Voice Pro®
Enterprise 7

Systemanforderungen und Architektur

Linguatec
Sprachtechnologien

Voice Pro Enterprise 7.0

Systemanforderungen und Architektur

Version 2.19.12051219

Inhaltsverzeichnis

1 Vorwort

Mit „Voice Pro Enterprise“ erhalten Sie professionelle Diktier- und Transkriptionswerkzeuge für die Umwandlung von gesprochener Sprache in geschriebene Sprache.

Die Spracherkennung „Voice Pro Enterprise“ ist eine Client-Server-Lösung mit flexibel einsetzbaren Client-Anwendungen. Sie besteht aus einem oder mehreren Spracherkennungsservern und mehreren Client-Anwendungen.

Die Client-Anwendung übermittelt ein Audiosignal oder eine Audio-Datei zusammen mit bestimmten Optionen an den Erkennungsserver. Der Server wandelt das Audiosignal in Text mit Zeitstempeln und weiteren Informationen um, bevor er es zurück an den Client übermittelt. Der Client formatiert den Text entsprechend den Benutzereinstellungen und integriert das Resultat in den Benutzer-Workflow.

Die Spracherkennungslösung „Voice Pro Enterprise“ kann sowohl als Einzelplatz-Version auf einem Rechner als auch als verteilte Architektur auf mehreren Clients und Servern in einem Netzwerk installiert und eingesetzt werden.

2 Architektur

Voice Pro Enterprise ist eine konsequent modulare Lösung, die jederzeit skaliert, erweitert und angepasst werden kann. Sie besteht aus den folgenden Komponenten:

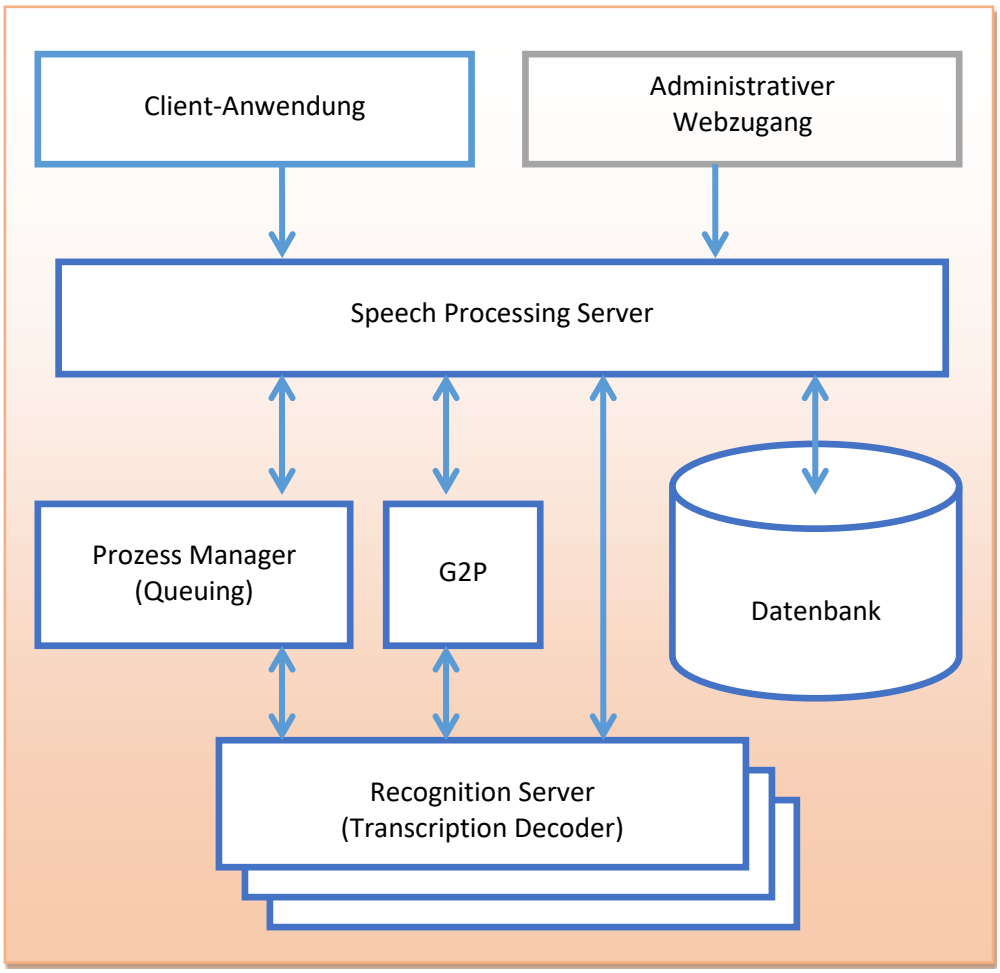
Client-Komponenten für die Endbenutzer und Administratoren:

- Client-Anwendung
- Administrativer Webzugang

Server-Komponenten:

- Speech Processing Server
- Process Manager (Queuing)
- Recognition Server (Transcription Decoder)
- Datenbank
- G2P

2.1 Abbildung: Architektur



2.2 Verbindungen und Ports

Die Kommunikation der Clients mit der gesamten Server-Lösung findet hauptsächlich

a) über HTTP- oder HTTPS-Ports unter Verwendung des WebSocket-Protokolls für das Transkribieren eines Diktates oder einer Datei in Echtzeit sowie
b) über HTTP- oder HTTPS-Protokolle für das Austauschen von benutzerspezifischen Wörterbüchern und Einstellungen statt.

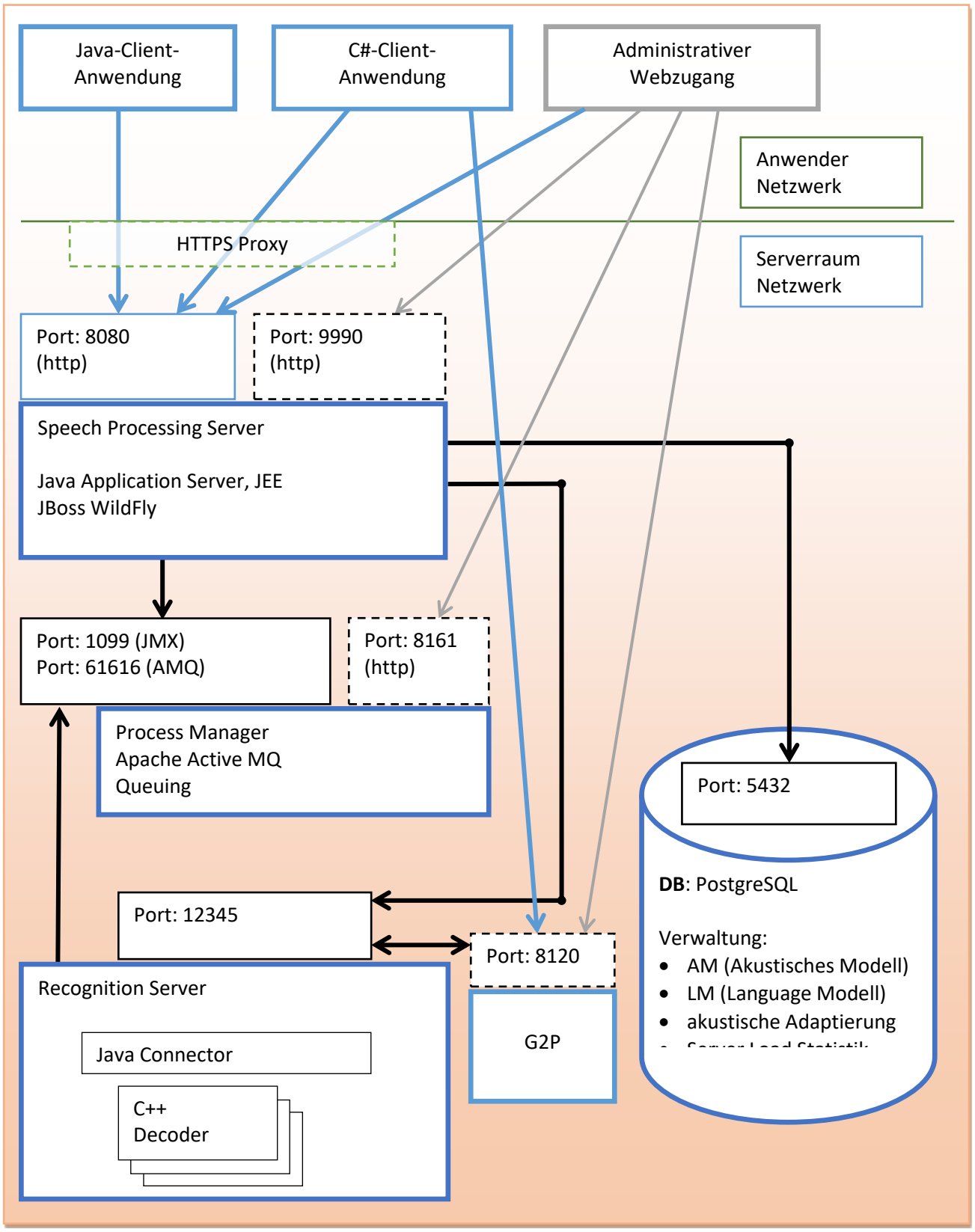
Standardmäßig wird dafür der Port 8080 auf dem Application Server verwendet.

Bei Bedarf können alle Ports während der Installation an die technischen Anforderungen ihrer Infrastruktur angepasst werden.

Für administrative Zwecke können die Ports 9990 und 8161 verwendet werden. Diese können bei Bedarf nur für bestimmte Personen oder Gruppen freigegeben werden. Sie dienen ausschließlich der Konfiguration und Beobachtung der einzelnen Server-Komponenten und werden für den produktiven Betrieb der Clientanwendung nicht benötigt.

Achtung: Die entsprechenden Ports müssen in der lokalen Firewall freigeschaltet sein.

2.3 Abbildung: Verbindungen und Ports



3 Serverkomponenten

3.1.1 Speech Processing Server

Diese Komponente implementiert die Schnittstellen zwischen der Erkennungsengine (Decoder) und den Client-Anwendungen.
Als Application Server wird JBoss Wildfly http://wildfly.org/ empfohlen und im Installationspaket mit ausgeliefert. Bei Bedarf kann die Installation und Konfiguration auf vorhandenen JBoss Application Servern durchgeführt werden.

3.1.2 Process Manager (Queuing)

Diese Komponente dient zur Verwaltung von mehreren Erkennungsengines, Sprachmodellen und Servern. Die Anwendung basiert auf dem Projekt „Apache ActiveMQ“ http://activemq.apache.org/

3.1.3 Datenbank

Die Datenbank wird für folgende Zwecke verwendet:

1. Verwaltung von sprachspezifischen Daten, die für die Erkennungsengine (Recognition Server) notwendig sind (z.B. Sprachmodelle)
2. Statistik der Erkennungsaufträge

Standardmäßig besteht die Datenbank aus ca. 20 Tabellen (~100 Spalten).
Die Größe der Datenbank ist stark von der Anzahl der Endanwender und der Anzahl an Erkennungsaufträgen abhängig:

1. Je nach Anwendung wird dafür ca. 1 MB pro Benutzer benötigt
2. Die notwendige Datengröße liegt bei ca. 1 MB
3. Dafür werden ca. 3 KB pro Erkennungsauftrag benötigt

3.1.4 G2P

Der Graphem-Zu-Phonem Server (im Weiteren G2P-Server genannt) ist notwendig für die Phonetisierung von benutzerspezifischen Wörterbüchern.
Damit werden die Sprachmodelle im laufenden Betrieb an die Benutzerlexika angepasst.

3.1.5 Recognition Server (Decoder)

Der Recognition Server wandelt den Audiostream in Text um.
Es können ein oder mehrere Recognition Server installiert und verwendet werden. Jeder davon führt mehrere Erkennungsprozesse aus. Die Anzahl der Erkennungsprozesse (Decoder) muss entsprechend der Anzahl von gleichzeitigen Diktaten und Dateierkennungen für Peak-Zeiten kalkuliert werden.
Ein Recognition Server besteht aus einem Java Konnektor und mehreren C++ basierten Decodern.

4 Systemvoraussetzungen für Server-Komponenten

Die Serverkomponenten können entweder auf Linux- oder auf Windows- 64-bit Betriebssystemen installiert werden. Die neuesten Betriebssystemversionen werden weder automatisch noch selbstverständlich unterstützt – jedes Major-Update muss vorher abgestimmt und getestet werden.

4.1 Windows Server

Die Installation der Server-Komponenten erfolgt mit nativen Installationen.

Für den Serverbetrieb werden die folgenden Betriebssysteme unterstützt:

- Windows Server 2016, 2019, 2022
- Windows 10 Desktop, Windows 11 Desktop

Für die Server-Installation sind folgende Pakete notwendig:

- Java Runtime 11 (Amazon Corretto, OpenJDK JRE, Oracle JRE, Azul Zulu)
- PostgreSQL
- ActiveMQ
- WildFly

Alle notwendigen Komponenten werden in der Installation mitgeliefert.

4.2 Linux Server

Die Installation der Server-Komponenten erfolgt mit Docker Containern und mit einer nativen Installation des Transcription Servers (Decoder). Der Transcription Server muss aus Performanz-Gründen immer nativ installiert werden.

Unterstützte Betriebssysteme:

- Ubuntu in den Versionen 16.4 LTS, 18.4 LTS, 20.4 LTS, 22.04. Die neuesten Versionen werden weder automatisch noch selbstverständlich unterstützt.
- CentOS 8, CentOS 7

Notwendige Pakete:

- libxml2
- libsndfile1
- libgomp1
- Java Runtime 11 (Amazon Corretto, OpenJDK JRE, Oracle JRE, Azul Zulu)
- Docker
- pwget
- unzip
- curl

5 Ressourcen und Skalierbarkeit für Server-Komponenten

Für die Echtzeiterkennung ist ein CPU-Kern pro Erkennungsprozess notwendig.
Für die Installation auf einem Rechner bzw. für die Offline-Version sind mindestens 2 Kerne CPU und 24 Gigabyte RAM erforderlich.
Für die Skalierung von Server und Ressourcen empfehlen wir die folgenden Werte zugrunde zu legen.

5.1 Prozessor (CPU)

- Application Server: 1 bis 2 Kerne je nach Anzahl von gleichzeitigen Diktaten
- Process Manager: 1 Kern
- Datenbank: 1 Kern
- Decoder: 1 Kern pro Erkennungsprozess

5.2 Arbeitsspeicher (RAM)

- Application Server
- Process Manager
- G2P
- Datenbank

Summe: ca. 8 Gigabyte
+

- Decoder 4 bis 8 GB pro geladenes Sprachmodel, und 1GB pro jede Echtzeit Erkennung

5.3 Beispiel: Server Kalkulation mit 16 Kernen, 24 GB RAM

CPU Kalkulation:

- Betriebssystem: 2 Kerne
- Application Server: 2 Kerne
- Process Manager: 1 Kern
- Datenbank: 1 Kern
- Decoder: 10 gleichzeitige Erkennungsprozesse

RAM Kalkulation:

- Application Server
- Process Manager
- G2P
- Datenbank

Summe: ca. 8 Gigabyte
+

- 10 gleichzeitige Erkennungsprozesse = 10 x 1 GB = 10 Gigabyte

---

Gesamtsumme: aufgerundet sind es 20 Gigabyte PLUS Betriebssystem

5.4 Festplatte (HDD bzw. SSD)

Für die Hauptkomponenten sollten 20 Gigabyte eingeplant werden.
Hinzu kommen 30 Gigabyte pro Sprachmodell.

Je nach getroffener Einstellung sind für eine Dual-Pass-Erkennung 10 Gigabyte pro Minute erforderlich.

Während eines Diktats (bei einer Echtzeiterkennung) werden die Sprachmodelle von der Festplatte gelesen. Die Lesegeschwindigkeit muss der Geschwindigkeit eines CPU-Kerns entsprechen. Der Einsatz von Raid mit SSD-Festplatten wird empfohlen.

5.5 Skalierbarkeit der verteilten Architektur

5.5.1 Recognition Server (= Transcription Decoder)

Die Hauptlast trägt der Recognition Server bzw. Decoder.
Es ist empfehlenswert, die Anwendung für den Einsatz mit mehr als 10 gleichzeitigen Benutzern auf mehrere Rechner zu verteilen. Dabei können die folgenden Komponenten auf derselben Maschine installiert sein:

- Application Server
- Process Manager
- G2P
- Datenbank

Es ist möglich, im laufenden Betrieb die Decoder Maschinen weg- oder hinzuzuschalten.
Der Einsatz von virtuellen Rechnern ist möglich.
Für den Recognition Server (Decoder) ist eine CPU mit **mindestens 2.4 GHZ** pro Kern notwendig.

5.5.2 Speech Processing Server

Diese Komponente implementiert die Schnittstellen zwischen der Erkennungsengine (Decoder) und den Client-Anwendungen.
Als Application Server wird JBoss Wildfly http://wildfly.org/ empfohlen und im Installationspaket mit ausgeliefert. Bei Bedarf kann die Installation und Konfiguration auf vorhandenen JBoss Application Servern durchgeführt werden.
Für die verteilte Architektur werden folgende Ports standardmäßig verwendet:

- 80
- 443
- 8080
- 8120
- 8443

Bei Bedarf können die Ports während der Installation an die technischen Anforderungen Ihrer Infrastruktur angepasst werden.

6 Systemanforderungen an Client-Komponenten

Da die Spracherkennung direkt auf dem Server erfolgt, sind die Systemanforderungen für den Client nur gering.
Die Systemmindestanforderungen für den Client sind wie folgt:

- Betriebssystem: Windows 7 (ab SP1), Windows 8.1, 10, 11 – mit aktuellen Windows-Updates
- Freier Festplattenspeicher: 350 MB
- Arbeitsspeicher: ab 2 GB RAM

Für Microsoft **Word** und **Outlook** in den folgenden Versionen bietet Voice Pro Enterprise Plug-Ins:

- Microsoft Office 2016 und 2019
- Microsoft Office 365: lokale Installation

Für die Client-Installation sind folgende Pakete/Komponenten notwendig:

- NET-Framework 4.5
- Lokal installiertes Audio-Aufnahmegerät

7 Anforderungen an Audioformate, Mikrofone und Diktiergeräte

Der Recognition Server benötigt ein Audiosignal im Format 16 kHz 16 Bit. Alle anderen Formate werden auf dem Client in dieses Format umgewandelt. Bei einem Diktat wird die Umwandlung in Echtzeit stattfinden.

Die besten Erkennungsergebnisse erzielen Sie mit Voice Pro Enterprise, wenn Ihr **Aufnahmegerät** mit **16-bit und 16 kHz** konfiguriert ist.

Die folgenden Konfigurationen werden eine Echtzeitkonvertierung auf dem Client erfordern:
Ohne Qualitätsverlust:

- 16-bit und 32 kHz
- 16-bit und 48 kHz

Mit geringem Qualitätsverlust:

- 16-bit und 22 kHz
- 16-bit und 44 kHz

Bei einer Aufnahmequalität von 8-bit oder Telefonie Qualität ist der Einsatz eines entsprechenden Sprachmodells erforderlich. Sprechen Sie uns in diesem Fall direkt an.

Für die Dateierkennung werden folgende Dateiformate unterstützt:
Audio Formate: WAV, MP3, WMA, AIFF, M4A, AC3, MP2, DSS, DSS Pro
Videoformate: AVI, M4V, MKV, MOV, MP4, WMV

Linguatec Sprachtechnologien
Gottfried-Keller-Str. 12
D- 81245 München
www.linguatec.de