Der Vocoder

Sprachübertragung bei geringer Bitrate

Thomas Emig, DL7TOM

Audiodaten und besonders Sprache werden heute vielerorts digital übertragen. Um Datenbandbreite zu sparen, kommen so genannte Vocoder zum Einsatz. Diese müssen den zu übertragenden Inhalt auf die reine Information reduzieren, dergestalt, dass sie beim Empfänger verstanden werden können.

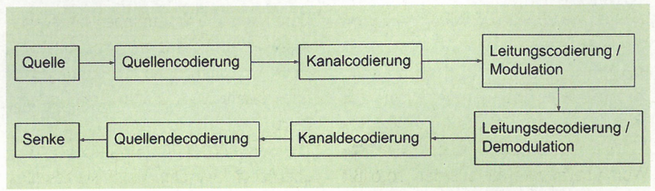

Bild 1: Grundbausteine eines Übertragungssystems

Die digitale Speicherung und Übertragung von Sprache ist ein wichtiger Bestandteil unseres alltäglichen Lebens. Ob wir telefonieren, Radio hören oder wie seit einiger Zeit üblich, an Konferenzen vom eigenen Heim aus teilnehmen. Auch im Amateurfunk finden sich zunehmend Betriebsarten, welche die Sprache digital aufnehmen und übertragen. Dabei können diese Betriebsarten zusätzlich zur Sprache oft viele weitere Informationen übermitteln. Eine kleine Auswahl umfasst unter anderem die Rufzeichenkennung, Textnachrichten und GPS-Standorte. Oft kann auch ein HF-Kanal in mehrere digitale Kanäle aufgeteilt werden, sodass dort mehrere Gespräche gleichzeitig stattfinden können. Um alle diese Funktionen realisieren zu können, ist es essenziell, die Datenrate für die Sprachübertragung möglichst gering zu halten.

Übertragungssystem

Üblicherweise besteht ein Übertragungssystem aus den in **Bild 1** dargestellten Bausteinen. In der oberen Zeile ist mit Quelle, Quellencodierung, Kanalcodierung und Modulation der Sender dargestellt. Dieser ist über den Kanal (hier als einfache Verbindung gezeichnet) mit dem Empfänger verbunden. Im Empfänger werden die im Sender vorgenommenen Codierungen durch die entsprechenden Decodierungen so weit wie möglich rückgängig gemacht. Dadurch entsteht ein Ausgangssignal, das je nach Anwendungsfall entweder eine möglichst exakte Repräsentation des Eingangssignals an der Quelle ist oder zumindest den Informationsgehalt des Eingangssignals möglichst gut wiedergibt.

Die Kanalcodierung dient in diesem System der Erkennung und Korrektur von Fehlern, die bei der Modulation, Demodulation und Übertragung über den Kanal passieren können. Dazu werden die physikalischen Eigenschaften des Kanals betrachtet und fließen in die Wahl des Codierungsverfahrens mit ein. Bei diesem Vorgang wird die Datenrate erhöht, um eine bessere Fehlersicherheit gewährleisten zu können.

Die Quellencodierung hingegen wird eingesetzt, um die Datenrate des Signals aus der Quelle zu verringern. Dies geschieht, indem die physikalischen Eigenschaften der Quelle oder der Senke betrachtet werden und daraus entsprechende Codierungsvorschriften abgeleitet werden. Werden Sprachsignale verarbeitet, wird der Quellen(de)codierer als Vocoder bezeichnet.

Für den hier vorgestellten Vocoder werden nur die Eigenschaften der Quelle betrachtet. In diesem Fall ist die Quelle der menschliche Sprechapparat. Bei anderen Codierverfahren, beispielsweise dem weit verbreiteten MP3-Verfahren werden auch die physikalischen Eigenschaften der Senke in Betracht gezogen (in diesem Fall das menschliche Ohr).

Datenrate

Wird ein Sprachsignal direkt mit einem Mikrofon und A/D-Wandler aufgenommen, so muss das Signal mit einer Abtastrate digitalisiert werden, die mindestens dem doppelten der maximal zu erwartenden Signalfrequenz entspricht.

Für ein typisches Sprachsignal kann von einer maximalen Signalfrequenz bis etwa 2,7 kHz ausgegangen werden. Theoretisch müsste die Abtastrate daher bei etwa 5,6 kHz liegen. Bei einer Quantisierung mit 8 Bit ergibt sich daraus eine Datenrate von etwa 44,8 kBit/s.

Wenn wir davon ausgehen, dass über einen typischen AWGN-Kanal bei einem SNR von 10 dB maximal etwa 3,5 Bit/(s · Hz) übertragen werden können, ergibt sich daraus eine geringste Signalbandbreite von 12,8 kHz. In praktischen Anwendungen wäre die Abtastrate sicherlich noch etwas größer und die Kanalkapazität nicht komplett ausgenutzt, sodass die Bandbreite für die reine Sprachübertragung noch höher liegen würde.

Um die Signalbandbreite möglichst gering zu halten und zusätzlich die oben genannten Zusatzfunktionen realisieren zu können, muss die Datenrate des ursprünglichen Audiosignals reduziert werden.

Redundanz

Wenn wir die Sprache aus einem anderen Blickwinkel betrachten und die Datenrate erneut berechnen, fällt etwas Erstaunliches auf.

Angenommen, natürliche Sprache bedient sich aus einem Wortschatz von 100 000 Wörtern. Um jedem dieser Wörter eine eindeutige Bitfolge zuzuordnen, werden ld(100 000) ≈ 16,6 Bit benötigt. Wenn wir weiterhin annehmen, dass 100 Wörter pro Minute gesprochen werden, so müssen wir zur Übertragung der Sprache 100 Mal pro Minute die Bitfolge für das jeweilige Wort übertragen, die aus den 16,6 Bit besteht. Insgesamt ergibt sich daraus eine Datenrate von

$$ld(100000)\frac{Bit}{Wort}\cdot\frac{100}{60}\frac{Wort}{s}\approx 28\frac{Bit}{s}$$

Diese Datenrate ist um drei Größenordnungen kleiner als die im vorhergehenden Abschnitt geschätzte Datenrate. Natürlich lässt diese Betrachtungsweise einige Informationen wie Intonation, Stimmlage und Lautstärke usw. außer Beachtung. Trotz allem legt sie jedoch nahe, dass wir bei direkter Abtastung und Digitalisierung von natürlicher Sprache „zu viele Daten produzieren". Dieses „zu viel" an Daten wird üblicherweise als Redundanz bezeichnet.

Den Sprung von der hohen Datenrate des direkt abgetasteten Signals hin zu einer Datenrate, die nur den eigentlichen Informationsgehalt der Sprache umfasst, muss die Quellencodierung, die in diesem Fall durch einen Vocoder ausgeführt wird, leisten. In der Praxis wird dieser Sprung natürlich nicht so extrem ausfallen, wie hier dargestellt ist. Je nach (Entwicklungs-)Aufwand und Rechenleistung, die in die Quellencodierung investiert wird, fallen die Ergebnisse unterschiedlich aus.

Spracherzeugung

Wird der menschliche Vokaltrakt stark vereinfacht betrachtet, so erkennt man zwei aufeinanderfolgende Systeme.

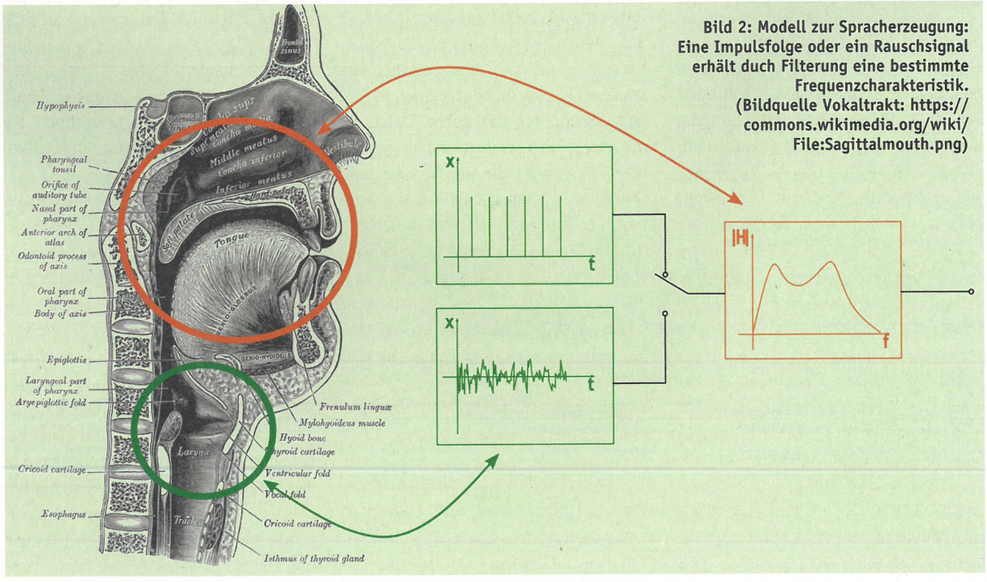

Bild 2: Modell zur Spracherzeugung: Eine Impulsfolge oder ein Rauschsignal erhält duch Filterung eine bestimmte Frequenzcharakteristik. (Bildquelle Vokaltrakt: https://commons.wikimedia.org/wiki/File:Sagittalmouth.png)

AWGN
Abk. für Additive White Gaussian Noise. Ein AWGN-Kanal fügt dem Signal auf der Strecke zwischen Sender und Empfänger ein gaußverteiltes weißes Rauschen hinzu. Dieses Kanalmodell wird in der Nachrichtentechnik häufig angewendet, um eine beliebige Übertragunsstrecke zu modellieren.

Zur Person

Thomas Emig, DL7TOM
Amateurfunkzulassung 2010; Studium der Elektrotechnik (Spezialisierung Hochfrequenztechnik und digitale Signalverarbeitung); seit 2016 Softwareentwickler zunächst für Funk-Lokalisierungssysteme, heute für Messempfänger.

Anschrift:
Friedenspromenade 98
81827 München
thomas.emig@posteo.de

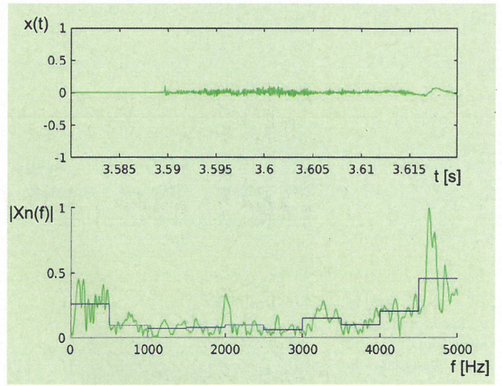

Bild 3: Ein stimmloses Signal im Zeitbereich (oben) und Frequenzbereich (unten). Im Frequenzbereich ist zusätzlich die mit Hilfe der Bänder angenäherte Filtercharakteristik in blau eingezeichnet

Zum einen die Stimmlippen hinter dem Kehlkopf (in **Bild 2** grün markiert) und daran anschließend der Rachen-, Mund- und Nasenraum (in Bild 2 orange markiert).

Die Stimmlippen können in zwei Modi „betrieben" werden. Wenn sie geschlossen sind und Luft durch den entstehenden Spalt von der Lunge kommend in Richtung zum Mund hindurchströmt, beginnen die Stimmlippen zu flattern. Durch das rhythmische Öffnen und Schließen entsteht eine Impulsfolge, die wir uns im Frequenzbereich wie eine Grundschwingung mit sehr vielen Oberwellen vorstellen können. Die Spannung der Stimmlippen beeinflusst dabei die Grundfrequenz und damit die Tonhöhe des erzeugten Lautes.

Im zweiten „Betriebsmodus" sind die Stimmlippen geöffnet. Aus der Lunge vorbeiströmende Luft wird jedoch an Engstellen verwirbelt, wodurch ein Rauschsignal entsteht.

Die im ersten Zustand erzeugten Laute werden als stimmhaft bezeichnet. Sie treten besonders deutlich in der Artikulation der Vokale hervor.

Im zweiten Fall werden die stimmlosen Laute erzeugt. Sie dienen der Artikulation der Laute s, ch usw.

Der anschließende Rachen-, Mund- und Nasenraum kann für den Vocoder als Filter mit einem bestimmten Frequenzgang modelliert werden. Durch dieses Filter erhält das neutrale Rausch- oder Impulssignal eine bestimmte Frequenzcharakteristik. Je nach Spannung der beteiligten Muskeln wird der entstehende Hohlraum und damit der Frequenzgang des Filters verändert. Diese Anschauung deckt einige Laute nicht ab, die wir als Menschen mit unserem Vokaltrakt

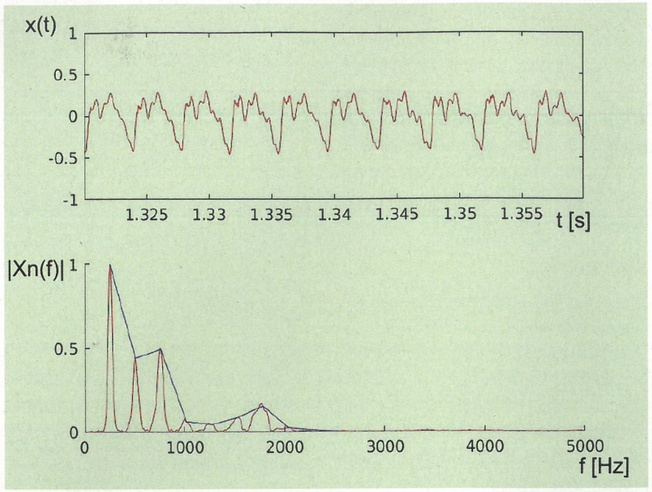

Bild 4: Ein stimmhaftes Signal im Zeit- und Frequenzbereich. In blau sind die Frequenz/Amplitudenpaare der Grundschwingung und Oberwellen eingezeichnet

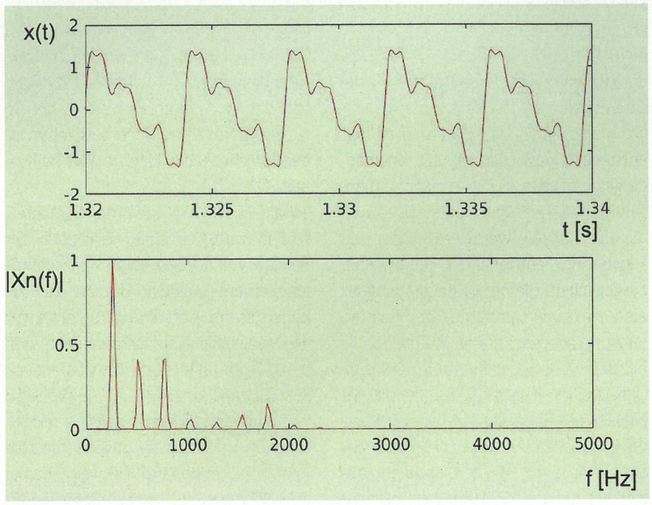

Bild 5: Synthese des stimmhaften Signals aus Bild 4. Es fällt deutlich der Unterschied zwischen dem originalen Sprachsignal und dem synthetisierten Signal auf

erzeugen können. Für eine ausführlichere Modellierung möchte ich auf [1] verweisen.

Für den hier vorgestellten Vocoder wird die Annahme getroffen, dass der Frequenzgang sowie der Zustand der Stimmlippen für kurze Zeitabschnitte von je 20 ms konstant bleibt.

Entfernung von Redundanz

Das Ziel des Vocoders ist es, den Zustand des Vokaltraktes für jeden Zeitabschnitt zu bestimmen und die so ermittelten Parameter anstatt des eigentlichen Audiosignals zu übertragen. Anschließend wird der im Empfänger arbeitende Teil des Vocoders aus den Parametern nach dem eben beschriebenen Modell wieder ein Audiosignal erzeugen.

Um den Zustand zu schätzen wird das Audiosignal zunächst in Blöcke von 20 ms Länge aufgeteilt, wobei die Blöcke um jeweils 10 ms überlappen. Die Überlappung wird eingeführt, um einen weicheren Übergang zwischen den einzelnen Blöcken zu erreichen.

Beispiel: Block 1 besteht aus dem Eingangssignal im Bereich [0 s, 0,02 s[, Block 2 aus [0,01 s, 0,03 s[, Block3 aus [0,02 s, 0,04 s[usw. Damit ist beispielsweise der Bereich [0,01 s, 0,02 s[sowohl in Block 1 als auch in Block 2 enthalten. Diese Überlappung dient dazu, später einen weicheren Übergang zwischen den einzelnen Blöcken herstellen zu können.

Für jeden Block wird nun entschieden, ob es sich um einen stimmhaften oder stimmlosen Laut oder um Stille handelt. Dazu wird der Signalpegel sowie das Signal im Frequenzbereich ausgewertet. Überwiegen Signalanteile in Frequenzen größer als 5 kHz, so wird von einem stimmlosen Laut ausgegangen, andernfalls von einem stimmhaften Laut. Für die stimmlosen Laute muss nun noch die Übertragungscharakteristik des Filters (aus Mund- und Rachenraum) ermittelt werden. Dazu wird das in den Frequenzbereich transformierte Signal in 10 gleiche Bänder zwischen 0 Hz und 5 kHz aufgeteilt und in jedem Band durch Mittelwertbildung die Signalleistung bis auf einen konstanten Faktor ermittelt. So entsteht eine stufige Frequenzcharakteristik, die sich an den tatsächlichen Frequenzgang des Filters annähert.

Für stimmhafte Laute wird nicht der komplette Frequenzgang des Filters geschätzt. Da das Signal aus der Grundschwingung und den dazugehörigen Oberwellen besteht, reicht es aus, zunächst die Frequenz der Grundschwingung und anschließend die Amplituden der Grundschwingung und ihrer Oberwellen zu bestimmen.

Die Grundschwingung kann bestimmt werden, indem im Frequenzbereich das Maximum zwischen 50 Hz und 250 Hz gesucht wird. In der Praxis hat sich gezeigt, dass diese Art der Analyse zwar fehleranfällig ist, da die Grundfrequenz gelegentlich eine geringere Amplitude aufweist als die erste Oberwelle. Dennoch sind die Ergebnisse für die Sprachcodierung gut brauchbar.

Im nächsten Schritt werden die Signalpegel der Oberwellen bestimmt, indem aus dem Signal im Frequenzbereich an den zuvor ermittelten Frequenzen die Pegelwerte abgelesen werden. Diese Methode ist durchaus ungenau und liefert nicht die tatsächlichen Pegelwerte des Signals, sondern immer einen Wert für die Signalleistung in der Umgebung des Signals. Zudem ist der Wert im Frequenzbereich sehr stark abhängig von den Parametern der verwendeten Fouriertransformation (beispielsweise, wie genau das Signal einen Messpunkt der verwendeten FFT trifft). Aber auch hier genügen die erzielten Ergebnisse den Anforderungen zur Sprachcodierung.

Das erste Ergebnis der Analyse ist demnach eine Entscheidung, ob es sich um einen stimmhaften oder einen stimmlosen Laut oder Stille handelt. Diese Entscheidung könnte mit 2 Bit codiert zum Empfänger übertragen werden.

Handelt es sich um einen stimmhaften Laut, müssen zusätzlich die Frequenz der Grundschwingung, zehn Amplitudenwerte der Grund- und Oberschwingungen sowie der Signalpegel des Gesamtsignals zum Empfänger übermittelt werden. Werden diese Werte mit jeweils 8 Bit quantisiert, so müssen weitere 12 · 8 Bit = 96 Bit übertragen werden.

Für einen stimmlosen Laut werden nur die Signalleistungen der zehn Bänder und der Signalpegel des Gesamtsignals an den Empfänger übermittelt. Bei 8 Bit Quantisierung müssen demnach weitere 11 · 8 Bit = 88 Bit übertragen werden.

Um die Datenrate des Systems zu schätzen, wird angenommen, dass jeder Block aus dem Audiosignal einen stimmhaften Laut enthält, da für stimmhafte Laute die meisten Daten übertragen werden müssen. Die Blöcke haben eine Länge von jeweils 20 ms, wobei sie um 50 % überlappen, sodass effektiv alle 10 ms ein neuer Block analysiert wird. Die Datenrate entspricht damit im schlimmsten Fall 98 Bit/10 ms = 9,8 kBit/s und ist damit deutlich geringer als die Datenrate eines direkt abgetasteten Signals (44,8 kBit/s) aber trotzdem noch deutlich größer als die Datenrate der reinen Sprachinformation (28 Bit/s).

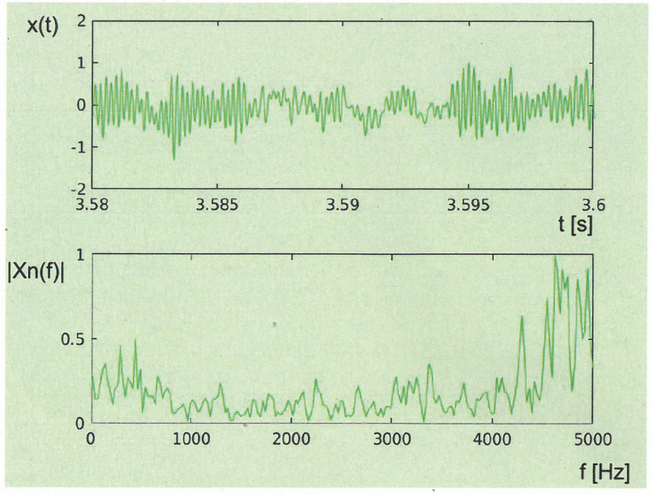

Bild 6: Synthese des stimmlosen Signals aus Bild 3

Sprachsynthese

Im Empfänger wird nun aus den übertragenen Parametern wieder ein Audiosignal generiert.

Für stimmhafte Laute wird das Signal aus zehn Sinusschwingungen mit den übertragenen Frequenz- und Amplitudenwerten zusammengesetzt.

Für stimmlose Laute ist die Synthese etwas komplexer. Hier wird zunächst weißes Rauschen erzeugt. Dieses Signal wird anschließend in den Frequenzbereich transformiert. Aus den übertragenen Signalleistungen der Frequenzbänder wird eine stufige Filtercharakteristik zusammengesetzt (wie in **Bild 3** gezeigt) und mit dem Rauschsignal im Frequenzbereich multipliziert. Zuletzt wird das Signal wieder in den Zeitbereich zurücktransformiert.

Die so entstandenen Blöcke von je 20 ms Länge werden nun mit entsprechender Fensterung versehen, um harte Signalsprünge am Beginn und Ende der jeweiligen Blöcke zu vermeiden.

Bei der Fensterung wird das Signal des Blocks mit einer bestimmten Fensterfunktion multipliziert. In der Mitte des Blocks hat die Fensterfunktion den Wert 1, sodass hier das Signal nicht verändert wird. An den Enden des Blocks fällt die Fensterfunktion auf 0 ab, sodass auch das mit der Fensterfunktion multiplizierte Signal zu den Enden des Blocks hin auf 0 abfällt. Die Fensterung verhindert Sprünge im Signal (und damit deutlich hörbare Störungen), wenn die Blöcke aneinandergefügt werden, da das Signal an den Grenzen der Blöcke garantiert 0 ist.

Nach der Blockbildung werden sie mit entsprechender Überlappung aneinandergefügt, sodass ein zusammenhängendes Ausgangssignal entsteht. Durch die Überlappung werden die durch die Fensterung entstehenden Nullstellen ausgeglichen, die ohne Überlappung als eine langsame Amplitudenmodulation im Ausgangssignal hörbar wären.

Bei Betrachtung der Ein- und Ausgangssignale fällt auf, dass das Ausgangssignal keineswegs identisch zum Eingangssignal ist. Der Vocoder überträgt also nicht das Signal „wie es ist", sondern reduziert es auf den eigentlichen Informationsgehalt, überträgt diesen und baut im Empfänger wieder ein Signal zusammen, das für den Zuhörer den gleichen Sprach-Informationsgehalt hat wie das ursprüngliche Signal.

Ausblick

Das in diesem Artikel vorgestellte Grundprinzip der Arbeitsweise eines Vocoders bietet noch viele Möglichkeiten der Verbesserung. Angefangen von den bereits beschriebenen Ungenauigkeiten in der Signalanalyse bis zur weiteren Verfeinerungen des zugrunde liegenden Sprachmodells.

Aufgrund der vielfältigen Anwendungen, die Sprachcodierung erforderlich machen, sind bereits einige sehr gute Vocoder sowohl im kommerziellen Bereich als auch in Form von quelloffener Software entwickelt worden. Diese Vocoder können Sprache verständlich bei extrem niedrigen Bitraten von beispielsweise 600 Bit/s übertragen und kommen damit dem reinen Informationsgehalt der Sprache bereits sehr nahe. Im Amateurfunk erfreut sich beispielsweise der Sprachcodec Codec2 großer Beliebtheit [2].

Sobald große Mengen an Sprachbeispielen verfügbar sind, wird auch der Einsatz von Methoden des maschinellen Lernens möglich. Um die zur Rekonstruktion notwendigen Parameter des Sprachsignals zu identifizieren, werden Algorithmen mit vielen Sprachaufzeichnungen trainiert und hinsichtlich verschiedener Eigenschaften des rekonstruierten Sprachsignals bewertet. So können Codierungen beispielsweise darauf optimiert werden, bei möglichst niedriger Bitrate besonders natürlich zu klingen [3].

Zu dem hier vorgestellten Konzept des Vocoders habe ich im Rahmen der HAM RADIO World einen Vortrag gehalten, der einige Hörbeispiele und praktische Demonstrationen enthält. Ein Mitschnitt des Vortrags kann unter [4] abgerufen werden. CQDL

Literatur und Bezugsquellen

- [1] Detaillierte Modellierung: https://www.phonetik.uni-muenchen.de/studium/skripten/AP/APKap2.html
- [2] Informationen zu Codec2: https://www.rowetel.com/?page_id=452
- [3] Training eines Low-Bit-Codecs: https://ai.googleblog.com/2021/02/lyra-new-very-low-bitrate-codec-for.html
- [4] Mitschnitt des Vortrags: https://thomas-emig.de/data/vocoder.mp4