NEUE ANSÄTZE ZUR SPRACHSYNTHESE MIT KODIERTEN SPRACHSEGMENTEN

Guntram Strecha

TU Dresden, Institut für Akustik und Sprachkommunikation
guntram.strecha@ias.et.tu-dresden.de

Abstract: Der Einsatz in kommerziellen Produkten (Embedded Systems) stellt an die Sprachverarbeitung spezielle Anforderungen. Neben einer hohen Akzeptanz beim Nutzer spielt beim Hersteller der Ressourcenverbrauch der Synthese eine wichtige Rolle. Im Vordergrund stehen dabei der Speicherbedarf und der Rechenaufwand. Im Bereich des Mobilfunks zeigt sich außerdem ein Trend zu höheren Bandbreiten als 8 kHz.
In dem Beitrag wird ein Synthesesystem vorgestellt, welches kodierte Sprachsegmente synthetisiert. Alle Segmente des (Diphon-)Inventars sind mit einem im Mobilfunkbereich häufig eingesetzten standardisierten Sprachkodierer komprimiert. Während der Synthese werden die geforderten Sprachsegmente dekodiert. Der Ansatz, der bei dem vorgestellten Synthesesystem verfolgt wird, ist die Integration der Prosodiemanipulation (f_0-, Dauersteuerung) in den Dekodierer. Dieser integrierte Ansatz basiert auf Gemeinsamkeiten des CELP-basierten Kodierers/Dekodierers mit Sprachsynthesetechniken.
Unter Verwendung verschiedener Kodierstufen werden Kompressionsraten von bis zu 18:1 (8 kHz) bzw. 26:1 (16 kHz) erreicht. Das entspricht Inventargrößen von 119 kByte bzw. 164 kByte.

1 Einleitung

Die Verbesserung der Natürlichkeit als qualitatives Merkmal von Text-To-Speech (TTS) Systemen ist einer der vordringlichsten Forschungsschwerpunkte. Zugleich ist der Ressourcenbedarf, der Rechenzeitaufwand der Algorithmen sowie der Speicherbedarf der Datenbasen und des Programmcodes, zu berücksichtigen, um Anwendungen in verschieden mobilen Multimediaprodukten zu ermöglichen.
Text-To-Speech Systeme welche zuvor aufgezeichnete Sprachsignalabschnitte verketten (Konkatenative Sprachsynthese) erreichen derzeit eine höhere Sprachqualität als parametrische Sprachsynthesesysteme. Dabei steigt die Qualität je länger die Sprachbausteine gewählt werden und je mehr Bausteinvarianten existieren. Längere Bausteine führen zu weniger Verkettungsstellen an den Signalunstetigkeiten entstehen können. Viele (prosodische) Varianten eines Bausteins ermöglichen die Auswahl eines geeigneteren Bausteins mit dem Ziel die prosodischen Signalmanipulationen (Dauer und Grundfrequenz) so gering wie möglich zu halten.
Solche korpusbasierten TTS-Systeme tendieren zu großen Ressourcenbedarf, hervorgerufen durch den Speicherverbrauch der Sprachsignalinventare und dem Rechenaufwand der Algorithmen [1]. Für die Anwendung in mobilen Multimediageräten, wie z. B. Mobiltelefone, PDAs (Personal Digital Assistants) oder Spielzeuge, sind diese Systeme nicht praktikable. Trotz des technischen Fortschritts stellen die Kosten des Speichers und der Rechenleistung der Geräte die Forderung nach sehr kleinen Sprachsynthesesystemen (embedded systems) [5].

Anknüpfend an vorangegange Arbeiten [2] wird in diesem Beitrag die akustische Synthese mit unterschiedlich kodierten Inventaren vorgestellt.

2 Der integrierte Ansatz

Die Aufgabe der akustischen Synthese ist einerseits die Auswahl der geeigneten Sprachsegmente (Bausteinauswahl) anhand der zu realisierenden Phonemfolge und andererseits die prosodische Manipulation, mit dem Ziel, die vorgegebenen Phonemdauern und die Grundfrequenzkontur auf die Sprachsegmente aufzuprägen. Bei der Sprachsynthese mit kodierten Inventaren lassen sich zwei prinzipielle Methoden unterscheiden. Der bereits in [3] vorgestellte zweistufige Ansatz dekodiert in einem ersten Schritt die ausgewählten Bausteine und prägt im zweiten Schritt die prosodischen Parameter auf (s. Abb. 1).

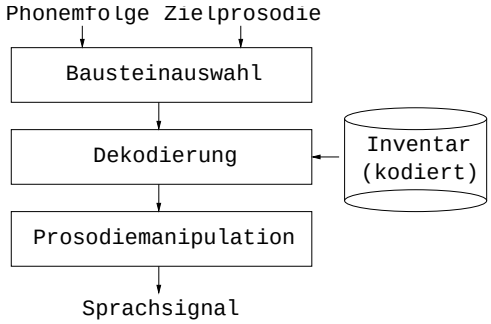

Abbildung 1 - Schema des zweistufigen Ansatzes.

Beim einstufigen oder integrierten Ansatz werden spezielle Sprachcodecs verwendet. Im vorgestellten System sind das der Adaptive Multi-Rate Narrowband (AMR-NB) und AMR Wideband (AMR-WB) Koder [4]. Diese Koder sind standardisiert und werden häufig im Mobilfunkbereich eingesetzt. Das zugrundeliegende Prinzip dieser Kodierer ist ACELP (Algebraic Code Excited Linear Prediction). Wie in Abbildung 2 dargestellt, nutzt der integrierte Ansatz das Sprachsyntheseprinzip der Koder, um während der Dekodierung die prosodischen Manipulationen der Sprachsegmente vorzunehmen.

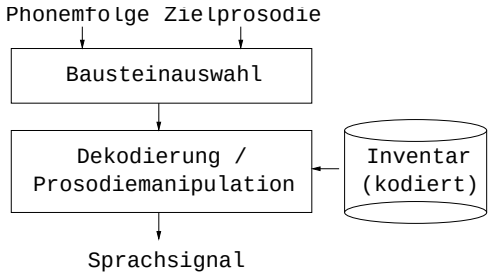

Abbildung 2 - Schema des einstufigen Ansatzes.

Im integrierten Ansatz ist die strikte Trennung von Kode und Daten fortgeführt, d. h. die von der akustischen Synthese verwendeten Datenbasen sind ohne Anpassung des Kodes austauschbar. Damit ist das System sprecherunabhängig und die Multilingualität gewährleistet. Der Kode ist vollständig in Festkommaarithmetik implementiert.

3 Die Akustische Synthese mit AMR-NB kodierten Inventaren

3.1 Der AMR-NB Koder

Der AMR-NB Koder verarbeitet Signale mit einer Bandbreite von 8 $^{kHz}/_{s}$ und besitzt die Möglichkeit mit verschiedenen Kompressionsstufen bzw. Datenraten zu arbeiten. Einstellbar sind acht Stufen (4.75 $^{kBit}/_{s}$, 5.15 $^{kBit}/_{s}$, 5.9 $^{kBit}/_{s}$, 6.7 $^{kBit}/_{s}$, 7.4 $^{kBit}/_{s}$, 7.95 $^{kBit}/_{s}$, 10.2 $^{kBit}/_{s}$, 12.2 $^{kBit}/_{s}$). Je nach Kompressionsstufe werden verschiedene Parameter vom Kodierer berechnet und an den Dekodierer paketweise übergeben. Ein Paket enthält die Information zur Dekodierung von

einem Signalabschnitt mit vier Blöcken (Subframes) fester Länge ($= 40\,\mathrm{Samples} = 20\,\mathrm{ms}$). Diese Parameter sind beispielsweise für den 12.2 kBits Modus:

- die Indizes der kodebuch-quantisierten Line Spectral Pairs (LSP), berechnet aus den Linear Predictive Coding (LPC) Koeffizienten,
- die Pitch-Verzögerung (pitch delay) p_0,
- die Indizes des adaptiven (p_i) und festen Kodebuchs (c_i) und
- die Verstärkungsfaktoren (g_p, g_c) des adaptiven und festen Kodebuchs.

Mittels dieser Parameter werden das Residualsignal (die Anregung) $e(n)$ und die Filterkoeffizienten a_k des LPC-Synthesefilters:

$$S'(z) = \frac{E(z)}{A(z)} \quad \multimap \quad \begin{aligned} s'(n) &= e(n) - \sum_{k=1}^{10} a_k s(n-k) \\ &= g_p v(n) + g_c c(n) - \sum_{k=1}^{10} a_k s(n-k) \end{aligned} \qquad n = 0, \ldots, 39 \qquad (1)$$

extrahiert und die 40 Samples des aktuellen Signalblocks $s'(n)$ synthetisiert (s. Abb. 3). Die Pitch-Verzögerung p_0 dient zur Berechnung des aktuellen adaptiven Kodebuchvektors $v(n)$ aus dem, um die Verzögerung p_0 verschobenen Kodebuchvektors. Eine abschließende Nachbearbeitung:

$$S(z) = H_t(z)\, H_f(z)\, S'(z) \qquad \text{mit} \qquad \begin{aligned} H_f(z) &= \frac{A(z/\gamma_n)}{A(z/\gamma_d)} \\ H_t(z) &= 1 - \mu z^{-1} \end{aligned} \qquad (2)$$

führt eine perzeptuale Gewichtung durch, um das Quantisierungsrauschen zu maskieren und den Signal-Rausch-Abstand der Formanten zu erhöhen.

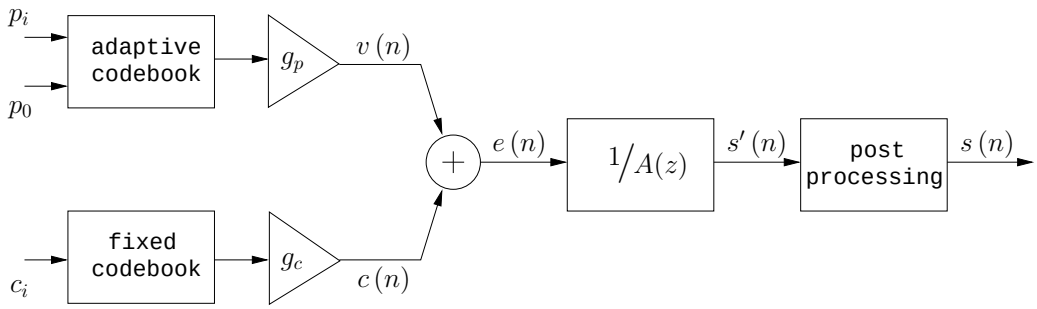

Abbildung 3 - Vereinfachtes Schema des AMR-NB Dekoders. p_0 - Pitch Verzögerung, p_i, c_i, g_p, g_c - Kodierparameter, $e(n)$ - Residualsignal, $A(z)$ - LPC Filterkoeffizienten, $s'(n), s(n)$ - Synthesesignale.

3.2 Die integrierte akustische Synthese

Wie in Abbildung 3 dargestellt, wird das Sprachsignal durch Filtern des Residualsignals $e(n)$ mit den LPC-Filterkoeffizienten $A(z)$ synthetisiert (s. Gl. 1). Trotz der unterschiedlichen Berechnung von $e(z)$ und $A(z)$ ist dieser Schritt für alle Modi äquivalent. Die prosodischen Modifikationen erfolgen direkt vor dem Filtern (s. Abb. 4), durch Manipulieren des Residualsignals.

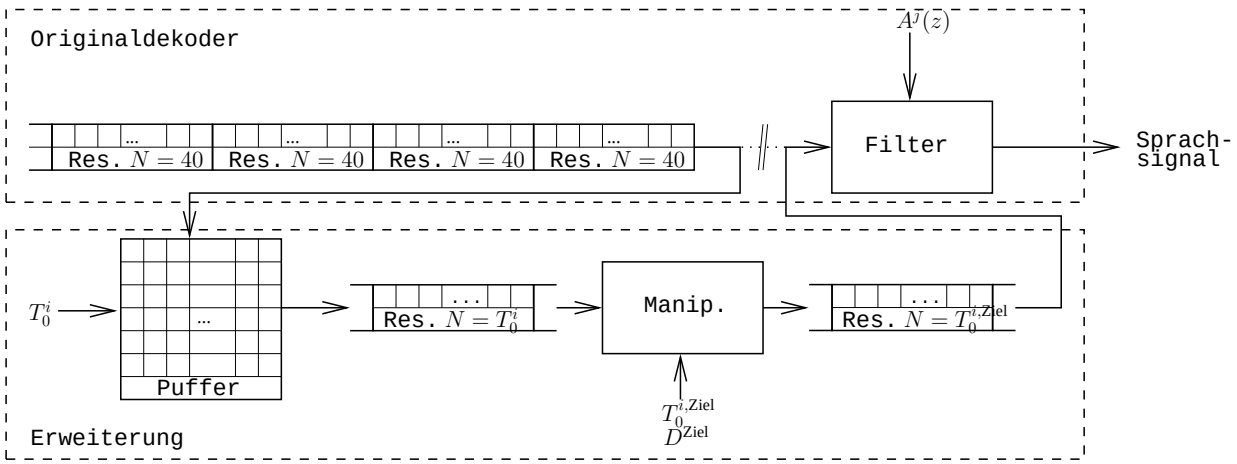

Abbildung 4 - Vereinfachte Verarbeitungskette der integrierten akustischen Synthese als Erweiterung des Originaldekoders. $A^j(z)$ - LPC Filterkoeffizienten, T_0^i - Periodenl¨ange, $T_0^{i,\text{Ziel}}$ - Zielperiodenl¨ange, D^{Ziel} - Zielphonemdauern.

Die Manipulationen erfolgen synchron zu den Periodenmarken T_0^i. Das ursprüngliche Residualsignal mit konstanter Blocklänge ($N = 40$) muß zwischengepuffert werden, um die variablen Abschnitte, korrespondierend zu den Periodenlänge T_0^i, zu erhalten.

Zur Aufprägung der Zielgrundfrequenz $T_0^{i,\text{Ziel}}$ wird die Periode i in ihrer Länge geändert. Eine Methode zur Streckung bzw. Stauchung der Periode sind Overlap-And-Add (OLA) Verfahren (z. B. TD-PSOLA, FD-PSOLA). Einfaches Auffüllen mit Nullen bzw. Beschneiden der Periode ist ebenfalls möglich. Die Steuerung der Phonemdauern erfolgt durch Verdopplung bzw. Löschung einzelner Perioden des Anregungssignals. Eine effiziente Methode dafür ist in [2] beschrieben.

Die Periodenlängen P_0^i des Inventars müssen zum Synthesezeitpunkt vorliegen. Methoden des Zugriffs auf P_0^i sind:

1. Berechnung der Periodenlängen aus der Pitch-Verzögerung p_0^j. Diese Verzögerung wird vom Kodierer offline berechnet und steht bei der Dekodierung für jeden Subframe j zur Verfügung. p_0^j korrespondiert zur Sprechergrundfrequenz, ist aber optimiert für die Rekonstruktion des Anregungssignals. Die auftretenden Abweichungen zur tatsächlichen Grundfrequenz resultieren in Fehlern bei der prosodischen Manipulation und damit in einer Verschlechterung des Synthesesignals.
2. Zufügen der Periodenmarken des unkodierten zum kodierten Inventar. Damit stehen zum Synthesezeitpunkt die exakten Periodenlängen zur Verfügung. Der Nachteil ist der Mehrbedarf an Speicher für das kodierte Inventar.
3. Vereinheitlichung aller Periodenlängen vor der Inventarkodierung. Die vorherige Monotonisierung des Inventars kann durch periodenweises Resamplen der Originalperioden erreicht werden. Die entstehenden Artefakte sind geringer je monotoner das Originalinventar bereits war. Ein besonderer Vorteil entsteht, falls die, nun konstante, Sprechergrundfrequenz mit der Subframelänge des Koders korrespondiert ($f_0 = 1/T_0 = 20\,\text{ms} = 40\,\text{Samples}$). In diesem Fall ist keine Pufferung des Residualsignals nötig.

4 Die Akustische Synthese mit AMR-WB kodierten Inventaren

4.1 Der AMR-WB Koder

Der AMR-WB Koder verarbeitet Signale mit einer Bandbreite von 16 kHz und funktioniert ähnlich dem AMR-NB Koder. Alle Verarbeitungsschritte des Dekoders vor der Nachbearbeitung (post processing, s. Abb. 3) sind auf eine Bandbreite von 12.8 kHz ausgelegt, d. h. das Anregungssignal $e(n)$ ist aufgeteilt in Abschnitte konstanter Länge von 256 Samples mit jeweils vier Unterabschnitten (64 Samples). Die LPC-Synthesefilterlänge und damit die Anzahl der LPC-Koeffizienten beträgt 16. Erst im Nachbearbeitungsprozeß wird das Signal auf 16 kHz umgetastet. Die fehlenden Frequenzanteile in den oberen Frequenzbändern werden, in Abhängigkeit von den unteren Frequenzbändern, geschätzt und das Signal mit dem entstehenden höherfrequenten Signal angereichert.

4.2 Die integrierte akustische Synthese

Die Integration der akustischen Synthese erfolgt in gleicher Weise wie beim integrierten Ansatz des AMR-NB (s. Abb. 5). Im Unterschied dazu ist beim AMR-WB der Nachverarbeitungs-

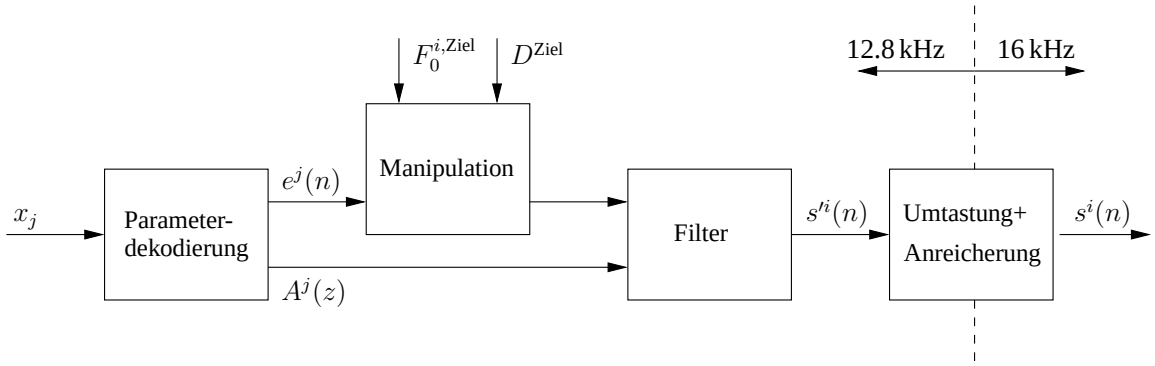

Abbildung 5 - Vereinfachtes Schema des AMR-WB Dekoders mit der integrierten akustischen Synthese. $F_0^{i,\text{Ziel}}$ - Zielgrundfrequenz des Sprechers, D^{Ziel} - Zielphonemdauer, $e^j(n)$ - Residualsignal, x_j - Kodierparameter, $A^j(z)$ - LPC-Koeffizienten, $s'^i(n)$ - 12.8 kHz Synthesesignal, $s^i(n)$ - 16 kHz Synthesesignal

schritt stärker abhängig vom Signal $s'(n)$ am Ausgang des LPC Synthesefilters. Manipulationen am Residualsignal und somit am 12.8 kHz-Synthesesignal resultieren in größeren Signalstörungen am 16 kHz-Synthesesignal. Die zahlreichen Filter des Nachverarbeitungsschrittes mit ihren Speicherzuständen korrespondieren weniger zu den vorhergehenden Verarbeitungsschritten des AMR-WB je stärker die prosodischen Manipulationen sind.

5 Die Inventargenerierung

In Hinblick auf die Begrenzung des Speichers für “embedded systems” ist klar ersichtlich, daß nur kleine Inventare die Chance haben durch eine Komprimierung die Restriktionen der Ressourcen zu erfüllen. Aus diesem Grund sind die Ausgangsinventare Diphoninventare. In Tabelle 1 sind für ein Inventar (deutsche, weibliche Stimme, 1176 Diphonbausteine) die erreichten Größen nach der Kodierung mit verschiedenen Kompressionsraten zusammengestellt.

	Kodierung ($\frac{kBit}{s}$)	monotonisiert (Byte)	mit Periodenmarken (Byte)	ohne Periodenmarken (Byte)
simone-nurdt 16 kHz	unkomprimiert	4491040	**4644333**	-
	AMR-WB 23.85 $\frac{kBit}{s}$	**564979**	592169	571351
	AMR-WB 23.05 $\frac{kBit}{s}$	546543	573519	552701
	AMR-WB 19.85 $\frac{kBit}{s}$	472799	498919	478101
	AMR-WB 18.25 $\frac{kBit}{s}$	435927	461619	440801
	AMR-WB 15.85 $\frac{kBit}{s}$	380619	405669	384851
	AMR-WB 14.25 $\frac{kBit}{s}$	343747	368369	347551
	AMR-WB 12.65 $\frac{kBit}{s}$	306875	331069	310251
	AMR-WB 8.85 $\frac{kBit}{s}$	219427	242599	221781
	AMR-WB 6.60 $\frac{kBit}{s}$	167251	189827	169009
simone-nurdt 8 kHz	unkomprimiert	2268303	**2336579**	-
	AMR-NB 12.20 $\frac{kBit}{s}$	**287255**	315413	297611
	AMR-NB 10.20 $\frac{kBit}{s}$	242657	269163	251361
	AMR-NB 7.95 $\frac{kBit}{s}$	192671	217289	199487
	AMR-NB 7.40 $\frac{kBit}{s}$	180215	204281	186479
	AMR-NB 6.70 $\frac{kBit}{s}$	164651	188173	170371
	AMR-NB 5.90 $\frac{kBit}{s}$	146811	169651	151849
	AMR-NB 5.15 $\frac{kBit}{s}$	130377	152557	134755
	AMR-NB 4.75 $\frac{kBit}{s}$	121311	143201	125399

Tabelle 1 - Übersicht der Inventargr¨oßen bei den m¨oglichen Kodierstufen des AMR. Die fett gedruckten Zahlen bezeichnen die Inventare, welche beim H¨ortest eingesetzt wurden.

Zur Erstellung des kodierten Inventars werden alle Bausteine einzeln kodiert. Aufgrund der Einschwingzeit des Dekodierers wird die erste Periode jedes Bausteins entsprechend der Länge der Einschwingzeit wiederholt. Während der Dekodierung werden die Signalwerte des Einschwingvorganges ignoriert.

6 Evaluation

Zur Evaluierung der integrierten akustischen Synthese wurde diese aus dem vollständigen TTS-System herausgelöst. Für einen Mean Opinion Score (MOS) Hörtest wurden drei natürlich gesprochene Sätze derselben Sprecherin ausgewählt, welche ihre Stimme für die Erstellung des Inventars bereitstellte. Um ausschließlich den Einfluß der akustischen Synthese zu evaluieren wurden die Phonemfolge und die prosodischen Parameter (Grundfrequenzkontur und Phonemdauern) aus diesen drei Sätzen extrahiert und zur Ansteuerung der Synthese verwendet. Neben der Originalsprache (unit[16]kHz und 8 kHz) wurden den Hörern folgende Synthesevarianten präsentiert:

1. 16 kHz Synthese mit unkodiertem Inventar,
2. 8 kHz Synthese mit unkodiertem Inventar,
3. Synthese mit AMR-WB (23.85 $\frac{kBit}{s}$) kodiertem Inventar (16 kHz),
4. Synthese mit AMR-NB (12.2 $\frac{kBit}{s}$) kodiertem Inventar (8 kHz) und
5. zusätzlich jeweils eine 16 und 8 kHz Cepstralsynthese (Inventargrößen: 556505 bzw. 274265 Byte).

In Abbildung 6 sind die Ergebnisse graphisch dargestellt. Die geringe Anzahl (21) an Hörern lassen nur einen ersten Eindruck von den Unterschieden der einzelnen Synthese zu. Auffällig ist der geringe Wert bei der AMR-WB Synthese, was auf die oben genannten Gründe zurückzuführen sein dürfte.

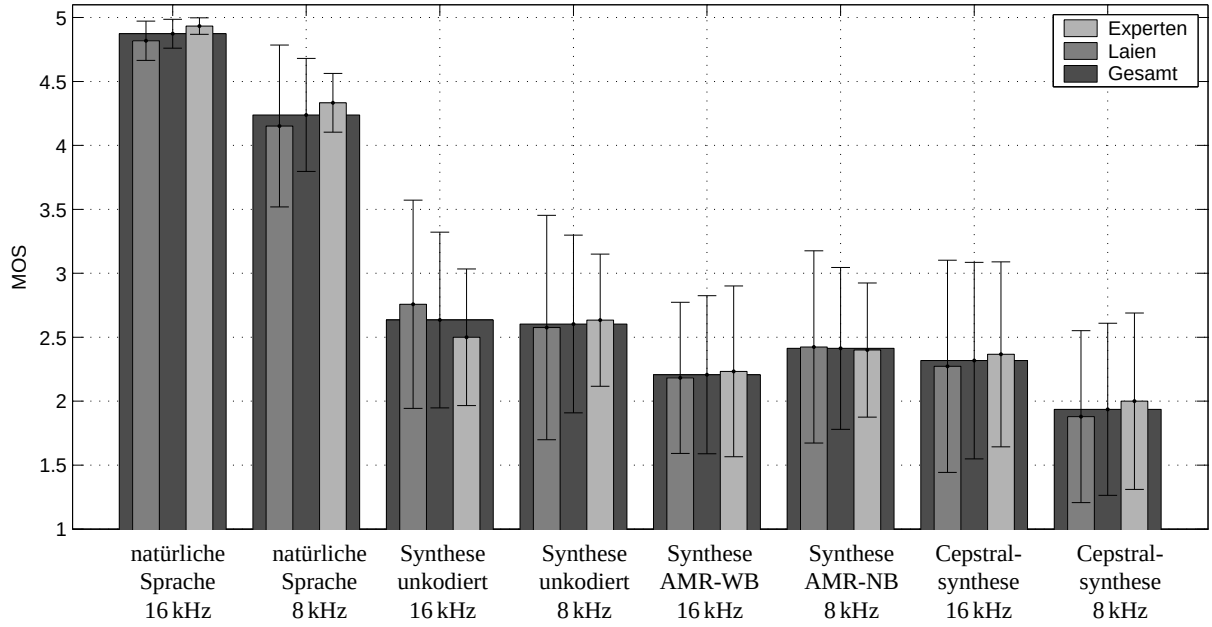

Abbildung 6 - Ergebnisse des MOS H¨ortests mit 3 beurteilten S¨atzen von jeweils 21 H¨orern (11 Laien, 10 Experten). Die senkrechten Linien geben die Varianz der gemessenen Werte an.

Literatur

[1] W. N. Campbell. Prosody and the selection of source units for concatenative synthesis. In *Proc. ESCA-Workshop on Speech Synthesis, Mohonk (NY)*, pages 61–64, 1994.

[2] R. Hoffmann, O. Jokisch, D. Hirschfeld, G. Strecha, H. Kruschke, and U. Kordon. A multilingual TTS system with less than 1 megabyte footprint for embedded applications. In *Proc. ICASSP*, Hong Kong, 2003.

[3] R. Hoffmann, O. Jokisch, H. Kruschke, and G. Strecha. microDRESS - a speech synthesis system with minimized footprint. In *Proc. 12th Czech-German Workshop Speech Processing*, Prague, 2002.

[4] R. L. R. and S. R. W.˙ *Digital processing of speech signals.* Prentice-Hall int., 1978.

[5] M. Schnell, O. Jokisch, R. Hoffmann, and M. Küstner. Text-to-speech for low-resource systems. In *Proc. 5th IEEE Workshop on Multimedia Signal Processing (MMSP)*, St. Thomas (US Virgin Is.), 2002.