



**SPEECHMATICS**

Real-time Virtual Appliance 4.0.0

## Table of Contents

- [Real-time Virtual Appliance](#)
  - [Important Notices](#)
  - [Important Notices](#)
  - [What's New](#)
    - [4.1.0](#)
    - [Known Limitations](#)
  - [Supported Platforms](#)
  - [Form Factors](#)
  - [Upgrade Path](#)
  - [Installation](#)
- [Real-time Virtual Appliance Installation and Admin Guide](#)
- [System requirements](#)
  - [Host requirements](#)
    - [AVX flags](#)
    - [Useful links](#)
  - [Virtual Appliance system requirements](#)
    - [Real-time Virtual Appliance](#)
    - [Batch Virtual Appliance](#)
    - [Important Message on IOPS](#)
- [Downloading the appliance](#)
- [Importing the appliance](#)
  - [Note on Performance Benefits on VMWare and VirtualBox](#)
  - [VMware ESXi](#)
  - [VMware Workstation Player](#)
  - [VirtualBox](#)
  - [Amazon Web Services](#)
    - [Prerequisites](#)
    - [Uploading the OVA file to S3](#)
    - [Importing the OVA as AMI instance](#)
      - [Creating an Import Service Role](#)
      - [Creating a Role Policy](#)
      - [Importing the OVA](#)
    - [Security](#)
      - [Real-time Virtual Appliance](#)
      - [Batch Virtual Appliance](#)
    - [Launching a Virtual Appliance](#)
- [Network Configuration](#)
  - [Network interface mapping](#)
    - [VMware ESXi](#)
    - [VMware Workstation Player](#)
    - [VirtualBox](#)
  - [IP Configuration](#)
    - [Configure static IP](#)
    - [Configure DHCP IP](#)
- [Licensing](#)
  - [Licensing with the enhanced model](#)
  - [Applying an Online License](#)
  - [Checking an Appliance License](#)
    - [Example Response \(unlicensed\)](#)
    - [Example Response \(licensed\)](#)
  - [Removing a License](#)

- [Using a Proxy Server](#)
- [Offline License Activation](#)
  - [Generating an Activation Certificate](#)
  - [Sending the Activation Certificate to Speechmatics](#)
  - [Applying the License Certificate](#)
- [Running an Appliance Offline](#)
- [Licensing Troubleshooting](#)
  - [Receiving Updates to a License](#)
  - [Invalid License](#)
  - [Appliance Offline](#)
  - [Offline Activation Error](#)
  - [Unable to Delete License when Offline](#)
  - [Virtual appliance is offline message when port 80 is blocked](#)
- [Verify and Go \(Real-time\)](#)
  - [Verify the service](#)
  - [Go!](#)
- [SSL Configuration](#)
  - [Default behaviour](#)
    - [Management API Examples](#)
    - [Monitoring API Example](#)
    - [Speech API Example](#)
  - [Using your own SSL certificate and private key](#)
    - [Uploading the certificate and key to the appliance](#)
    - [Disabling HTTP access](#)
    - [Enable Basic Authentication for Admin](#)
  - [FAQs](#)
    - [How do I reset the SSL settings?](#)
    - [What if I forget the admin password?](#)
    - [What versions of SSL/TLS do you support?](#)
      - [What cipher suites do you support?](#)
- [Networking](#)
  - [Network Requirements](#)
  - [Configure Static IP](#)
  - [Configure DHCP](#)
  - [Firewall Ports](#)
  - [Using Proxies](#)
- [Virtual Appliance Scaling](#)
  - [Real-time Virtual Appliance Scaling](#)
    - [Worker Limits](#)
    - [View Maximum Workers](#)
    - [Setting Maximum Workers](#)
  - [Batch Virtual Appliance Scaling](#)
    - [Worker Limits](#)
    - [View Maximum Workers](#)
    - [Setting Maximum Workers](#)
- [Monitoring](#)
- [Services](#)
  - [Batch Virtual Appliance](#)
  - [Service status](#)
  - [Real-time Virtual Appliance](#)
  - [Service status](#)
  - [Service restart](#)
  - [Access Logs](#)

- [System restart](#)
- [System shutdown](#)
- [Troubleshooting](#)
  - [Transcription job failure](#)
  - [Illegal instruction errors](#)
  - [AVX2 Warning](#)
- [Console for Advanced Troubleshooting](#)
  - [License](#)
  - [Networking](#)
  - [Reboot and Shutdown](#)
  - [Security](#)
  - [Services](#)
  - [Tools](#)
  - [Workers](#)
- [Security](#)
  - [Overview](#)
  - [Ports and Protocols](#)
  - [Custom Dictionary Cache](#)
    - [Size available](#)
    - [Size of cache entries](#)
    - [Cache life cycle](#)
  - [Administering the Cache](#)
    - [View Cache Usage](#)
    - [Purge Cache Contents](#)
- [Introduction](#)
  - [Overview](#)
    - [Terms](#)
  - [Getting Started](#)
  - [Input Formats](#)
  - [Transcription Output Format](#)
    - [Final transcripts](#)
    - [Partial transcripts](#)
  - [The WebSocket Protocol](#)
- [Real-time API](#)
  - [Getting Started](#)
    - [Messages](#)
      - [StartRecognition](#)
      - [Explaining Max Delay Mode](#)
      - [SetRecognitionConfig](#)
      - [AddAudio](#)
    - [AudioAdded](#)
      - [Implementation details](#)
      - [AddTranscript](#)
      - [AddPartialTranscript](#)
      - [EndOfStream](#)
      - [EndOfTranscript](#)
    - [Supported audio types](#)
  - [Example communication](#)
  - [Configuring Additional Features](#)
    - [Transcription config](#)
    - [Requesting an enhanced model](#)
  - [Advanced punctuation](#)
    - [Additional words](#)

- [Output locale](#)
- [Punctuation overrides](#)
- [Errors, Warnings and Info Messages](#)
  - [Error messages](#)
    - [Error types](#)
  - [Warning messages](#)
    - [Warning types](#)
  - [Info messages](#)
    - [Info message types](#)
- [Example Connection to the API](#)
  - [WebSocket URI](#)
  - [Session Configuration](#)
    - [TranscriptionConfig](#)
    - [AddAudio](#)
    - [Final and Partial Transcripts](#)
- [Example Usage](#)
  - [JavaScript](#)
  - [Python Libraries](#)
- [Formatting Common Entities](#)
  - [Overview](#)
  - [Supported Languages](#)
  - [Using the enable\\_entities parameter](#)
  - [Configuration example](#)
  - [Different entity classes](#)
  - [Output locale styling](#)
  - [Example output](#)

# Real-time Virtual Appliance

## Important Notices

:::warning Removal Note The legacy V1 API that the Real-time Virtual Appliance supported is **now removed**. We recommend all customers move to using the V2 API. Please see the section [How to use the V2 API](#). :::

## Important Notices

Speechmatics now supports exclusively `speechmatics-python` for use in both our Real-time Container and our Real-time Virtual Appliance. The older library `smwebsocket-py` will still work, but is not compatible with the new enhanced model and is no longer supported. Please see [here](#) for access to `speechmatics-python`.

*The new enhanced model has increased compute requirements and new recommended AVX flags. Each concurrent worker will require at least 3GB of memory and up to 5GB if using other features such as Custom Dictionary. Please check the updated system requirements in the installation guide and ensure your hardware meets Speechmatics' recommendations. Otherwise you may see a slow down in processing speed when using the enhanced model. It is also now necessary to run the appliance on processors that support AVX2 in order to take advantage of latest performance optimisations for both the standard and enhanced model for all language packs.*

*If you are importing an appliance through VirtualBox, and AVX flags are not automatically enabled, you can also take advantage of the the performance benefits from AVX 2 following [these guidelines](#).*

## What's New

### 4.1.0

- Deprecation of the V1 API for Real-time Virtual Appliance
- 16 Languages updated with additional punctuation marks for improved readability
  - The following languages now support ( . ? , ! ): Bulgarian, Catalan, Czech, Greek, Finnish, Croatian, Hungarian, Lithuanian, Latvian, Norwegian, Polish, Romanian, Slovak, Slovenian, Korean
- Improved accuracy for French, including more data for Canadian French (fr-ca)
- Improved accuracy for Portuguese, including more data for Brazilian Portuguese (pt-br)
- Standard operating point improved accuracy for Romanian, Hungarian, Danish, Slovakian, Croatian, Bulgarian, Finnish, Slovenian, Lithuanian
- Updated Danish, Norwegian and Swedish to remove undesired character sets
- Improved accuracy in localised spelling for English output locale feature
- Fixes for English and Italian written form numeric entities
- Improved accuracy of percentage symbol recognition in French
- Fixed issue where occasional end times of words could be before the start time

## Known Limitations

The following are known issues in this release:

Issue ID	Summary	Detailed Description and Possible Workarounds
REQ-1409	Proteus HCL with <unk> causes out of memory error	A custom dictionary list that contains the word '' causes the worker to crash.
REQ-7549	Memory leak affecting gRPC	There is a small memory leak in the gRPC Python server <a href="https://github.com/grpc/grpc/issues/5913">https://github.com/grpc/grpc/issues/5913</a> .
REQ-10160	Advanced punctuation for Spanish (es)	Inverted marks [ ¿ ¡ ] are not currently available for Spanish advanced punctuation.

	does not contain inverted marks.	
REQ-10627	Double full stops when acronym is at the end of the sentence	If there is an acronym at the end of the sentence, then a double full stop will be output, for example: "team G.B.."
REQ-11792	Speaker change token positioning is incorrect	We are aware of a consistent mis-placing of the speaker change token after the first word of the new speakers' sentence rather than before it.
REQ-12202	High memory usage when using custom dictionary	It has been observed that when using custom dictionary an additional 800-1700MB of memory is required (depending on the size of the wordlist used).
REQ-16256	Heavy usage of RAM when swapping between 8kHz and 16kHz input	Where multiple persistent workers are configured with Custom Dictionary that swap between 8kHz and 16kHz input, this can cause a memory leak that causes the container to crash. If this starts to impact services it is recommended to restart all the services with the management API or drop the worker count to 1 and then increase it again

## Supported Platforms

Virtual Appliance image (OVA) for installation on:

- VMware ESXi 6.5+ or VMware Workstation Player.
- VirtualBox 5.2+
- Amazon EC2

See the Installation and Admin Guide for details on the minimum specifications for the VM. The maximum number of concurrent jobs (maxworkers) that you can run on a single appliance is 30.

## Form Factors

There are five variants of the Real-time Virtual Appliance.

Variant	Image Size	Max. Disk Space	Languages
nano	9GB	40GB	en
mini	13GB	40GB	en, de, es
midi	23GB	60GB	en, de, es, fr, ko, ja, nl, pt
maxi	38GB	80GB	en, de, es, fr, ko, ja, nl, pt, it, da, pl, ca, hi, ru, sv
plus	45GB	80GB	en, cmn, no, ar, bg, cs, el, fi, hu, hr, lt, lv, ro, sk, sl, tr, ms, id, yue

## Upgrade Path

Remove the license from your old appliance (see the Admin Guide), then re-import the new OVA and configure networking as per the Installation and Admin guide. You will need to re-apply the license code you have once the OVA has imported.

## Installation

Upload the OVA to VMWare ESX, VMWare Workstation Player, or VirtualBox. See the Installation and Admin Guide for more information.

# Real-time Virtual Appliance Installation and Admin Guide

This guide explains how to install and administer the Real-time Virtual Appliance using the Management REST API.

The Speechmatics virtual appliance is available in two modes: real-time and batch. For the most part installation and administration are identical for both modes. Where differences exist this is explicitly noted in this guide.

**Note:** Most of the code examples in these docs use HTTP rather than HTTPS to communicate with the appliance. However, we recommend using HTTPS for your production deployments. For information on SSL/HTTPS configuration, see the 'SSL Configuration' section of the docs.

## System requirements

The Speechmatics Real-time Virtual Appliance operates on a hypervisor host system. For this version of the appliance, the following hypervisors are supported:

- VMware®
- VirtualBox
- AWS EC2

For the virtual appliance to operate as required, the host must meet the requirements and have the resources available as defined below.

## Host requirements

The host machine requires a processor with following microarchitecture specification:

- If using the standard model offering at least the Broadwell Class is required
- If using the enhanced model a chip that is at least the CascadeLake class is required, as this is the minimum spec that will support AVX512\_VNNI - for more information see below.
- It is also recommended if using the enhanced model that the hardware supports the AVX512\_VNNI flag, as this will greatly improve transcription processing speed
  - Examples of this among popular hosting providers include the Microsoft Azure DSV-4 class, and the Amazon M5n EC2 server class
  - If you are using the enhanced model and running on VMWare, you will have to upgrade to `hardware_version` 18 to take advantage of the AVX512\_VNNI flag. Please note this is only supported by ESXi version 7.0 onwards
  - If you are using VMWare and the enhanced model, and encounter performance issues, we recommend allocating dedicated memory and/or processors to the appliance. How to apply dedicated processors in VMWare is documented [here](#), setting memory is documented [here](#)
- If you encounter performance issues when running the enhanced model, disabling hyperthreading when running the enhanced model can also improve transcription speed. How to do so when running on Amazon Web Services is shown [here](#), and for Microsoft Azure please see [here](#)

## AVX flags

The hardware you run the appliance on must support Advanced Vector Extensions (AVX). Advanced Vector Extensions are necessary to allow Speechmatics to carry out transcription:

- For the standard model, it is necessary to use at least a processor that supports at least Advanced Vector Extensions 2 (AVX2).
  - You should also ensure your hypervisor is enabled to use AVX2.



- For the enhanced model, it is recommended to run the appliance on hardware that supports the AVX512\_VNNI flag in addition to AVX2, which will substantially improve transcription processing speed.

To see what AVX flags are supported by the CPU of your host system, you can run the following query via the Management API of the appliance:

```
curl -X GET "https://{HOSTAPPLIANCE}/v1/management/cpuinfo" -H "accept: application/json"
```

You will receive information about the host CPU. Supported AVX flags will be returned as flags in the Management API response. An example is below:

```
{
  "usage_percentage": 2.5,
  "architecture": "X86_64",
  "model_name": "Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz",
  "cpus": "2",
  "vendor": "GenuineIntel",
  "hyperthreading": false,
  "flags": "3dnowprefetch abm adx aes apic arat arch_capabilities arch_perfmon avx avx2
avx512_vnni bmi1 bmi2 clflush cmov constant_tsc cpuid cpuid_fault cx16 cx8 de f16c flush_l1d fma
fpu fsgsbase fxsr hypervisor ibpb ibrs invpcid invpcid_single lahf_lm lm mca mce md_clear mmx
movbe msr mtrr nonstop_tsc nopl nx pae pat pcid pclmulqdq pdpe1gb pge pni popcnt pse pse36 pti
rdrand rdseed rdtscp sep smap smep ss ssbd sse sse2 sse4_1 sse4_2 ssse3 stibp syscall tsc
tsc_adjust tsc_deadline_timer tsc_reliable vme x2apic xsave xsaveopt xtopology"
}
```

## Useful links

See below for minimum Real-time Virtual Appliance VM (guest) specifications; the host machine must have enough resources (processor, memory and storage) to run the hypervisor, the guest VMs you intend to host on it, plus any other processes you expect to run on it. Vendor guidelines should be followed for other host requirements and installation process.

For VMWare, the document Performance Best Practices for VMware vSphere® 6.0 contains a comprehensive overview of hardware considerations and recommendations on how to optimize your host platform. See <https://www.vmware.com/support.html> for up-to-date technical information on VMWare.

For VirtualBox, please consult the online documentation: <https://www.virtualbox.org/wiki/Documentation>

For Amazon EC2, the following link explains how to setup a VM using an Amazon S3 to store the OVA file: <https://docs.aws.amazon.com/vm-import/latest/userguide/vmimport-image-import.html>.

## Virtual Appliance system requirements

### Real-time Virtual Appliance

The Speechmatics Real-time Virtual Appliance must be allocated the following *minimum* specification:

- 2 vCPU
- 8GB RAM
- Up to 38GB hard disk space

For each concurrent input stream using the standard model the appliance requires an additional 1 vCPU and at least 1.5GB RAM.

If you are using the custom dictionary (additional words) feature then each concurrent input stream that is configured to use it will require up to 3GB RAM.

If you are using the enhanced model, then each concurrent input stream that is configured to use it will require up to 3GB RAM. If the enhanced model is used in conjunction with other features like Custom Dictionary and encountering performance issues, then up to 5GB may be required.

## Batch Virtual Appliance

For operation in batch mode, the following *minimum* specifications are required:

- 2 vCPUs
- 8GB RAM
- Up to 44GB hard disk space

## Important Message on IOPS

Heavy usage of the appliance at scale can sometimes result in very high percentage usage of volume throughput. If this is the case, we recommend increasing the maximum IOPs supported by your hardware to a value between 8,000-12,000. This is not necessary in all circumstances, but may result in better performance if you are running more than 10 concurrent workers. Increasing the IOPS also will result in an increase in cost for resource usage. If you use AWS, setting the `volume type` to `io2` is also recommended in this scenario. How to change the maximum IOPS supported by your hardware is documented [here for AWS](#), [here for Microsoft Azure](#), and [here for VMWare](#). You may need to do this if:

- You are using close to or the maximum number of workers supported by that appliance size
- The jobs being processed are all long files, and diarization is requested

## Downloading the appliance

A download link will be provided by Speechmatics through the solutions section of the support portal (<https://support.speechmatics.com>). The latest version of the appliance can be located within the solutions section. Select the required version number within the "Real-time Virtual Appliance" area that you wish to download to view the download link and all associated documentation for the virtual appliance. Once the download link is selected the download will begin, or a save file prompt will appear, enabling the file to be saved (the exact method will depend on the web browser being used). After the download a file with an ".ova" extension will be stored on the computer.

An account is required to access the documents and download link in the support portal. If an account is not available or the "Real-time Virtual Appliance" section is not visible in the support portal, please contact Speechmatics Support [support@speechmatics.com](mailto:support@speechmatics.com) for help.

## Importing the appliance

Once the **.ova** file has been downloaded, it is ready to be imported into the host you have already prepared. Please ensure that the host meets the requirements stated earlier in this guide, then based on the hypervisor environment follow the instructions below.

## Note on Performance Benefits on VMWare and VirtualBox

To take advantage of recent Speechmatics improvements in performance using AVX2, the `hardware_version` of the Appliance has been upgraded from 9 to 11. If you are running VMWare ESXi host 6.5 or later, this should not affect any system behaviour. If you are on an earlier version, you can downgrade the hardware version as documented [here](#); however please note that this will mean you cannot take advantage of more recent optimisations in performance from using Advanced Vector Extensions 2 (AVX2).

If you are running your Appliance in VirtualBox please follow [the following guidelines](#) to enable AVX2 if it is not done automatically during the importing process.

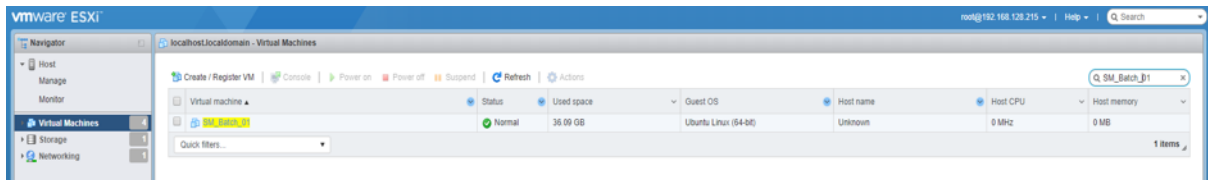
## VMware ESXi

The following steps can be used to import the virtual appliance into VMWare ESXi 6.5:

- Open the vSphere web console on the host
- Choose "Virtual Machines" from the Navigator
- Select "Create/Register VM" option
- A wizard will appear:

- Choose "Deploy a virtual machine from an OVF or OVA file" and click "Next"
  - Enter a VM name e.g. "SM\_Batch\_01", and drag the downloaded .ova file onto the window and click "Next"
  - Select a datastore that has enough capacity to store the virtual appliance and click "Next"
  - From the "VM network" dropdown box, select a network
  - Choose Thin or Thick disk provisioning (the Speechmatics Real-time Virtual Appliance supports either. Choose the options that is right for the hosting environment refer to VMWare documentation for help and click "Next"
  - Check the details are correct and click "Finish"
- The virtual appliance will import. This can take a few minutes depending on the datastore chosen.

Once the VM has imported it should be visible on the vSphere web console:



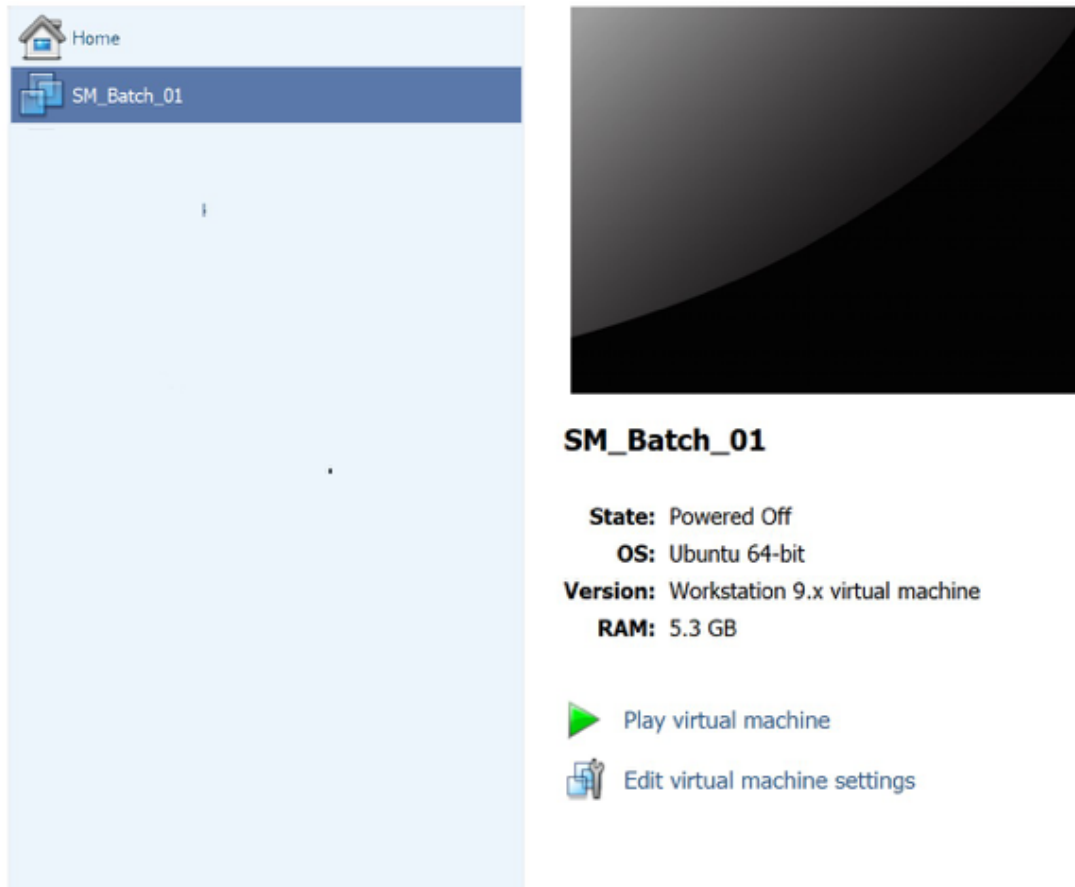
### Important Notice

If you are running VMWare ESXi version 6.5 version or above, change the `hardware_version` of the appliance to 11 to take advantage of recently implemented Speechmatics performance improvements. How to do so is documented [here](#)

## VMware Workstation Player

- Open VMware Workstation Player
- From the top options bar select "Player", then "File" and "Open..."
- The "Open Virtual Machine" window will appear. Navigate to the ".ova" file you downloaded earlier, select it, click "Open"
- Enter a VM name e.g. "SM\_Batch\_01"
- A default storage location for the virtual appliance will be shown, the can be changed if required. Click "Import".
- Dropdown box from the top options bar, click on "File"
- The virtual appliance will import. This can take a few minutes depending on the hard disk chosen

Once the VM has imported it should be visible on the Workstation player:

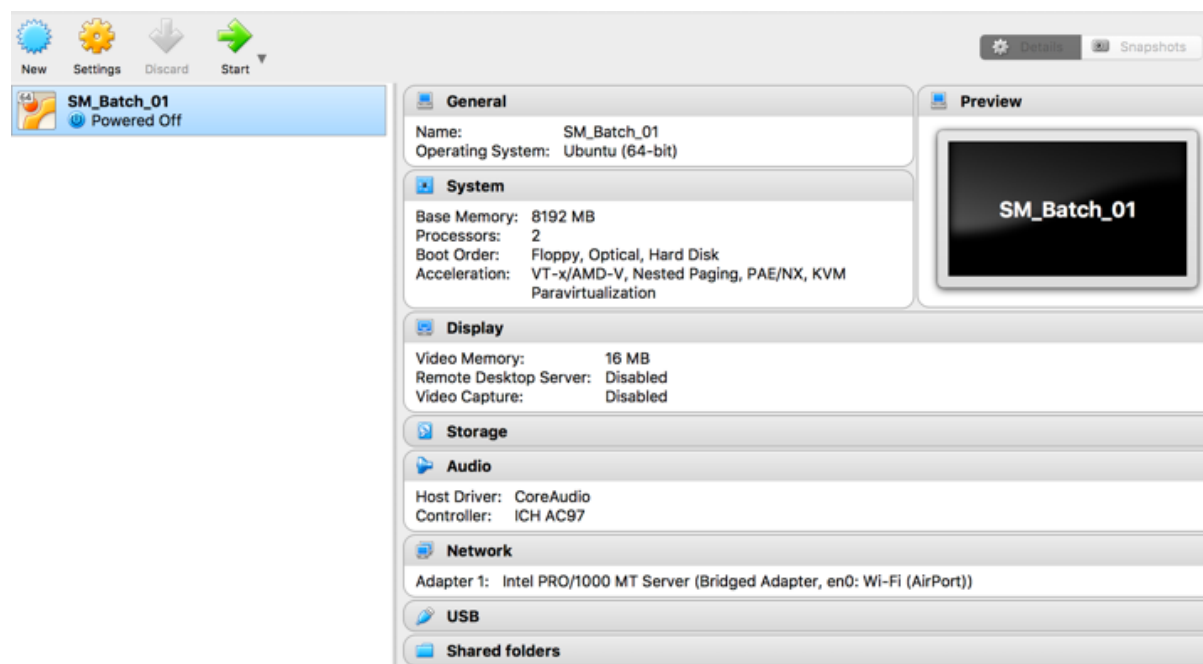


## VirtualBox

The following steps can be used to import the virtual appliance into VirtualBox 5.2 or above.

- Open VirtualBox
- From the Manager window select "File", then "Import Appliance..."
- In the Name field, name the Real-time Virtual Appliance e.g. "SM\_Batch\_01"
- Browse to the OVA file and click on the "Import" button

Once the VM has imported it should be visible on the VirtualBox Manager:



## Amazon Web Services

This section explains how to create a Real-time Virtual Appliance EC2 instance on the Amazon Web Services (AWS) platform by using the AWS VM Import/Export tool. This tool is designed for importing VM images from the OVA file format provided by Speechmatics. You will import the image as an Amazon Machine Image (AMI), from which you can then launch machine instances.

The information in this section is taken from the official AWS documentation and parts of it have been extracted to focus more on the particulars of the Speechmatics Real-time Virtual Appliance. For more details of the Amazon VM image import process, please refer to <https://docs.aws.amazon.com/vm-import/latest/userguide/vmimport-image-import.html>

### Prerequisites

There are a few pre-requisites that you will need to have setup before you can follow the instructions in this section:

- [AWS Command Line Interface](#) (CLI)
- Python 2.6.5 or higher

Please follow the recommendations on configuration of the AWS CLI by referring to the [Getting Started](#) guide.

### Uploading the OVA file to S3

This section describes the process of uploading the Speechmatics OVA file to an Amazon S3 bucket from where it can be imported as an AMI instance. We recommend using a bucket in the same region where you want the AMI to be created and made available.

Once you've identified or created the S3 bucket on your account where the Speechmatics Real-time Virtual Appliance OVA will be uploaded to, you can use any of the tools below to help with the upload of the OVA file.

- The following [AWS SDK libraries](#) support S3 multipart upload (which is the recommended method given the large size of the OVA file):
  - AWS SDK for Java
  - AWS SDK for .NET
  - AWS SDK for PHP
  - AWS SDK for Python (Boto)
  - AWS SDK for Ruby

- You can also use the [Multipart Upload API](#) directly
- User interface tools, for instance:
  - [S3 Browser](#)
  - [CloudBerry S3 Explorer](#)

For more information about the multipart uploads, see the [AWS documentation](#).

## Importing the OVA as AMI instance

After the Virtual Appliance OVA file has been successfully uploaded to an S3 bucket, it's time to import the image.

See the AWS documentation that covers [uploading an image](#) for full details.

The steps that you will perform in this section include (in order):

- Creating a Service Role on your AWS account
- Assigning a Role Policy to this Service Role
- Importing the OVA for the Real-time Virtual Appliance from the S3 bucket file

### Creating an Import Service Role

First of all, a **service role** needs to be created on your AWS account. This allows certain operations, including downloading images from an S3 bucket.

Create a file named `trust-policy.json` with the following policy:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": { "Service": "vmie.amazonaws.com" },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": {
          "sts:Externalid": "vmimport"
        }
      }
    }
  ]
}
```

Then use the `create-role` command from the AWS CLI to create a role named `vmimport`. You need to specify the full path of the `trust-policy.json` file:

```
aws iam create-role --role-name vmimport --assume-role-policy-document file://trust-policy.json
```

You need to ensure that the `file://` prefix is prepended to the filename.

### Creating a Role Policy

Create a file named `role-policy.json` with the following policy. Where you see `ova-bucket` it will need to be replaced with the name of the S3 bucket where the OVA file is stored.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetBucketLocation",

```

```

        "s3:GetObject",
        "s3:ListBucket"
    ],
    "Resource": [
        "arn:aws:s3:::ova-bucket",
        "arn:aws:s3:::ova-bucket/*"
    ]
},
{
    "Effect": "Allow",
    "Action": [
        "ec2:ModifySnapshotAttribute",
        "ec2:CopySnapshot",
        "ec2:RegisterImage",
        "ec2:Describe*"
    ],
    "Resource": "*"
}
]
}

```

Use the `put-role-policy` command to attach the policy to the role. You must specify the full path to the location of the `role-policy.json`:

```

aws iam put-role-policy --role-name vmimport --policy-name vmimport --policy-document
file://role-policy.json

```

### Importing the OVA

Importing the virtual appliance image (OVA) to Amazon EC2 as an Amazon Machine Image (AMI) is the next step.

Create a file named `containers.json` with the following content. Where you see `ova-bucket` it will need to be replaced with the name of the S3 bucket where the OVA file is stored and where you see `example-virtual-appliance.ova` it will need to be replaced with the name of the OVA file to be imported (e.g. `batch-appliance-<version>-maxi-<build-number>.ova` or `rt-appliance-<version>-maxi-<build-number>.ova`).

```

[
  {
    "Description": "Virtual Appliance OVA",
    "Format": "ova",
    "UserBucket": {
      "S3Bucket": "ova-bucket",
      "S3Key": "example-virtual-appliance.ova"
    }
  }
]

```

Use the `import-image` command to create an import task (Specify the full path to the location of the `containers.json`):

```

aws ec2 import-image --description "Virtual Appliance OVA" --disk-containers
file://containers.json

```

The resulting JSON output will show an `ImportTaskId` which you can use to check the status of the import task. You do this by running the `describe-import-image-tasks` command:

```

aws ec2 describe-import-image-tasks --import-task-ids import-ami-abcd1234

```

You need to replace the task identifier with the `ImportTaskId` for your import task ( `import-ami-abcd1234` in this example).

When the status is in the `completed` state the AMI is ready to use.

## Security

For more background on creating security groups refer to the official [AWS documentation](#). See the [Ports and Protocols](#) section for a list of the ports that are used. These ports should be opened so that you can submit jobs and manage and monitor the Speechmatics Virtual Appliance.

## Real-time Virtual Appliance

If you setup HTTPS as described in the 'SSL Configuration' section of these docs then you only need to expose port 443, **unless** you require use of the v1 WebSockets API, which requires use of port 9000. We recommend use of our updated v2 API unless you are a customer who has already implemented code against our v1 API.

## Batch Virtual Appliance

If you setup HTTPS as described in the 'SSL Configuration' section of these docs then you only need to expose port 443.

## Launching a Virtual Appliance

Now that the Virtual Appliance has been imported, it will be available as an AMI which can be launched as an instance. To launch a Speechmatics Virtual Appliance, do the following:

- Login to the AWS console and find your image under **EC2 Service | Images**
- Right-click the image and choose **Launch**
- Refer to the **System requirements** section of the Speechmatics Quick Start Guide or Admin Guide to identify how much system resources is required for your set up. Choose the instance type that meets your requirement
- Choose **Review and Launch** from the console. Setup the Key Pair if required and choose **Launch** again.

Full instructions for launching instances can be found here:

<https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/launching-instance.html>

# Network Configuration

Before starting the virtual appliance for the first time, it is important to consider the network settings that will be used. The section below describes the options.

## Network interface mapping

Whilst the virtual appliance is powered off, the virtual network adaptor should be mapped to the correct physical adaptor on the host. The virtual interface must be mapped to a physical adaptor on which the Speechmatics Real-time Virtual Appliance will be contacted. Steps are provided below for the supported hypervisors.

### VMware ESXi

There is nothing to configure here. The network as specified during the import stage described above will be used.

### VMware Workstation Player

Speechmatics recommends using bridged network mode. To ensure bridged networking is selected:

- Open VMware Workstation Player
- Right click on the virtual appliance e.g. "SM\_App\_01", and select "Settings..."
- Select the "Network Adapter" in the devices list
  - Select "Bridged: Connected directly to the physical network"
- Click "OK"

This will result in the VM using an IP Address for its use that is independent from that of the host.



## VirtualBox

Speechmatics recommends using bridged network mode. To ensure bridged networking is selected:

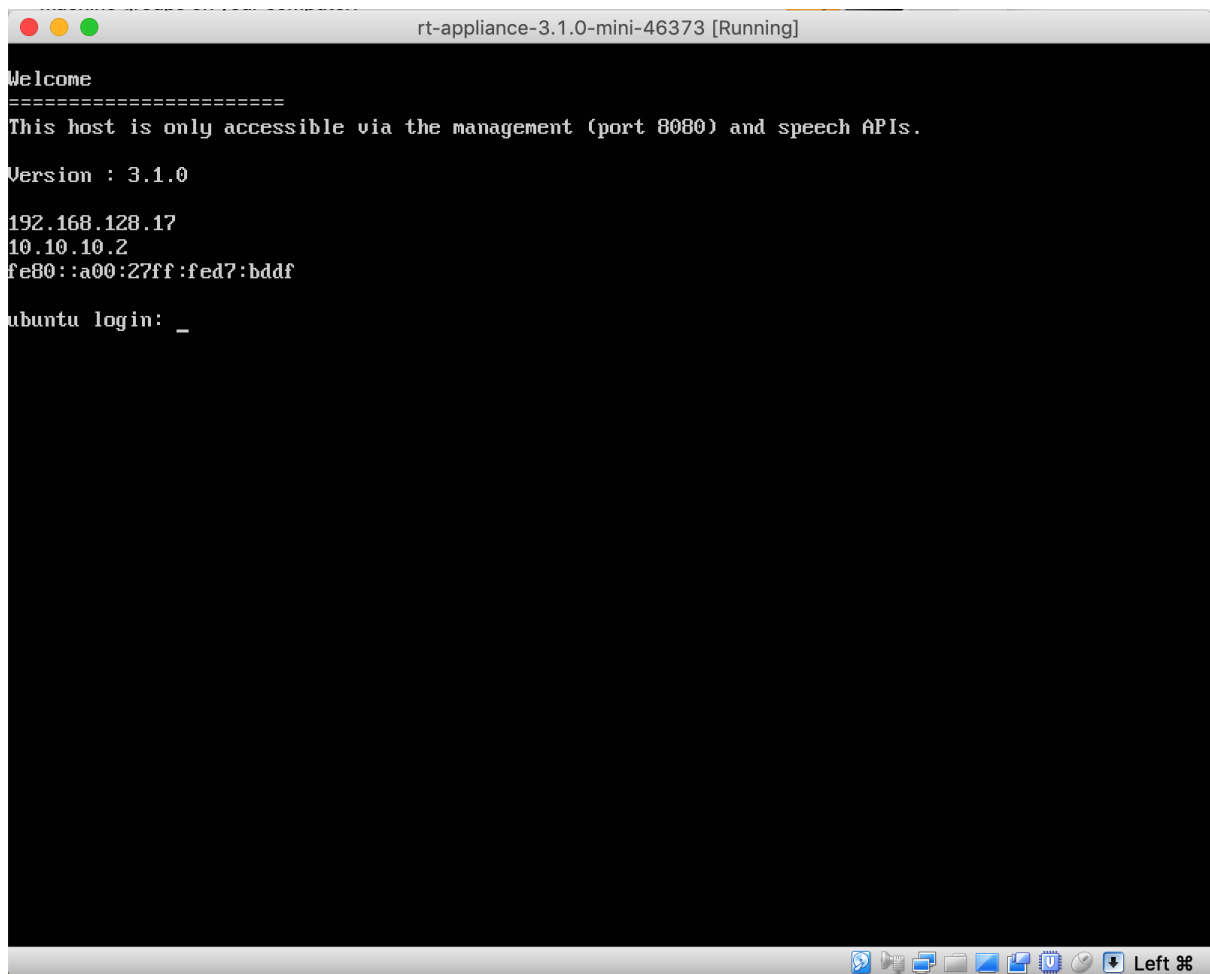
- Open VirtualBox
- Right click on the virtual appliance and select "Settings..."
- Select "Network" and from the "Attached to:" dropdown box, select "Bridged Adaptor"
- Click "OK"

This will result in the VM using an IP Address for its use that is independent from that of the host.

## IP Configuration

When the Speechmatics Real-time Virtual Appliance is started, the default behavior will be to dynamically acquire an IP address. If there is no DHCP service available on the network, it will fall back to an IP address automatically assigned.

The IP address information can be viewed by opening the virtual appliance console once it has booted, as shown below.



```
rt-appliance-3.1.0-mini-46373 [Running]
Welcome
=====
This host is only accessible via the management (port 8080) and speech APIs.
Version : 3.1.0
192.168.128.17
10.10.10.2
fe80::a00:27ff:fed7:bddf
ubuntu login: _
```

The screen shot above shows the 10.10.10.2 IP address as the fallback address. The other address shown was allocated by DHCP and should be used for all communication.

If DHCP cannot be used, a static IP address can be configured as described below.

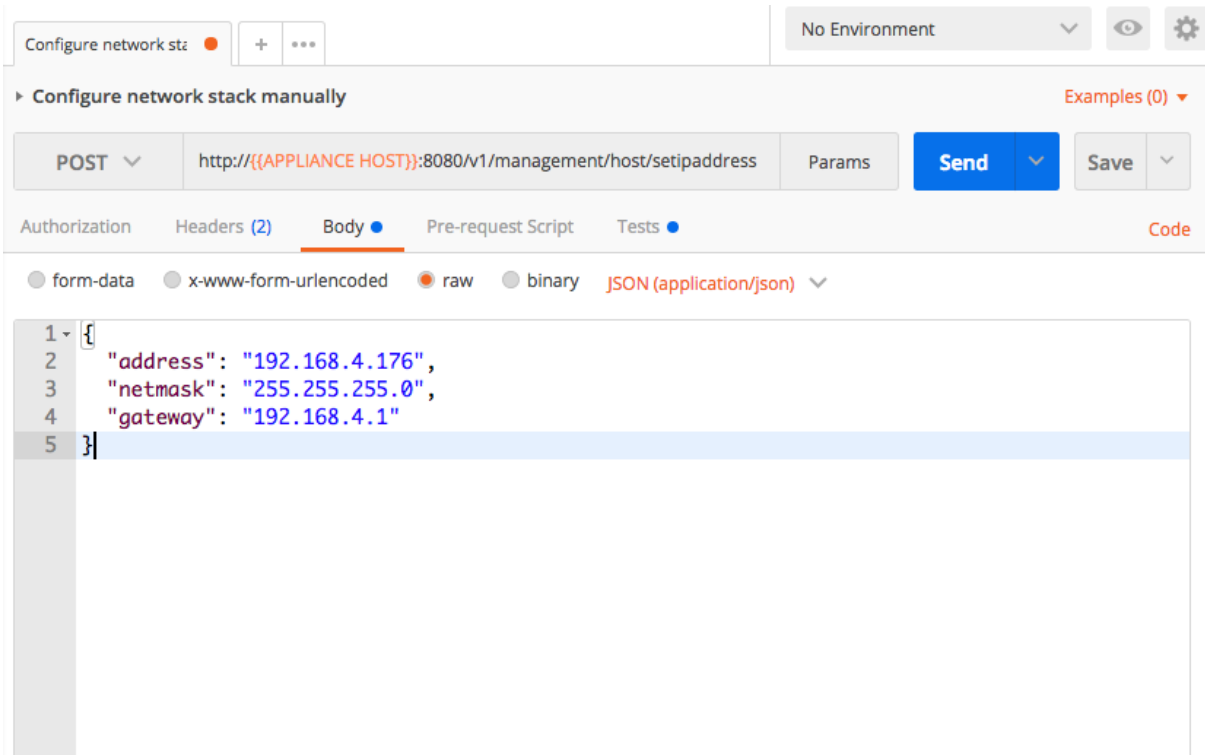
### Configure static IP

To configure a static IP address, the Management REST API for the virtual appliance is used. The following information is required:

- **Method:** POST
- **URL:**  
http://\${APPLIANCE\_HOST}:8080/v1/management/host/setipaddress
- **Body Format:** JSON
- **Body:** address, netmask, gateway, nameservers

Where \${APPLIANCE\_HOST} is the hostname or IP address of your Real-time Virtual Appliance.

The example below shows use of [Postman](#) (available for free from the Chrome web store) to POST new IP settings.



You can optionally specify a list of nameservers to use (if none are specified then, 8.8.8.8 is used), for example this time using [curl](#) from the command-line to make the POST request:

```
curl -L -X POST 'http://${APPLIANCE_HOST}:8080/v1/management/host/setipaddress' \
  -H 'Accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '@network-config.json'
```

In this example, a local file `network-config.json` is used for the JSON configuration:

```
{
  "address": "192.168.128.96",
  "netmask": "255.255.255.0",
  "gateway": "192.168.4.1",
  "nameservers": ["208.67.222.222", "208.67.220.220"]
}
```

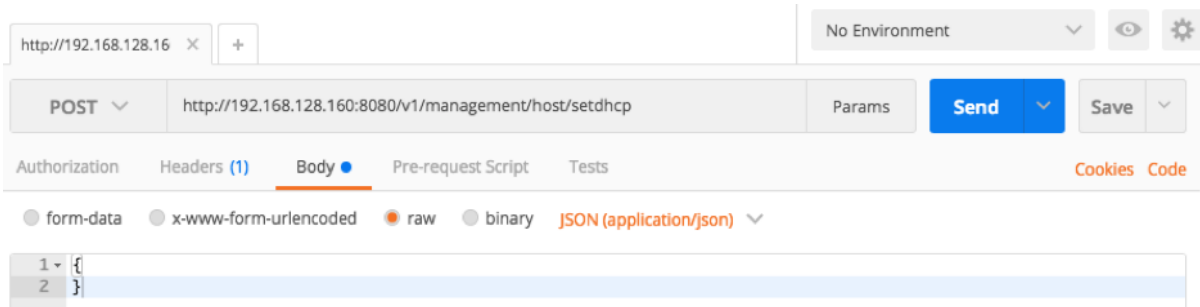
**NOTE:** once the POST is sent, the virtual appliance will automatically reboot. Check the console to verify the new IP address has been applied.

## Configure DHCP IP

To configure a dynamic IP address using DHCP, the admin REST API is used as follows:

- **Method:** POST
- **URL:**  
http://\${APPLIANCE\_HOST}:8080/v1/management/host/setdhcp
- **Body format:** JSON

The example below shows how to use Postman to POST to the REST API in order to configure a DHCP address.



```
curl -L -X POST 'http://${APPLIANCE_HOST}:8080/v1/management/host/setdhcp' \
-H 'Accept: application/json'
```

**NOTE:** once submitted, the virtual appliance will automatically reboot. Check the console to verify the new IP address has taken affect.

## Licensing

The Speechmatics Real-time Virtual Appliance uses two licensing options: an online cloud-based licensing mechanism, and offline licensing. The online license requires that the appliance must be connected to the Internet in order to activate the license, and then while running will need connection to the external internet via Port 80.

An offline license allows a user to apply a license without ever connecting an appliance to the external internet, even to initially license the appliance. Users generate an Activation Certificate via the Management API, which is then sent to Speechmatics Support to generate a separate license certificate. The license certificate can be used to then successfully generate transcription from the offline appliance. How to do so is described in more detail below.

Your appliance must have been activated with a valid license before the Speech API can be used. Use of the Management API does not require a license. Please contact Speechmatics support [support@speechmatics.com](mailto:support@speechmatics.com) if you do not have a license.

You can only apply to one license to an appliance at a time. If you want to apply a new license, you must first remove the old license, and then apply a new license as shown [here](#).

### Licensing with the enhanced model

If you are using both the standard and the enhanced model interchangeably, please note that you will need two separate licenses, one for standard appliances which will be entitled to use a standard model, and one for appliances that will be entitled to use a standard and enhanced model. You should ensure in your routing logic that jobs using the standard model are sent to the appliance which only uses the standard model. Otherwise, there is a risk of overbilling. You will also require a new license to use the enhanced model.

If you are not certain whether you are entitled to use the enhanced model, please check with your account manager.

### Applying an Online License

To apply a license that you have received from Speechmatics you use a POST request to the Management API. If your license supports fully offline activation and your appliance has no route to the Internet, then you should follow the instructions in the section on [Offline License Activation](#) later on. Otherwise keep reading this section.

Assuming your appliance is deployed in a network that has a route to the Internet you can make the activations request to the `/v1/management/license` endpoint as follows:

- **Method:** POST
- **URL:**  
`http://${APPLIANCE_HOST}:8080/v1/management/license`
- **Body format:** JSON
- **Body:** `license`, `username`, `email_address`, `company_name`

You must supply the `license` value. The other fields ( `username`, `email_address` and `company_name` ) are optional, but we recommend that you fill them in with your details to help in case of support issues.

**Note:** make sure when applying the license, that all the appliance services are running; otherwise the activation will fail.

The example below shows how to make an activation call using the Management REST API:

```
curl -L -X POST "http://${APPLIANCE_HOST}:8080/v1/management/license" \  
-H 'Accept: application/json' \  
-H 'Content-Type: application/json' \  
-d '{  
  "license": "494953679762904933",  
  "username": "Amy Liu",  
  "email_address": "a-liu@example.com",  
  "company_name": "Example Pty"  
}' \  
| jq
```

The response should indicate that the licensed status is true. The licensing activation requires a connection to the Internet (using TCP port 80). Blocking this port with an online license can cause transcription issues. If you are behind a corporate firewall that does not allow a direct connection to the Internet then you can configure the appliance to use a proxy server to allow you to license the appliance as shown in the documentation [here](#).

## Checking an Appliance License

You can check whether the appliance is licensed by using a GET request to the `/license` endpoint on the Management API. For example:

```
curl -L -X GET "http://${APPLIANCE_HOST}:8080/v1/management/license" \  
-H 'Accept: application/json' \  
| jq
```

### Example Response (unlicensed)

If the appliance has *not* been licensed then you will see something like this:

```
{  
  "licensed": false,  
  "product": "103",  
  "subscription_expiry": "1970-01-01T00:00:00Z",  
  "status": "-113 License status server trial expired",  
  "message": "",  
  "transcription_minutes_allowed": "0",  
  "license_type": "",  
  "activation_type": "",  
  "transcription_secs_allowed": 0,  
  "transcription_secs_allocated": 1800000,  
  "connected": true,  
  "customer_id": 4920,  
}
```

```
"license_code": ""
}
```

The `licensed` property is `false`, and the `license_code` property is empty, indicating that the appliance has not yet been activated with a valid license code. The appliance can be managed through the Management API whilst in this state, but any attempt to use the Speech API to transcribe speech will return a `not authorised` error with the reason "You are not authorised to perform this action: License is invalid".

### Example Response (licensed)

If the appliance *has* been licensed then you will see a return like this:

```
{
  "licensed": true,
  "product": "103",
  "subscription_expiry": "2021-10-16T12:26:37Z",
  "status": "3 License status concurrent license",
  "message": "",
  "transcription_minutes_allowed": "60",
  "license_type": "6 License is concurrent subscription",
  "activation_type": "1 License was activated online",
  "transcription_secs_allowed": 3600,
  "transcription_secs_allocated": 1800000,
  "connected": true,
  "customer_id": 4949,
  "license_code": "494913586168289666"
}
```

This shows that the appliance has been licensed with code 494913586168289666. The license is due to expire on the 16th October 2021. 500 hours (1800000 seconds) have been allocated for use. A local cache of 1 hour (3600 seconds) is maintained on the appliance to allow it to operate offline for limited periods of time; both of these values will be defined depending on your license requirements.

## Removing a License

If you no longer wish to use the appliance, or you need to perform an upgrade to a newer version then you should first remove the license before powering down the virtual appliance. This will return any unused audio hours that you may have cached on the appliance. You must be online to perform this action. Once the license is removed the appliance will no longer be able to transcribe speech. You use an HTTP DELETE to remove the license:

```
curl -L -X DELETE "http://${APPLIANCE_HOST}:8080/v1/management/license" \
  -H 'Accept: application/json'
| jq
```

The response shows the number of unused seconds that were returned for use by future activations, for instance:

```
{
  "transcription_secs_returned": 60780
}
```

## Using a Proxy Server

The appliance needs to talk to the cloud licensing service using HTTP (TCP port 80). The license credentials are encrypted over this link. If the network the appliance is installed on uses a proxy server to access the Internet, then you will need to configure the appliance to use that proxy. This is a pre-requisite before attempting to apply the license.

To configure the appliance, use a POST to the `/v1/management/license/network` endpoint:

```
curl -X POST "http://${APPLIANCE_HOST}:8080/v1/management/license/network" \
-H 'Accept: application/json' \
-H 'Content-Type: application/json' \
-d '{ "http_configuration": {
    "ip": "${PROXY_HOST}",
    "port": "${PROXY_PORT}",
    "user": "${PROXY_USERNAME}",
    "password": "${PROXY_PASSWORD}"
  }
}' \
| jq
```

Where `${PROXY_HOST}` is the IP address or hostname of your proxy server, and `${PROXY_PORT}` is the port number it uses. If you use username and password authentication for the proxy server, then these also need to be specified using the `${PROXY_USERNAME}` and `${PROXY_PASSWORD}` options. If the proxy server does not require authentication then they should be left out.

Once you have configured the Real-time Virtual Appliance to use your proxy server you will be able to activate the appliance – see the [section above](#) on how to apply a new license.

## Offline License Activation

This section explains how to license your appliance if it is in a completely offline environment (ie. there is no route to the Internet), *and* you are not able to connect to the Internet even during initial activation of the license. If this is the case then you need to generate an *activation certificate*, send this to [support@speechmatics.com](mailto:support@speechmatics.com), and then apply the *license certificate* that is sent back. Follow the steps in this section to do this.

It is recommended as part of best practice where the appliance cannot connect to the internet to activate the license offline using the method below.

### Generating an Activation Certificate

The process is similar to online activation: you will receive a license code from Speechmatics support. However, additional steps are required, and different endpoints on the Management API are used.

Offline activations require a POST request to the `/v1/management/license/offlineactivation` endpoint:

- **Method:** POST
- **URL:**  
`http://${APPLIANCE_HOST}:8080/v1/management/license/offlineactivation`
- **Body format:** JSON
- **Body:** `license`, `username`, `email_address`, `company_name`

**Note:** make sure when applying the license, that all the appliance services are running; otherwise the activation will fail.

The example below shows how to make an offline activation call using the a POST request to the Management REST API:

```
curl -L -X POST "http://${APPLIANCE_HOST}:8080/v1/management/license/offlineactivation" \
-H 'Accept: application/json' \
-H 'Content-Type: application/json' \
-d '{
  "license": "494853989762904933",
  "username": "Fiona Kelly",
  "email_address": "fjk@example.com"
}' \
| jq
```

### Sending the Activation Certificate to Speechmatics

The response contains a long string of alphanumeric characters. This is the activation certificate. You should save this as a text file and send to Speechmatics support [support@speechmatics.com](mailto:support@speechmatics.com), along with the license code that you used.

Once this has been done, the support team will use the activation certificate to generate a license certificate. They will then send this back to you by reply of email.

**Note:** You should make sure, when in the process of applying an offline license, that you do not reboot the appliance between generating the activation certificate and applying the license certificate. The reason being that the computer identifier which is used as a component of the certificates will be different between reboots.

## Applying the License Certificate

Once you have been sent the license certificate by Speechmatics support you can use this to activate the appliance by making a **PUT** request to the `/v1/management/license/offlineactivation` endpoint:

- **Method:** PUT
- **URL:**  
`http://${APPLIANCE_HOST}:8080/v1/management/license/offlineactivation`
- **Body format:** JSON
- **Body:** `license , certificate`

Where `certificate` is the license certificate that Speechmatics support have provided to you, and `license` is the original activation license code you received.

**Note:** It is important that you use the **PUT** method to send the license certificate to the appliance. If you use the **POST** method you will end up with an error and will need to repeat the license activation process.

The example below shows how to make this activation call:

```
curl -L -X PUT "http://${APPLIANCE_HOST}:8080/v1/management/license/offlineactivation" \  
-H 'Accept: application/json' \  
-H 'Content-Type: application/json' \  
-d '{  
  "license": "494853989762904933",  
  "certificate": "7bc3c6684...f46d9781cfbae3c8129505e"  
}' \  
| jq
```

**Note:** The certificate is a very long string of alphanumeric characters. We shorten it here for brevity.

Once the appliance has been licensed in this way you will see a return like this:

```
{  
  "licensed": true,  
  "product": "100",  
  "subscription_expiry": "2019-04-07T14:35:09Z",  
  "status": "3 License status concurrent license",  
  "message": "",  
  "transcription_minutes_allowed": "2999",  
  "license_type": "6 License is concurrent subscription",  
  "activation_type": "1 License was activated online",  
  "transcription_secs_allowed": 179998,  
  "transcription_secs_allocated": 2147483647,  
  "connected": false,  
  "customer_id": 4948,  
  "license_code": "494853989762904933",  
  "computer_id": "bc8e7e32-7cc6-4292-83e5-555c726ae8d8",  
  "error_message": ""  
}
```

## Running an Appliance Offline

If you have activated your license offline, and have already processed the activation certificate using the steps described above, your virtual appliance will go into offline mode automatically, and will no longer need to talk to an external server.

If you have activated the appliance online, but are then expecting to run it in a completely 'dark' network (that is, your appliance will not have any route to the Internet), then after you have activated the license online, then, *after* licensing you should put it into an *offline* mode. This is a quicker way of activating the license where the appliance can be online for license activation only. Running the appliance without an internet connection Activating and deactivating your license must still be done when your appliance is online. We recommend that you run licensing in offline mode if the appliance will not have any route to the internet as best practice.

Enabling offline mode can be done through the Management API by setting the `offline` state to `true`, like this:

```
curl -X POST "http://${APPLIANCE_HOST}:8080/v1/management/license/offlinemode" \
-H 'Accept: application/json' \
-H 'Content-Type: application/json' \
-d '{ "offline": true}' \
| jq
```

**Note:** You can only use this mode if your license allows offline mode to be used. If offline mode is not supported by your license you will see an error message returned:

```
{
  "error": "Error happened in the license_management service: ModelException: Unknown Rpc Error.
Status code=StatusCode.UNKNOWN: <_Rendezvous of RPC that terminated with:\n\tstatus =
StatusCode.UNKNOWN\n\tetails = \"Offline mode not enabled for this
license\"\n\tdebug_error_string = \"
{\\\"created\\\":\\\"@1544116759.516901870\\\",\\\"description\\\":\\\"Error received from
peer\\\",\\\"file\\\":\\\"src/core/lib/surface/call.cc\\\",\\\"file_line\\\":1099,\\\"grpc_message\\\":\\\"Offline
mode not enabled for this license\\\",\\\"grpc_status\\\":2}\\\"\\n>\",
  \"code\": 13
}
```

**Note:** Putting the appliance into offline mode like this stops any online license checking, which means that any changes to your license will not be picked up until you disable offline mode and go back online.

If you think that you may need to run your appliance in offline mode then please contact [support@speechmatics.com](mailto:support@speechmatics.com).

## Licensing Troubleshooting

### Receiving Updates to a License

In the case where Speechmatics has provided an update to an already activated license (for example, to increase the number of license activations, or the amount of audio hours allowed), then you will need to restart the services on your appliance, and ensure that the appliance is online when you do so, in order for the license updates to take effect.

### Invalid License

If you attempt to activate your virtual appliance by applying a license code and you see this error message, then it means that the license code is invalid.

```
Input exception: Not activated
```

If this occurs please contact [support@speechmatics.com](mailto:support@speechmatics.com), sending back the full output from the activate license POST request.

### Appliance Offline

If you see the following error when attempting to activate the appliance:



```
Input exception: Not activated - Cannot activate when offline
```

Then the appliance is unable to contact the cloud license service. Make sure that you are able to reach the licensing service by pinging the following hostname: my.nalpeiron.com. If you use a proxy server to connect to the Internet, ensure that the appliance has been configured to use the proxy (making sure that you specify at least the IP address or hostname of the proxy, and the correct port number). Look in the logs of your proxy server to check that the appliance is using the correct proxy server. To check whether you are running online or not you can run the following:

```
curl -L -X GET "http://${APPLIANCE_HOST}:8080/v1/management/license" \
-H 'Accept: application/json' \
| jq '.connected'
```

### Offline Activation Error

If you are carrying out activation offline and you use the **POST** method rather than a **PUT** to send the license certificate to the appliance you will see a `"code": 13` error with a description message of `"Unable to request activation certificate: -1121"`. The full error response will look like this:

```
{
  "error": "Error happened in the license_management service: ModelException: Unknown Rpc Error.
Status code=StatusCode.UNKNOWN: <_Rendezvous of RPC that terminated with:\n\tstatus =
StatusCode.UNKNOWN\n\tetails = \"Unable to request activation certificate:
-1121\"\n\tdebug_error_string = \"{\n\t\t\"created\": \"@1738245024.927287214\", \"description\": \"Error
received from
peer\", \"file\": \"src/core/lib/surface/call.cc\", \"file_line\": 1036, \"grpc_message\": \"Unable to
request activation certificate: -1121\", \"grpc_status\": 2}\n\t\",
  \"code\": 13
}
```

### Unable to Delete License when Offline

If you have activated a license online, but you then go into offline mode, you will get an error if you attempt to remove the license (`DELETE /v1/management/license`).

```
{
  "error": "Error happened in the license_management service: ModelException: Unknown Rpc Error.
Status code=StatusCode.UNKNOWN: <_Rendezvous of RPC that terminated with:\n\tstatus =
StatusCode.UNKNOWN\n\tetails = \"Cannot remove license in offline mode\"\n\tdebug_error_string =
\"{\n\t\t\"created\": \"@1738245024.927287214\", \"description\": \"Error received from
peer\", \"file\": \"src/core/lib/surface/call.cc\", \"file_line\": 1036, \"grpc_message\": \"Cannot
remove license in offline mode\", \"grpc_status\": 2}\n\t\",
  \"code\": 13
}
```

In order to remove the license you will need to exit offline mode, and make sure that there is a route to the Internet before trying to remove the license.

### Virtual appliance is offline message when port 80 is blocked

Communication with the cloud license service relies on port 80 being open. If there is a firewall in your network that blocks port 80 then you will see error messages like this when attempting to make a licensing call:

```
{
  "error": "Error happened in the license_management service: ModelException: Unknown Rpc Error.
Status code=StatusCode.CANCELLED: <_Rendezvous of RPC that terminated with:\n\tstatus =
StatusCode.CANCELLED\n\tetails = \"Received RST_STREAM with error code 8\"\n\tdebug_error_string
= \"{\n\t\t\"created\": \"@1546953904.251876385\", \"description\": \"Error received from
peer\", \"file\": \"src/core/lib/surface/call.cc\", \"file_line\": 1099, \"grpc_message\": \"Received
```

```
RST_STREAM with error code 8\", \"grpc_status\":1)\"\\n>\",
  \"code\": 13
}
```

In such cases make sure that port 80 is open, or use configure the appliance to use a proxy server.

## Verify and Go (Real-time)

This section explains how to verify the correct operation of the Real-time Virtual Appliance using the Websockets Speech API.

The first time that the Real-time Virtual Appliance is started up there are no **persistent workers** configured. This means that workers will be spun up dynamically to the limit of the maxworkers limit. If you want to pre-allocate workers so that they are ready for incoming streams, you can select a language to configure as a persistent worker. You can find out the list of available languages on your appliance using the REST API:

- **Method:** GET
- **URL:**  
`http://${APPLIANCE_HOST}:8080/v1/management/persistentworkers`

This will return as JSON output the list of persistent workers (by language code) and the number of instances; initially they will all be zero. The list of supported languages are available on the Speechmatics website <https://www.speechmatics.com/language-support/>, or you can consult the release notes.

Use the Management API to set `persistent_workers` with a count of at least 1 and the language code to use. For example, to set French as a persistent worker, use the following method:

- **Method:** POST
- **URL:**  
`http://${APPLIANCE_HOST}:8080/v1/management/persistentworkers`
- **Body Format:** JSON
- **Body:** `{ "persistent_workers": [ { "count": "1", "id": "fr" } ] }`

This will return an updated list of persistent workers with entry for French (fr) updated to 1.

```
curl -s -L -X POST "http://${APPLIANCE_HOST}:8080/v1/management/persistentworkers" \
  -H 'Accept: application/json' \
  -d '{ "persistent_workers":
    [ { "count": "1", "id": "fr" } ]
  }'
```

## Verify the service

Check that all the Speechmatics services within the appliance are up and running before trying to open a WebSockets connection. The Management API can be used for this.

```
curl -s -L -X GET "http://${APPLIANCE_HOST}:8080/v1/management/services" \
  -H 'Accept: application/json'
```

## Go!

The Real-time Virtual Appliance is now ready to use.

The Speech API Guide provides details of how to use WebSockets to stream audio to the Speechmatics engine in real time and obtain transcripts. For all WebSocket communication, ensure that the IP address identified in the steps above is used.

## SSL Configuration

When the appliance is imported it contains a default self-signed certificate, so you can use HTTPS to access the appliance via the Management, Monitoring and Speech APIs. However, we recommend replacing this default SSL certificate with your own certificate, signed by your organisation or a trusted third-party certificate authority (CA).

## Default behaviour

By default, our appliances allow connections over HTTP. The services on the appliance expose several ports for HTTP access, such as 8080 for the management API and 3000 for the monitoring API.

Since version 3.4.0 of the appliances, we also support HTTPS access to these services over port 443. To use HTTPS simply change the protocol used for API calls from 'http' to 'https', and remove the port from the URL. If you are copying the examples from this document you can set the `$APPLIANCE_HOST` environment variable like this: `export APPLIANCE_HOST=localhost`.

## Management API Examples

```
curl -L -X GET "http://${APPLIANCE_HOST}:8080/v1/management/services" \  
-H 'Accept: application/json'
```

To modify this to use a secure API call, change `http://` to `https://` and remove the port number `:8080` from the URL:

```
curl -L -X GET "https://${APPLIANCE_HOST}/v1/management/services" \  
-H 'Accept: application/json'
```

**Note:** If you are using a self-signed certificate (your own, or the Speechmatics certificate that is used by default), then you will see a warning like this when using the above curl command:

```
curl: (60) SSL certificate problem: self signed certificate
```

**Warning:** The default SSL certificate on the appliance is a self-signed certificate created by Speechmatics, which is not signed by any certificate authority. Your HTTP client or web browser may warn that this is insecure. This warning can be suppressed, for example with cURL by adding the `--insecure` flag, however customers who are serious about security should not be using the self-signed certificate. We recommend uploading your own SSL certificate to the appliance. Instructions for doing this can be found below.

**Important:** We have added `--insecure` to some of them cURL examples in this document so that the command trusts the self signed certificate. You won't need this option once you've uploaded your own certificate and configured your own system to trust it.

## Monitoring API Example

With access to the Monitoring API (available on port 3000 if you are using HTTP) you will need to prefix the endpoint with `/monitor`. For example:

```
curl --insecure -L -X GET "https://${APPLIANCE_HOST}/monitor/api/3/mem"
```

## Speech API Example

Access to the REST Speech API (available on port 8082 using HTTP), is also possible via HTTPS:

```
curl --insecure -L -X GET "https://${APPLIANCE_HOST}/v1.0/user/1/jobs/"
```

## Using your own SSL certificate and private key

To use your own SSL certificate you'll need to upload your *certificate* file as well as the associated *private key* file.

- The **private key** file normally has a '.key' extension and should look similar to the example below.

```
-----BEGIN RSA PRIVATE KEY-----
xqgLwi4gJ9+9Qkavpk3WpPFTTYUfVrCJNviKEn5wAltutqLQkRTcxJtrEk8trKI
fCxeZo35yVhYmDGUIuAdAcPRTPj0XZkXQRhkITmD8TYMc/sVlJpFr+TAssGzute8
... 21 lines redacted ...
+bLv4aqI9tZrwpyeziaOuyQRhYodpAjhCyCFMkJjY59BKv/cqMHx8FPDQmaZ9Xs0
SmE9JAKnDgF5yLHm1Q6WZ1/L/M4SkgIqEglF7ifLd5M3wskpmHia6/f8Fa2KwbBJ
-----END RSA PRIVATE KEY-----
```

**Note:** We do not currently support encrypted/password protected private key files.

- The **certificate** file should be PEM encoded and normally has a '.crt' or '.pem' extension. It should look similar to this:

```
-----BEGIN CERTIFICATE-----
MIIGuzCCBaOgAwIBAgIIIHlfyznYUA8wDQYJKoZIhvcNAQELBQAwgbcwCzAJBgNV
BAYTA1VTMRAwDgYDVQQIEwdBcm16b25hMRMwEQYDVQQHEwptY290dHNkYXlMR0w
... 32 lines redacted ...
P4LMbjCA4mqQvlipeSAN1E4OrFL47zLcy+H9M0+Rw2CUiwL8QZFq+TAiIZ34tC
UVCh52xpB9/BhO++QbGd1zObqDhcGEg8pJpJIycej9t4GN1eqNSudn0ibsQWew8=
-----END CERTIFICATE-----
```

Both files should be in [PKCS8](#) format. If you have to upload a certificate chain, then the file you upload should contain the individual certificates concatenated, with your organisation's certificate first.

## Uploading the certificate and key to the appliance

To upload your own certificate to the appliance you will need to make a POST request to the `/v1/security/sslcertificate` endpoint. This can be done using an HTTP client on the command line or with the management interface in a browser.

With the example shown here set `APPLIANCE_HOST` as appropriate (e.g. `export APPLIANCE_HOST=localhost` if your appliance is running locally):

```
curl --insecure -X POST "https://${APPLIANCE_HOST}/v1/security/sslcertificate" \
-F "keyfile=@appliance.key" -F "certfile=@appliance.crt"
```

**Warning:** Do not upload these files over HTTP, or you risk leaking the private key for your certificate.

If the upload succeeds then you should receive an HTTP 200 response with a success message:

```
{
  "success": true,
  "message": "certificate and private_key applied successfully"
}
```

Be aware that setting a new certificate will cause the web server in the appliance to restart which can take around five seconds. During this period, requests will still be served, however the old certificate will be used. Existing connections such as job uploads or WebSocket streams will not be interrupted.

You can check the certificate on the appliance by using the `openssl` tool:

```
$ openssl s_client -connect ${APPLIANCE_HOST}:443
```

## Disabling HTTP access

If desired, HTTP access may be disabled, which will cause any requests to the appliance using HTTP to fail. To do this, make a POST request to the `/v1/security/insecureports` endpoint, with a JSON body containing

```
{"enable_insecure_ports": false}:
```

```
curl -X POST "https://${APPLIANCE_HOST}/v1/security/insecureports" \  
-H "Content-Type: application/json" \  
-d "{ \"enable_insecure_ports\": false}"
```

If the request succeeded then you should receive an HTTP 200 response. The web server in the appliance will take around five seconds to restart. Now, when attempting to make an HTTP request to the appliance you should see that no response is returned:

```
curl -X GET "http://${APPLIANCE_HOST}:8080/v1/management/services"  
  
curl: (52) Empty reply from server
```

## Enable Basic Authentication for Admin

An admin password can be set to enable [HTTP basic authentication](#) for an `admin` user. Note that **authentication is only enforced when using HTTPS**. If you set an admin password then you **must** also disable HTTP access as described in the previous section. If you do not do this then it will be possible for someone else to override the admin password by making an unauthorized HTTP request.

To set a password, make a POST request to the `/v1/security/adminpassword` endpoint. The username for basic auth is always `admin`.

```
curl -X POST "https://${APPLIANCE_HOST}/v1/security/adminpassword" \  
-H "Content-Type: application/json" \  
-d "{ \"password\": \"example\" }"  
  
{ "success": true, "message": "nginx_restart" }
```

If this request was successful then you should receive an HTTP 200 response with a success message. The web server in the appliance will take around five seconds to restart. All requests to HTTPS endpoints will now require a valid `Authorization` header as specified by [RFC7617](#). Unauthenticated requests will fail, for example:

```
$ curl -X GET "https://${APPLIANCE_HOST}/v1/management/services"  
<html>  
<head><title>401 Authorization Required</title></head>  
<body>  
<center><h1>401 Authorization Required</h1></center>  
<hr><center>nginx/1.17.6</center>  
</body>  
</html>
```

Authenticated requests should succeed. If you are using curl then the `--user` flag can be used to set the username and password (separated with a colon). If you're using the Management UI in a browser than a prompt will appear for a username and password.

```
$ curl --insecure -X GET --user "admin:example"  
"https://${APPLIANCE_HOST}/v1/management/services"
```

If you have disabled HTTP access then it should now be impossible to make requests to the appliance without knowing the admin password. Please be aware that plain HTTP access does **not** require the admin password, and should be disabled if you are using a password.

## FAQs

### How do I reset the SSL settings?

If you have made a mistake in your SSL configuration, it is possible to reset the appliance to it's default settings. This will return it to using the self-signed certificate from Speechmatics, and will delete any configured admin password. If you

have disabled HTTP access then you need to know the existing admin password in order to do this.

To do this, make a DELETE request to the `/v1/security/reset` endpoint:

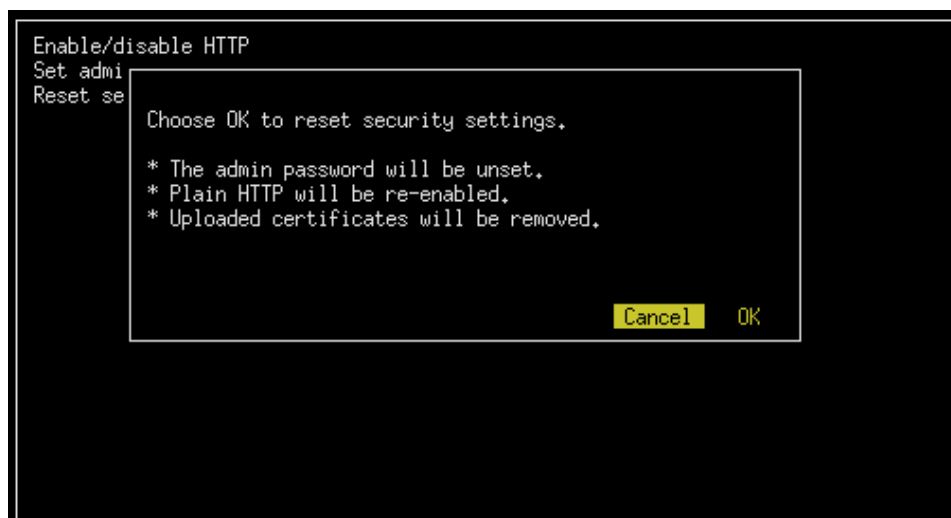
```
$ curl -X DELETE --user "admin:$PWD" "https://${APPLIANCE_HOST}/v1/security/reset"
{"success": true, "message": "nginx_restart"}
```

### What if I forget the admin password?

If you have forgotten the admin password you have set, and have disabled HTTP access to the appliance then it will not be possible to interact with the appliance over HTTP/HTTPS. Fortunately there is a way to reset the SSL configuration if you have direct access to the appliance's console (through the hypervisor that you use).

See the 'Administration -> Services -> Console for Advanced Troubleshooting' section for instructions on how to access the console.

Once you have opened the console open the 'Security' menu and select the 'Reset security' option to reset all security settings. It is also possible to toggle HTTP access and set the admin password using this interface.



### What versions of SSL/TLS do you support?

We support TLS 1.2 and TLS 1.3. We do not support earlier versions of TLS/SSL as these are considered weak. In general we would recommend you keep your client frameworks up to date with the latest security patches and try to use the strictest TLS configuration that you can.

#### What cipher suites do you support?

For TLS 1.3 we support the following cipher suites that are considered strong (in server-preferred order):

- TLS\_AES\_256\_GCM\_SHA384
- TLS\_CHACHA20\_POLY1305\_SHA256
- TLS\_AES\_128\_GCM\_SHA256

For TLS 1.2 we support the following cipher suites that are considered strong (in server-preferred order):

- TLS\_ECDHE\_RSA\_WITH\_AES\_256\_GCM\_SHA384
- TLS\_ECDHE\_RSA\_WITH\_CHACHA20\_POLY1305\_SHA256
- TLS\_ECDHE\_RSA\_WITH\_ARIA\_256\_GCM\_SHA384
- TLS\_ECDHE\_RSA\_WITH\_AES\_128\_GCM\_SHA256
- TLS\_ECDHE\_RSA\_WITH\_ARIA\_128\_GCM\_SHA256

Other cipher suites are available for TLS 1.2, but they are considered to be weak. Our recommendation is that you select one of the above cipher suites.

# Networking

## Network Requirements

When the virtual appliance is started for the first time it will automatically try to acquire an IP address using DHCP. If it is able to successfully acquire an address, it will be displayed on the VM console along with the fallback IP address: 10.10.10.2. However, if there is no DHCP server available on the network only the 10.10.10.2 IP address will be displayed.

The 10.10.10.2 address is a fallback address enabling communication with the virtual appliance when no DHCP services are available. This address should be used temporarily to set a static IP address if no DHCP is available. To do this, ensure that the client connecting to this address is on the same network by assigning it a suitable IP address (e.g. 10.10.10.3/24).

**Note:** The appliance uses three internal networks:

- docker\_gwbridge - 10.254.0.0/22
- ingress - 10.254.4.0/25
- docker0 - 10.254.4.128/25

You need to ensure that any network you use does not have an IP address conflict with anything in the range: 10.254.0.0 to 10.254.4.255.

## Configure Static IP

The virtual appliance can be configured to work on any IP network.

Setting a static IP requires three parameters: the IP address, subnet mask and default gateway. You set the static IP address like this:

```
curl -L -X POST "http://${APPLIANCE_HOST}:8080/v1/management/host/setipaddress" \
-H 'Accept: application/json' \
-H 'Content-Type: application/json' \
-d '{
  "address": "192.168.128.160",
  "netmask": "255.255.255.0",
  "gateway": "192.168.128.1"
}' \
| jq
```

**Note:** Once the POST is sent, the virtual appliance will automatically reboot. Check the console (or make an API call) to verify the new IP address has taken affect.

## Configure DHCP

You can also change back to using DHCP. Before undertaking this, ensure the network the virtual appliance is on has DHCP enabled.

```
curl -L -X POST "http://${APPLIANCE_HOST}:8080/v1/management/host/setdhcp" \
-H 'Accept: application/json'
```

**NOTE:** once submitted, the virtual appliance will automatically reboot. Check the console to verify the new IP address has taken affect.

## Firewall Ports

There are several firewall rules that may need to be enabled to ensure the communication can be made to the virtual appliance:

- 8080/TCP - Used for the Management API to manage the virtual appliance
- 3000/TCP - Monitoring API
- 8082/TCP - Speech API for submitting jobs (Batch Appliance only)
- 9000/TCP - WebSockets Speech API for submitting jobs (Realtime Appliance only)
- 443/TCP - HTTPS access to the above APIs

## Using Proxies

If the network that you are deploying your appliance into does not have a direct route to the Internet, you may need to use a proxy server in order to talk to the cloud-based license service. See the relevant section in Licensing (below) for details on how to set this up.

# Virtual Appliance Scaling

## Real-time Virtual Appliance Scaling

This section explains how to scale the Real-time Virtual Appliance, and gives advice on how to make sure you've allocated enough resources for your workload.

### Worker Limits

The number of concurrent workers can be restricted using the Management API. This can be used to ensure that the system resources do not get exhausted by clients starting more sessions than expected. The maximum number of concurrent workers is set for the entire system, irrespective of which language packs are being used. The default number of maximum concurrent workers is 1.

### View Maximum Workers

Use a GET request to the `maxworkers` endpoint to view the maximum number of workers:

```
curl -L -X GET 'http://${APPLIANCE_HOST}:8080/v1/management/maxworkers' \
  -H 'Accept: application/json' \
  | jq
```

This shows the maximum number of workers that can run concurrently on the appliance. If more sessions are opened by clients using the Speech API then you will receive the job error: `No worker can be scheduled because the service is at capacity.`

### Setting Maximum Workers

Before changing the maximum number of concurrent workers for real-time transcription, it is important that the virtual appliance has enough system resources (CPU and RAM) to support the new requirement (see the Real-time Virtual Appliance system requirements). This example shows how to set the maximum number of concurrent workers to 5:

```
curl -L -X POST 'http://${APPLIANCE_HOST}:8080/v1/management/maxworkers' \
  -H 'Accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{ "count": "5" }'
```

As a rule of thumb, each concurrent worker will require 1 vCPU and up to 2GB RAM.

## Batch Virtual Appliance Scaling

This section explains how to scale the Batch Virtual Appliance, and gives advice on how to make sure you've allocated enough resources for your workload.

### Worker Limits



The number of concurrent workers (jobs) can be restricted using the Management API. This can be used to ensure that the system resources do not get exhausted by clients starting more transcriptions than expected. The maximum number of concurrent workers is set for the entire system, irrespective of which language packs are being used. The default number of maximum concurrent workers is 1.

## View Maximum Workers

Use a GET request to the maxworkers endpoint to view the maximum number of workers:

```
curl -L -X GET 'http://${APPLIANCE_HOST}:8080/v1/management/maxworkers' \
  -H 'Accept: application/json' \
  | jq
```

The response will indicate the maximum number of workers that can run concurrently on the appliance. If more jobs are submitted by clients using the Speech API then these will be queued up and processed once there is spare capacity on the appliance.

## Setting Maximum Workers

Before changing the maximum number of concurrent workers, it is important that the virtual appliance has enough system resources (CPU and RAM) to support the new requirement (see the Batch Virtual Appliance system requirements).

This example shows how to set the maximum number of concurrent workers to 5:

```
curl -L -X POST 'http://${APPLIANCE_HOST}:8080/v1/management/maxworkers' \
  -H 'Accept: application/json' \
  -H 'Content-Type: application/json' \
  -d'{ "count": "5" }'
```

As a rule of thumb, each concurrent worker will require 1 vCPU and up to 5GB of RAM (depending on the quality of the audio).

If the number of jobs submitted exceeds the maximum number of concurrent workers then jobs will start to be queued, and the real-time factor (RTF) will increase, meaning you will wait longer for your transcripts to be made available.

# Monitoring

Appliance resources can be monitored at a system-wide level. Exhaustion of any of the resources can have a negative impact on the speed of the transcription.

The following resources that can be monitored:

Resource ID (rID)	Description
cpu	Provides the CPU usage across all the vCPU assigned
mem	Provides the total RAM usage of the appliance

Here is an example GET request for the `mem` (RAM) resource:

```
curl -L -X GET "http://${APPLIANCE_HOST}:8080/v1/management/resource/mem" \
  -H 'Accept: application/json' \
  | jq
```

Here is an example response:

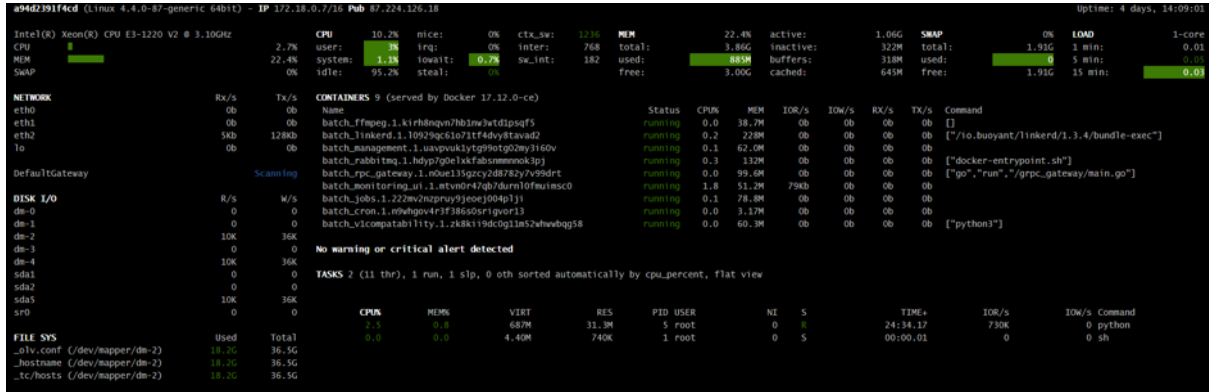
```
{
  "rId": "mem",
```

```

"percentage": 13.4
}

```

For advanced monitoring, a utility called [Glances](#) is available that runs on TCP port 3000. It allows real-time resource stats to be monitored on the Real-time Virtual Appliance. The easiest way to access this is via a web browser using the link `http://${APPLIANCE_HOST}:3000/` in the address bar.



It is also possible to access the Glances API using XML-RPC or HTTP REST (for JSON output), for example:

```

curl -L -X GET "http://${APPLIANCE_HOST}:3000/api/3/mem/percent" \
-H 'Accept: application/json' \
| jq

```

For more information on the HTTP REST interface, consult the [Glances documentation](#).

## Services

The virtual appliance has internal services that are required for operation.

There are system-wide services, and services specific to transcription workers for a given language.

## Batch Virtual Appliance

For the Batch Virtual Appliance, this table lists the services:

Service Name (Begins with)	Description	Required Status
batch_bja...	V2 REST API	Running.
batch_rpc_gateway...	RPC endpoint	Running
batch_license...	Licensing service	Running
batch_linkerd...	Internal Networking	Running
batch_management...	Management functions	Running
batch_ba_worker...	Job Queue management	Running
batch_monitoring_ui...	Monitoring Web GUI	Running
batch_batch-cron...	Completed job clean-up	Running
batch_v1compatibility...	V1 REST API	Running
jobs...	Used to perform ASR and transcription	Running

batch_swaggerui...	Swagger UI for certain APIs	Running
batch_nginxlb...	HTTP gateway	Running
batch_postgres...	Jobs Database	Running

The service will always have a current state, these states include:

Service Status	Description
running	Service has started and is running
created	Service is in the process of starting
exited	Service has stopped and is no longer running

## Service status

This can be used to ensure all services have the required status to operate (see table above). Example: GET to list services and corresponding status:

```
curl -L -X GET 'http://${APPLIANCE_HOST}:8080/v1/management/services' \
-H 'Accept: application/json' \
| jq
```

If the appliance has been licensed then you will see a return like this (for the Batch Virtual Appliance):

```
{
  "service_status": [
    {
      "service": "job-50",
      "status": "running"
    },
    {
      "service": "batch_bja.1.qegys910pamsduryf9tujm2db",
      "status": "running"
    },
    {
      "service": "batch_swaggerui.1.0limj506dokkscu4mvy00gt70",
      "status": "running"
    },
    {
      "service": "batch_rpc_gateway.1.10aoi8f9cvkcko8s5jhrio8b6",
      "status": "running"
    },
    {
      "service": "batch_batch-cron.1.uahr5xz4edjx11fm06bflhthx",
      "status": "running"
    },
    {
      "service": "batch_vlcompatibility.1.5t9hbwk30zqt2cnx5xzjf9zkt",
      "status": "running"
    },
    {
      "service": "batch_nginxlb.1.p2mq6ho4k5hho180zkog2maej",
      "status": "running"
    },
    {
      "service": "batch_license.1.urx4qlzru7430lhv9669h9xxy",

```

```

    "status": "running"
  },
  {
    "service": "batch_management.1.5r92dvzwu0021g7mc9pb7qtg0",
    "status": "running"
  },
  {
    "service": "batch_postgres.1.yvef8y8g8tq8nt62bc6ow987z",
    "status": "running"
  },
  {
    "service": "batch_monitoring_ui.1.m29c6ne7621y6dapq5fjojxj3",
    "status": "running"
  },
  {
    "service": "batch_linkerd.1.30ng6rrqiar7fqgkb9tesn9uw",
    "status": "running"
  },
  {
    "service": "batch_ba_worker.1.yliwg0uynenv2jcno9x423brc",
    "status": "running"
  }
]
}

```

## Real-time Virtual Appliance

For the Real-time Virtual Appliance, this table lists the services:

Service Name (Begins with)	Description	Required Status
rt_rt-server...	Load-balancing handling job requests	Running
rt_linkerd...	Proxy	Running
rt_management...	MGMT API Calls	Running
appliance_autoscaler...	required only during OVA build	Exited
rt_redis...	Handles worker availability	Running
rt_rpc_gateway...	Internal service management	Running
rt_monitoring_ui...	Monitoring Web GUI	Running
rt_nginx...	Proxying requests	Running
rt_rt-janitor...	Completed job clean-up	Running
rt_license...	Licensing	Running
rt_autoscaler...	Used to perform ASR and transcription	Running

The service will always have a current state, these states include:

Service Status	Description
running	Service has started and is running
created	Service is in the process of starting
exited	Service has stopped and is no longer running

## Service status

```
curl -L -X GET 'http://${APPLIANCE_HOST}:8080/v1/management/services' \  
-H 'Accept: application/json' \  
| jq
```

This can be used to ensure all services have the required status. If successful you will see the following response

```
{  
  "service_status": [  
    {  
      "service": "rt_rt-server.1.jgwwfsybbxmdq8205dqdz2r4",  
      "status": "running"  
    },  
    {  
      "service": "rt_linkerd.1.tetkum9u3iowqn2w7lok2nfp",  
      "status": "running"  
    },  
    {  
      "service": "rt_management.1.wk2kse9inpaie5nnby57zgjck",  
      "status": "running"  
    },  
    {  
      "service": "appliance_autoscaler-bootstrap-task_run_f92039b26280",  
      "status": "exited"  
    },  
    {  
      "service": "rt_redis.1.osd52r5esip3cvpsa3bsyfa3o",  
      "status": "running"  
    },  
    {  
      "service": "rt_rpc_gateway.1.mhb1yk8i50qxqs50jmu573u2o",  
      "status": "running"  
    },  
    {  
      "service": "rt_monitoring_ui.1.qzir2168b0lzroej5khlgac0x",  
      "status": "running"  
    },  
    {  
      "service": "rt_nginxlb.1.z9uwrh458ttct6mg2iilcp427",  
      "status": "running"  
    },  
    {  
      "service": "rt_rt-janitor.1.leqrp4vre3eqg213uceye41zm",  
      "status": "running"  
    },  
    {  
      "service": "rt_license.1.jeop3k5hscque3vw9qo24jmtu",  
      "status": "running"  
    },  
    {  
      "service": "rt_autoscaler.1.jbpngc1rokzf7zs7i7r97uxij",  
      "status": "running"  
    }  
  ]  
}
```

## Service restart

**Note:** After a service is restarted it will have a random string identifier post fixed to its name.

If required for troubleshooting you may need to restart all the services. During the restart, all transcription will stop. The following command performs a service restart:

```
$ curl -X DELETE 'http://<APPLIANCE_HOST>:8080/v1/management/services' \  
-H 'Accept: application/json'
```

## Access Logs

The individual services on the system provide log files that can be collected to help with troubleshooting. The service name will need to be provided when retrieving logs. See above for instructions on how to view the names of the running services

The following parameters are available when accessing logs:

Name	Description	Required Status
name	Name of the service to collect the logs for	Required
count	Number of log lines wanted, defaults to 100; if all lines are to be returned set to -1	Optional

Example: GET to retrieve logs for batch\_monitoring\_ui service:

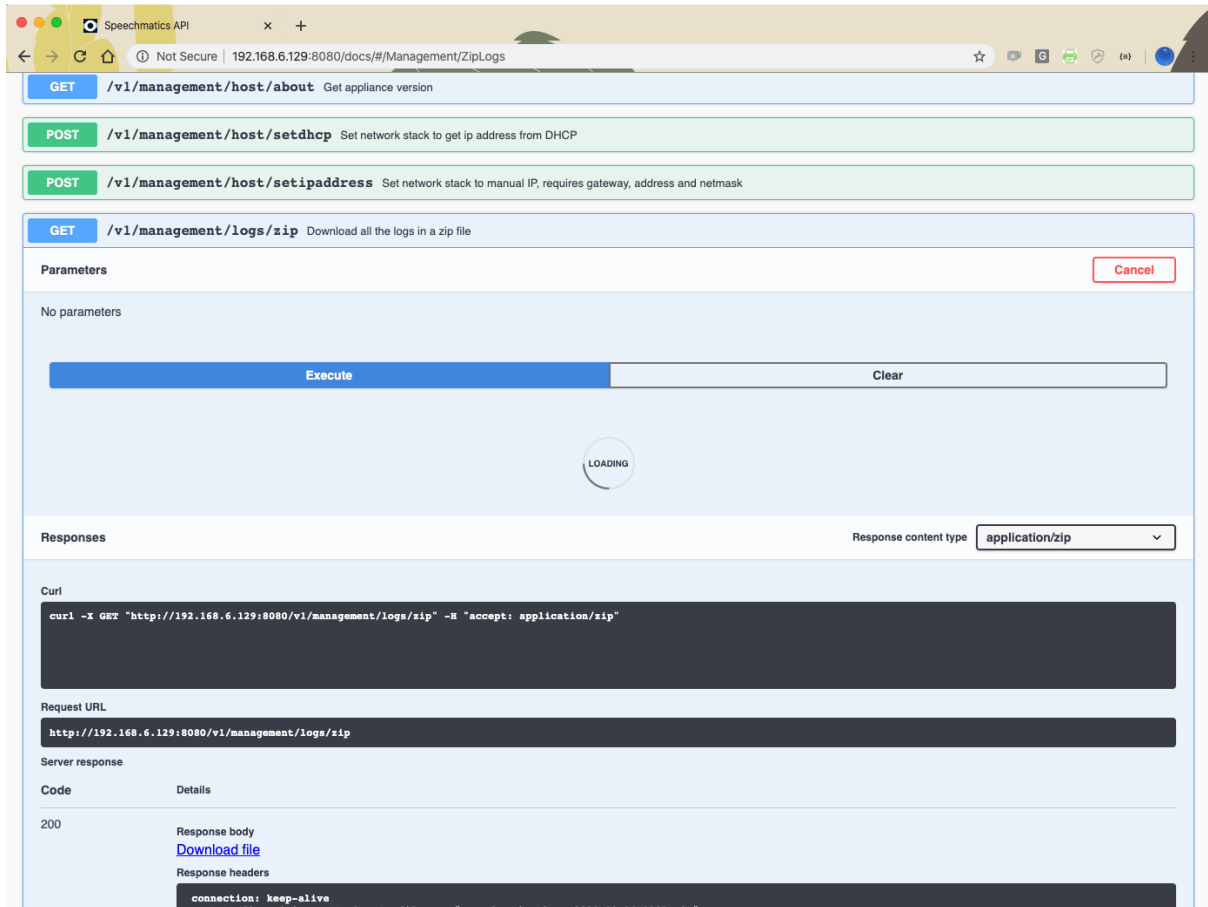
```
curl -L -X GET  
'http://${APPLIANCE_HOST}:8080/v1/management/logs/batch_monitoring_ui.1.mtvn0r47qb7durn10fmuimsc0'  
\  
-H 'Accept: application/json' \  
| jq -r '.log_lines'
```

If you want to download *all* the logs (in order to provide information for a support ticket for instance) as a ZIP file, then it is possible to do this using the following command:

```
curl -L -X GET 'http://${APPLIANCE_HOST}:8080/v1/management/logs/zip' \  
-H 'Accept: application/json' \  
-o ./speechmatics.zip
```

It is also possible to do this directly from the Swagger UI by entering in the following URL to your browser:

[http://\\${APPLIANCE\\_HOST}:8080/docs/#/Management/ZipLogs](http://${APPLIANCE_HOST}:8080/docs/#/Management/ZipLogs), and then clicking on the download link when the ZIP file is ready.



## System restart

If the virtual appliance becomes unresponsive, there might be a need to restart it. If this is the case, it's recommended that the system is restarted through the management API, like this:

```
curl -L -X DELETE 'http://${APPLIANCE_HOST}:8080/v1/management/reboot'
```

If the Management API is not available, then you should reboot the appliance from the hypervisor console. For further information on how to restart the virtual machine via the console, please follow the manufacturers advice.

## System shutdown

You may wish to shut down the appliance. If so, it's recommended that the system is shut down through the management API, like this:

```
curl -L -X DELETE 'http://${APPLIANCE_HOST}:8080/v1/management/shutdown'
```

If the Management API is not available, then you should shut down the appliance from the hypervisor console. For further information on how to shut down the virtual machine via the console, please follow the manufacturers advice.

## Troubleshooting

There may be times unexpected behavior is observed with the Real-time Virtual Appliance. If this is the case the following should be performed/checked:

- Check the license is valid (see licensing)
- Check the worker services are running

- Check the resources (CPU, memory & disk) to ensure they are not exhausted
- Restart all the services
- Restart the virtual appliance
- Collect logs and contact Speechmatics support: [support@speechmatics.com](mailto:support@speechmatics.com).

### Transcription job failure

If your transcription job fails with an `error` job status, more information can be found by looking at the logs from the `jobs` container (using the Management API, as previously described). Search the logs for the job id corresponding with your failure. If you see a `SoftTimeLimitExceeded` exception, this indicates that the job took longer than anticipated and as such was terminated. This is typically caused by poor VM performance, in particular slow disk IO operations (IOPS). If issues persist it may be necessary to improve the disk IO performance on the underlying host, or you may need to increase the RAM available to the VM such that memory caches can be taken advantage of. Please consult the section above on Host requirements, and the optimization advice specific to your hypervisor to ensure that you are not over-committing your compute resources.

### Illegal instruction errors

If jobs fail repeatedly and you see `Illegal instruction` errors in the log information for these jobs then it is likely that the host hardware you are running on does not support AVX. The host machine requirements for the Real-time Virtual Appliance must meet the following minimum specification: Intel® Xeon® CPU E5-2630 v4 (Sandy Bridge) 2.20GHz (or equivalent). This is important because these chipsets (and later ones) support Advanced Vector Extensions (AVX). The machine learning algorithms used by Speechmatics ASR require the performance optimizations that AVX provides.

You can check this by looking in the management log when the appliance starts up. If you see a message like this:

```
2019-03-26 16:53:07,136    sm_management.app    ERROR    Processor not AVX capable. Tensorflow
language models cannot run.
```

Then it means that your host's CPU does not support AVX, or that your hypervisor does not have AVX support.

A console is available to help with advanced troubleshooting in the event that the Management API is unavailable. It is described in the next section.

### AVX2 Warning

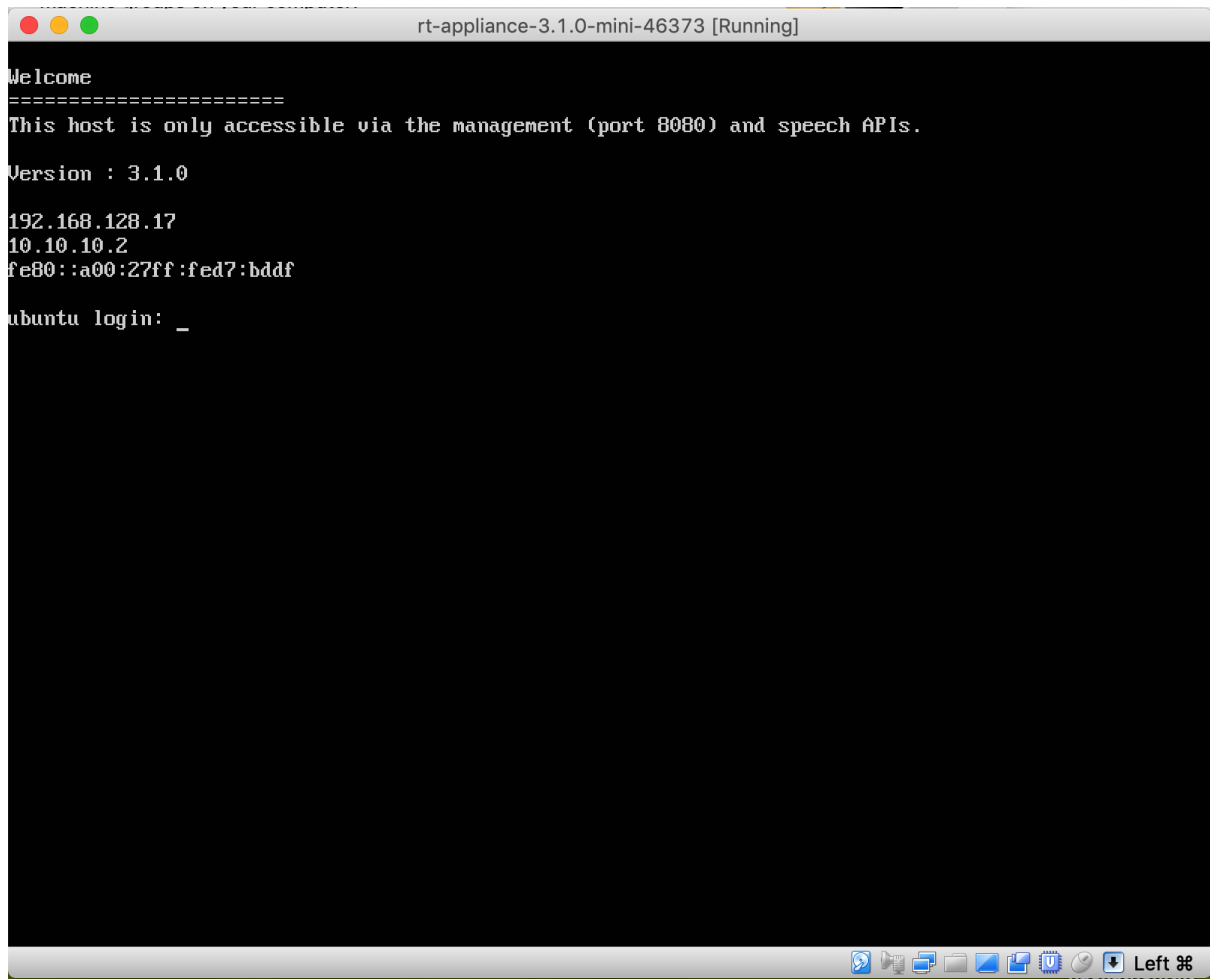
Speechmatics Appliance is optimised for running on hardware that supports the AVX2 flag. If you see the below message, your hardware is not optimised, and you may see slower performance of jobs

```
WARNING ([5.5.675~1-0c22]:SetupMathLibrary():asengine/asengine.cc:356) Unable to set CNR mode
to 10 (AVX2); falling back to 9. The transcription might be slower and/or use more CPU resource.
```

## Console for Advanced Troubleshooting

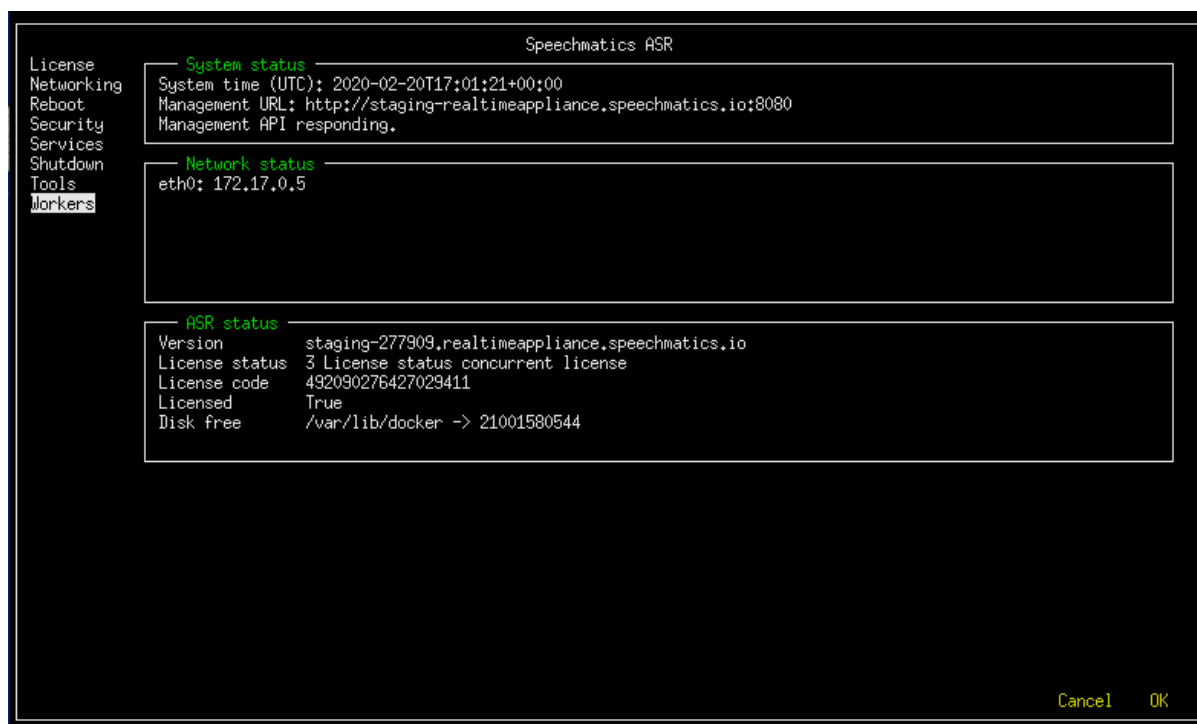
In the event that the Management API is unavailable (it is unresponsive, or there is no network connectivity) you can use the console to restore network connectivity, restart the appliance, or view information about services. To use this you need to use your hypervisor's GUI to access the logon screen for the appliance.





From this screen use the CTRL+ALT+F5 key combination to get to the console. Once you are in the console you have the following menu options available:

- License
- Networking
- Reboot
- Services
- Shutdown
- Tools
- Workers



The home screen shows high-level information about the appliance: IP addressing, software version and license status.

In the **System status** panel the **API responding** indicator shows the state of the Management API. **Network status** shows the IP address the appliance is currently configured with, and **ASR status** shows the license state and available storage space on the appliance.

In the event that you need to provide information to Speechmatics support you may be asked to connect to the console and provide this information. This section provides some tips on how to use the console to perform basic troubleshooting yourself.

**Note:** We recommend that you use the Management API for most troubleshooting tasks as it is easier to use. The console can be used in the event that the Management API is unavailable, but it does not provide all the features of the Management API.

## License

The [Licensing Troubleshooting](#) section provides detailed instructions on how to use the Management API to resolve common licensing issues. If you cannot use the Management API then you can still use console to check the license status and perform basic licensing steps.

## Networking

You can use the networking option to configure a static IP address, or use DHCP.

## Reboot and Shutdown

Reboot and Shutdown options exist to allow you to restart or shutdown the appliance from the console. You will be asked to select OK to confirm.

## Security

From this menu you can manage the security settings on the appliance, such as disabling HTTP access, changing the admin password for HTTP basic authentication, and resetting the SSL configuration.

## Services

From this menu you can access the list of services that are running on the appliance. Selecting a service shows the log entries for that service.

## Tools

This menu allows you to access a number of useful Unix utilities that can be used for advanced troubleshooting. In order to help progress a support ticket you may be asked to provide the output (ie. a screenshot) from running one of these commands.

## Workers

This allows you to view and change the maximum number of workers allowed to run concurrently.

# Security

The appliance is designed to be installed within your own security perimeter. It has its own firewall installed to only allow ingress to ports that are required for its management, monitoring and Speech APIs.

## Overview

The appliance uses a microservices architecture running on a customized Ubuntu machine. [AppArmor](#) default security policies are used to protect the OS and running applications on the appliance.

Data on the appliance (including audio and video data that is submitted via the Speech API, logs, and output transcripts) are encrypted on disk.

## Ports and Protocols

There are several firewall rules that may need to be enabled to ensure the communication can be made to the virtual appliance. If you setup HTTPS as described in the 'SSL Configuration' section of these docs then you only need to expose port 443.

Port/Protocol	Description
8080/TCP	Used for the Management API to manage the virtual appliance
3000/TCP	Monitoring (Glances)
8082/TCP	REST Speech API for batch ASR
9000/TCP	Websocket Speech API for real-time ASR
443/TCP	Used for HTTPS communication with all of the above services

## Custom Dictionary Cache

:::note Cache availability The custom dictionary cache is only available in the Real-time Virtual Appliance. :::

The Speechmatics Real-time Virtual Appliance includes a cache mechanism for custom dictionaries. By using this cache mechanism, the appliance is able to reduce the time needed for processing a custom dictionary before starting the recognition of an audio stream.

Once the Real-time Virtual Appliance has started a recognition session with a given custom dictionary, any subsequent streams with the identical custom dictionary sent to the appliance by any client will benefit from a reduced setup time. Transcription requests using an already cached custom dictionary will start recognition in less than 3 seconds if the custom dictionary is up to our recommended limit of 1000 words.

:::note V1 API When using V1 API the setup time spent processing a cached entry is about 3 seconds longer compared to V2 API. :::

## Size available

The size available for storing Custom Dictionary Cache entries depends on the variant of the Real-time Virtual Appliance.

Variant	Cache Space
nano	100MB
mini	150MB
midi	200MB
maxi	250MB
plus	300MB

## Size of cache entries

The entry size varies depending on the amount of information included in the custom dictionary. For guidance, the table below displays some values for different custom dictionaries.

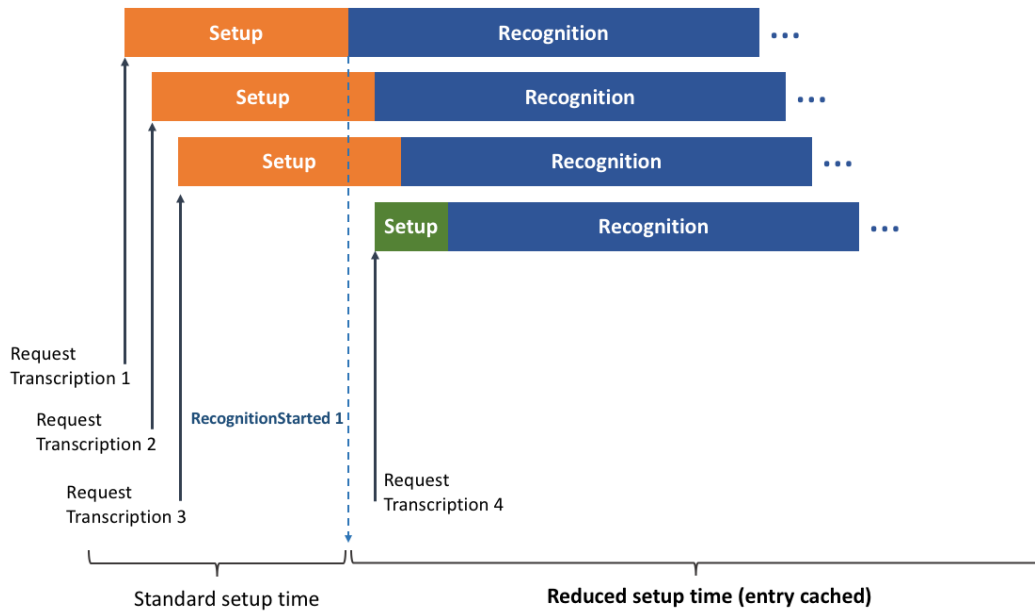
Custom Dictionary	Cache Used Bytes	Cache Entry Size
None (empty cache)	14KiB	NA
1000 words	120KiB	106KiB
1000 words + 1 sounds like each	188KiB	174KiB

note Size of empty cache The custom dictionary cache needs to keep certain metadata files in order to function. For this reason the used\_bytes reported will never be 0, even if the cache is not storing any entry. The amount of storage used by an empty cache is typically around 14KiB.

## Cache life cycle

When a custom dictionary is used for a transcription, it is automatically cached by the Real-time Virtual Appliance. This allows for a reduced setup time on any further transcriptions using the same custom dictionary. Custom dictionary cache is persisted in disk, making it available between reboots. If there is no space left in the cache for a new entry, entries are deleted in order of when they were last used when submitting a job.

Cache entries are ready to be consumed by any subsequent transcription request, from any client, after a `RecognitionStarted` message is emitted by the Real-time Appliance. For this reason, transcription requests made in parallel, each with the same custom dictionary that is new to the appliance, won't benefit from a reduced setup time.



A custom dictionary could be cached beforehand by requesting a transcription for an empty audio file. After receiving the `RecognitionStarted` message from the appliance, other requests using the same custom dictionary will benefit from a reduced startup time.

## Administering the Cache

When the Real-time Virtual Appliance is started for the first time the cache will be empty. The management API allows you to retrieve cache usage data and to purge the cache contents.

### View Cache Usage

Cache usage reports the maximum cache size and the used number of bytes in the cache.

In order to retrieve usage statistics, send a GET request to the `/v1/management/cache` endpoint:

```
curl -L -X GET http://${APPLIANCE_HOST}:8080/v1/management/cache \
-H 'Accept: application/json' \
| jq
```

Here is an example response:

```
{
  "total_bytes": "105188352",
  "used_bytes": "192512"
}
```

### Purge Cache Contents

It is possible to remove all contents in the cache.

In order to purge the cache contents, send a DELETE request to the `/v1/management/cache` endpoint:

```
curl -L -X DELETE http://${APPLIANCE_HOST}:8080/v1/management/cache \
-H 'Accept: application/json' \
| jq
```

Here is an example response:

```
{
  "confirmation": "Custom dictionary cache purged successfully",
  "usage": {
    "total_bytes": "105188352",
    "used_bytes": "14336"
  }
}
```

## Introduction

### Overview

The WebSocket Speech API allows communication from a client application over a WebSocket connection to the Speechmatics ASR engine (as implemented in the Real-time Virtual Appliance or the standalone Real-time Container). This connection provides the ability to convert a stream of audio into a transcript providing the words and timing information as the live audio is processed.

The WebSocket API can be used directly as described in this document; client libraries and frameworks that support WebSockets are available for most popular programming languages. Speechmatics provides reference Python libraries that can be used to wrap the WebSocket interface, and provide the ability to connect directly to a microphone or RTSP feed.

This document will guide you through setting up and configuring the WebSocket connection, and explain which features are available via the API.

### Terms

For the purposes of this guide the following terms are used.

Term	Description
Client	An application connecting to the Real-time Virtual Appliance using the Speech API. The client will provide audio containing speech, and process the transcripts received as a result.
Server	The Real-time Container or Appliance providing the ASR engine which processes human speech and returns transcripts in real-time.
Management API	The REST API that allows administrators to manage the virtual appliance over port 8080. To access the documentation you can use the following URI: <code>http://\${APPLIANCE_HOST}:8080/help/</code> , where <code>\${APPLIANCE_HOST}</code> is the IP address or hostname of your appliance.
Speech API	The WebSocket API that allows users to submit ASR jobs over server port 9000. The endpoint <code>wss://\${APPLIANCE_HOST}:9000/v2</code> is used for the Speech API. This is the API that is described in this document.
Real-time Container	A Docker container that provides real-time ASR transcription.
Real-time Virtual Appliance	An appliance (VM) that provides real-time ASR transcription.

## Getting Started

In order to use the Websocket Speech API you need access to a Real-time Virtual Appliance. See the Speechmatics Virtual Appliance Installation and Admin Guide on how to install, configure, and license the appliance.

## Input Formats

A wide variety of input sources are supported, including:

- Raw audio (microphone)
- The following file formats:
  - aac
  - amr
  - flac
  - m4a
  - mp3
  - mp4
  - mpg
  - ogg
  - wav

If you attempt to use an audio file format that is not supported, then you will see the following error message:

```
Error / job_error: An internal error happened while processing your file. Please check that your audio format is supported.
```

## Transcription Output Format

The transcript output format from the Speech API is JSON. It is described in detail in the [API Reference](#). There are two types of transcript that are provided: final transcripts and partial transcripts. Which one you decide to consume will depend on your use case, and your latency and accuracy requirements.

### Final transcripts

Final transcripts are sentences or phrases that are provided based on the Speechmatics ASR engine automatically determining the best point at which to provide the transcript to the client. These transcripts occur at irregular intervals. Once output, these transcripts are considered final, they will not be updated after output. The timing of the output is determined by Speechmatics based on the ASR algorithm. This is affected by pauses in speech and other parameters resulting in a latency between audio input and output of up to 10 seconds. This 10 second default can be changed with the `max_delay` property in `transcription_config` when starting the recognition session. Final transcripts provide the most accurate transcription.

### Partial transcripts

A partial transcript is a transcript that can be updated at a later point in time. It is believed to be correct at the time of output, but once further audio data is available, the Speechmatics ASR engine may use the additional context that is available to update parts of the transcript that have already been output. These transcripts are output as soon as any transcript is available, regardless of accuracy, and are therefore available at very low latency. These are the fastest way to consume transcripts but at the cost of needing to accept updates at a later point. Partial transcripts provide latency values between audio input and initial output of less than 1 second. This is the least accurate transcription method, but can be used in conjunction with the final transcripts to enable fast display of the transcript, adjusting over time. Partial transcripts must be explicitly enabled (using the `enable_partials` setting) in the config for the session, otherwise only final transcripts will be output.

## The WebSocket Protocol

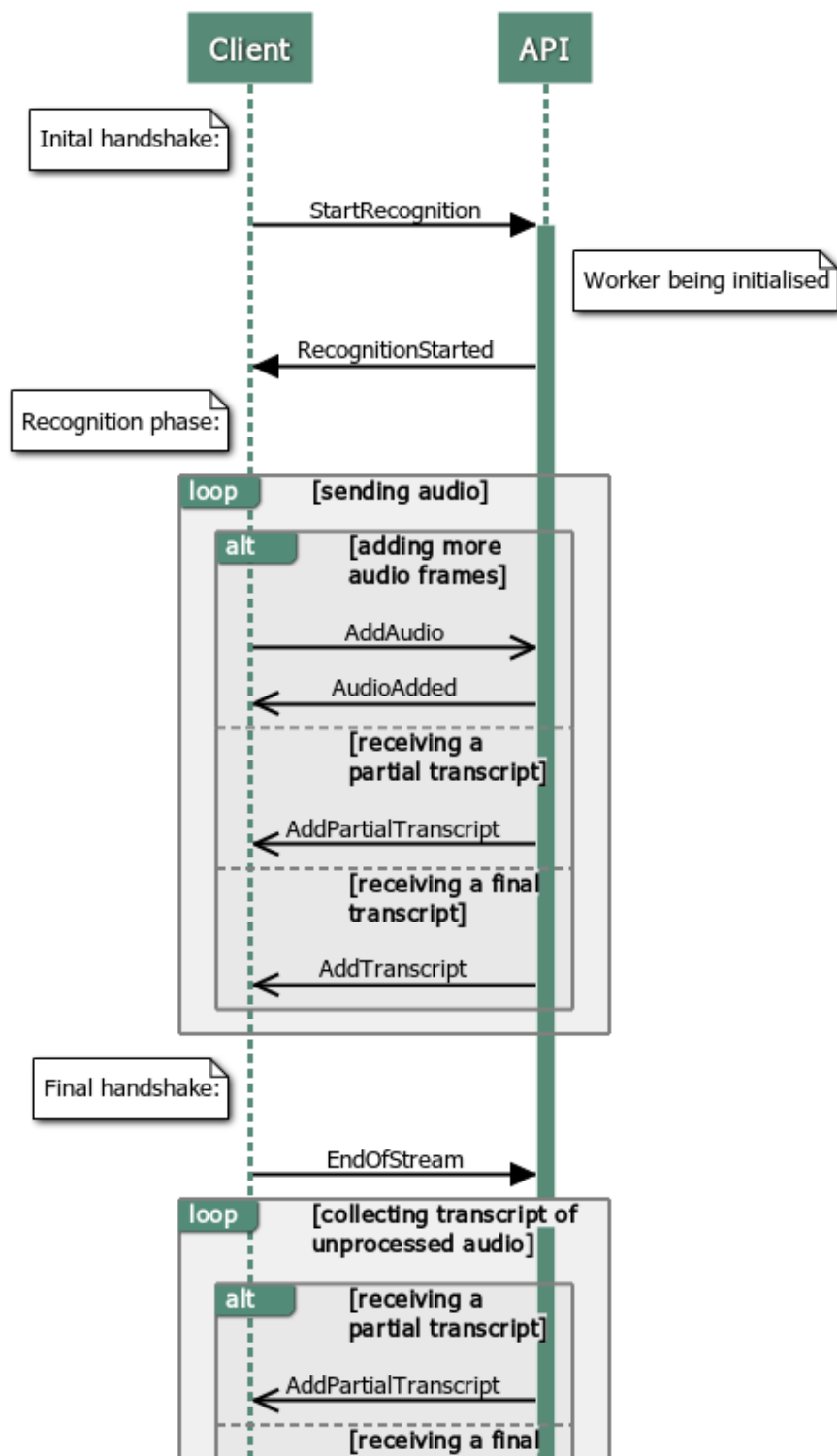
WebSockets are used to provide a two-way transport layer between your client and the Real-time Appliance or Container, enabling use with most modern web-browsers, and programming languages. See [RFC 6455](#) for the detailed specification of the WebSocket protocol.

The wire protocol used with the WebSocket consists mostly of packets of stringified JSON objects which comprise a message name, plus other fields that are message dependant. The only exception is that a binary message is used for transmitting the audio.

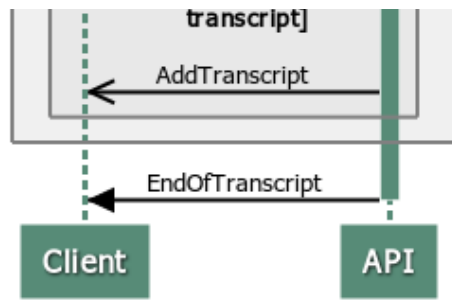
You can develop your real time client using any programming language that supports WebSockets. This document provides a list of the messages that are required for the client and server communication. Some of the messages are required to be sent in a particular order (outlined below) whilst others are optional. There are some examples provided at the end of this document on how to access the Speech API using JavaScript.

When implementing your own websocket client, we recommend using a ping/pong timeout of 60 seconds. More details about ping/pong messages can be found in the WebSocket RFC here: <https://tools.ietf.org/html/rfc6455#page-37>.

For a working Python example, please refer to our reference Python client implementations. The library is available from [here](#)







## Real-time API

This page specifies the Real-time API at its current state. The basic elements in the communication are the following:

- **Client** - An application connecting to the API, providing the audio and processing the transcripts received from the **Server**.
- **Server** (also called **API**) - An entry point of the API, allows external connections and provides the transcripts back.
- **Worker** - An internal speech recognizer. This is an internal entity that actually runs the heavy speech recognition.

## Getting Started

The communication is done using WebSockets, which are implemented in most of the modern web-browsers, as well as in many common programming languages (namely C++ and Python, for instance using <http://autobahn.ws/>).

### Messages

Each message that the **Server** accepts is a stringified JSON object with the following fields:

- `message` (String): The name of the message we are sending. Any other fields depend on the value of the `message` and are described below.

The messages sent by the **Server** to a **Client** are stringified JSON objects as well.

The only exception is a binary message sent from the **Client** to the **Server** containing a chunk of audio which will be referred to as `AddAudio`.

The following values of the `message` field are supported:

### StartRecognition

Initiates recognition, based on details provided in the following fields:

- `message: "StartRecognition"`
- `audio_format` (Object:AudioType): Required. Audio stream type you are going to send: see [Supported audio types](#).
- `transcription_config` (Object:TranscriptionConfig): Required. Set up configuration values for this recognition session, see [Transcription config](#).

A `StartRecognition` message must be sent exactly once after the WebSocket connection is opened. The client must wait for a `RecognitionStarted` message before sending any audio.

In case of success, a message with the following format is sent as a response:

- `message: "RecognitionStarted"`
- `id` (String): Required. A randomly-generated GUID which acts as an identifier for the session. e.g. "807670e9-14af-4fa2-9e8f-5d525c22156e".

In case of failure, an [error message](#) is sent, with `type` being one of the following: `invalid_model`, `invalid_audio_type`, `not_authorized`, `insufficient_funds`, `not_allowed`, `job_error`

An example of the `StartRecognition` message:

```
{
  "message": "StartRecognition",
  "audio_format": {
    "type": "raw",
    "encoding": "pcm_f32le",
    "sample_rate": 16000
  },
  "transcription_config": {
    "language": "en",
    "output_locale": "en-US",
    "diarization": "speaker_change",
    "max_delay": 3.5,
    "max_delay_mode": "flexible",
    "enable_partials": true,
  }
}
```

The example above starts a session with the Global English model ready to consume raw PCM encoded audio with float samples at 16kHz. It also includes an `additional_vocab` list containing the names of different types of pasta. `speaker_change` diarization is enabled, and partials are enabled meaning that `AddPartialTranscript` messages will be received as well as `AddTranscript` messages. Punctuation is configured to restrict the set of punctuation marks that will be added to only commas and full stops.

### Explaining Max Delay Mode

Users can specify the latency of the Real-time Speechmatics engine using the `max_delay` parameter. If a value of '5' was chosen, this would mean that transcripts would always be returned in 5 seconds from the word first being spoken. This happens even if a word is detected that may overrun that threshold. In some cases this can lead to inaccuracies in recognition and in finalised transcripts. This can be especially noticeable with key entities such as numerals, currencies, and dates.

`max_delay_mode` allows a greater flexibility in this latency only when a potential entity has been detected. Entities are common concepts such as numbers, currencies and dates, and can be seen in more detail [here](#).

There are two potential options for `max_delay_mode`: `fixed` and `flexible`. If no option is chosen, the default is `fixed`. Where an entity is detected with `flexible`, the latency of a transcript may exceed the threshold specified in `max_delay`, however the recognition of entities will be more accurate. If a user specifies `fixed`, the transcript will be returned in segments that will never exceed the `max_delay` threshold, even if this causes inaccuracies in entity recognition.

### SetRecognitionConfig

Allows the **Client** to configure the recognition session even after the initial `StartRecognition` message without restarting the connection. **This is only supported for certain parameters.**

- `message`: "SetRecognitionConfig"
- `transcription_config` (Object:TranscriptionConfig): A TranscriptionConfig object containing the new configuration for the session, see [Transcription config](#).

The following is an example of such a configuration message:

```
{
  "message": "SetRecognitionConfig",
  "transcription_config": {
```

```
"language": "en",
"max_delay": 3.5,
"enable_partials": true
}
}
```

Note: The `language` property is a mandatory element in the `transcription_config` object; however it is not possible to change the language mid-way through the session (it will be ignored if you do). It is only possible to modify the following settings through a **SetRecognitionConfig** message after the initial `StartRecognition` message:

- `max_delay`
- `max_delay_mode`
- `enable_partials`

If you wish to alter any other parameters you must terminate the session and restart with the altered configuration. Attempting otherwise will result in an error.

### AddAudio

Adds more audio data to the recognition job started on the WebSocket using `StartRecognition`. The server will only accept audio after it is initialized with a job, which is indicated by a `RecognitionStarted` message. Only one audio stream in one format is currently supported per WebSocket (and hence one recognition job). `AddAudio` is a binary message containing a chunk of audio data and no additional metadata.

### AudioAdded

If the `AddAudio` message is successfully received, an `AudioAdded` message is sent as a response. This message confirms that the **Server** has accepted the data and will make a corresponding **Worker** process it. If the **Client** implementation holds the data in an internal buffer to resubmit in case of an error, it can safely discard the corresponding data after this message. The following fields are present in the response:

- `message`: "AudioAdded"
- `seq_no` (Int): Required. An incrementing number which is equal to the number of audio chunks that the server has processed so far in the session. The count begins at 1 meaning that the 5th `AddAudio` message sent by the client, for example, should be answered by an `AudioAdded` message with `seq_no` equal to 5.

Possible errors:

- `data_error`, `job_error`, `buffer_error`

When sending audio faster than real time (for instance when sending files), make sure you don't send too much audio ahead of time. For large files, this causes the audio to fill out networking buffers, which might lead to disconnects due to WebSocket ping/pong timeout. Use `AudioAdded` messages to keep track what messages are processed by the engine, and don't send more than 10s of audio data or 500 individual `AddAudio` messages ahead of time (whichever is lower).

### Implementation details

Under special circumstances, such as when the client is sending the audio data faster than real time, the **Server** might read the data slower than the **Client** is sending it. The **Server** will not read the binary `AddAudio` message if it is larger than the internal audio buffer on the **Server**. Note that for each **Worker**, there is a separate buffer. In that case, the server will read any messages coming in on the WebSocket, until enough space is made in the buffer by passing the data to a corresponding **Worker**. The **Client** will only receive the corresponding `AudioAdded` response message once the binary data is read. The WebSocket might eventually fill all the TCP buffers on the way, causing a corresponding WebSocket to fail to write and close the connection [with prejudice](#). The **Client** can use the [bufferedAmount](#) attribute of the WebSocket to prevent this.

### AddTranscript

This message is sent from the **Server** to the **Client**, when the **Worker** has provided the **Server** with a segment of transcription output. It contains the transcript of a part of the audio the **Client** has sent using `AddAudio` - the **final transcript**. These messages are also referred to as **finals**. Each message corresponds to the audio since the last

`AddTranscript` message. The transcript is final - any further `AddTranscript` or `AddPartialTranscript` messages will only correspond to the newly processed audio. An `AddTranscript` message is sent when we reach an endpoint (end of a sentence or a phrase in the audio), or after 10s if we haven't reached such an event. This timeout can be further configured by setting `transcription_config.max_delay` in the `StartRecognition` message.

- `message`: "AddTranscript"
- `metadata` (Object): Required.
  - `start_time` (Number): Required. An approximate time of occurrence (in seconds) of the audio corresponding to the beginning of the first word in the segment.
  - `end_time` (Number): Required. An approximate time of occurrence (in seconds) of the audio corresponding to the ending of the final word in the segment.
  - `transcript` (String): Required. The entire transcript contained in the segment in text format. Providing the entire transcript here is designed for ease of consumption; we have taken care of all the necessary formatting required to concatenate the transcription results into a block of text. This transcript lacks the detailed information however which is contained in the `results` field of the message - such as the timings and confidences for each word.
- `results` (List:Object):
  - `type` (String): Required. One of 'word', 'entity', 'punctuation' or 'speaker\_change'. 'word' results represent a single word. 'punctuation' results represent a single punctuation symbol. 'word' and 'punctuation' results will both have one or more `alternatives` representing the possible alternatives we think the word or punctuation symbol could be. 'entity' is only a possible type if `enable_entities` is set to `true` and indicates a formatted entity. 'speaker\_change' results however will have an empty `alternatives` field. 'speaker\_change' results will only occur when using `speaker_change` diarization.
  - `start_time` (Number): Required. The start time of the result **relative to** the `start_time` of the whole segment as described in `metadata`.
  - `end_time` (Number): Required. The end time of the result **relative to** the `start_time` of the segment in the message as described in `metadata`. Note that punctuation symbols and `speaker_change` results are considered to be zero-duration and thus for those results `start_time` is equal to `end_time`.
  - `is_eos` (Boolean): Optional. Only present for 'punctuation' results. This indicates whether or not the punctuation mark is considered an end-of-sentence symbol. For example full-stops are an end-of-sentence symbol in English, whereas commas are not. Other languages, such as Japanese, may use different end-of-sentence symbols.
  - `alternatives` (List:Object): Optional. For 'word' and 'punctuation' results this contains a list of possible alternative options for the word/symbol.
    - `content` (String): Required. A word or punctuation mark. When `enable_entities` is requested this can be multiple words with spaces, for example "17th of January 2022".
    - `confidence` (Number): Required. A confidence score assigned to the alternative. Ranges from 0.0 (least confident) to 1.0 (most confident).
    - `display` (Object): Optional. Information about how the word/symbol should be displayed.
      - `direction` (String): Required. Either 'ltr' for words that should be displayed left-to-right, or 'rtl' vice versa.
    - `language` (String): Optional. The language that the alternative word is assumed to be spoken in. Currently this will always be equal to the language that was requested in the initial `StartRecognition` message.
    - `tags` (array): Optional. Only `[disfluency]` and `[profanity]` are displayed. This is a set list of profanities and disfluencies respectively that cannot be altered by the end user. `[disfluency]` is only present in English, and `[profanity]` is present in English, Spanish, and Italian.
  - `entity_class` (String): Optional. If `enable_entities` is requested in the `startTranscriptionConfig` request, and an entity is detected, `entity_class` will represent the type of entity the word(s) have been formatted as.
  - `spoken_form` (List:Object): Optional. If `enable_entities` is requested in the `startTranscriptionConfig` request, and an entity is detected, this is a list of words without formatting

applied. This follows the `results` list API formatting.

- `written_form` (List:Object): Optional. If `enable_entities` is requested in the `startTranscriptionConfig` request, and an entity is detected, this is a list of formatted words that matches the entity `content` but with individual estimated timing and confidences. This follows the `results` list API formatting.

### AddPartialTranscript

A partial-transcript message. The structure is the same as `AddTranscript`. A partial transcript is a transcript that can be changed and expanded by a future `AddTranscript` or `AddPartialTranscript` message and corresponds to the part of audio since the last `AddTranscript` message. For `AddPartialTranscript` messages the `confidence` field for `alternatives` has no meaning and will always be equal to 0.

Partials will only be sent if `transcription_config.enable_partials` is set to `true` in the `StartRecognition` message.

### EndOfStream

This message is sent from the Client to the API to announce that it has finished sending all the audio that it intended to send. No more `AddAudio` message are accepted after this message. The Server will finish processing the audio it has received already and then send an `EndOfTranscript` message. This message is usually sent at the end of file or when the microphone input is stopped.

- `message`: "EndOfStream"
- `last_seq_no` (Int): Required. The total number of audio chunks sent (in the `AddAudio` messages).

### EndOfTranscript

Sent from the API to the Client when the API has finished all the audio, as marked with the `EndOfStream` message. The API sends this only after it sends all the corresponding `AddTranscript` messages first. Upon receiving this message the Client can safely disconnect immediately because there will be no more messages coming from the API.

### Supported audio types

An `AudioType` object always has one mandatory field `type`, and potentially more mandatory fields based on the value of `type`. The following types are supported:

`type`: "raw"

Raw audio samples, described by the following additional mandatory fields:

- `encoding` (String): Encoding used to store individual audio samples. Currently supported values:
  - `pcm_f32le` - Corresponds to 32 bit float PCM used in the WAV audio format, little-endian architecture. 4 bytes per sample.
  - `pcm_s16le` - Corresponds to 16 bit signed integer PCM used in the WAV audio format, little-endian architecture. 2 bytes per sample.
  - `mulaw` - Corresponds to 8 bit  $\mu$ -law (mu-law) encoding. 1 byte per sample.
- `sample_rate` (Int): Sample rate of the audio

Please ensure when sending raw audio samples in real-time that the samples are undivided. For example, if you are sending raw audio via `pcm_f32le`, the sample should always contain 4 bytes. Here, if a sample did not contain 4 bytes, and then an `EndOfStream` message were sent, this would then cause an error. Required byte sizes per sample for each type of raw audio are listed above.

`type`: "file"

Any audio/video format supported by GStreamer. The `AddAudio` messages have to provide all the file contents, including any headers. The file is usually not accepted all at once, but segmented into reasonably sized messages.

Example `audio_format` field value: `audio_format: {type: "raw", encoding: "pcm_s16le", sample_rate: 44100}`

## Example communication

The communication consists of 3 stages - initialization, transcription and a disconnect handshake.

On **initialization**, the `StartRecognition` message is sent from the Client to the API and the Client must block and wait until it receives a `RecognitionStarted` message.

Afterwards, the **transcription** stage happens. The client keeps sending `AddAudio` messages. The API asynchronously replies with `AudioAdded` messages. The API also asynchronously sends `AddPartialTranscript` (if Partials enabled) and/or `AddTranscript` messages, depending on whether Partials were enabled.

Once the client doesn't want to send any more audio, the **disconnect handshake** is performed. The Client sends an `EndOfStream` message as its last message. No more messages are handled by the API afterwards. The API processes whatever audio it has buffered at that point and sends all the `AddTranscript` and `AddPartialTranscript` messages accordingly. Once the API processes all the buffered audio, it sends an `EndOfTranscript` message and the Client can then safely disconnect.

Note: In the example below, `->` denotes a message sent by the Client to the API, `<-` denotes a message send by the API to the Client. Any comments are denoted `"[like this]"`.

```
-> {"message": "StartRecognition", "audio_format": {"type": "file"},
    "transcription_config": {"language": "en", "enable_partials": true}}

<- {"message": "RecognitionStarted", "id": "807670e9-14af-4fa2-9e8f-5d525c22156e"}

-> "[binary message - AddAudio 1]"
-> "[binary message - AddAudio 2]"

<- {"message": "AudioAdded", "seq_no": 1}
<- {"message": "Info", "type": "recognition_quality", "quality": "broadcast", "reason": "Running
recognition using a broadcast model quality."}
<- {"message": "AudioAdded", "seq_no": 2}

-> "[binary message - AddAudio 3]"

<- {"message": "AudioAdded", "seq_no": 3}

"[asynchronously received transcripts:]"

<- {"message": "AddPartialTranscript", "metadata": {"start_time": 0.0, "end_time":
0.5399999618530273, "transcript": "One"},
    "results": [{"alternatives": [{"confidence": 0.0, "content": "One"}],
                 "start_time": 0.47999998927116394, "end_time": 0.5399999618530273, "type":
"word"}
                ]}

<- {"message": "AddPartialTranscript", "metadata": {"start_time": 0.0, "end_time":
0.7498992613545260, "transcript": "One to"},
    "results": [{"alternatives": [{"confidence": 0.0, "content": "One"}],
                 "start_time": 0.47999998927116394, "end_time": 0.5399999618530273, "type":
"word"},
                {"alternatives": [{"confidence": 0.0, "content": "to"}],
                 "start_time": 0.6091238623430891, "end_time": 0.7498992613545260, "type":
"word"}
                ]}

]]
```

```

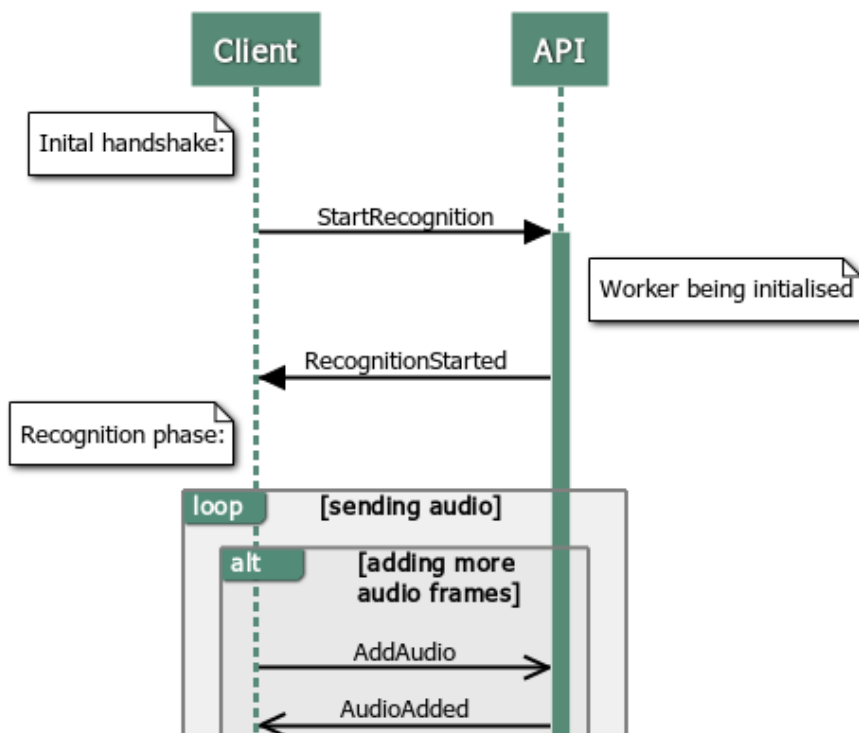
<- {"message": "AddPartialTranscript", "metadata": {"start_time": 0.0, "end_time":
0.9488123643240011, "transcript": "One to three"},
  "results": [{"alternatives": [{"confidence": 0.0, "content": "One"}],
    "start_time": 0.47999998927116394, "end_time": 0.53999999618530273, "type":
"word"},
    {"alternatives": [{"confidence": 0.0, "content": "to"}],
    "start_time": 0.6091238623430891, "end_time": 0.7498992613545260, "type":
"word"}
    {"alternatives": [{"confidence": 0.0, "content": "three"}],
    "start_time": 0.8022338627780892, "end_time": 0.9488123643240011, "type":
"word"}
  ]}
<- {"message": "AddTranscript", "metadata": {"start_time": 0.0, "end_time": 0.9488123643240011,
"transcript": "One two three."},
  "results": [{"alternatives": [{"confidence": 1.0, "content": "One"}],
    "start_time": 0.47999998927116394, "end_time": 0.53999999618530273, "type":
"word"},
    {"alternatives": [{"confidence": 1.0, "content": "to"}],
    "start_time": 0.6091238623430891, "end_time": 0.7498992613545260, "type":
"word"}
    {"alternatives": [{"confidence": 0.96, "content": "three"}],
    "start_time": 0.8022338627780892, "end_time": 0.9488123643240011, "type":
"word"}
    {"alternatives": [{"confidence": 1.0, "content": "."}],
    "start_time": 0.9488123643240011, "end_time": 0.9488123643240011, "type":
"punctuation", "is_eos": true}
  ]}

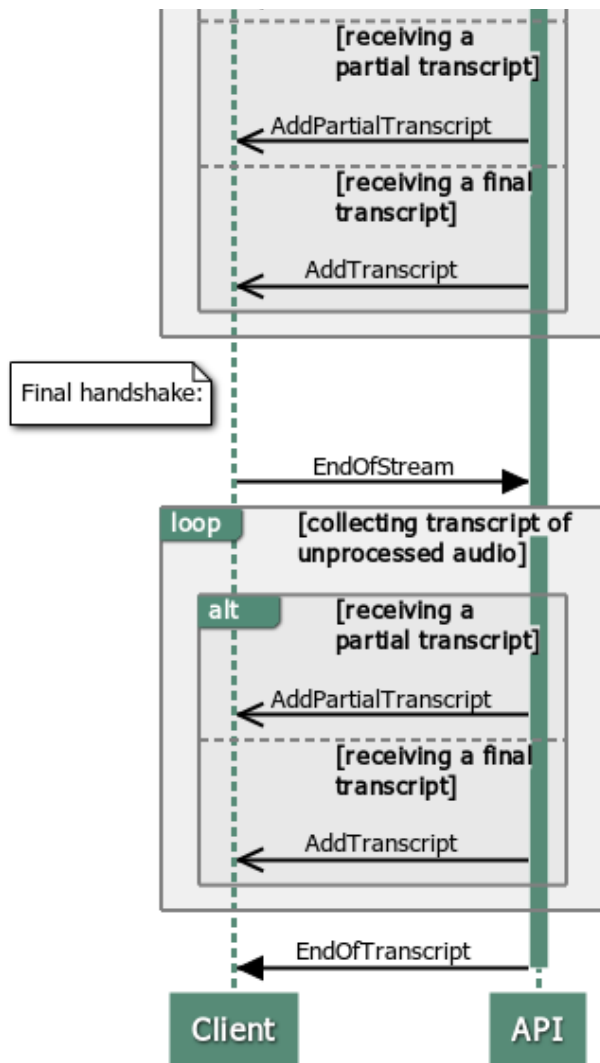
"[closing handshake]"

-> {"message": "EndOfStream", "last_seq_no": 3}

<- {"message": "EndOfTranscript"}

```





## Configuring Additional Features

### Transcription config

A `TranscriptionConfig` object specifies various configuration values for the recognition engine. All the values are optional, using default values when not provided.

- `language` (String): Required. Language model to process the audio input, normally specified as an ISO language code e.g. 'en'.
- `additional_vocab` (List:AdditionalWord): Optional. Configure **additional words**. See [Additional words](#). Default is an empty list. You should be aware that there is a performance penalty (latency degradation and memory increase) from using `additional_vocab`, especially if you intend to load in a large word list. When initialising a session that uses `additional_vocab` in the config you should expect a delay of up to 15 seconds, and an additional 800MB to 1700MB of memory (depending on the size of the list).
- `diarization` (String): Optional. The speaker diarization method to apply to the audio. The default is "none" indicating that no diarization will be performed. An alternative option is "speaker\_change" diarization in which the ASR system will attempt to detect any changes in speaker. Speaker changes are indicated in the results using an object with a `type` set to `speaker_change`.
- `enable_partials` (Boolean): Optional. Whether or not to send partials (i.e. `AddPartialTranscript` messages) as well as finals (i.e. `AddTranscript` messages). The default is `false`.



- `max_delay` (Number): Optional. Maximum delay in seconds between receiving input audio and returning final transcription results. The default is 10. The minimum and maximum values are 2 and 20.
- `max_delay_mode` (String): Optional. There are two options: `fixed` and `flexible`. `flexible` will return segments of transcripts according to exact `max_delay` parameter specified in the `startTranscriptionConfigRequest`. `flexible` will vary the exact length of a segment only when a potential entity has been detected to ensure the best possible accuracy. The default is `flexible`.
- `output_locale` (String): Optional. Configure **output locale**. See [Output locale](#). Default is an empty string.
- `punctuation_overrides` (Object:PunctuationOverrides): Optional. Options for controlling punctuation in the output transcripts. See [Punctuation overrides](#).
- `speaker_change_sensitivity` (Number): Optional.: Controls how responsive the system is for potential speaker changes. The value ranges between zero and one. High value indicates high sensitivity, i.e. prefer to indicate a speaker change if in doubt. The default is 0.4. This setting is only applicable when using `"diarization": "speaker_change"`.
- `operating_point` (String): Optional. Which model within the language pack you wish to use for transcription with a choice between `standard` and `enhanced`. See API How-to Guide for more details
- `enable_entities` (Boolean): Optional. Whether a user wishes for entities to be identified with additional spoken and written word format. Supported values `true` or `false`. The default is `false`.

## Requesting an enhanced model

Speechmatics supports two different models within each language pack; a standard or an enhanced model. The standard model is the faster of the two, whilst the enhanced model provides a higher accuracy, but a slower turnaround time.

The enhanced model is a premium model. Please contact your account manager or Speechmatics if you would like access to this feature.

An example of requesting the enhanced model is below

```
{
  "message": "StartRecognition",
  "audio_format": {
    "type": "raw",
    "encoding": "pcm_f32le",
    "sample_rate": 16000
  },
  {
    "transcription_config": {
      "language": "en",
      "operating_point": "enhanced"
    }
  }
}
```

Please note: `standard`, as well as being the default option, can also be explicitly requested with the `operating_point` parameter.

## Advanced punctuation

All Speechmatics language packs support Advanced Punctuation. This uses machine learning techniques to add in more naturalistic punctuation, improving the readability of your transcripts.

The following punctuation marks are supported for each language:

Language(s)	Supported Punctuation	Comment
Cantonese, Mandarin	, . ? ! 、	Full-width punctuation supported
Japanese	。 、	Full-width punctuation supported
Hindi	। ? , !	

All other languages	.,!?	
---------------------	------	--

If you do not want to see any of the supported punctuation marks in the output, then you can explicitly control this through the `punctuation_overrides` settings, for example:

```
"transcription_config": {
  "language": "en",
  "punctuation_overrides": {
    "permitted_marks": [ ".", ", " ]
  }
}
```

This will exclude exclamation and question marks from the returned transcript.

Note that changing the punctuation setting from the default can take a couple of seconds, which means if the user is using non-default neural punctuation sensitivity, after they send the `StartRecognition` message, there will be a slight delay (2-3 seconds) before the `RecognitionStarted` message is sent back.

All Speechmatics output formats support Advanced Punctuation. JSON output places punctuation marks in the results list marked with a `type` of `"punctuation"`.

**Note:** Disabling punctuation may slightly harm the accuracy of speaker diarization. Please see the ["Speaker diarization post-processing"](#) section in these docs for more information.

## Additional words

**Additional words** expand the standard recognition dictionary with a list of words or phrases called **additional words**. An **additional word** can also be a phrase, as long as individual words in the phrase are separated by spaces. This is the **custom dictionary** supported in other Speechmatics products. A pronunciation of those words is generated automatically or based on a provided `sounds_like` field. This is intended for adding new words and phrases, such as domain-specific terms or proper names. Better results for domain-specific words that contain common words can be achieved by using phrases rather than individual words (such as `action plan`).

`AdditionalWord` is either a `String` (the **additional word**) or an `Object`. The object form was introduced in 0.7.0. The object form has the following fields:

- `content` (`String`): The **additional word**.
- `sounds_like` (`List:String`): A list of words with similar pronunciation. Each word in this list is used as one alternative pronunciation for the additional word. These don't have to be real words - only their pronunciation matters. This list must not be empty. Words in the list must not contain whitespace characters. When `sounds_like` is used, the pronunciation automatically obtained from the `content` field is not used.

The `String` form `"word"` corresponds with the following `Object` form: `{"content": "word", "sounds_like": ["word"]}`.

Full example of `additional_vocab`:

```
"additional_vocab": [
  "speechmatics",
  {"content": "gnocchi", "sounds_like": ["nyohki", "nokey", "nochi"]},
  {"content": "CEO", "sounds_like": ["seeoh"]},
  "financial crisis"
]
```

To clarify, the following ways of adding the word `speechmatics` are equivalent with all using the default pronunciation:

1. `"additional_vocab": ["speechmatics"]`
2. `"additional_vocab": [{"content": "speechmatics"}]`
3. `"additional_vocab": [{"content": "speechmatics", "sounds_like": ["speechmatics"]}]`

## Output locale

Change the spellings of the transcription according to the output locale language code. If the selected language pack supports a different output locale, this config value can be used to provide spelling for the transcription in one of these locales. For example, the English language pack currently supports locales: `en-GB`, `en-US` and `en-AU`. The default value for `output_locale` is an empty string that means the transcription will use its default configuration (without spellings being altered in the transcription).

The following locales are supported for Chinese Mandarin. The default is simplified Mandarin.

- Simplified Mandarin (cmn-Hans)
- Traditional Mandarin (cmn-Hant)

## Punctuation overrides

This object contains settings for configuring punctuation in the transcription output.

- `permitted_marks` (List:String) Optional. The punctuation marks which the client is prepared to accept in transcription output, or the special value 'all' (the default). Unsupported marks are ignored. This value is used to guide the transcription process.
- `sensitivity` (Number) Optional. Ranges between zero and one. Higher values will produce more punctuation. The default is 0.5.

## Errors, Warnings and Info Messages

### Error messages

Error messages have the following fields:

- `message`: "Error"
- `code` (Int): Optional. A numerical code for the error. See below. TODO: This is not yet finalised.
- `type` (String): Required. A code for the error message. See the list of possible errors below.
- `reason` (String): Required. A human-readable reason for the error message.

### Error types

- `type: "invalid_message"`
  - The message received was not understood.
- `type: "invalid_model"`
  - Unable to use the model for the recognition. This can happen if the language is not supported at all, or is not available for the user.
- `type: "invalid_config"`
  - The config received contains some wrong/unsupported fields.
- `type: "invalid_audio_type"`
  - Audio type is not supported, is deprecated, or the `audio_type` is malformed.
- `type: "invalid_output_format"`
  - Output format is not supported, is deprecated, or the `output_format` is malformed.
- `type: "not_authorized"`
  - User was not recognised, or the API key provided is not valid.
- `type: "not_allowed"`
  - User is not allowed to use this message (is not allowed to perform the action the message would invoke).
- `type: "job_error"`
  - Unable to do any work on this job, the **Worker** might have timed out etc.
- `type: "data_error"`
  - Unable to accept the data specified - usually because there is too much data being sent at once

- `type: "buffer_error"`
  - Unable to fit the data in a corresponding buffer. This can happen for clients sending the input data faster than real-time.
- `type: "protocol_error"`
  - Message received was syntactically correct, but could not be accepted due to protocol limitations. This is usually caused by messages sent in the wrong order.
- `type: "unknown_error"`
  - An error that did not fit any of the types above.

Note that `invalid_message`, `protocol_error` and `unknown_error` can be triggered as a response to any type of messages.

The transcription is terminated and the connection is closed after any error.

## Warning messages

Warning messages have the following fields:

- `message: "Warning"`
- `code` (Int): Optional. A numerical code for the warning. See below. TODO: This is not yet finalised.
- `type` (String): Required. A code for the warning message. See the list of possible warnings below.
- `reason` (String): Required. A human-readable reason for the warning message.

### Warning types

- `type: "duration_limit_exceeded"`
  - The maximum allowed duration of a single utterance to process has been exceeded. Any `AddAudio` messages received that exceed this limit are confirmed with `AudioAdded`, but are ignored by the transcription engine. Exceeding the limit triggers the same mechanism as receiving an `EndOfStream` message, so the Server will eventually send an `EndOfTranscript` message and suspend.
  - It has the following extra field:
    - `duration_limit` (Number): The limit that was exceeded (in seconds).

## Info messages

Info messages denote additional information sent from the **Server** to the **Client**. Those are similar to `Error` and `Warning` messages in syntax, but don't actually denote any problem. The **Client** can safely ignore these messages or use them for additional client-side logging.

- `message: "Info"`
- `code` (Int): Optional. A numerical code for the informational message. See below. TODO: This is not yet finalised.
- `type` (String): Required. A code for the info message. See the list of possible info messages below.
- `reason` (String): Required. A human-readable reason for the informational message.

### Info message types

- `type: "recognition_quality"`
  - Informs the client what particular quality-based model is used to handle the recognition.
  - It has the following extra field:
    - `quality` (String): Quality-based model name. It is one of `"telephony"`, `"broadcast"`. The model is selected automatically, for high-quality audio (12kHz+) the broadcast model is used, for lower quality audio the telephony model is used.
- `type: "model_redirect"`
  - Informs the client that a deprecated language code has been specified, and will be handled with a different model. For example, if the `model` parameter is set to one of `en-US`, `en-GB`, or `en-AU`, then the

request may be internally redirected to the Global English model (en).

## Example Connection to the API

The WebSocket Speech API aligns with other Speechmatics platforms such as the Batch Virtual Appliance and Speechmatics SaaS.

### WebSocket URI

To use the V2.7 API you use the '/v2' endpoint for the URI, for example:

```
wss://rt-asr.example.com:9000/v2
```

### Session Configuration

The V2 API is configured by sending a `StartRecognition` message initially when the WebSocket connection begins. We have designed the format of this message to be very similar to the `config.json` object that has been used for a while now with the Speechmatics batch mode platforms (Batch Virtual Appliance, Batch Container and SaaS). The `transcription_config` section of the message should be almost identical between the two modes. There are some minor differences (for example batch features a different set of diarization options, and real-time features some settings which don't apply to batch such as `max_delay`).

### TranscriptionConfig

A `transcription_config` structure is used to specify various configuration values for the recognition engine when the `StartRecognition` message is sent to the server. All values apart from `language` are optional. Here's an example of calling the `StartRecognition` message with this structure:

```
{
  "message": "StartRecognition",
  "transcription_config": {
    "language": "en"
  },
  "audio_format": {
    "type": "raw",
    "encoding": "pcm_f32le",
    "sample_rate": 16000
  }
}
```

### AddAudio

Once the websocket session is setup and you've successfully called `StartRecognition` you'll receive a `RecognitionStarted` message from server. You can then just to send the binary audio chunks, which we refer to as `AddAudio` messages.

You would replace this in the V2 API with much simpler code:

```
// NEW V2 EXAMPLE
function addAudio(audioData) {
  ws.send(audioData);
  seqNoIn++;
}
```

Speechmatics real-time Speech API will tolerate no more than 10 seconds of audio data or 500 individual AddAudio messages ahead of time. If you send more than this amount you will not receive an `AudioAdded` response until there is capacity in the buffer. This is to prevent any slowdown in latency and system performance

If you have implemented your own client-side solution and/or wrapper, one possible blocking implementation of the rate-limiting is a semaphore of size 500, acquired before sending each AddAudio message, and released after receiving any AudioAdded message. Make sure receiving messages runs in another thread or uses some other mechanism to avoid getting blocked by the semaphore.

## Final and Partial Transcripts

The `AddTranscript` and `AddPartialTranscript` messages from the server output a JSON format which aligns with the JSON output format used by other Speechmatics products. There is now a `results` list which contains the transcribed words and punctuation marks along with timings and confidence scores. Here's an example of a final transcript output:

```
{
  "message": "AddTranscript",
  "results": [
    {
      "start_time": 0.11670026928186417,
      "end_time": 0.4049381613731384,
      "alternatives": [
        {
          "content": "gale",
          "confidence": 0.7034434080123901
        }
      ],
      "type": "word"
    },
    {
      "start_time": 0.410246878862381,
      "end_time": 0.6299981474876404,
      "alternatives": [
        {
          "content": "eight",
          "confidence": 0.670033872127533
        }
      ],
      "type": "word"
    },
    {
      "start_time": 0.6599999666213989,
      "end_time": 1.0799999237060547,
      "alternatives": [
        {
          "content": "becoming",
          "confidence": 1.0
        }
      ],
      "type": "word"
    },
    {
      "start_time": 1.0799999237060547,
      "end_time": 1.6154180765151978,
      "alternatives": [
        {
          "content": "cyclonic",

```

```

        "confidence":1.0
      }
    ],
    "type":"word"
  },
  {
    "start_time":1.6154180765151978,
    "is_eos":true,
    "end_time":1.6154180765151978,
    "alternatives":[
      {
        "content":".",
        "confidence":1.0
      }
    ],
    "type":"punctuation"
  }
],
"metadata":{
  "transcript":"gale eight becoming cyclonic.",
  "start_time":190.65994262695312,
  "end_time":194.46994256973267
},
"format":"2.7"
}

```

You can use the `metadata.transcript` property to get the complete final transcript as a chunk of plain text. The `format` property describes the exact version of the transcription output format. This may change in future releases if the output format is updated.

## Example Usage

This section provides some client code samples that show simple usage of the V2 WebSockets Speech API. It shows how you can test your Real-time Appliance or Container using a minimal WebSocket client.

### JavaScript

The basic usage of the WebSockets interface from a JavaScript client is shown in this section. In the first instance you setup the connection to the server and define the various event handlers that are required:

```

var ws = new WebSocket('wss://rta:9000/v2');
ws.binaryType = "arraybuffer";
ws.onopen = function(event) { onOpen(event) };
ws.onmessage = function(event) { onMessage(event) };
ws.onclose = function(event) { onClose(event) };
ws.onerror = function(event) { onError(event) };

```

In the above example, the hostname of the Real-time Appliance or Container is `rta` – change this to match the IP address or hostname of your Real-time Appliance or Container. The port used is 9000 and you need to make sure that you add `/v2` to the WebSocket URI. Note that the Real-time Appliance only supports the secure WebSocket (wss) protocol. On the other hand the Real-time Container only supports WebSocket (ws) protocol. You should also ensure that the `binaryType` property of the WebSocket object is set to `"arraybuffer"`.

In the `onopen` handler you initiate the session by sending the **StartRecognition** message to the server, for example:

```

function onOpen(evt) {
  var msg = {

```

```

    "message": "StartRecognition",
    "transcription_config": {
      "language": "en",
      "output_locale": "en-GB"
    },
    "audio_format": {
      "type": "raw",
      "encoding": "pcm_s16le",
      "sample_rate": 16000
    }
  };

  ws.send(JSON.stringify(msg));
}

```

An `onmessage` handler is where you will respond to the *server-initiated messages* sent by the appliance or container, and decide how to handle them. Typically, this involves implementing functions to display or process data that you get back from the server.

```

function onMessage(evt) {
  var objMsg = JSON.parse(evt.data);

  switch (objMsg.message) {
    case "RecognitionStarted":
      recognitionStarted(objMsg); // TODO
      break;

    case "AudioAdded":
      audioAdded(objMsg); // TODO
      break;

    case "AddPartialTranscript":
    case "AddTranscript":
      transcriptOutput(objMsg); // TODO
      break;

    case "EndOfTranscript":
      endTranscript(); // TODO
      break;

    case "Info":
    case "Warning":
    case "Error":
      showMessage(objMsg); // TODO
      break;

    default:
      console.log("UNKNOWN MESSAGE: " + objMsg.message);
  }
}

```

Once the WebSocket is initialized, the **StartRecognition** message is sent to the appliance or container to setup the audio input. It is then a matter of sending audio data periodically using the **AddAudio** message.

Your **AddAudio** message will take audio from a source (for example microphone input, or an audio stream) and pass it to the Real-time Appliance or Container.



```
// Send audio data to the API using the AddData message.
function addAudio(audioData) {
  ws.send(audioData);
  seqNoIn++;
}
```

In this example we use a counter `seqNoIn` to keep track of the `addAudio` messages we've sent.

A set of server-initiated transcript messages are triggered to indicate the availability of transcribed text:

- `AddTranscript`
- `AddPartialTranscript`

See above for changes to the JSON output schema in the V2 API. For full details of the output schema refer to the [AddTranscript](#) section in the API reference.

Finally, the client should send an **EndOfStream** message and close the WebSocket when it terminates. This should be done in order to release resources on the appliance or container and allow other clients to connect and use resources.

The [Mozilla developer network](#) provides a useful reference to the WebSocket API.

## Python Libraries

For all Speechmatics' supported Real-time products, you can use a Python library called `speechmatics-python`. The library is available [here](#) if you require it.

The `speechmatics-python` library can be incorporated into your own applications, used as a reference for your own client library, or called directly from the command line (CLI) like this (to pass a test audio file to the appliance or container):

```
speechmatics transcribe --url ws://rtc:9000/v2 --lang en --operating-point enhanced --ssl-mode none test.mp3
```

Note that configuration options are specified on the command-line as parameters, with a '\_' character in the configuration option being replaced by a '-'. The CLI option accepts an audio stream on standard input, meaning that you can stream in a live microphone feed. To get help on the CLI use the following command:

```
speechmatics transcribe --help
```

The library depends on Python 3.7 or above, since it makes use of some of the newer `asyncio` features introduced with Python 3.7.

## Formatting Common Entities

### Overview

Entities are commonly recognisable classes of information that appear in languages, for example numbers and dates. Formatting these entities is commonly referred to as Inverse Text Normalisation (ITN). Using ITN, Speechmatics will output entities in a predictable, consistent written form, reducing post-processing work required aiming to make the transcript more readable.

The language pack will use these formatted entities by default in the transcription. Additional metadata about these entities can be requested via the API including the spoken words without formatting and the entity class that was used to format it.

### Supported Languages

Entities are supported in the following languages:

- Cantonese
- Chinese Mandarin (Simplified and Traditional)
- English
- French
- German
- Hindi
- Italian
- Japanese
- Portuguese
- Russian
- Spanish

## Using the `enable_entities` parameter

Speechmatics now includes an `enable_entities` parameter. This can be requested via the API. By default this is `false`.

Changing `enable_entities` to `true` will enable a richer set of metadata in the JSON output only. Customers can choose between the default written form, spoken form, or a mixture, for their own workflows.

The changes are as following:

- A new `type - entity` in the JSON output in addition to `word` and `punctuation`. For example: "1.99" would have a `type` of `entity` and a corresponding `entity_class` of `decimal`
- The `entity` will contain the formatted text in the `content` section, like other words and punctuation
  - The `content` can include spaces, non-breaking spaces, and symbols (e.g. \$/£/%)
- A new output element `entity`, `entity_class` has been introduced. This provides more detail about how the entity has been formatted. A full list of entity classes is provided below.
- The start and end time of the entity will span all the words that make up that entity
- The entity also contains two ways that the content will be output:
  - `spoken_form` - Each individual `word` within the entity, written out in words as it was spoken. Each individual word has its own start time, end time, and confidence score. For example: "one", "million", "dollars"
  - `written_form` - The same output as within `entity` content, with a `type` of `word` instead. If there are spaces in the content it will be split into individual words. For example: "\$1", "million"

## Configuration example

Please see an example configuration file that would request entities:

```
{
  "message": "StartRecognition",
  "transcription_config": {
    "language": "en",
    "enable_entities": true
  }
}
```

## Different entity classes

The following `entity_classes` can be returned. Entity classes indicate how the numerals are formatted. In some cases, the choice of class can be contextual and the class may not be what was expected (for example "2001" may be a "cardinal" instead of "date"). The number of `entity_classes` may grow or shrink in the future.

N.B. Please note existing behaviour for English where numbers from zero to 10 (excluding where they are output as a decimal/money/percentage) are output as **words** is unchanged.

--	--	--	--

Entity Class	Formatting Behaviour	Spoken Word Form Example	Written Form Example
alphanum	A series of three or more alphanumerics, where an alphanumeric is a digit less than 10, a character or symbol	triple seven five four	77754
cardinal	Any number greater than ten is converted to numbers. Numbers ten or below remain as words. Includes negative numbers	nineteen	19
credit card	A long series of spoken digits less than 10 are converted to numbers. Support for common credit cards	one one one one two two two two three three three three four four four four	1111222233334444
date	Day, month and year, or a year on its own. Any words spoken in the date are maintained (including "the" and "of")	fifteenth of January twenty twenty two	15th of January 2022
decimal	A series of numbers divided by a separator	eighteen point one two	18.12
fraction	Small fractions are kept as words ("half"), complex fractions are converted to numbers separated by "/"	three sixteenths	3/16
money	Currency words are converted to symbols before or after the number (depending on the language)	twenty dollars	\$20
ordinal	Ordinals greater than 10 are output as numbers	forty second	42nd
percentage	Numbers with a per cent have the per cent converted to a % symbol	duecento percento	200%
span	A range expressed as "x to y" where x and y correspond to another entity class	one hundred to two hundred million pounds	100 to £200 million
time	Times are converted to numbers	eleven forty a m	11:40 a.m.
word	Entities that do not match a specific class	hundreds	hundreds

## Output locale styling

Each language has a specific style applied to it for thousands, decimals and where the symbol is positioned for money or percentages.

For example

- English contains commas as separators for numbers above 9999 (example: "20,000"), the money symbol at the start (example: "\$10") and full stops for decimals (example: "10.5")
- German contains full stops as separators for numbers above 9999 (example: "20.000"), the money symbol comes after with a non-breaking space (example: "10 \$") and commas for decimals (example: "10,5")
- French contains non-breaking spaces as separators for numbers above 9999 (example: "20 000"), the money symbol comes after with a non-breaking space (example: "10 \$") and commas for decimals (example: "10,5")

## Example output

Here is an example of a transcript requested with `enable_entities` set to `true`:

- An `entity` that is "17th of January 2022", including spaces
  - The start and end times span the entire entity

- o An `entity_class` of `date`
- o The `spoken_form` is split into the following individual words: "seventeenth", "of", "January", "twenty", "twenty", "two". Each word has its own start and end time
- o the `written_form` split into the following individual words: "17th", "of", "January", "2022". Each word has its own start and end time

```
[{
  "message": "AddTranscript",
  "format": 2.7,
  "results": [{
    "alternatives": [{
      "confidence": 1,
      "content": "17th of January 2022",
      "language": "en"
    }],
    "end_time": 3.0899999141693115,
    "entity_class": "date",
    "spoken_form": [{
      "alternatives": [{
        "confidence": 1,
        "content": "Seventeenth",
        "language": "en"
      }],
      "end_time": 1.3799999952316284,
      "start_time": 0.8399999737739563,
      "type": "word"
    }],
    {
      "alternatives": [{
        "confidence": 1,
        "content": "of",
        "language": "en"
      }],
      "end_time": 1.4399999380111694,
      "start_time": 1.3799999952316284,
      "type": "word"
    }],
    {
      "alternatives": [{
        "confidence": 1,
        "content": "January",
        "language": "en"
      }],
      "end_time": 1.9199999570846558,
      "start_time": 1.4399999380111694,
      "type": "word"
    }],
    {
      "alternatives": [{
        "confidence": 1,
        "content": "twenty",
        "language": "en"
      }],
      "end_time": 2.25,
      "start_time": 1.9199999570846558,
      "type": "word"
    }],
  }],
}
```

```

    {
      "alternatives": [{
        "confidence": 1,
        "content": "twenty",
        "language": "en"
      }],
      "end_time": 2.549999952316284,
      "start_time": 2.25,
      "type": "word"
    },
    {
      "alternatives": [{
        "confidence": 0.9504331946372986,
        "content": "two",
        "language": "en"
      }],
      "end_time": 3.0899999141693115,
      "start_time": 2.549999952316284,
      "type": "word"
    }
  ],
  "start_time": 0.8399999737739563,
  "type": "entity",
  "written_form": [{
    "alternatives": [{
      "confidence": 1,
      "content": "17th",
      "language": "en"
    }],
    "end_time": 1.1999999682108562,
    "start_time": 0.8399999737739563,
    "type": "word"
  }],
  {
    "alternatives": [{
      "confidence": 1,
      "content": "of",
      "language": "en"
    }],
    "end_time": 1.559999962647756,
    "start_time": 1.1999999682108562,
    "type": "word"
  },
  {
    "alternatives": [{
      "confidence": 1,
      "content": "January",
      "language": "en"
    }],
    "end_time": 1.9199999570846558,
    "start_time": 1.559999962647756,
    "type": "word"
  },
  {
    "alternatives": [{
      "confidence": 1,
      "content": "2022",

```

```

        "language": "en"
      }
    ],
    "end_time": 3.0899999141693115,
    "start_time": 1.9199999570846558,
    "type": "word"
  }
]
}],
"metadata": {
  "end_time": 5.16,
  "start_time": 0,
  "transcript": "17th of January 2022 "
}
}]

```

If `enable_entities` is set to `false`, the output is as below:

```

[
  {
    "message": "AddTranscript",
    "format": 2.7,
    "results": [
      {
        "alternatives": [
          {
            "confidence": 1,
            "content": "17th",
            "language": "en"
          }
        ],
        "end_time": 1.1999999682108562,
        "start_time": 0.8399999737739563,
        "type": "word"
      },
      {
        "alternatives": [
          {
            "confidence": 1,
            "content": "of",
            "language": "en"
          }
        ],
        "end_time": 1.559999962647756,
        "start_time": 1.1999999682108562,
        "type": "word"
      },
      {
        "alternatives": [
          {
            "confidence": 1,
            "content": "January",
            "language": "en"
          }
        ],
        "end_time": 1.9199999570846558,
        "start_time": 1.559999962647756,
        "type": "word"
      },
      {
        "alternatives": [
          {
            "confidence": 1,
            "content": "2022",
            "language": "en"
          }
        ],
        "end_time": 3.0899999141693115,
        "start_time": 1.9199999570846558,

```

```
    "type": "word"
  }
],
"metadata": {
  "end_time": 5.16,
  "start_time": 0,
  "transcript": "17th of January 2022 "
}
}]
```