M14 - Multimodel Audio

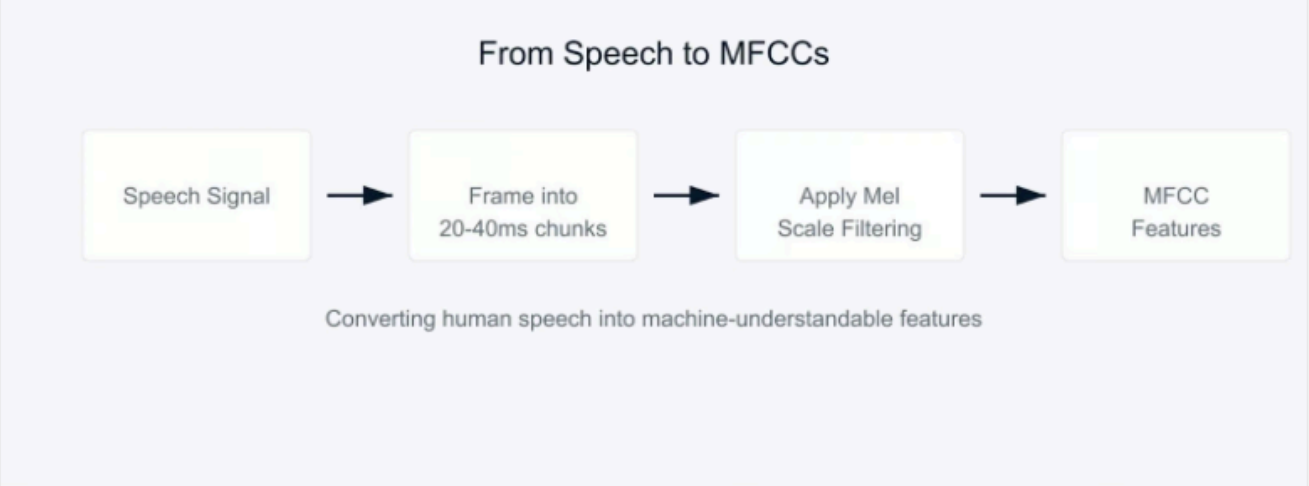
Stand: 03.2025

1 Grundlagen Audioerkennung

Ein Computer "hört" ein Audiosignal nicht wie ein Mensch. Für ihn stellt ein Audiosignal lediglich eine Sequenz von Zahlenwerten dar, wobei jeder Wert die Amplitude des Schalldrucks zu einem bestimmten Zeitpunkt repräsentiert. Diese Zahlenwerte werden von Algorithmen verarbeitet, um Muster und Strukturen zu erkennen, die für die Audioerkennung essenziell sind.

Der Prozess der Audioerkennung umfasst folgende Schritte:

1. **Audioerfassung**: Ein Audiosignal wird in eine numerische Darstellung umgewandelt. Dabei werden Audioformate wie WAV oder MP3 in eine Sequenz von Amplitudenwerten umgerechnet.
2. **Vorverarbeitung**: Das Signal wird normalisiert, gefiltert und aufbereitet. Dies kann Schritte wie Rauschunterdrückung, Amplitudenanpassung oder Frequenzfilterung umfassen.
3. **Merkmalsextraktion**: Wichtige Merkmale wie Frequenzspektren, Tonhöhen oder Rhythmusmuster werden identifiziert. Hier können Methoden wie Fast Fourier Transform (FFT), **Mel-Frequency Cepstral Coefficients (MFCCs)** oder Spektrogramme angewendet werden.



Mel-Frequency Cepstral Coefficients

Bild: Merkmalsextraktion beim maschinellen Lernen: Ein vollständiger Leitfaden | DataCamp

4. **Klassifikation**: Basierend auf den extrahierten Merkmalen erfolgt eine Klassifikation des Audiosignals. Dies geschieht mithilfe von maschinellen Lernmodellen oder regelbasierten Systemen. MediaPipe)

2 Methoden

2.1 Traditionelle Methoden

Bei traditionellen Methoden müssen explizit Merkmale definiert werden, die als relevant gelten (z.B. Frequenzspektrum, Rhythmus, Tonhöhe). Klassische Verfahren beinhalten Methoden wie:

- **Frequenzanalyse** mittels Fourier-Transformation oder Wavelet-Transformation
- **Merkmalsvektoren** wie MFCCs (Mel-Frequency Cepstral Coefficients) oder Chroma-Features
- **Dynamic Time Warping (DTW)**, um zeitliche Muster in Audiosignalen zu vergleichen

Diese Verfahren erfordern umfassendes domänenspezifisches Wissen und sind oft anfällig für Variationen in Lautstärke, Tonhöhe oder Hintergrundgeräuschen.

2.2 Deep Learning

Moderne Ansätze setzen auf neuronale Netze, insbesondere Convolutional Neural Networks (CNNs) und Recurrent Neural Networks (RNNs), die eigenständig lernen, welche Merkmale relevant sind. Diese Netzarchitekturen bestehen aus mehreren Schichten, die folgende Aufgaben erfüllen:

- **CNNs mit 1D-Faltung oder 2D-Faltung auf Spektrogrammen**: Extrahieren lokale Merkmale im Zeit- oder Frequenzbereich
- **RNNs (LSTM, GRU)**: Modellieren zeitliche Abhängigkeiten im Audiosignal
- **Attention-Mechanismen**: Fokussieren auf relevante Teile des Audiosignals
- **Transformer-Architekturen**: Verarbeiten lange Sequenzen mit globaler Kontexterfassung

Deep-Learning-Modelle werden auf große Datensätze trainiert, wodurch sie eine hohe Generalisierungsfähigkeit erreichen und in der Lage sind, komplexe Muster autonom zu lernen.

3 Audio-Modelle

3.1 Text-zu-Audio-Modelle

In diesem Abschnitt wird der Fokus auf Text-zu-Audio-Modelle gelegt, einem innovativen Bereich der künstlichen Intelligenz, der es Maschinen ermöglicht, Audioinhalte basierend auf Textbeschreibungen zu erzeugen. Diese Modelle überbrücken die Kluft zwischen Sprache

und akustischen Inhalten, indem sie eine natürliche Spracheingabe in eine vollständige Audiodarstellung umsetzen. Die Generierung von Audio aus Text basiert auf Deep-Learning-Methoden, insbesondere durch die Kombination von natürlicher Sprachverarbeitung (NLP) und Audio-Signalverarbeitung.

Ein zentrales Merkmal dieser Modelle ist ihre Fähigkeit, Zusammenhänge zwischen sprachlichen und akustischen Elementen zu erfassen. Durch das Training mit umfangreichen Datensätzen, die Texte mit den dazugehörigen Audiodaten verknüpfen, lernen sie, sprachliche Beschreibungen wie „ein Gewitter mit starkem Regen" mit relevanten akustischen Eigenschaften wie Klangfarbe, Rhythmus, Tonhöhe und zeitlichen Mustern zu verbinden. Modelle wie AudioLM, MusicLM und AudioGen haben das kreative Potenzial dieser Technologie eindrucksvoll demonstriert, indem sie sowohl realistische Umgebungsgeräusche als auch musikalische Kompositionen direkt aus textlichen Vorgaben generiert haben.

Hier wird veranschaulicht, wie ein KI-Modell ein Audiobeispiel mit Naturgeräuschen eines Waldes generiert:

3.2 Multimodale Modelle

In diesem Abschnitt wird die spannende Welt multimodaler Modelle untersucht, die verschiedene Datentypen wie Text, Bilder, Audio und Video verarbeiten und zu einem ganzheitlichen Verständnis verknüpfen. Diese Modelle markieren einen bedeutenden Fortschritt in der KI, da sie Eingaben aus unterschiedlichen Quellen kombinieren und dadurch komplexere Aufgaben bewältigen können – ähnlich wie der Mensch Informationen aus verschiedenen Sinneseindrücken zusammenführt.

Während spezialisierte Modelle, etwa Text-zu-Audio-Systeme, bereits einzelne Verbindungen zwischen Modalitäten ermöglichen, gehen multimodale Modelle einen Schritt weiter. Sie analysieren und verstehen die Beziehungen zwischen verschiedenen Informationsarten, was vielfältige Anwendungen ermöglicht. Dazu zählen das Generieren von Audiobeschreibungen zu Bildern, das Synchronisieren von Lippenbewegungen in Videos mit Audioinhalten oder die Erstellung von Soundtracks auf Basis von visuellen und textuellen Eingaben.

Ein anschauliches Beispiel für multimodale Fähigkeiten ist die Analyse eines Videos mit einem musikalischen Auftritt. Ein solches Modell kann die visuellen Elemente (Instrumente, Spielbewegungen), den Audioinhalt (Melodie, Rhythmus) und kontextuelle Informationen (Musikstil, Umgebung) erfassen und diese zu einem umfassenden Verständnis der Szene verbinden.

Durch die Kombination verschiedener Datentypen ermöglichen multimodale Modelle eine fortschrittlichere und intuitivere Interaktion zwischen Mensch und KI. Dies macht sie zu einem vielseitigen Werkzeug mit breiten Anwendungsmöglichkeiten – von Gesundheitswesen (Erkennung von Krankheitssymptomen in Sprache) und Barrierefreiheit

(automatische Audiodeskription) bis hin zu Unterhaltung (adaptive Soundtracks) und Musikproduktion (KI-unterstützte Komposition).

4 Audio-Embeddings

Ähnlich wie bei Text-Embeddings und Image-Embeddings, die Textinhalte oder Bilder in einer Weise kodieren, dass semantische Ähnlichkeiten erhalten bleiben, transformieren Audio-Embeddings akustische Merkmale in eine für Maschinen lernbare Form.

Mithilfe neuronaler Netze – typischerweise Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) oder Transformer-Modelle wie Wav2Vec2 oder HuBERT – werden komplexe Audiodaten in kompakte Vektoren umgewandelt. Diese Embeddings ermöglichen Aufgaben wie Audioähnlichkeitssuche, Clustering oder die Kombination von Audio- und anderen Daten für multimodale Modelle.

Audio-Embeddings finden Anwendung in:

- Spracherkennung und Sprecheridentifikation
- Musikgenre-Klassifikation und Musikempfehlungssystemen
- Umgebungsgeräuscherkennung
- Audiosuche und -retrieval
- Multimodaler Verarbeitung mit Text und visuellen Daten

Die Qualität dieser Embeddings hängt maßgeblich von der Trainingsdatendiversität und der Modellarchitektur ab. Moderne Ansätze nutzen selbstüberwachtes Lernen, um auch aus ungelabelten Audiodaten robuste Repräsentationen zu extrahieren.

Audio_Viz