HIS ▸ HIJP ▸ AIGP

Harmonisierung der Informatik in der Strafjustiz
Harmonisation de l'informatique dans la justice pénale
Armonizzazione dell'informatica nella giustizia penale

softfakt
Potenzial effektiv entfalten

Softfakt GmbH
Rosenbergstrasse 75
8498 Gibswil

+41 55 245 11 66

info@softfakt.ch
www.softfakt.ch

Speech-to-Text in der Justiz

Bericht

Autor: Ralph Wildhaber
Version: 1.0
Datum: 31.08.2021

Inhaltsverzeichnis

Tabellen

Abbildungen

1 Grundlagen zu Speech-to-Text

Speech-to-Text gewinnt vielerorts an Beliebtheit und Bedeutung. Oft begegnet man dem Thema im Zusammenhang von Digitalisierungsaktivitäten oder Prozessoptimierungen in Unternehmungen. Doch auch im privaten Alltag sind Systeme mit Spracheingabe nicht mehr wegzudenken. Entsprechend bunt präsentiert sich auf dem Markt der Angebotsstrauss solcher Lösungen.

1.1 Begriffe und Funktionsweise

Die Begriffe automatische Spracherkennung, Speech-to-Text, Speech Recognition, Voice Recognition, Transkriptionssystem, Audiotranskription und ähnliche stehen im Grundsatz immer für das Gleiche: Die computergestützte Transkription von gesprochener Sprache.

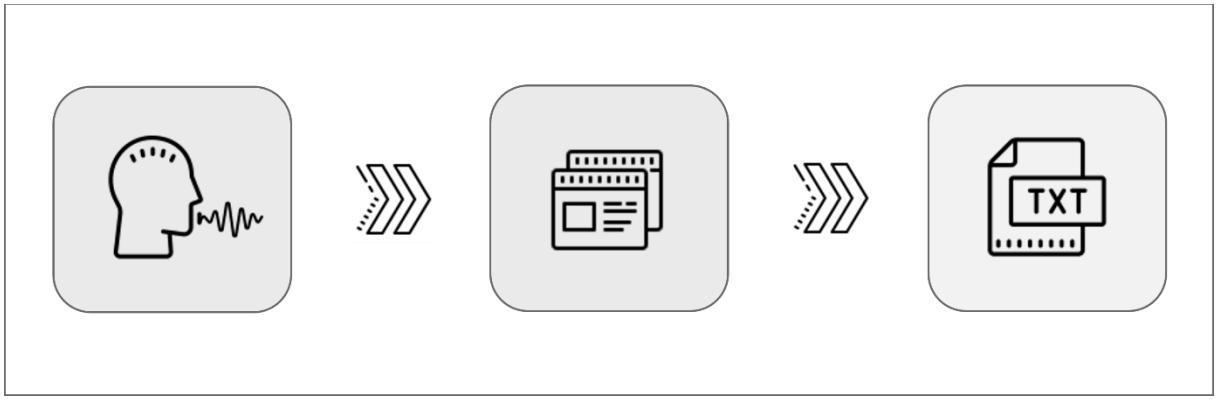

Abbildung 1: Ablauf der automatischen Spracherkennung

Hierbei wird ein Audiosignal (Gesprochenes) mithilfe eines Computers aufgezeichnet und durch eine Spracherkennungssoftware zu lesbarem Text verarbeitet. Im Kern eines solchen Programms werden mehrstufigen Verfahren unter Anwendung unterschiedlicher Algorithmen ausgeführt. Die Analyse durchläuft in der Regel nacheinander die drei folgenden Modelle:

- **Akustisches Modell** zur Zerlegung des Audiosignals
- **Lexikon** zur Erkennung von Wörtern
- **Sprachmodell** zum Aufbau eines Satzes anhand von Dreiwortfolgen

Obschon sich die Informatik bereits seit Jahrzehnten mit der Spracherkennung befasst, hat die Bedeutung und Verbreitung in den vergangenen Jahren stark zugenommen. Die Gründe dazu liegen einerseits in der gestiegenen Leistungsfähigkeit der Computer, welche die oben aufgelisteten Analysen zeitnah ausführen können. Andererseits trägt die weite Verbreitung der mobilen Technologien (Smartphone, Tablet) einiges dazu bei, dass die Entwicklungen in Speech-to-Text intensiviert werden.

Je nach eingesetzter Lösung erfolgt die Transkription direkt beim Diktieren (simultan) oder steht einige Minuten nach der Einspeisung der Audiodatei zur Verfügung.

1.2 Abgrenzungen der Spracherkennung

Speech-to-Text-Systeme gilt es klar von biometrischen Verfahren für Identifikationszwecke (Sprechererkennung) oder Stimmanalysen zu differenzieren. Ebenso werden solche Spracherkennungslösungen (noch) nicht für Sprachübersetzungen eingesetzt.

1.3 Infrastruktur-Komponenten

Wenn man betrachtet, was ein Benutzer einer Spracherkennungssoftware an Infrastruktur benötigt, präsentiert sich das äusserst einfach. Je nach Art des Setups ergeben sich kleine Unterschiede.

1.3.1 Hardware

Hardwareseitig wird neben einem einigermassen zeitgemässen Computer ein Mikrofon benötigt. Hier empfiehlt sich eine Headset-Lösung einzusetzen. Diese Geräte sind auf Sprachaufzeichnungen optimiert und von der Positionierung her wenig anfällig auf störende Umgebungsgeräusche. Dennoch bleibt anzumerken, dass sich mit fix verbauten Mikrofonen in Computern (Notebooks) oder Webcams ebenfalls hervorragende Transkriptionsresultate erzielen lassen.

Betrachtet man das Ganze von Betreiberseite aus, stellt sich die Situation ein bisschen anders dar. Viele Lösungen, gerade im Bereich der lokalen Diktiersoftware, ermöglichen es, einen Server im eigenen Rechenzentrum aufzubauen. Die Anforderungen an eine entsprechende Maschine genügen den üblichen Standards. Dadurch lassen sich diese problemlos in virtuellen Umgebungen betreiben.

1.3.2 Software

Abhängig vom Einsatzgebiet und der Art des Services wird unterschiedliche Software für die Aufzeichnung beziehungsweise Transkription der Sprachinformation benötigt. Unterschieden wird zwischen Weblösungen und lokal zu installierenden Programmen (Client-Applikationen).

Der grosse Vorteil von Weblösungen besteht darin, dass die Programme in den gängigen Webbrowsern verwendet werden können und somit unabhängig vom zugrunde liegenden Betriebssystem sind.

Des Weiteren gibt es Lösungen auf dem Markt, welche eine Installation der Spracherkennungssoftware auf den Zielrechner nötig machen. Es gilt zu beachten, dass die verschiedenen Produkte jeweils nicht zwingend in Versionen für die alle klassischen Betriebssysteme (Windows, MacOS, Linux) zur Verfügung stehen.

1.4 Entwicklungen und Technologien

Die Entwicklungen im Bereich der Spracherkennung laufen bereits seit vielen Jahrzehnten. Erst mit der gestiegenen Leistungsfähigkeit moderner Computer wurde jedoch die Möglichkeit geschaffen, neue Methoden in der Sprachtranskription zu integrieren. Dieser Entwicklungsschritt ist mitunter ein Grund, weshalb die automatische Spracherkennung an Bedeutung gewonnen hat und immer mehr gewinnen wird. Wurden früher vor allem statische Modelle eingesetzt, sind es heute rechenintensivere Ansätze, welche auf künstlicher Intelligenz basieren.

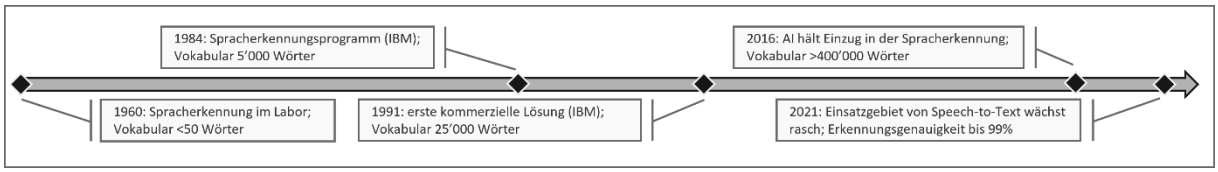

Abbildung 2: Fortschritt der automatischen Spracherkennung

Dies hat einen grossen Vorteil: KI-basierte Systeme stehen dem Benutzer ohne Lernphase der Spracherkennungssoftware zur Verfügung. Früher mussten sich Sprecher und Maschine mithilfe eines Trainings «kennenlernen». Das heisst, das System lernte anhand vorgegebener Wörter und Phrasen zu verstehen, wie sich der Sprecher anhört. Dies hatte eine weitere Einschränkung zur Folge. Die Software konnte nur mit dem angelernten Sprecher gute Resultate erzielen. Hat ein anderer Benutzer (ohne Training) die gleiche Konfiguration benutzt, sind die Transkriptionsresultate viel schlechter ausgefallen, weil das System auf diesen Sprecher nicht vorbereitet wurde.

Nebst dem Wegfall des initialen Trainings zur Konfiguration des Systems, bieten die modernen Lösungen einen weiteren Vorteil. Sie lernen permanent, wie sich der Benutzer ausdrückt (bevorzugte Wörter und Wendungen). Jede Transkription fliesst in das persönliche Sprach-Profil eines Sprechers ein und kann in der Folge wieder angewendet werden. Auf diese Weise verbessert sich die Erkennungsrate kontinuierlich. Erkennungsraten von bis zu 99% werden heute von diversen Lösungen erreicht. Dieser Wert hängt verständlicherweise von verschiedenen Faktoren ab: Thema, Länge, Komplexität des Inhalts etc.

1.5 Funktionsumfang

Spracherkennung ist nicht gleich Spracherkennung. Aus technischer Sicht ist diese Aussage tendenziell falsch. Immer häufiger werden die gleichen Konzepte umgesetzt. Anders sieht es bei der Betrachtung der Einsatzgebiete aus. Diese unterscheiden sich hinsichtlich diverser Aspekte wie beispielsweise Bedienung, Funktionsumfang, Komplexität und Vokabular.

1.5.1 Grundfunktion

Grundsätzlich geht es immer darum, **gesprochene Sprache in lesbaren Text umzuwandeln**. Der produktgerechte Einsatz solcher Systeme führt heutzutage zu hervorragenden Transkriptionsresultaten. Will heissen, dass Sprachassistenten (siehe Abschnitt 2.1.1) passende Antworten auf einfache Fragen liefern. Beim Versuch, beispielsweise einen medizinischen Fachartikel mit einem Sprachassistenten wie Siri aufzuzeichnen, würde sich hingegen ein Resultat ergeben, welches viel Nacharbeit nach sich ziehen würde.

1.5.2 Zusatzfunktionen

Ein Blick auf den Markt der Speech-to-Text-Lösungen zeigt, dass die Funktionspaletten stetig umfangreicher werden. Die einstigen Differenzierungsmerkmale einzelner Produkte finden sich mehr und mehr in einer Vielzahl von Produkten wieder. Dennoch gibt es einige nützliche Funktionen, welche einerseits nicht in jeder Lösung zu finden sind und andererseits bei der Evaluation eines Produkts durchaus eine entscheidende Rolle spielen könnten. Nachfolgende Auflistung zeigt einige solcher Features:

- Einbindung fachspezifischer Vokabulare (z.B. Medizin, Justiz)
- Möglichkeit, bestehende Vokabulare zu ergänzen (Individualisierung)
- Anlegen eines oder mehrerer Benutzerprofile (z.B. für den Einsatz unterschiedlicher Sprachen)
- Verarbeitung von Audiodateien (mp3, wav etc.)
- Verarbeitung von Videodateien (z.B. für die Erzeugung von Untertiteln)
- Unterschiedliche Ausgabeformate (rft, txt, doc, docx etc.)
- Automatische Erkennung von Satz- und Formatierungszeichen
- Individuelle Sprachbefehle für eigene Textbausteine und Signaturen
- Individuelle Formatierungsbefehle
- Programmsteuerung per Sprache

Die Liste ist keineswegs abschliessend. Sie soll darauf hinweisen, dass inzwischen bei der Suche nach passenden Speech-to-Text-Lösungen viele Anforderungsparameter berücksichtigt werden können.

2 Marktübersicht

2.1 Produktelandschaft

Die Verbreitung der Spracherkennungssysteme lässt sich grob in drei Bereiche aufteilen. Es gilt jedoch anzumerken, dass bei dieser Kategorisierung keine messerscharfe Abgrenzung möglich ist. Mittlerweile gibt es Produkte, die in mehreren Kategorien auftauchen und Einsatz finden.

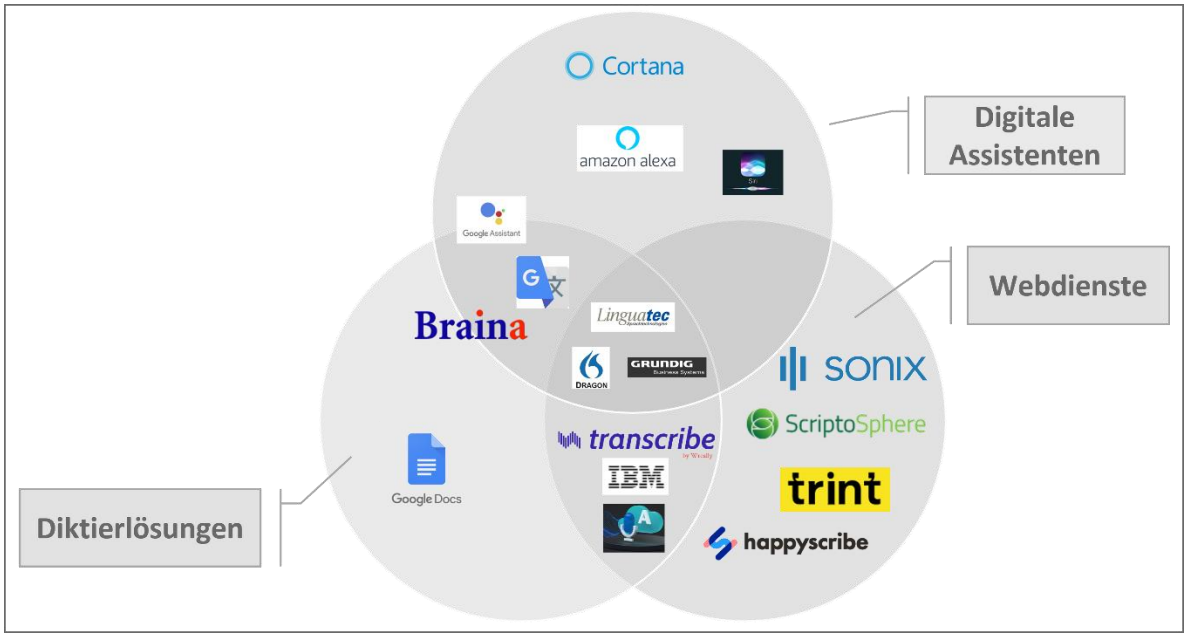

Abbildung 3: Produktelandschaft im Bereich der Spracherkennung

2.1.1 Digitaler Assistent

In die Kategorie der digitalen Assistenten fallen beispielsweise Siri von Apple oder Alexa von Amazon etc., aber auch das sprachgesteuerte Navigationssystem im Fahrzeug. Diese Dienste sind darauf ausgerichtet, einfachste Anweisungen auszuführen. Solange es darum geht, einen Wetterbericht vorzulesen, eine Zugverbindung vorzuschlagen, einen Kalendereintrag zu erfassen oder einen Kontakt im Telefon aufzurufen, funktionieren diese Services sehr gut. Zweifellos weniger oder gar nicht geeignet sind sie, um längere Texte aufzuzeichnen oder gar sauber formatierte Dokumente zu generieren.

Zudem ist es in der Regel nicht möglich, das Vokabular (Menge der Sprachbefehle) eines solchen Assistenten zu erweitern. Dies im Gegensatz zu den beiden anderen Kategorien.

2.1.2 Webdienst

Webdienste sind dann hilfreich, wenn es darum geht, Vorträge, Vorlesungen, Interviews oder dergleichen zu verschriftlichen. Der Fokus liegt hier klar bei der schriftlichen Ablage des Gesprochenen. Anforderungen wie eine saubere

Formatierung oder orthografisch korrekte Texte zu erhalten, sind von geringer Relevanz und werden aktuell nur von wenigen Diensten angeboten.

Ein ebenfalls oft genanntes Kriterium für den Einsatz von online Speech-to-Text-Lösungen ist die ortsunabhängige Verfügbarkeit.

Heute gibt es Webdienste, bei denen es möglich ist, das Vokabular zu erweitern. Hingegen fehlt die Bereitstellung von fachspezifischen Wort- und Ausdruckssammlungen.

2.1.3 Diktierlösung

Gerade in den Berufsgruppen Medizin und Justiz, aber auch in anderen administrativen Arbeitsbereichen, werden heute Diktierlösungen sehr verbreitet eingesetzt. Bei entsprechenden Produkten besteht der Anspruch an eine hohe Texterkennung inklusive der Anwendung von fachspezifischen Vokabularen. Zudem wird erwartet, dass die Texte per Sprachsteuerung formatieren werden können.

Im Hinblick auf die Individualisierung des Systems bieten diese Lösungen grosse Vorteile. Üblicherweise lässt sich das Vokabular mit eigenen Ausdrücken erweitern. Zudem lassen sich neue Sprachbefehle definieren, welche dazu eingesetzt werden können, um persönliche Standardtexte oder präferierte Formatierungen zu verwenden.

Diese Systeme lernen durch den Einsatz von Machine Learning-Methoden die Sprachgewohnheiten des Sprechers kennen und anwenden.

2.2 Produktevergleich

Produkt	happyscribe	Trint	Transcribe	Dragon	Watson	Azure	Braina
Hersteller	Happy Scribe	Trint	Wreally	Nuance	IBM	Microsoft	Brainasoft
URL	happyscribe.com	trint.com	transcribe.wreally.com	nuance.com	ibm.com/cloud/watson-speech-to-text	azure.microsoft.com	brainasoft.com
Systemanforderung	Webbrowser	Webbrowser	Webbrowser	Client-Applikation	Webbrowser	Webbrowser	Client-Applikation
Audiotranskription							
Sprachen	DE, FR, IT, EN (>60 Sprachen)	DE, FR, IT, EN (ca. 50 Sprachen)	DE, FR, IT, EN (ca. 70 Sprachen)	DE, FR, IT, EN (ca. 8 Sprachen)	DE, FR, IT, EN (ca. 10 Sprachen; mehrere Dialekte)	DE, FR, IT, EN (ca. 40 Sprachen)	DE, FR, IT, EN (ca. 100 Sprachen)
Sprachauswahl	manuell	manuell	manuell / automatisch	manuell	manuell	manuell	manuell
Fachvokabulare	nein	nein	nein	ja, Justiz & Medizin	nein	nein	nein
Vokabular erweiterbar	ja, max. 100 Wörter	ja, max. 100 Wörter	nein	ja	ja	ja	ja
Simultan-Transkription	nein	ja (via Streaming)	ja (eigener Editor)	ja (eigener Editor oder direkt in Zielapplikation)	ja	ja	nein
Datei-Upload	ja	Ja	ja	ja	ja	ja	nein
Inputformate (Audio)	AAC, AIFF, FLAC, M4A, MP3, Ogg Vorbis, WAV, WMA u.a.	AAC, M4A, MP3, WAV, WMA	AAC, AIFF, FLAC, M4A, MP3, Ogg Vorbis, WAV, WMA u.a.	MP3, WAV, M4A, WMA, DSS, DS2, AIFF, M4V	MP3, MPEG, WAV, FLAC, OPUS	MP3, OPUS/OGG, FLAC, ALAW, MULAW	-
Outputformate	TXT, DOC, PDF, JSON, STL, SRT, VTT u.a.	DOCX, SRT, VTT, TXT, STL, EDL, HTML, XML, CSV	DOC	RTF	Text	Text	direkt in Zielapplikation Human Language Interface (Braina nativ)
Textformatierung	ja	ja	nein	ja	teilweise	nein	nein
Weitere Funktionen		autom. Übersetzung	Ansätze Mehrsprechererkennung (Experimental Mode)	Batchverarbeitung mehrerer Audioaufnahmen	Ansätze Mehrsprechererkennung (nur Englisch)	autom. Satzzeichenerkennung	
Betriebsmöglichkeiten							
SaaS / Cloud	ja	ja	ja	nein	ja	ja	nein
On Premise	nein	nein	nein	nein	ja	ja	nein
Client-Installation	-	-	-	ja	nein	nein	ja
Preise / Abrechnung							
Lizenz	-	-	-	Home: ca. CHF 220 Professional: ca. CHF 750	-	-	ca. $ 300 (lifetime)
Nutzung	€ 12 pro Stunde	ab € 55 pro Monat	$ 20 pro Jahr und $ 6 pro Stunde	-	ca. CHF 1.15 pro Stunde	Std: CHF 0.99 pro Stunde Cust: CHF 1.4 pro Stunde	ca. $ 60 pro Jahr (alternativ)

Tabelle 1: Gegenüberstellung ausgewählter Spracherkennungslösungen

2.2.1 Grundig Business Systems

Einen interessanten Ansatz verfolgt Grundig Business Systems. Diese auf Spracherkennung spezialisierte Unternehmung stellt eine Service-Palette zu Verfügung, welche auf dem weitverbreiteten und branchenführenden Produkt Dragon der Firma Nuance aufbaut. Funktional ist die Lösung identisch wie Dragon. Auf konfigurativer Seite stellt der Anbieter jedoch optimierte Konfigurationen zur Verfügung. Für den Einsatz in der Deutsch-Schweiz gibt es beispielsweise angepasste Vokabulare (ss statt ß).

2.2.2 Abgrenzungen der Betrachtung

In der Tabelle 1 sind ausgewählte Produkte aus den Bereichen Webdienste und Diktierlösungen gegenübergestellt. Reine digitale Assistenten wurden bewusst nicht weiter untersucht, da sie im vorliegenden Kontext nicht dem eigentlichen Einsatzgebiet entsprechen.

3 Betriebliche Aspekte

Der Abschnitt über betriebliche Aspekte befasst sich primär mit den Betriebs- und Kostenmodellen. Zudem werden einige Einflussfaktoren aufgelistet, welche die Qualität der Spracherkennung positiv beeinflussen.

3.1 Betriebsmodelle

Das gesamte Spektrum der denkbaren Betriebsmöglichkeiten für Spracherkennungssysteme wird heute abgedeckt. Von lokalen Installationen auf Arbeitsplatzrechnern oder Verteilung über Citrix-Umgebungen über Cloud Services, welche eigen- oder fremdbetrieben werden, bis hin zu Mobile Apps findet sich alles auf dem Markt.

Webdienste können in der Regel nicht in eigenen Rechenzentren betrieben werden. Aktuell bilden die bekannteren Namen auf dem Markt hier sicherlich noch die Ausnahme. Denn sowohl IBM als auch Microsoft bieten den sogenannten on premise-Betrieb ihrer Spracherkennungslösungen an.

Bei Diktatlösungen, welche nebst der lokalen Installation auch einen Server zu Verfügung stellen, ist es in der Regel möglich, diese Infrastruktur in einem eigenen Rechenzentrum aufzubauen. Der Vorteil einer solchen Lösung liegt sicherlich darin, dass die Datenschutzanforderungen eingehalten werden können. Zudem ermöglicht der eigene Betrieb auch eine lieferantenunabhängige Pflege der System- und Benutzer-Konfigurationen sowie der Zugangsrechte. Üblicherweise ist der Sinn solcher Lösungen, dass auf dem Server gemeinsame Vokabulare und zentrale Konfiguration abgelegt sind.

3.2 Lizenzmodelle und Kosten

In den beiden näher betrachteten Kategorien der Speech-to-Text-Lösungen kommen unterschiedliche Lizenzmodelle und Preispläne zum Einsatz. Dabei ergeben sich kostenseitig mitunter massive Unterschiede (siehe Tabelle 1, Abschnitt ‘Preis / Abrechnung’).

3.2.1 Lizenzmodelle

Im Bereich der extern betriebenen Webdienste ähneln sich die Lizenzmodelle. Hier wird üblicherweise über die Diktatdauer (Zeit) abgerechnet. Teilweise erheben Anbieter eine zusätzliche Grund- oder Jahresgebühr (z.B. transcribe der Firma Wereally).

Für eigenbetriebene Lösungen kommen herkömmliche Lizenzierungsmodelle zum Einsatz. Einmallizenzen (lifelong) sind weit verbreitet. Ergänzt werden diese Preispläne öfters mit Gruppen- oder Jahreslizenzen. In diesem Betriebssetup ist es zudem üblich, Support- und Wartungsverträge abzuschliessen.

3.2.2 Kosten

Aufgrund der Abrechnungsmodelle und den resultierenden eher tiefen Kosten pro Zeiteinheit eignen sich Webdienste hervorragend, um Speech-to-Text-Lösungen auszuprobieren und kennenzulernen. Ebenso fallen für einen nur gelegentlichen Gebrauch eines solchen Services verhältnismässig tiefe Kosten an. Dies im Vergleich zu Lösungen mit Einmal- oder Jahreslizenzen. Solche Produkte zahlen sich bei häufigem Einsatz (täglich mehrere Stunden) aus. Fairerweise gilt es anzumerken, dass es sich hierbei um Spracherkennungssysteme handelt, welche einen grösseren Funktionsumfang bieten. Typischerweise sind diese in der Kategorie der Diktierlösungen angesiedelt.

In der Tabelle 1 ist das breite Kostenspektrum ersichtlich, welches sich nur schon bei der Gegenüberstellung weniger Produkte ergibt.

3.3 Transkriptionsqualität

Die mehrheitlich ohnehin bereits erstaunlichen Resultate der Speech-to-Text-Lösungen lassen sich unter Berücksichtigung einiger einfacher Tipps weiter verbessern:

- Gute Ergebnisse werden durch eine klare und deutliche Aussprache erreicht. Es empfiehlt sich, das Diktieren zu üben.
- In natürlichem Tonfall und in üblicher Sprechgeschwindigkeit diktieren. Weder extra langsames noch speziell lautes Diktieren verbessert die Resultate; im Gegenteil, die Qualität wird dadurch eher verschlechtert, da die Analysemodelle auf einer natürlichen Art der Sprechweise basieren.
- Satz- und Formatierungszeichen müssen stets diktiert werden.
- Es empfiehlt sich, die wichtigsten Formatierungsbefehle und Tastenkürzel (auswendig) zu lernen.
- Geringe oder konstante Umgebungsgeräusche helfen die Resultate zu verbessern. Wie eingangs im Abschnitt 1.3.1 erwähnt, kann ein Headset-Mikrofon beispielsweise in nicht ganz stillen Umgebungen deutliche Verbesserungen bringen.

Ergänzend zu den obigen Punkten an dieser Stelle weitere Erkenntnisse aus einer mehrmonatigen Testphase:

- Viele Produkte sind so konzipiert, dass sie durch den Benutzer erweitert und angepasst werden. Im Bereich der Vokabulare sollte dies unbedingt genutzt werden.
- Optimierungen bei bestehenden Steuerungsbefehlen, Tastaturkürzeln, Textblöcken etc. erleichtern die Arbeit um einiges.
- Einarbeitung ist keine verschwendete Zeit.

4 Einsatzmöglichkeiten in der Justiz

4.1 Kanzlei-Arbeit und Administration

In der Justiz lassen sich Speech-to-Text-Lösungen (Diktierlösungen) sehr gut als Ersatz für die heute oft eingesetzten Diktiergeräte einsetzen. Sie ermöglichen beispielsweise einem Richter fallspezifische Korrespondenz, Verfügungen, Entscheide und ähnliches direkt zu diktieren und verfassen. Der herkömmliche Weg, das aufgezeichnete Diktat durch Sekretariatsmitarbeitende abzutippen, entfällt.

Selbstverständlich lässt sich ein Spracherkennungssystem auch in die tägliche Sekretariatsarbeit einbinden. Mit der Möglichkeit vordefinierte Texte über Sprachbefehle zu abzurufen, entstehen Briefe in Sekundenschnelle. Die gleiche Vorgehensweise lässt sich auch auf die Verfassung von E-Mails anwenden.

4.2 Einsatz bei Einvernahmen und Gerichtsverfahren

Gerade in der Justiz gibt es Anwendungsfälle, bei denen der Einsatz einer automatischen Spracherkennung auf der Hand zu liegen scheint. So könnte beispielsweise die Protokollierung von Einvernahmen oder Gerichtsverhandlungen um einiges vereinfacht werden, wenn eine Speech-to-Text-Lösung zum Einsatz käme. Da sich bei diesen Anwendungsfällen weitere Anforderungen stellen, gilt es zu prüfen, ob diese durch den heutigen Entwicklungsstand entsprechender Systeme abgedeckt wird.

Bei den beiden erwähnten Anwendungsfällen gilt es zwei Dinge zu beachten:

1. **Sprechersprache / Dialekt**: Sowohl in Gerichtsverhandlungen als auch bei Einvernahmen wird häufig nicht die Standardsprache (Schriftsprache) verwendet, sondern führt diese im persönlichen Dialekt. Bekanntlich unterscheiden sich Dialekte regional teilweise erheblich. Aus systemischer Sicht wird die Sprachvielfalt dadurch noch grösser und die Erkennung von dialektspezifischen Nuancen um einiges komplexer.
2. **Mehrsprechererkennung**: Spracherkennungssysteme arbeiten heute oft basierend auf Benutzerprofilen und sind darum auf einen Sprecher ausgelegt. Zwar kann ein System die Eingaben von verschiedenen Personen erkennen und verschriftlichen, aber eine zuverlässige Zuordnung zum Sprecher (falls eine solche überhaupt vorgesehen ist), ist noch nicht in der gewünschten Zuverlässigkeit möglich.

Beide der oben angesprochenen Punkte sind den Produkteanbietern von Spracherkennungssystemen bekannt. Darum investieren diese in die Forschung und Entwicklung zur besseren Erkennung von Sprechern und Dialekten. Grundig Business Systems arbeitet diesbezüglich mit dem Fraunhofer-Institut zusammen.

Zurzeit sind auf dem Markt solche Lösungen zur automatischen Mehrsprechererkennung, die in der Praxis eingesetzt werden können, jedoch noch nicht verfügbar. IBM Watson scheint in diesem Feld dennoch recht weit fortgeschritten, zumindest für Dialoge in englischer Sprache. Minderheitensprachen wie das Schweizerdeutsch und seine Dialekte sind aufgrund ihrer Verbreitung auf Nischenanbieter wie Recapp oder Spitch angewiesen, welche sich genau diesem Thema verschrieben haben.

4.3 Zusammenfassung der Einsatzmöglichkeiten

Die folgende Tabelle fasst die wesentlichen Aussagen zum Einsatz von Spracherkennungssystemen in der Justiz zusammen:

Geeignet	(noch) nicht geeignet
Administration • Briefe, E-Mails • Anwendung von Standardtexten und -dokumenten	**Mehrsprechersituationen** • Einvernahmen (Dialog) • Gerichtsverhandlungen
Kanzleiarbeit • Entscheide, Anordnungen etc. • fallspezifische Korrespondenz	**Minderheitensprachen / Dialekte** • Rätoromanisch • Schweizerdeutsch • starke Akzente
Überall, wo heute das Diktiergerät eingesetzt wird	Fremdsprachenübersetzungen

Tabelle 2: Einsatzmöglichkeiten in der Justiz

Der Hauptnutzen des Einsatzes von Speech-to-Text-Systemen in der Justiz liegt **zurzeit** sicherlich bei der direkten Erfassung von Texten innerhalb der Zielapplikationen.

4.4 Einfluss auf Arbeitsprozesse

Wie im vorigen Abschnitt beschrieben, lassen sich mithilfe von Diktiersoftware gewisse Arbeitsschritte vereinfachen, beispielsweise das direkte Erfassen eines Entscheides durch einen Richter. Das bedeutet auf der anderen Seite, dass herkömmliche Tätigkeiten für Administrations- oder Sekretariatsmitarbeitende entfallen. Anders ausgedrückt, können sich dadurch Tätigkeitsfelder verändern, was Chancen mit sich bringt, jedoch auch gewisse Risiken birgt. In der Tabelle 3 sind einige Aspekte dazu aufgelistet:

Chancen	Risiken
Effizientere Prozess • Direkte Erstellung des Dokuments • Keine manuelle Transkription • «sprechen geht schneller als tippen»	Anpassungen bestehender und funktionierender Abläufe
Qualitätsverbesserung	Fehlende Akzeptanz
Moderne Technologien und Methoden	Nutzen nicht ersichtlich
Abbau von Spezialhardware • Diktiergeräte • Fusspedale	Falsche Erwartungen (an das System)
Anpassung der Job-Profile • neue Aufgaben • mehr Verantwortung	Technische Herausforderungen
	Zusatzaufwand in der Einführungsphase

Tabelle 3: Chancen und Risiken durch den Einsatz von Spracherkennungssystemen

5 Speech-to-Text-Lösung beschaffen und einführen

5.1 Zentrale Fragestellungen

Sobald die Beschaffung eines Produkts – im vorliegenden Fall einer Speech-to-Text-Lösung – im Raum steht, gilt es einige grundlegende Fragestellungen zu klären. Dabei wird das Ziel verfolgt, die Rahmenbedingungen an den Betrieb und die Eckdaten zum Produkt ins Bewusstsein zu rufen. Auf der Basis dieser Fragen werden im Anschluss die funktionalen und nicht funktionalen Anforderungen abgeleitet und formuliert. Mögliche Fragestellungen hierzu finden sich in der folgenden Tabelle:

Fragen im Hinblick auf nicht funktionale Anforderungen
Wo muss/soll die Lösung betrieben werden? (inhouse / extern)
Wer muss/soll die Lösung pflegen / warten? (intern / extern)
Wie viele Nutzer wird das System haben?
Wie setzen sich die Benutzergruppen zusammen?
Wie viele Nutzer wird das System haben?
Welche Datenschutzaspekte müssen berücksichtigt werden?
Wie sieht der Kostenrahmen aus? (Budget)
Wie präsentiert sich die Systemlandschaft auf Benutzerseite?
Welche betrieblichen Restriktionen bzw. Anforderungen bestehen auf Benutzerseite?
Welche Vorgaben gibt es, in Bezug auf den Einsatz von Software-Produkten?
Fragen im Hinblick auf funktionale Anforderungen
Welche Transkriptionssprachen müssen unterstützt werden?
Welche Funktionen muss die Lösung bereitstellen? (Konkretisierung mithilfe Tabelle 1)
Welche Datenformate (Input / Output) müssen unterstützt werden?
Welche Eingabemöglichkeiten muss die Lösung bereitstellen?
Welche Konfigurationsmöglichkeiten muss die Lösung anbieten?
Welche konkreten Probleme werden durch den Einsatz einer Speech-to-Text-Lösung behoben oder reduziert?

Tabelle 4: Auswahl grundlegender Fragestellungen zur Produktebeschaffung

5.2 Vorgehensplanung

Wie bei jedem Projekt, gilt es auch bei der Beschaffung und Einführung eines Spracherkennungssystems eine klare Vorgehensplanung zu erstellen. Diese lässt sich grob in zwei Hauptphasen aufteilen: Vorbereitung sowie Beschaffung und Inbetriebnahme.

5.2.1 Vorbereitung

Die Vorbereitungsphase beginnt mit der vertieften Auseinandersetzung mit den funktionalen und nicht funktionalen Anforderungen. Diese können entlang von grundlegenden Fragestellungen (siehe beispielsweise Abschnitt 5.1) erhoben und

verfeinert werden. Sobald die Anforderungen geklärt und erfasst sind, kann mit der Evaluation eines geeigneten Produkts begonnen werden. Dazu gilt es mögliche Lösungen anhand von vordefinierten Kriterien (gemäss den Anforderungen) zu bewerten.

Ein wesentlicher Teil der Anforderungsbeschreibung ist die Auflistung der jeweiligen Akzeptanzkriterien. Diese werden nebst der Evaluation des Produkts auch für die Beschreibung der Testfälle herangezogen. Eine Tätigkeit, welche unbedingt in der Vorbereitungsphase zu erfolgen hat.

Planerische Aspekte wie die Klärung der Zeit- und Ressourcenverhältnisse und die Ableitung eines Umsetzungsplan inklusive Festlegung der Projektmethodik bilden einen dritten wichtigen Teil in der Vorbereitungsphase.

5.2.2 Beschaffung und Inbetriebnahme

Die Installation des evaluierten (und beschafften) Produkts steht am Anfang der Inbetriebnahme. Dabei wird das Speech-to-Text-System in der Zielumgebung bzw. in den Zielumgebungen installiert und konfiguriert.

Nach einer Benutzerschulung in der nötigen Tiefe, sollten erst einmal im Rahmen eines Pilotbetriebs erste Erkenntnisse und Erfahrungen gesammelt werden. Es empfiehlt sich, den Nutzerkreis zu Beginn ziemlich eingeschränkt zu halten, damit die Feedbacks überschaubar bleiben und rasch in die Optimierung und Konfiguration einfliessen können. Gegebenenfalls gilt es auch die Anforderungen aufgrund der gewonnenen Informationen zu schärfen.

Davon ausgehend, dass die Pilotphase erfolgreich verläuft (positiver Entscheid zum Einsatz der evaluierten und getesteten Lösung), folgt anschliessend das Ausrollen an den endgültigen Benutzerkreis. Dies sollte ebenfalls schrittweise und in überschaubaren Gruppengrössen erfolgen. Nach welchen Kriterien (Sprache, Kanton, Amt etc.) dabei vorgegangen wird, bleibt zu bestimmen.

Die Erfahrung zeigt, dass für die Einführung und Kennenlernphase genügend Zeit eingeplant werden sollte. In einer beispielsweise zweimonatigen Testperiode haben die Benutzer genügend Zeit, die unterschiedlichen Facetten des Systems kennenzulernen.

6 Empfehlungen

In Besprechungen mit der HIS-Projektleitung hat sich gezeigt, dass vor allem der Einsatz von Speech-to-Text-Lösungen **bei Einvernahmen und Gerichtsverhandlungen von grossem Interesse ist**. Wie erläutert, stehen zurzeit noch keine praxistauglichen Systeme hierfür zur Verfügung. Aus diesem Grund macht der Aufbau einer Testumgebung im HIS-Umfeld zum aktuellen Zeitpunkt (noch) keinen Sinn.

Nichtsdestotrotz empfiehlt es sich aufgrund der Entwicklungsaktivitäten im Bereich der Mehrsprechererkennung und der Transkription von Dialekten das **Thema** wenigstens am Rande **weiterzuverfolgen** und mit Anbietern von potenziellen Lösungen im Kontakt zu bleiben.

Zudem ist es empfehlenswert, über die **konkreten Anforderungen an ein Spracherkennungssystem nachzudenken** und diese zu sammeln. Dabei spielen funktionale und nicht funktionale Anforderungen eine zentrale Rolle für die Auswahl einer passenden Speech-to-Text-Lösung (vgl. Tabelle 1).