vCon
Internet-Draft
Intended status: Standards Track
Expires: 5 November 2026

T. McCarthy-Howe
VCONIC

4 May 2026

vCon World Transcription Format Extension
draft-howe-vcon-wtf-extension-02

Abstract

This document defines the World Transcription Format (WTF) extension for Virtualized Conversations (vCon). The WTF extension provides a standardized analysis framework for representing speech-to-text transcription data from multiple providers within vCon containers. This extension defines a comprehensive analysis format that enables consistent transcription processing, quality assessment, and interoperability across different transcription services while preserving provider-specific features through extensible fields.

The WTF extension is designed as a Compatible Extension that introduces a new analysis type for transcription data without altering existing vCon semantics, ensuring backward compatibility with existing vCon implementations.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://vcon-dev.github.io/draft-howe-vcon-wtf-extension/draft-howe-vcon-wtf-extension-02.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-howe-vcon-wtf-extension/.

Discussion of this document takes place on the vCon Working Group mailing list (mailto:vcon@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/vcon/. Subscribe at https://www.ietf.org/mailman/listinfo/vcon/.

Source for this draft and an issue tracker can be found at https://github.com/vcon-dev/draft-howe-vcon-wtf-extension.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on 5 November 2026.

Copyright Notice

Table of Contents

1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC2119 [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Introduction

Virtualized Conversations (vCon) [I-D.draft-ietf-vcon-vcon-core] provide a standardized container format for conversation data, enabling interoperability across different communication platforms and modalities. An overview of the vCon ecosystem and its use cases is provided in [I-D.draft-ietf-vcon-overview]. As speech-to-text technology becomes increasingly important for conversation analysis, compliance, and accessibility, there is a growing need for standardized transcription analysis representation within vCon containers.

Current transcription services each use proprietary formats, making it difficult to:

* Switch between transcription providers without data conversion
* Perform comparative analysis across different providers
* Maintain consistent data processing pipelines
* Preserve transcription metadata during system migrations
* Ensure long-term accessibility and archival compliance

This document defines the World Transcription Format (WTF) extension for vCon, which addresses these challenges by providing a comprehensive analysis framework that includes:

* A unified analysis schema for transcription data from any speech-to-text provider
* Hierarchical organization from words to segments to complete transcripts
* Extensible fields for provider-specific analytical features
* Built-in quality metrics and confidence scoring for analysis assessment
* Support for real-time and batch transcription analysis workflows
* Consistent export capabilities to standard subtitle and caption formats

The WTF extension defines a new category of vCon analysis specifically for speech-to-text transcription. This analysis type enables organizations to standardize their transcription analysis workflows while maintaining the flexibility to use multiple providers and preserve provider-specific analytical enhancements.

3. Conventions and Definitions

3.1. Core Terms

World Transcription Format (WTF): A standardized analysis framework and JSON schema for representing speech-to-text transcription data from any provider in a consistent, interoperable format. WTF defines a comprehensive analysis type for vCon that includes transcription content, quality metrics, confidence scoring, and provider-specific analytical features.

Transcription Analysis: The structured representation of speech-to-text conversion results, including the transcribed text, timing information, confidence metrics, speaker identification, and quality assessments.

Transcription Provider: A service or system that converts audio or video content to text, such as Whisper(TM), Deepgram(TM), AssemblyAI(TM), Google Cloud Speech-to-Text(TM), Amazon Transcribe(TM), or Azure Speech Services(TM).

Segment: A logical chunk of transcribed content, typically representing sentence or phrase boundaries with associated timing information and analytical metadata.

Speaker Diarization: The process of partitioning an audio stream into homogeneous segments according to the speaker identity, enabling "who spoke when" analysis.

Compatible Extension: A vCon extension that introduces additional data without altering the meaning or structure of existing elements, as defined in [I-D.draft-ietf-vcon-vcon-core].

4. World Transcription Format Overview

The World Transcription Format defines a comprehensive analysis framework using a hierarchical JSON structure designed to capture transcription analysis results at multiple levels of granularity while maintaining consistency across different providers.

4.1. Design Principles

Completeness: WTF captures all essential transcription analysis data including text, timing, confidence scores, speaker information, and quality assessments.

Consistency: Provides uniform analysis structure regardless of the underlying transcription provider.

Extensibility: Supports provider-specific analytical features through structured extension fields.

Validation: Includes built-in data integrity checking and quality metrics for analysis assessment.

Hierarchical Organization: Natural progression from words to segments to complete transcripts enables multi-level analysis.

4.2. Core Structure

WTF uses a hierarchical JSON structure with three required sections and multiple optional enrichment layers:

```
{
  "transcript": { /* Required: Core transcript information */ },
  "segments": [ /* Required: Time-aligned text segments */ ],
  "metadata": { /* Required: Processing metadata */ },
  "words": [ /* Optional: Word-level details */ ],
  "speakers": { /* Optional: Speaker diarization */ },
  "alternatives": [ /* Optional: Alternative transcriptions */ },
  "enrichments": { /* Optional: Analysis features */ },
  "extensions": { /* Optional: Provider-specific data */ },
  "quality": { /* Optional: Quality metrics */ },
  "streaming": { /* Optional: Streaming information */ }
}
```

5. vCon WTF Extension Definition

5.1. Extension Classification

The WTF extension is a *Compatible Extension* as defined in Section 2.5 of [I-D.draft-ietf-vcon-vcon-core]. This extension:

* Introduces a new analysis type for transcription data without altering existing vCon semantics
* Defines a comprehensive analysis framework for speech-to-text results
* Can be safely ignored by implementations that don’t support transcription analysis
* Does not require listing in the critical parameter
* Maintains backward compatibility with existing vCon implementations

5.2. Extension Registration

This document defines the "wtf_transcription" extension token for registration in the vCon Extensions Names Registry:

* *Extension Name*: wtf_transcription

* *Extension Description*: World Transcription Format - A standardized analysis framework for speech-to-text transcription with multi-provider support, quality metrics, and confidence scoring

* *Change Controller*: IESG

* *Specification Document*: This document

5.3. Extension Usage

vCon instances that include WTF transcription analysis SHOULD include "wtf_transcription" in the extensions array to indicate support for the WTF analysis framework:

```
{
  "uuid": "01234567-89ab-cdef-0123-456789abcdef",
  "extensions": ["wtf_transcription"],
  "created_at": "2025-01-02T12:00:00Z",
  "parties": [...],
  "dialog": [...],
  "analysis": [
    {
      "type": "wtf_transcription",
      "start": "2025-01-02T12:15:30Z",
      "party": 0,
      "dialog": 0,
      "encoding": "json",
      "body": {
        // WTF transcription analysis data structure defined below
      }
    }
  ]
}
```

6. WTF Analysis Structure

6.1. Analysis Storage

Transcription analysis results MUST be stored as vCon analysis using the standard analysis object structure defined in Section 4.5 of [I-D.draft-ietf-vcon-vcon-core]. The analysis mechanism provides the association between the analysis results and the corresponding dialog elements.

The WTF transcription analysis object MUST include:

* *type*: MUST be set to "wtf_transcription" to identify this as a WTF analysis
* *encoding*: MUST be set to "json" for structured analysis data
* *body*: MUST contain the WTF transcription analysis data structure as defined below

The WTF transcription analysis SHOULD include:

* *start*: ISO 8601 timestamp when the transcription analysis was performed
* *party*: Index of the party in the vCon parties array (for single-speaker transcription analysis)
* *dialog*: Index of the associated dialog in the vCon dialog array

6.2. WTF Analysis Schema

The body field of the WTF transcription analysis object MUST contain a JSON object conforming to the WTF analysis schema with the following structure:

6.2.1. Required Fields

6.2.1.1. Transcript Object

```
"transcript": {
  "text": "string",       // Complete transcription text
  "language": "string",   // BCP-47 language code (e.g., "en-US")
  "duration": "number",   // Total audio duration in seconds
  "confidence": "number"  // Overall confidence score \[0.0-1.0\]
}
```

The transcript object provides the high-level summary of the entire transcription:

* *text*: The complete, concatenated transcription text

* *language*: MUST use BCP-47 format [BCP47] (examples: "en-US", "es-MX", "fr-CA")

* *duration*: Floating-point seconds, MUST be >= 0

* *confidence*: Normalized to [0, 1] range regardless of provider scale

6.2.1.2. Segments Array

```
"segments": [
  {
    "id": "integer",              // Sequential segment identifier
    "start": "number",            // Start time in seconds
    "end": "number",              // End time in seconds
    "text": "string",             // Segment text content
    "confidence": "number",       // Segment-level confidence \[0.0-1.0\]
    "speaker": "integer|string",  // Optional: Speaker identifier
    "words": ["integer"]          // Optional: Array of word indices
  }
]
```

Segments represent logical chunks of transcribed content, typically sentence or phrase boundaries:

* *id*: MUST be unique within the document, typically sequential

* *start*/*end*: Floating-point seconds, where end > start

* *text*: SHOULD be trimmed of leading/trailing whitespace

* *speaker*: Can be integer (0, 1, 2) or string ("Speaker A")

* *words*: References indices in the words array

6.2.1.3. Metadata Object

```
"metadata": {
  "created_at": "string",        // ISO 8601 timestamp
  "processed_at": "string",      // ISO 8601 timestamp
  "provider": "string",          // Provider name (lowercase)
  "model": "string",             // Model/version identifier
  "processing_time": "number",   // Optional: Processing duration in seconds
  "audio": {
    "duration": "number",        // Source audio duration in seconds
    "sample_rate": "integer",    // Optional: Sample rate in Hz
    "channels": "integer",       // Optional: Number of channels
    "format": "string",          // Optional: Audio format
    "bitrate": "integer"         // Optional: Bitrate in kbps
  },
  "options": "object"            // Provider-specific options used
}
```

The metadata object captures processing and source information:

* *created_at*/*processed_at*: MUST use ISO 8601 format

* *provider*: Lowercase identifier for supported providers

* *model*: Provider's model identifier (e.g., "whisper-large-v3", "nova-2")

* *options*: Preserves provider-specific configuration

6.2.2. Optional Fields

6.2.2.1. Words Array

```
  "words": [
    {
      "id": "integer",              // Sequential word identifier
      "start": "number",            // Word start time in seconds
      "end": "number",              // Word end time in seconds
      "text": "string",             // Word text
      "confidence": "number",       // Word-level confidence \[0.0-1.0\]
      "speaker": "integer|string",  // Optional: Speaker identifier
      "is_punctuation": "boolean"   // Optional: Punctuation marker
    }
  ]
```

6.2.2.2. Speakers Object

```
"speakers": {
  "speaker_id": {
    "id": "integer|string",       // Speaker identifier
    "label": "string",            // Human-readable speaker name
    "segments": ["integer"],      // Array of segment IDs for this speaker
    "total_time": "number",       // Total speaking time in seconds
    "confidence": "number"        // Diarization confidence \[0.0-1.0\]
  }
}
```

6.2.2.3. Quality Object

```
   "quality": {
     "audio_quality": "string",      // high, medium, low
     "background_noise": "number",   // Noise level \[0.0-1.0\]
     "multiple_speakers": "boolean",
     "overlapping_speech": "boolean",
     "silence_ratio": "number",      // Percentage of silence
     "average_confidence": "number",
     "low_confidence_words": "integer",
     "processing_warnings": ["string"]
   }
```

6.2.2.4. Extensions Object

```
   "extensions": {
     "provider_name": {
       // Provider-specific fields preserved during conversion
     }
   }
```

6.2.2.5. Alternatives Array

The alternatives array captures multiple transcription hypotheses when a provider returns ranked results.

```
  "alternatives": [
    {
      "rank": "integer",       // 1-based rank (1 = highest confidence)
      "confidence": "number",  // Alternative-level confidence [0.0-1.0]
      "transcript": {
        "text": "string",
        "confidence": "number"
      },
      "segments": []           // Optional: segment-level alternatives
    }
  ]
```

6.2.2.6. Enrichments Object

The enrichments object contains NLP analysis results layered on top of the base transcription.

```
"enrichments": {
  "sentiment": {
    "overall": "string",     // positive, negative, neutral
    "score": "number"        // [-1.0, 1.0]
  },
  "entities": [
    {
      "text": "string",      // Entity text as it appears
      "type": "string",      // PERSON, ORG, LOCATION, DATE, etc.
      "start": "number",     // Start time in seconds
      "end": "number"        // End time in seconds
    }
  ],
  "topics": ["string"],      // Detected topic labels
  "summary": "string"        // Optional: abstractive summary
}
```

6.2.2.7. Streaming Object

The streaming object captures metadata specific to real-time transcription sessions.

```
"streaming": {
  "session_id": "string",    // Streaming session identifier
  "is_final": "boolean",     // Whether this is a final (not interim) result
  "stability": "number",     // Interim result stability [0.0-1.0]
  "latency_ms": "integer"    // Processing latency in milliseconds
}
```

7. Provider Integration Guidelines

7.1. Supported Providers

The WTF extension supports integration with major transcription providers:

* *Whisper(TM)*: OpenAI’s open-source speech recognition system

* *Deepgram(TM)*: Real-time speech-to-text API

* *AssemblyAI(TM)*: AI-powered transcription and audio intelligence

* *Google Cloud Speech-to-Text(TM)*: Google's speech recognition service
* *Amazon Transcribe(TM)*: AWS speech-to-text service
* *Azure Speech Services(TM)*: Microsoft's speech recognition platform
* *Rev.ai(TM)*: Automated and human transcription services
* *Speechmatics(TM)*: Real-time and batch speech recognition
* *Wav2Vec2(TM)*: Facebook's self-supervised speech recognition model
* *Parakeet(TM)*: NVIDIA's speech recognition toolkit

7.2. Conversion Requirements

When converting from provider-specific formats to WTF:

1. *Normalize confidence scores* to [0.0, 1.0] range
2. *Convert timestamps* to floating-point seconds
3. *Standardize language codes* to BCP-47 format
4. *Preserve provider-specific features* in extensions field
5. *Validate output* against WTF schema requirements

7.3. Provider-Specific Mappings

7.3.1. Whisper Integration

```
{
  "extensions": {
    "whisper": {
      "tokens": ["array of token IDs"],
      "temperature": "number",
      "compression_ratio": "number",
      "avg_logprob": "number",
      "no_speech_prob": "number"
    }
  }
}
```

7.3.2. Deepgram Integration

```
{
  "extensions": {
    "deepgram": {
      "utterances": ["array of utterance objects"],
      "paragraphs": ["array of paragraph objects"],
      "search_terms": ["array of detected search terms"]
    }
  }
}
```

8. Quality and Confidence Metrics

8.1. Confidence Score Normalization

All confidence scores MUST be normalized to the [0.0, 1.0] range:

* *1.0*: Highest confidence (perfect recognition)
* *0.9-1.0*: High confidence
* *0.7-0.9*: Medium confidence
* *0.5-0.7*: Low confidence
* *0.0-0.5*: Very low confidence

Provider-specific scales are converted during import:

* Percentage (0-100) -> divide by 100
* Log probability -> exponential transformation
* Custom scales -> linear normalization

8.2. Quality Metrics

The quality object provides assessment metrics:

* *audio_quality*: Categorical assessment (high/medium/low)
* *background_noise*: Noise level [0.0-1.0]
* *multiple_speakers*: Boolean indicator of multi-speaker content
* *overlapping_speech*: Boolean indicator of speaker overlap

* *silence_ratio*: Percentage of audio that is silence
* *average_confidence*: Mean confidence across all words/segments
* *low_confidence_words*: Count of words below 0.5 confidence
* *processing_warnings*: Array of processing issues or notices

9. Security Considerations

9.1. Data Privacy

Transcription data often contains sensitive personal information. Developers implementing WTF SHOULD consult [I-D.draft-ietf-vcon-privacy-primer] for comprehensive guidance on privacy considerations when handling vCon data. Implementations SHOULD:

* Apply appropriate access controls to WTF analysis objects
* Consider encryption requirements for transcription data at rest and in transit
* Implement data retention policies consistent with privacy regulations
* Provide mechanisms for transcription data redaction or anonymization

9.2. Provider-Specific Security

When integrating with external transcription providers:

* Validate provider credentials and API security
* Implement secure communication channels (TLS 1.2 or higher)
* Consider data residency requirements for audio processing
* Audit provider data handling practices and compliance certifications

9.3. Integrity Protection

WTF transcription analysis objects SHOULD be integrity protected using vCon signing mechanisms as defined in [I-D.draft-ietf-vcon-vcon-core] to prevent unauthorized modification of transcription data.

9.4. Temporal Validation

Implementations SHOULD validate that transcription timestamps are consistent with the associated dialog timing information to detect potential tampering or synchronization issues.

10. WTF as an Analysis Framework

The World Transcription Format is fundamentally an *analysis framework* that defines how speech-to-text transcription results should be structured, validated, and stored within vCon containers. This section clarifies the distinction between WTF as an analysis type and its storage mechanism.

10.1. Analysis vs. Storage

WTF defines an analysis type: The core contribution of this specification is the definition of a standardized analysis framework for transcription data. This includes:

* Structured representation of transcription content (words, segments, complete transcripts)
* Quality assessment metrics (confidence scores, audio quality indicators)
* Speaker diarization analysis
* Provider-specific analytical features
* Processing metadata and provenance information

Analysis provides storage: The vCon analysis mechanism is used as the storage container for WTF analysis results. The analysis type "wtf_transcription" identifies the stored data as conforming to the WTF analysis schema.

10.2. Analysis Processing Workflow

The typical workflow for WTF transcription analysis is:

1. *Audio Processing*: Dialog audio is processed by a transcription provider
2. *Analysis Generation*: Provider results are converted to WTF analysis schema

3. *Quality Assessment*: Analysis includes confidence scores and quality metrics

4. *Storage*: WTF analysis is stored as a vCon analysis object with type "wtf_transcription"

5. *Retrieval*: Consumers read the analysis object to access transcription results

6. *Utilization*: Analysis data is used for search, compliance, accessibility, or further processing

11. IANA Considerations

11.1. vCon Extensions Names Registry

This document requests IANA to register the following extension in the vCon Extensions Names Registry established by [I-D.draft-ietf-vcon-vcon-core]:

* *Extension Name*: wtf_transcription

* *Extension Description*: World Transcription Format - A standardized analysis framework for speech-to-text transcription with multi-provider support, quality metrics, and confidence scoring

* *Change Controller*: IESG

* *Specification Document*: This document

11.2. WTF Analysis Type Values Registry

This document requests IANA to establish a new registry for WTF analysis type values with the following initial registration:

* *Type Value*: wtf_transcription

* *Description*: Structured speech-to-text transcription analysis using the World Transcription Format framework, including transcription content, quality metrics, confidence scoring, and provider-specific analytical features

* *Change Controller*: IESG

* *Specification Document*: This document

11.2.1. Registration Template

* *Type Value*: The string value used as the analysis type identifier in vCon analysis objects

* *Description*: Brief description of the analysis type and its analytical capabilities

* *Change Controller*: For Standards Track RFCs, list "IESG". For others, give the name of the responsible party.

* *Specification Document(s)*: Reference to defining documents with URIs where available

11.3. WTF Provider Registry

This document requests IANA to establish a new registry for WTF transcription analysis providers with initial registrations for supported providers:

* *Provider Name*: whisper

* *Description*: OpenAI Whisper(TM) speech recognition system

* *Change Controller*: IESG

* *Specification Document*: This document (Additional provider registrations would be added for each supported provider)

12. Examples

12.1. Basic Two-Party Call Transcription Analysis

```
{
  "uuid": "01928e10-193e-8231-b9a2-279e0d16bc46",
  "extensions": ["wtf_transcription"],
  "created_at": "2025-01-02T12:00:00Z",
  "parties": [
    {
      "tel": "+1-555-123-4567",
      "name": "Alice"
    },
    {
      "tel": "+1-555-987-6543",
      "name": "Bob"
    }
  ],
  "dialog": [
```

```
    {
      "type": "recording",
      "start": "2025-01-02T12:15:30Z",
      "duration": 65.2,
      "parties": [0, 1],
      "mediatype": "audio/wav",
      "filename": "call-recording.wav"
    }
  ],
  "analysis": [
    {
      "type": "wtf_transcription",
      "start": "2025-01-02T12:16:35Z",
      "dialog": 0,
      "vendor": "deepgram",
      "product": "nova-2",
      "encoding": "json",
      "body": {
        "transcript": {
          "text": "Hello, this is Alice from customer service. How can I help you today?
Hi Alice, I’m having trouble with my account. Can you help me reset my password?",
          "language": "en-US",
          "duration": 65.2,
          "confidence": 0.92
        },
        "segments": [
          {
            "id": 0,
            "start": 0.5,
            "end": 4.8,
            "text": "Hello, this is Alice from customer service. How can I help you today
?",
            "confidence": 0.95,
            "speaker": 0,
            "words": [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
          },
          {
            "id": 1,
            "start": 5.2,
            "end": 9.1,
            "text": "Hi Alice, I’m having trouble with my account. Can you help me reset
my password?",
            "confidence": 0.88,
            "speaker": 1,
            "words": [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28,
                      29, 30]
          }
        ],
        "words": [
          {
            "id": 0,
```

```
            "start": 0.5,
            "end": 0.8,
            "text": "Hello",
            "confidence": 0.98,
            "speaker": 0,
            "is_punctuation": false
          },
          {
            "id": 1,
            "start": 0.9,
            "end": 1.1,
            "text": ",",
            "confidence": 0.95,
            "speaker": 0,
            "is_punctuation": true
          }
          // Additional words...
        ],
        "speakers": {
          "0": {
            "id": 0,
            "label": "Alice (Customer Service)",
            "segments": [0],
            "total_time": 4.3,
            "confidence": 0.95
          },
          "1": {
            "id": 1,
            "label": "Bob (Customer)",
            "segments": [1],
            "total_time": 3.9,
            "confidence": 0.88
          }
        },
        "metadata": {
          "created_at": "2025-01-02T12:15:30Z",
          "processed_at": "2025-01-02T12:16:35Z",
          "provider": "deepgram",
          "model": "nova-2",
          "processing_time": 12.5,
          "audio": {
            "duration": 65.2,
            "sample_rate": 8000,
            "channels": 1,
            "format": "wav",
            "bitrate": 128
          },
          "options": {
```

```
            "punctuate": true,
            "diarize": true,
            "language": "en"
          }
        },
        "quality": {
          "audio_quality": "high",
          "background_noise": 0.1,
          "multiple_speakers": true,
          "overlapping_speech": false,
          "silence_ratio": 0.15,
          "average_confidence": 0.92,
          "low_confidence_words": 0,
          "processing_warnings": []
        },
        "extensions": {
          "deepgram": {
            "utterances": [
              {
                "start": 0.5,
                "end": 4.8,
                "confidence": 0.95,
                "channel": 0,
                "transcript": "Hello, this is Alice from customer service. How can I help
 you today?"
              }
            ]
          }
        }
      }
    }
  ]
}
```

12.2. Multi-Provider Transcription Analysis Comparison

```
   {
     "analysis": [
       {
         "type": "wtf_transcription",
         "dialog": 0,
         "vendor": "openai",
         "product": "whisper-large-v3",
         "encoding": "json",
         "body": {
           "transcript": {
             "text": "The quick brown fox jumps over the lazy dog.",
             "language": "en-US",
             "duration": 3.5,
```

```
            "confidence": 0.96
          },
          "segments": [
            {
              "id": 0,
              "start": 0.0,
              "end": 3.5,
              "text": "The quick brown fox jumps over the lazy dog.",
              "confidence": 0.96,
              "speaker": 0
            }
          ],
          "metadata": {
            "created_at": "2025-01-02T12:00:00Z",
            "processed_at": "2025-01-02T12:00:15Z",
            "provider": "whisper",
            "model": "whisper-large-v3",
            "processing_time": 15.2
          },
          "extensions": {
            "whisper": {
              "temperature": 0.0,
              "compression_ratio": 2.1,
              "avg_logprob": -0.25,
              "no_speech_prob": 0.01
            }
          }
        }
      },
      {
        "type": "wtf_transcription",
        "dialog": 0,
        "vendor": "deepgram",
        "product": "nova-2",
        "encoding": "json",
        "body": {
          "transcript": {
            "text": "The quick brown fox jumps over the lazy dog.",
            "language": "en-US",
            "duration": 3.5,
            "confidence": 0.94
          },
          "segments": [
            {
              "id": 0,
              "start": 0.0,
              "end": 3.5,
              "text": "The quick brown fox jumps over the lazy dog.",
```

```
              "confidence": 0.94,
              "speaker": 0
            }
          ],
          "metadata": {
            "created_at": "2025-01-02T12:00:00Z",
            "processed_at": "2025-01-02T12:00:08Z",
            "provider": "deepgram",
            "model": "nova-2",
            "processing_time": 8.1
          },
          "extensions": {
            "deepgram": {
              "model_info": {
                "name": "nova-2",
                "version": "2024-01-09",
                "uuid": "4d892fb6-7cc1-4e7a-a1b3-1c2e3f4a5b6c"
              }
            }
          }
        }
      }
    ]
  }
```

13. References

13.1. Normative References

[I-D.draft-ietf-vcon-vcon-core]
Petrie, D. G., "The JSON format for vCon - Conversation Data Container", Work in Progress, Internet-Draft, draft-ietf-vcon-vcon-core-02, January 2026, <https://datatracker.ietf.org/doc/draft-ietf-vcon-vcon-core/>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC3339] Klyne, G., "Date and Time on the Internet: Timestamps", July 2002, <https://www.rfc-editor.org/rfc/rfc3339.html>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

[RFC8949] Bormann, C., "Concise Binary Object Representation (CBOR)", December 2020, <https://www.rfc-editor.org/rfc/rfc8949.html>.

13.2. Informative References

[BCP47] Phillips, A., "Tags for Identifying Languages", September 2009, <https://www.rfc-editor.org/rfc/rfc5646.html>.

[I-D.draft-ietf-vcon-overview] McCarthy-Howe, T., "The vCon - Conversation Data Container - Overview", Work in Progress, Internet-Draft, draft-ietf-vcon-overview, 2025, <https://datatracker.ietf.org/doc/draft-ietf-vcon-overview/>.

[I-D.draft-ietf-vcon-privacy-primer] James, D. and T. McCarthy-Howe, "Privacy Primer for vCon Developers", Work in Progress, Internet-Draft, draft-ietf-vcon-privacy-primer, 2025, <https://datatracker.ietf.org/doc/draft-ietf-vcon-privacy-primer/>.

[WHISPER] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and I. Sutskever, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision", September 2022, <https://openai.com/research/whisper>.

Acknowledgements

* Appreciation to the transcription provider community for their input on standardization requirements and analysis framework design.

* Thanks to the vCon working group for their feedback and guidance on extension design patterns and the distinction between analysis types and attachment mechanisms.

Trademark Notice

All trademarks mentioned in this document are the property of their respective owners. The use of these trademarks does not imply endorsement by the IETF or the authors of this document. The following trademarks are referenced:

* Whisper(TM) is a trademark of OpenAI, Inc.

* Deepgram(TM) is a trademark of Deepgram, Inc.

* AssemblyAI(TM) is a trademark of AssemblyAI, Inc.
* Google Cloud Speech-to-Text(TM) is a trademark of Google LLC.
* Amazon Transcribe(TM) is a trademark of Amazon.com, Inc.
* Azure Speech Services(TM) is a trademark of Microsoft Corporation.
* Rev.ai(TM) is a trademark of Rev.com, Inc.
* Speechmatics(TM) is a trademark of Speechmatics Limited.
* Wav2Vec2(TM) is a trademark of Meta Platforms, Inc.
* Parakeet(TM) is a trademark of NVIDIA Corporation.

Author's Address

Thomas McCarthy-Howe
VCONIC
Email: ghostofbasho@gmail.com