PERFORMANZUNTERSUCHUNGEN ZUR STIMMKONVERTIERUNG

Oliver Jokisch, Hamurabi Gamboa Rosales

TU Dresden, Institut für Akustik und Sprachkommunikation
oliver.jokisch@tu-dresden.de

Kurzfassung: Es haben sich unterschiedliche Verfahren zur Stimmkonvertierung etabliert, um die angestrebte Charakteristik eines Zielsprechers zu modellieren und dabei eine möglichst natürliche Sprachsignalqualität zu gewährleisten. Der Beitrag stellt Hörexperimente für vier Stimmkonvertierungsmethoden vor, bei denen die Höranstrengung, die allgemeine Sprachqualität und die Ähnlichkeit zur Zielstimme bewertet werden. Die subjektive Ähnlichkeitsbewertung wird mittels einer objektiven Abstandsmessung auf Basis der logarithmierten Spektralverzerrung überprüft. Der praktische Einsatz des Konvertierungsverfahrens erfordert darüber hinaus eine geeignete Performanz bezüglich des Laufzeitverhaltens sowie der Speichernutzung. Der Beitrag diskutiert das Laufzeitverhalten auf Basis verschiedener Parametrisierungen einer ausgewählten Stimmkonvertierungsmethode im Kontext typischer Einsatzbedingungen. Dabei wird der Einfluss der Rechenressourcen, der Konvertierungsparameter sowie der Trainingseinstellungen getestet. Der ermittelte Echtzeitfaktor der nicht-optimierten Konvertierungsmethode ist für viele kommerzielle Anwendungen ungeeignet.

1 Einführung

In der Literatur werden sehr unterschiedliche Verfahren zur Stimmkonvertierung (engl. Voice Conversion, VC) beschrieben. Die Stimmkonvertierung – oft auch als Sprechertransformation bezeichnet – strebt die zielgerichtete Veränderung der Stimme eines Quellsprechers an. Das Konvertierungsergebnis soll dabei möglichst exakt der Stimme eines bestimmten Zielsprechers entsprechen. Damit grenzt sich die Stimmkonvertierung von Methoden der allgemeinen Stimmumwandlung (engl. Voice Transformation, VT) oder des Voice Morphing (VM) ab.

Das Konzept der Stimmumwandlung stammt ursprünglich aus der Text-to-Speech-Synthese und zielte auf die Generierung zusätzlicher Synthesestimmen durch meist regelbasierte Modifikationen von Parametern vorhandener Sprecherdatenbasen.

Je nach Einsatzbereich und Konvertierungsziel werden Algorithmen zur Sprecheranpassung, z. B. die Vokaltraktlängen-Normalisierung, sowie Methoden zur Anpassung weiterer Stimmqualitäts- oder prosodischer Parameter angewendet bzw. kombiniert. Teilweise werden dialektale oder fremdsprachliche Merkmale manipuliert. Bei den meisten Stimmkonvertierungsverfahren ist ein vorheriges Training auf Basis von Referenzdaten oder mit Stimmbeispielen der Quell- und Zielstimme erforderlich. Die algorithmischen Entwicklungen und Experimente konzentrieren sich in der Regel darauf, die angestrebte Charakteristik der normalisierten Stimme oder eines Referenzsprechers zu modellieren bzw. eine hohe perzeptive Sprachsignalqualität zu erzielen, da diese Faktoren über eine erfolgreiche Anwendung von Stimmkonvertierung entscheiden. Der praktische Einsatz in der Spracherkennung und -synthese, im Medienbereich sowie im Spielesektor erfordert darüber hinaus eine geeignete Performanz bezüglich des Laufzeit-

verhaltens und der Speichernutzung. Aufgrund der rechenintensiven Mapping- und Filteralgorithmen – auch beim Training der Stimmkonvertierung – ist die zeitliche Performanz ebenfalls anwendungskritisch. Eine entsprechende algorithmische Optimierung ist nicht trivial.

2 Verfahren zur Stimmkonvertierung

Einsatzszenarien, verwendete Algorithmen und Bewertungskriterien für Stimmkonvertierung sind vielfältig. In diesem Beitrag werden exemplarisch vier VC-Methoden und ihre Performanz aus sprachqualitativer Sicht erläutert. Anschließend wird das beste Verfahren (VC-Methode 1) einer Laufzeit-Performanzanalyse unterzogen, wobei die Einflussfaktoren Quellsignallänge, etwaige Trainingsphasen, Konvertierungsparameter (warping factors) sowie PC-System Berücksichtigung finden.

2.1 Zielrichtung

Das Konvertierungsergebnis auf Basis des Eingangssignals (Quellstimme) soll möglichst exakt der Stimme eines Zielsprechers entsprechen. Sowohl für die Quell- als auch die Zielstimme existieren Sprachbeispiele (PCM, 16 kHz, 16 Bit). In typischen Medienanwendungen treten potentielle Quellsignale im Minuten- bzw. Stundenbereich auf. Im Gegensatz dazu sind in der Regel für die intendierte Zielstimmencharakteristik nur kurze Beispiele im Sekundenbereich verfügbar.

Die zielgerichtete Manipulation der Stimmcharakteristik soll die Signalgrundqualität möglichst gering beeinflussen (verschlechtern). Abgesehen von der Stimmlagenanpassung (Modifikation der mittleren Sprechergrundfrequenz) bleiben prosodische Modelle [1] oder dialektale Aspekte in der ausgewählten VC-Methode 1 unberücksichtigt.

2.2 Algorithmen

Folgende VC-Methoden werden im Beitrag untersucht und u. a. in der angegebenen Literatur näher erläutert:

- VC mit linearer Transformation (u. a. in [2]),
- Vocal Tract Length Normalization (VTLN) im Zeit- oder Frequenzbereich [3, 4, 5, 6],
- VC mittels Hidden Markov Model (HMM) [7].

Die entsprechenden vier Methoden benötigen eine mehrstufige Signalverarbeitung und teils rechenintensive Filteroperationen. Die Ergebnisse resultieren aus verschiedenen Vorprojekten – teils mit kommerziellen Partnern – und werden deshalb anonymisiert dargestellt.

2.3 Bewertungskriterien

Die Literatur konzentriert sich auf die qualitative Bewertung und Gegenüberstellung verschiedener VC-Methoden (Ähnlichkeit des konvertierten Signals zur Zielstimme, Signalartefakte usw.). Dabei werden sowohl subjektive Kriterien (Hörtests) als auch objektive Kriterien (Fehlermaße) angesetzt und im vorliegenden Beitrag kurz vorgestellt.

Der Beitrag untersucht darüber hinaus das Laufzeitverhalten einer ausgewählten VC-Methode, wobei folgende Einflussfaktoren untersucht werden:

- verfügbare Rechenressourcen,
- VC-Parametrisierung,
- Trainingskonstellation.

3 Sprachqualitäts- und Ähnlichkeitsbewertung

Die Bewertung der qualitativen Performanz erfolgte anhand der Ähnlichkeit von konvertierter und Referenzstimme sowie auf Basis der allgemeinen Sprachqualität und Höranstrengung. Über die Hörexperimente hinaus wurden objektive und reproduzierbare Abstandsmaße untersucht („instrumentelle Bewertung“). Die Ergebnisse werden lediglich exemplarisch vorgestellt, da der Fokus auf der Analyse der Laufzeitperformanz einer VC-Methode liegt.

3.1 Hörexperimente

An dem Hörtest nahmen 14 Probanden mit Vorkenntnissen in Sprachtechnologie teil. Es wurden 12 Beispieläußerungen je VC-Methode sowie die zugehörigen 12 Referenzäußerungen der Zielstimme bewertet.

Die Abbildung 1 stellt den Mean Opinion Score (MOS) der „Höranstrengung“ für vier ausgewählte VC-Methoden auf einer Skala von 1 (mangelhaft) bis 5 (ausgezeichnet) dar. Zum Vergleich wird der durchschnittliche MOS-Wert für das originale Referenzsignal mit 4,67 angegeben.

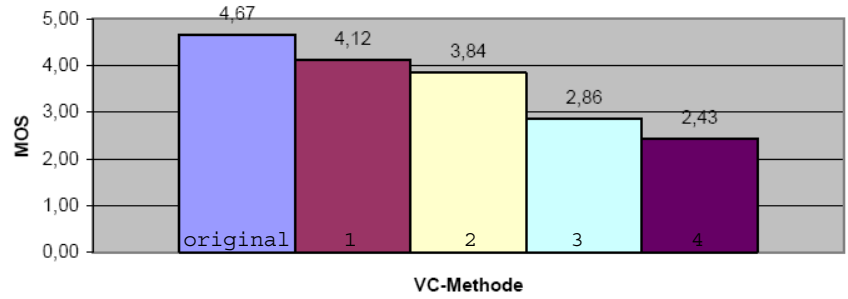

Abbildung 1: MOS-Bewertung der „Höranstrengung“.

Die Abbildung 2 veranschaulicht die “allgemeine Sprachqualität“ auf der MOS-Skala für die gleiche Methoden-Auswahl und das jeweilige Referenzsignal (MOS = 4,55).

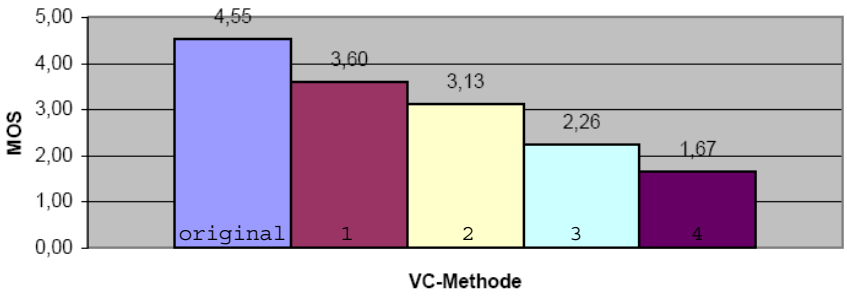

Abbildung 2: MOS-Bewertung der allgemeinen Sprachqualität.

Die Abbildung 3 bezieht sich auf eine zu Abbildung 1 bzw. 2 identische Auswahl der VC-Methoden und stellt die Hörpräferenz der jeweiligen Methode im Paarvergleich („Welches

Hörbeispiel ist der Referenzstimme ähnlicher?“) bezogen auf den erreichbaren Maximalwert (100 % Präferenzurteile über alle Paarvergleiche, die auf diese VC-Methode entfallen) dar.

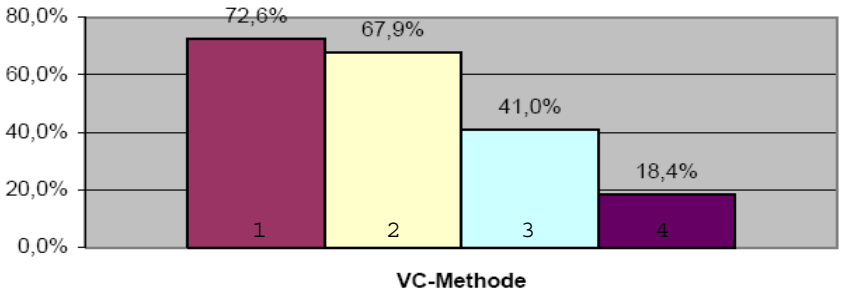

Abbildung 3: Hörpräferenz normiert auf den Maximalwert.

3.2 Instrumentelle Bewertung

Das Ziel der Quellstimmenkonvertierung betrifft die Ähnlichkeit des resultierenden Sprachsignals mit der betreffenden Referenz (Zielstimme). Um die Ähnlichkeit von zwei spektralen Vektorfolgen zu bewerten, eignet sich als Abstandsmaß die log. Spektralverzerrung (Log-Spectral Distortion, LSD) nach [8] (vgl. auch [6]). Dabei werden textlich identische Sprachäußerungen vorausgesetzt, die mittels Dynamic Time Warping (DTW) synchronisiert wurden. Die LSD ist als Distanz zwischen den Cepstralkoeffizienten der konvertierten Quelle $\tilde{x}_1^K$ sowie des Ziels y_1^K definiert. Für die Vergleichbarkeit mit früheren Distanzmaßen wird eine zusätzliche Konstante eingeführt:

$$D_{\mathrm{LSD}} = \frac{10\sqrt{2}}{K \ln 10} \sum_{k=1}^{K} |\tilde{x}_k - y_k|.$$

Je geringer das LSD-Maß ausfällt, desto ähnlicher sind die spektralen Vektorfolgen, was mit einer erfolgreichen VC assoziiert wird. Die LSD soll den subjektiven Höreindruck adäquat widerspiegeln, wobei gewisse Einschränkungen bestehen. Die Abbildung 4 stellt die gemittelte log. Spektralverzerrung für vier ausgewählte VC-Methoden dar. Dafür wurden jeweils 21 Beispieläußerungen konvertiert und die LSD zu den zugehörigen Zielstimmenbeispielen berechnet.

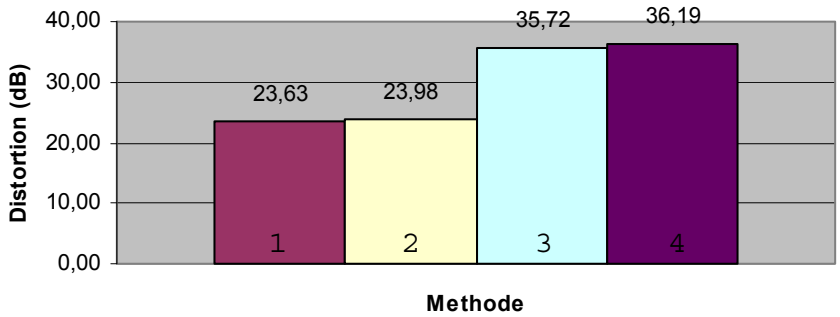

Abbildung 4: Log. Spektralverzerrung (log-spectral distortion).

Die VC-Methoden 1 und 2 führen zu ähnlichen LSD-Werten von 23,6 dB bzw. 24,0 dB. Die Methoden 3 und 4 bewirken eine stärkere spektrale Verzerrung von 35,7 dB bzw. 36,2 dB. Das instrumentell ermittelte Ranking entspricht der Bewertungsreihenfolge aus den Hörexperimenten.

4 Laufzeit-Performanz

VC-Methoden sind in der Regel rechenintensiv und werden oft mittels einer mathematischen Toolbox wie MATLAB entwickelt und dabei bezüglich ihrer qualitativen Zielparameter getestet und optimiert. Aus der Laborvariante können die benötigten Softwarealgorithmen (z. B. als C-Code) anschließend automatisch generiert werden.

Viele kommerzielle Anwendungen benötigen eine aufwendige Optimierung des Laufzeitverhaltens – oftmals verbunden mit der Softwareportierung auf einen digitalen Signalprozessor (DSP). Dabei stehen Aspekte wie Echtzeitfähigkeit, große Signallängen oder blockweise Verarbeitung sowie ggf. Mehrkanaligkeit im Vordergrund. Um die potentiellen Laufzeitprobleme zu illustrieren, werden im Folgenden Testergebnisse für die VC-Methode 1 vorgestellt.

4.1 Einfluss des Rechnersystems

Im Testszenario wird die Konvertierungszeit t für eine weibliche Quellstimme gemessen, wobei die Signallänge τ von 1,6 s bis 180 s variiert wird. Die Einstellparameter dieser VC-Methode werden willkürlich gewählt: $\alpha = 1,13$ s und $\rho = 0,842$ s (mittlere Einstellparameter für eine Konvertierung Frau 1 $\rightarrow$ Frau 2), d. h., es findet kein Training statt. Die Abbildung 5 vergleicht die manuell gemessenen (über drei Wiederholungen gemittelten) Zeiten für zwei PC-Systeme. Als langsamer Test-PC fungiert ein typisches Laptop-Modell aus dem Jahr 2007. Der „schnelle“ Rechner ist ein aktuelles PC-Modell.

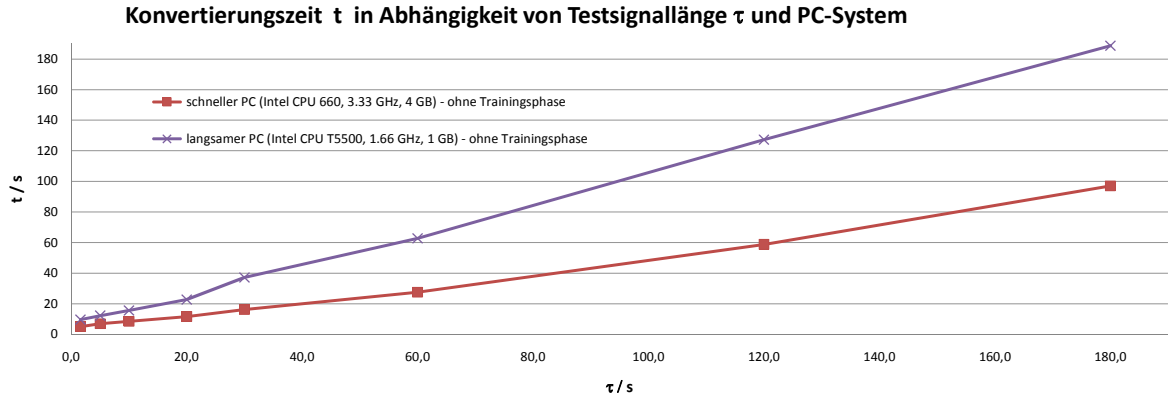

Abbildung 5: Konvertierungsdauer für verschiedene Signallängen und zwei PC-Beispiele.

Der schnelle PC unterschreitet den Echtzeitfaktor (Verhältnis von Rechenzeit und verarbeiteter Signallänge) von 1 etwa ab einer Signallänge von 10 s – vorher fällt der Overhead der Datenvorbereitung stärker in das Gewicht. Der Echtzeitfaktor im eingeschwungenen Zustand beträgt 0,48 … 0,53. Oberhalb des Minutenbereichs ist das Verhältnis von Rechenzeit und Signallänge weitgehend linear. Im Sub-Minutenbereich ergeben sich Abweichungen, da die Signalverarbeitung und -speicherung blockweise erfolgt. Bereits das einfache Testszenario (u. a. ohne Training) beansprucht umfangreiche Rechenressourcen und ist bezüglich des Echtzeitfaktors grenzwertig, v. a., wenn weitere Algorithmen parallel laufen. Der Echtzeitfaktor des langsamen Systems beträgt 1,05 … 1,15.

4.2 Einfluss der VC-Parametrisierung

Der Einfluss der VC-Parametrisierung (Anpassungsfaktoren α und ρ) wird mit der gleichen Quellstimme ($\tau = \text{const} = 60$ s) und für den langsamen PC getestet, um ein realistisches Test-

szenario abzubilden. Die Abbildung 6 stellt die Konvertierungsdauer für unterschiedliche α-Werte dar ($\rho = \text{const} = 1{,}0$). Der Echtzeitfaktor variiert dabei um ca. 6 % (0,99 … 1,05).

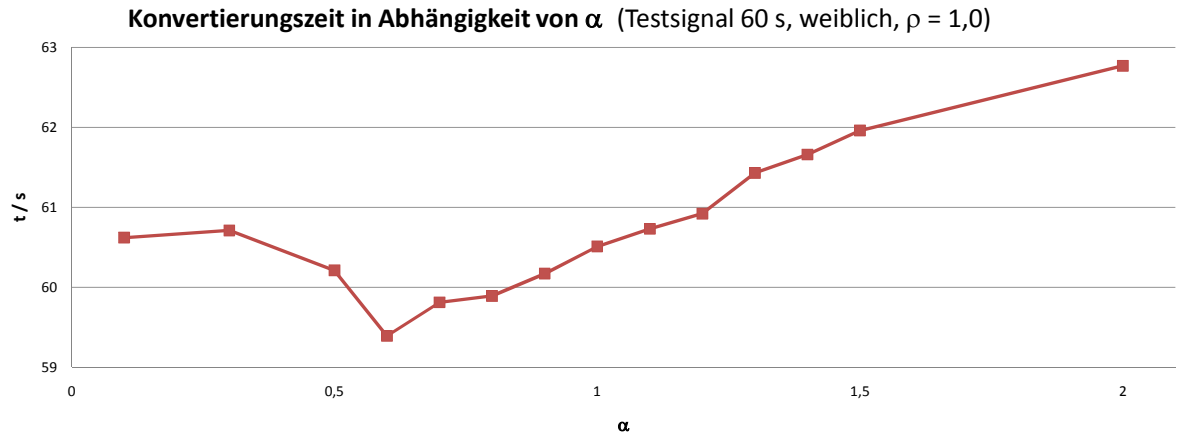

Abbildung 6: Konvertierungsdauer für verschiedene α-Werte.

Die Abbildung 7 zeigt die Konvertierungsdauer in Abhängigkeit von ρ ($\alpha = \text{const} = 1{,}0$). Die Schwankungsbreite des Echtzeitfaktors beträgt max. 15 % (0,95 … 1,10).

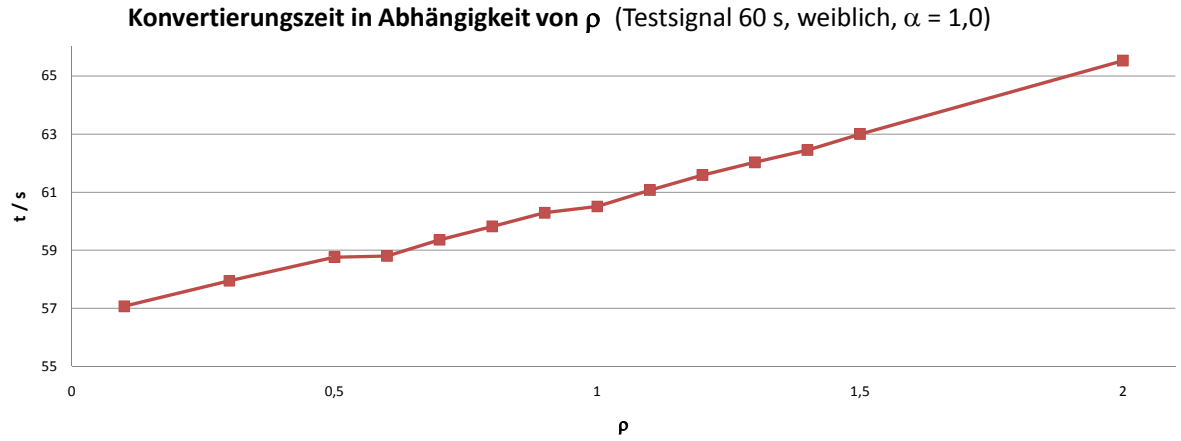

Abbildung 7: Konvertierungsdauer für verschiedene ρ-Werte.

Ein weiteres Experiment betrifft vier Standardparameter-Sätze für die Konvertierung Mann / Frau $\rightarrow$ Mann / Frau (vgl. Tabelle 1) sowie den Test mit dem neutralen Parametersatz $\alpha = \rho = 1{,}0$. Die Testergebnisse werden in der Abbildung 8 dargestellt. Der Echtzeitfaktor liegt in der gleichen Größenordnung wie zuvor (ca. 0,99 … 1,08). Die Standardabweichung für wiederholte Messungen innerhalb einer Kategorie beträgt max. 541 ms (0,9 %, bezogen auf die Signallänge) und für den neutralen Parametersatz 75 ms (0,1 %, bezogen auf die Signallänge).

4.3 Einfluss der Trainingsphase

In den Vorexperimenten wurden Konvertierungszeiten ohne vorheriges Training zur Bestimmung der VC-Parameter ermittelt. Ein realistisches Anwendungsszenario erfordert in der Regel ein entsprechendes Training. Die Abbildung 9 stellt die Rechenzeiten des langsamen PCs ohne Training (α und ρ manuell gewählt), mit einem kurzen Training (Zielsignal 1,8 s) und mit einem normalen Training (Zielsignal 25,6 s) gegenüber.

Tabelle 1: VC-Standardparameter (ohne Training oder manuelle Justage).

Quelle	Ziel	α	ρ
Mann	Mann	1,210	0,928
Mann	Frau	0,890	1,623
Frau	Mann	1,130	0,615
Frau	Frau	1,130	0,842

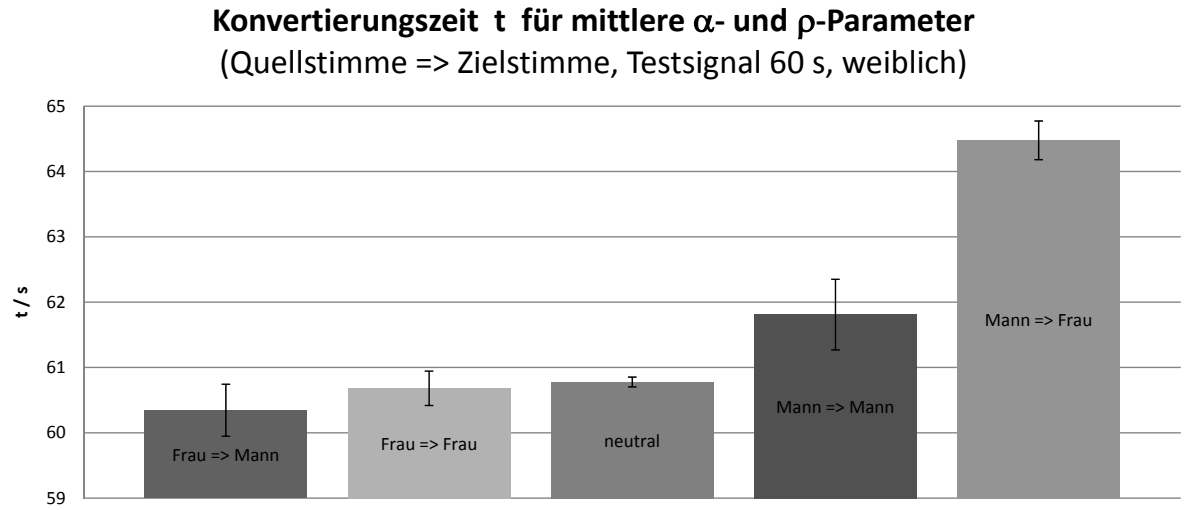

Abbildung 8: Konvertierungsdauer für Standardparameter (Mittelwerte).

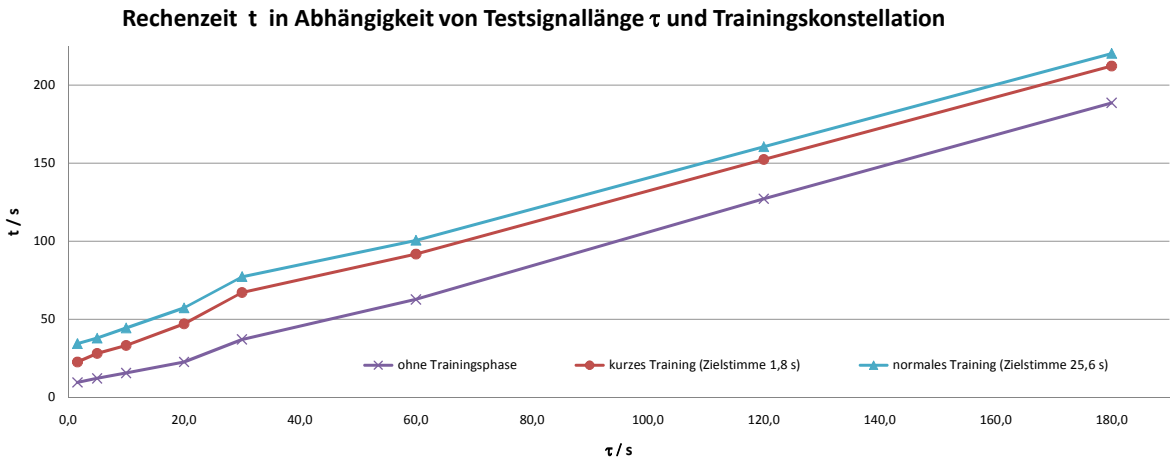

Abbildung 9: Konvertierungsdauer für verschiedene Signallängen und Trainingskonstellationen.

Bereits ein „kurzes Training“ führt zu einer zusätzlichen mittleren Rechenzeit von 22,3 s. Der resultierende Echtzeitfaktor beträgt bei kurzen Signalen (< 10 s) 3 …4 und stagniert für Signallängen ≥ 180 s bei ca. 1,2. Ein „normales Training“ verlängert die mittlere Rechenzeit um weitere 10,0 s.

5 Zusammenfassung

Bei der Entwicklung von VC-Methoden stehen die Qualitätsaspekte (Ähnlichkeit zur Zielstimme, Signalartefakte etc.) im Vordergrund. Die mit subjektiven Paarvergleichen (Präferenz-

Hörtest) ermittelte Ähnlichkeitsreihenfolge der VC-Methoden entspricht dem Ranking auf Basis objektiver Abstandsmessungen zwischen den VC-Beispielen und ihrer jeweiligen Zielstimmenreferenz. Die von den Hörern empfundene Höranstrengung sowie die Sprachqualität korrelieren ebenfalls mit der Ähnlichkeitsbewertung.

Der Beitrag untersuchte darüber hinaus das Laufzeitverhalten einer ausgewählten VC-Methode als Voraussetzung für kommerzielle Implementierungen, z. B. im Medienbereich. Selbst unter günstigen Voraussetzungen (schnelles Rechnersystem, lange Signalabschnitte, ohne Training, keine Parallelprozesse etc.) wird nur ein Echtzeitfaktor von ca. 0,5 erreicht.

Auf einem langsamen PC ist die Laufzeitperformanz inakzeptabel. Der Echtzeitfaktor liegt deutlich oberhalb von 1 – gegebenenfalls noch überlagert von dynamischen Betriebssystemeinflüssen. Die Laufzeitabhängigkeit von den gewählten Konvertierungsparametern α und ρ ist mit einer Schwankungsbreite von max. 15 % als weniger kritisch einzuschätzen.

Bedingt durch Overhead bei der Datenvorbereitung, führt bereits ein kurzes Training – verbunden mit einer fehlerbehafteten Bestimmung der Konvertierungsparameter – zu einer erheblichen Rechenzeitverlängerung. Der Echtzeitfaktor verschlechtert sich für kurze Signale auf über 3 und stagniert für lange Signale bei ca. 1,2. Die zusätzliche Laufzeit einer verlängerten Trainingsphase ist hingegen akzeptabel und bewirkt eine robustere Bestimmung der Konvertierungsparameter.

Literatur

[1] SCHWARZ, J.: *Statistische Stimmenumwandlung in Kombination mit prosodischen Modellen*. Christian-Albrechts-Universität zu Kiel, 2010. Dissertation.

[2] PITZ, M. und H. NEY: *Vocal Tract Normalization Equals Linear Transformation in Cepstral Space*. In: *Proc. IEEE Trans. on Speech and Audio Processing*, Band 13, Seiten 930 – 944, 2005.

[3] EICHNER, M., M. WOLFF und R. HOFFMANN: *Voice Characteristics Conversion for TTS Using Reverse VTLN*. In: *Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal*, Band 1, Seiten 17 – 20, 2004.

[4] SÜNDERMANN, D., G. STRECHA, A. BONAFONTE, H. HÖGE und H. NEY: *Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis*. In: *Proc. Interspeech, Lissabon*, 2005.

[5] STRECHA, G., O. JOKISCH, M. EICHNER und R. HOFFMANN: *Codec Integrated Voice Conversion for Embedded Speech Synthesis*. In: *Proc. Interspeech, Lissabon*, Seiten 2589 – 2592, 2005.

[6] SÜNDERMANN, D.: *Text-Independent Voice Conversion*. Universität der Bundeswehr München, 2008. Dissertation.

[7] STRECHA, G., M. WOLFF, F. DUCKHORN, S. WITTENBERG und C. TSCHÖPE: *The HMM Synthesis Algorithm of an Embedded Unified Speech Recognizer and Synthesizer*. In: *Proc. Interspeech, Brighton*, Seiten 1763–1766, 2009.

[8] HAGEN, R.: *Spectral Quantization of Cepstral Coefficients*. In: *Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing (ICASSP), Adelaide*, Band 1, Seiten 509 – 512, 1994.