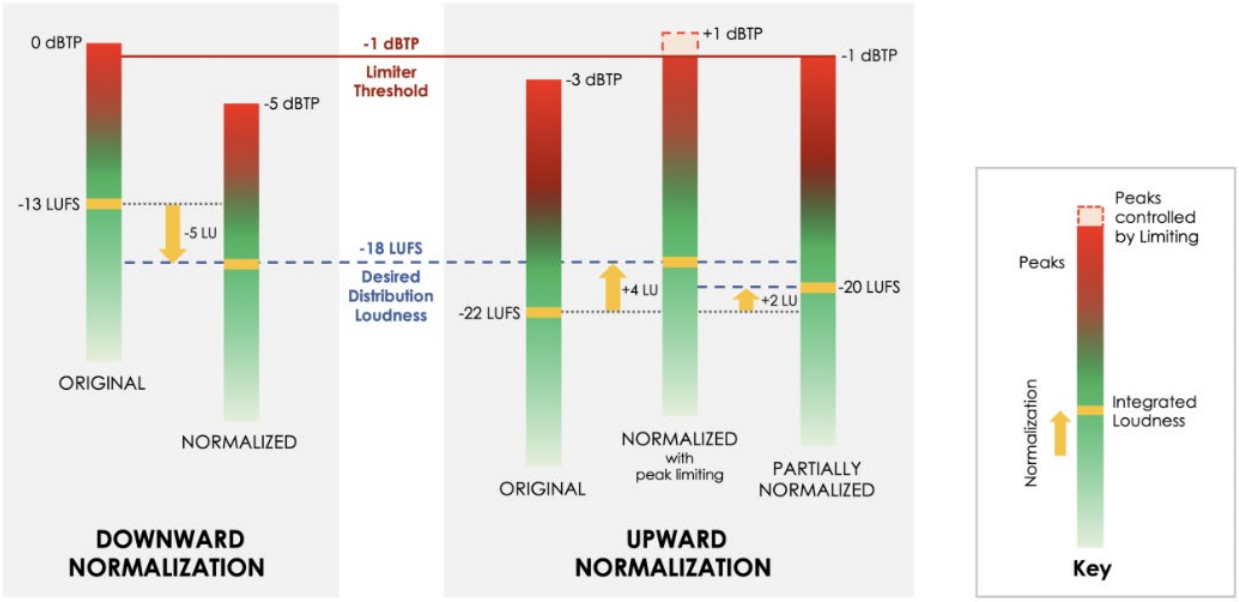
https://engineering.fb.com/2023/04/11/video-engineering/high-quality-audio-xhe-aac-codec-meta/

Normalization Technique

Normalization is the active level matching of file-based audio content, such as a record album or radio program, to a defined "target loudness." Normalizing avoids listener annoyance by matching loudness from one content asset to the next but it does not affect the quality of the content. At its most basic, file-based audio content is normalized by these steps:

- The full length of the audio content is measured for its Integrated Loudness, in LUFS. See the Loudness Basics section for measurement
- The amplitude of the entire audio content is then corrected so that the Integrated Loudness matches the target loudness. For example, if the target loudness is -24 LUFS (which is often used for audio content creation), and the content measures -27 LUFS, a gain offset of +3 LU to the content produces the target loudness.

https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/



https://aes2.org/resources/audio-topics/loudness-project/loudness-normalization/

"[i] processing the time-aligned machine transcribed media by the server processor to extract time-aligned textual metadata associated with the source media; and"

40. The Accused Instrumentality processes the time-aligned machine transcribed media (e.g., automatically transcribed audio and the auto-generated captions, etc.) by the Meta server processor to extract time-aligned textual metadata (e.g., label certain sensitive contents within the auto-generated captions, etc.) associated with the source media.

41. For example, as shown below, the Accused Instrumentality uses a convolutional neural network architecture to provide context to the transcribed media. The timing information is maintained to sync the audio with the auto-generated captions. In addition, it uses AI to detect and extract time-aligned sensitive contents within the captions so those contents can be replaced in the captions and the corresponding audio. To achieve this, captions have to be analyzed and the sensitive contents have to be labelled (e.g., metadata) and time aligned with audio so they can be replaced in both the captions and the corresponding audio.

For wav2vec, we created an architecture consisting of two multilayer convolutional neural networks stacked on top of each other. The encoder network maps raw audio input to a representation, where each vector covers about 30 milliseconds (ms) of speech. The context network uses those vectors to generate its own representations, which cover a larger span of up to a second.

The model then uses these representations to solve a self-supervised prediction task. Within each 10-second audio clip that the model is trained on, wav2vec generates a number of distractor examples, which swap out 10 ms of the original audio with sections from elsewhere in the clip. The model must then determine which version is correct. And this selection process is repeated multiple times for each 10-second training clip, essentially quizzing the model to discern accurate speech sounds from distractor samples hundreds of times per second.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

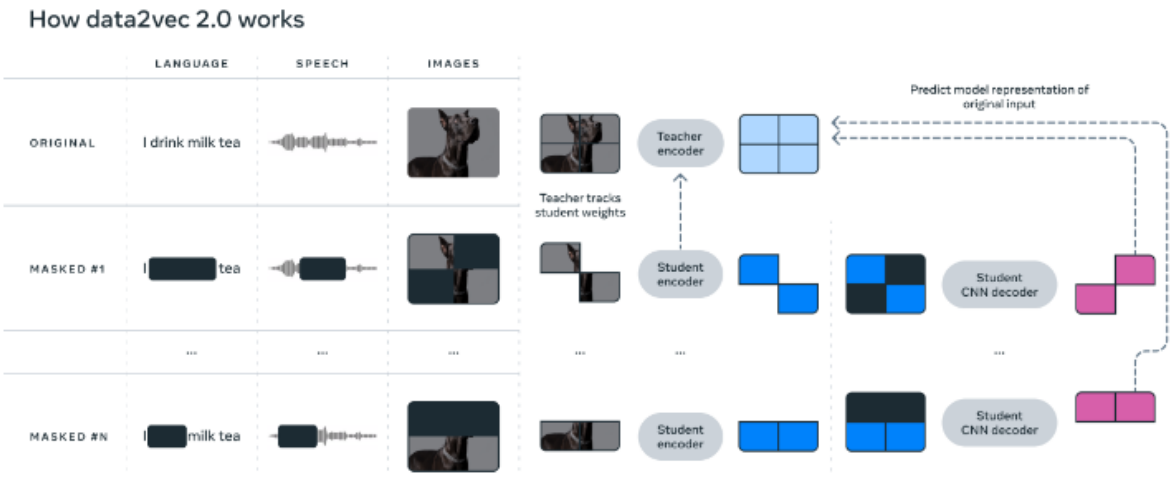
How it works:

We built Libri-light using more than 60,000 hours of unlabeled speech in English from LibriVox, a large repository of public domain audiobooks. In addition to filtering corrupted and duplicated data and adding speech activity, speaker, and genre metadata to make Libri-light useful in the context of ASR training, we built baselines systems and evaluation metrics on top of the popular LibriSpeech ASR benchmark.

https://ai.meta.com/blog/a-new-open-benchmark-for-speech-recognition-with-limited-or-no-supervision/

How data2vec 2.0 works

The general idea of self-supervised learning is for machines to learn the structure of images, speech, and text simply by observing the world. Advances in this area have led to many breakthroughs in speech (e.g., wav2vec 2.0), computer vision (e.g., masked autoencoders), and natural language processing (e.g., BERT). But modern systems can be computationally demanding, as training very large models requires many GPUs.



https://ai.meta.com/blog/ai-self-supervised-learning-data2vec/

- We're releasing our code for wav2vec, an algorithm that uses raw, unlabeled audio to train automatic speech recognition (ASR) models.
- This self-supervised approach beats traditional ASR systems that rely solely on transcribed audio, including a 22 percent accuracy improvement over Deep Speech 2, while using two orders of magnitude less labeled data.
- Wav2vec trains models to learn the difference between original speech examples and modified versions, often repeating this task hundreds of times for each second of audio, and predicting the correct audio milliseconds into the future.
- Reducing the need for manually annotated data is important for developing systems that understand non-English languages, particularly those with limited existing training sets of transcribed speech. Wav2vec is also part of our ongoing commitment to self-supervised training, which could accelerate the development of AI systems across the field.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

Although self-supervision has shown promise in natural language processing (NLP) tasks — including RoBERTa, Facebook AI's optimized pretraining method that recently topped the leaderboard for a major NLP benchmark — wav2vec applies the approach specifically to speech. Our algorithm does not require transcriptions, and our model learns from unlabeled audio data.

Most current ASR models train on the log-mel filter bank features of speech data, meaning audio that's been processed to make vocal features stand out. Our approach instead turns raw speech examples into a representation — specifically, a code — that can be fed into an existing ASR system. Using wav2vec's representations as inputs enables the algorithm to work with a wide variety of existing speech recognition models, making unlabeled audio data more widely useful for speech-related AI research.

One of the primary challenges in building wav2vec was dealing with the continuous nature of speech data, which makes it difficult to directly predict the data. We addressed this issue by using a pretraining regime inspired in part by the popular NLP algorithm word2vec. This algorithm learns representations by training a model to distinguish between the true data and a set of distractor samples.

https://ai.meta.com/blog/wav2vec-state-of-the-art-speech-recognition-through-self-supervision/

We preprocessed the data to improve quality and to make it usable by our machine learning algorithms. To do so, we trained an alignment model on existing data in over 100 languages and used this model together with an efficient forced alignment algorithm that can process very long recordings of about 20 minutes or more. We applied multiple rounds of this process and performed a final cross-validation filtering step based on model accuracy to remove potentially misaligned data. To enable other researchers to create new speech datasets, we added the alignment algorithm to PyTorch and released the alignment model.

Thirty-two hours of data per language is not enough to train conventional supervised speech recognition models. This is why we built on wav2vec 2.0, our prior work on self-supervised speech representation learning, which greatly reduced the amount of labeled data needed to train good systems. Concretely, we trained self-supervised models on about 500,000 hours of speech data in over 1,400 languages — this is nearly five times more languages than any known prior work. The resulting models were then fine-tuned for a specific speech task, such as multilingual speech recognition or language identification.

https://ai.meta.com/blog/multilingual-model-speech-recognition/

What are Instagram Auto-Generated Captions?

Instagram's auto-generated captions are the text transcriptions of any speech found in an Instagram video or Reel. They appear on the screen in coordination with the speech in the video, meaning that they synchronize with the visual content for context. As the name suggests, these captions are automatically generated by the app without you having to manually input them.

https://influencermarketinghub.com/instagram-auto-generated-captions/

Instagram has been criticised by deaf people online after it was revealed that its new automatic captions feature for Stories censors profanities spoken by users.

Videos on social media have shown the tool, announced on Tuesday, replacing the audio with a bleep effect, while the caption itself sees the curse word switched for '$@#%&'.

Commenting on the revelation on Thursday, Charlotte Hyde, a Deaf accessibility advocate, tweeted: "Just learned via [YouTuber] Daniel J Layton's Instagram Story that the new Captions sticker not only censors swear words on the captions, but also puts a literal beep sound over you saying them.

https://limpingchicken.com/2021/05/07/deaf-news-frustration-as-instagram-stories-automatic-captions-feature-censors-swear-words/

"We understand the impact that offensive, derogatory language can have on people and think nobody should have that experience on Instagram," an Instagram spokesperson told Mashable. "That's why we use AI to detect this sort of language and actively block from captions. We do this across a range of features on Instagram like comment controls and hidden words, which automatically hide a list of offensive words."

https://mashable.com/article/instagram-stories-automatic-captions

"[j] storing the time-aligned machine transcribed media, the time-aligned audio frames, the time-aligned aural amplitudes, time-aligned textual metadata and the normalized amplitude value of each frame of the audio stream in a database."

42. The Accused Instrumentality stores the time-aligned machine transcribed media (e.g., automatic transcribed audio, etc.), the time-aligned audio frames (e.g., audio frames, etc.), the time-aligned aural amplitudes (e.g., time-aligned LUFS values for the audio, etc.), time-aligned textual metadata (e.g., labeled sensitive contents within the auto-generated captions) and the normalized amplitude value (e.g., normalized LUFS values, etc.) of each frame of the audio stream in a Meta database.

43. As shown below, Meta datacenters store all data and metadata related to contents uploaded over its servers. For processing and transmission, the Accused Instrumentality stores data related to transcribed audio (time-aligned machine transcribed media), audio frames of an uploaded video (the time-aligned audio frames), the original LUFS values of an audio (time-aligned aural amplitudes), time-aligned labelled sensitive contents (the time-aligned textual metadata associated with non-transcribed source media), and normalized LUFS values (normalized amplitude value of each frame of the audio stream), etc.

On our Products, you can send messages, take photos and videos, buy or sell things and much more. We call all of the things you can do on our Products "activity." We collect your activity across our Products and information you provide, such as:

- Content you create, like posts, comments or audio
- Content you provide through our camera feature or your camera roll settings, or through our voice-enabled features. Learn more about what we collect from these features, and how we use information from the camera for masks, filters, avatars and effects.
- Messages you send and receive, including their content, subject to applicable law. We can't see the content of end-to-end encrypted messages unless users report them to us for review. Learn more.
- Metadata about content and messages, subject to applicable law
- Types of content, including ads, you view or interact with, and how you interact with it
- Apps and features you use, and what actions you take in them. See examples.

https://www.facebook.com/privacy/policy?subpage=1.subpage.1-YourActivityAndInformation

MTIA has been deployed in our data centers and is now serving models in production. We are already seeing the positive results of this program as it's allowing us to dedicate and invest in more compute power for our more intensive AI workloads.

The results so far show that this MTIA chip can handle both low complexity and high complexity ranking and recommendation models which are key components of Meta's products. Because we control the whole stack, we can achieve greater efficiency compared to commercially available GPUs (graphics processing units).

https://about.fb.com/news/2023/05/metas-infrastructure-for-ai/

Facebook's services rely on fleets of servers in data centers all over the globe — all running applications and delivering the performance our services need. This is why we need to make sure our server hardware is reliable and that we can manage server hardware failures at our scale with as little disruption to our services as possible.

https://engineering.fb.com/2020/12/09/data-center-engineering/how-facebook-keeps-its-large-scale-infrastructure-hardware-up-and-running/

Meta is the latest major **hyperscale cloud company** that has adopted AMD EPYC CPUs to power its data centers. Both companies worked together to define an open, cloud-scale, single-socket server designed for performance and power efficiency, based on the 3rd Gen EPYC processor.

https://techhq.com/2021/11/amd-strikes-chip-deal-to-power-metas-data-centers/

44. As demonstrated, Meta via its datacenters has directly infringed and continues to directly infringe one or more claims of the '547 Patent, including at least representative claim 1, in violation of 35 U.S.C. §§ 271(a) by, without using the Accused Instrumentalities within the United States.

45. Meta's infringing activities at Meta datacenters have and continue to be without authority or license under the '547 Patent.

46. Datascription has and continues to suffer damages as a direct and proximate result of Meta's direct infringement of the '547 Patent at Meta datacenters.

PRAYER FOR RELIEF

Datascription respectfully requests that the Court enter judgment in its favor and grant the following relief:

A. A judgment that Meta infringes one or more claims of the '547 Patent;

B. Damages under 35 U.S.C. § 284 adequate to compensate Datascription for Meta's infringement, but in no event less than a reasonable royalty;

C. Enhancements for willful infringement as permitted by law;

D. Pre- and post-judgment interest;

E. A permanent injunction or, in the alternative, an ongoing royalty;

F. Attorneys' fees and costs;

G. Any other relief the Court deems just and proper.

DEMAND FOR JURY TRIAL

47. Pursuant to Federal Rule of Civil Procedure 38(b), Datascription demands a trial by jury of all issues so triable.

Dated May 4, 2026

Respectfully submitted,

/s/ Timothy F. Dewberry

Timothy Dewberry TX Bar. No. 24090074
Joseph M. Abraham, TX Bar No. 24088879
FOLIO LAW GROUP PLLC
13492 Research Blvd., Suite 120, No. 177
Austin, TX 78750
Tel: (737) 234-0201
Email:timothy.dewberry@foliolaw.com
joseph.abraham@foliolaw.com

Alexandra Fellowes, CA Bar No. 261929
Cliff Win, Jr., CA Bar No. 270517
FOLIO LAW GROUP PLLC
1200 Westlake Ave. N., Suite 809
Seattle, WA 98109
Tel: (206) 880-1802
Email: alexandra.fellowes@foliolaw.com
cliff.win@foliolaw.com

Attorneys for Plaintiff Datascription, LLC