Sprechertransformation auf Basis der Line-Spectrum-Frequencies

Ulrich Balss, Herbert Reininger

Institut für Angewandte Physik der J.W.Goethe-Universität Frankfurt a.M.
Robert-Mayer-Straße 2–4, D–60054 Frankfurt am Main, BRD
e-mail: Balss@iap.uni-frankfurt.de

1 Einleitung

In verschiedenen Anwendungsbereichen ist die Möglichkeit der Adaption synthetischer Sprache an die Stimmcharakteristik eines vorgegebenen realen Sprechers ein für die Akzeptanz relevantes Leistungsmerkmal. In diesem Beitrag wird eine auf einer Matrixtransformation beruhende Methode der Stimmtransformation vorgestellt. Ausgehend von der Darstellung eines Vorversuchs, dessen Ergebnisse die Verwendung von Line-Spectrum-Frequencies (LSF) motivieren, wird zunächst die hier verwandte Transformationsvorschrift entwickelt. Nachfolgend wird ein sprecherspezifisches Training der Transformationsparameter diskutiert. Anhand der Ergebnisse subjektiver Hörtests wird schließlich die Leistungsfähigkeit der neuen Methode mit der eines VQ-basierten Transformationsverfahrens verglichen.

2 Untersuchungen zur Sprecherabhängigkeit von Prädiktorparametern

Als Vorversuch wurde zunächst für verschiedene Darstellungen der Parameter eines linearen Prädiktors ermittelt, ob und auf welche Weise eine zur Sprechertransformation heranziehbare Relation zwischen den Parametersätzen einander entsprechender Sprachsignale verschiedener Sprecher besteht. Dazu werden jeweils ausgehend von Sprachsignalen eines einzelnen Sprechers durch Resynthese anhand einer Datenbasis korrespondierende Signale erzeugt, die abgesehen von den – zuvor durch ein automatisches Verfahren [1] aus der Originalsprache ermittelten und zur Ansteuerung der konkatenativen Synthese [2] verwandten – Intonationsparameter Lautdauer und Grundfrequenz die Sprechercharakteristik der Datenbasisstimme aufweisen. Die Prädiktorparameter der paarweise betrachteten Original- und Synthesesignale lassen sich in grundfrequenzsynchronen Analyserahmen bestimmen und deren einander jeweils korrespondierende Werte graphisch gegeneinander auftragen.

Für die LSF zeigt sich dabei, insbesondere wenn zu jedem gesprochenen Laut eine separate Graphik erstellt wird, ein gut auswertbarer Zusammenhang. Bild 1 auf der folgenden Seite illustriert diesen am Beispiel eines kurzen Testsignals bestehend aus drei aufeinanderfolgenden, von einem männlichen Originalsprecher gesprochenen langen Vokalen. Wie man sieht, sind die Datenpunkte innerhalb eines Lautes mit nur geringen Abweichungen entlang einer Kurve angeordnet. Diese verläuft zwar nahe der Diagonalen, unterscheidet sich aber doch sichtbar von ihr, was insofern relevant ist, als daß die Diagonale durch die für LSF geltende Ordnungsrelation hervorgehoben ist. Somit repräsentiert gerade die Abweichung der Kurvenform von der Diagonalen die sprecherspezifischen Unterschiede in der Signalcharakteristik.

3 Entwurf der Transformationsvorschrift

Als Transformationsvorschrift auf Basis des beobachteten Zusammenhangs wurden im folgenden verschiedene Methoden untersucht. Für diese erweist sich die Eigenschaft der LSF, sich gut interpolieren zu lassen [3], als Vorteil. Ein numerisches Verfahren stellt die Approximation des gemittelten Kurvenverlaufs durch Splinefunktionen dar. Aufwandsgüstiger als dieses ist hingegen die Transformation der analyserahmenweise zu Vektoren zusammengefaßten LSF-Werte über eine Diagonalmatrix, mit der eine nahezu identische Synthesesprache resultiert. Bei einer hinreichend geringen Abweichung der einzelnen LSF-Werte zu ihren Mittelwerten, entspricht dieses Verfahren näherungsweise einer Abtastung der gesuchten Funktion.

Solange nicht der empfindliche Einfluß des Abstandes nahe beieinanderliegender LSF-Werte auf das LPC-Spektrum berücksichtigt wird, treten jedoch zusätzliche Störsignale in der Synthesesprache auf. Daher müssen in eine erweiterte Matrixmethode die Korrelationen zwischen benachbarten LSF-Werten einbezogen werden. Indem der spektrale Einfluß der einzelnen LSF-Werte lokal begrenzt ist, genügt es dabei, in der Transformation eines Wertes jeweils nur seine unmittelbaren Nachbarn mitzubetrachten, so daß eine Tridiagonalstruktur der Matrizen resultiert. Ausgehend von den, mit einen Prädiktor der Ordnung P bestimmten, LSF $l_i^D(n)$ der Datenbasisstimme im n-ten Analyserahmen und den Matrixelementen $\alpha_{i,i+j}$ mit $i = 1, 2, \ldots, P\,; \quad j = -1, 0, 1\,; \quad l_0^D := 0\,; \quad l_{P+1}^D := \pi$ folgen somit die zugehörigen transformierten LSF als

$$l_i^T(n) = \sum_{j=-1}^{1} \alpha_{i,i+j} \cdot l_{i+j}^D(n) \quad . \tag{1}$$

Zur Adaption der $\alpha_{i,i+j}$ wird der über die N Analyserahmen eines Lautes aufakkumulierte quadratische Fehler

$$\varepsilon := \sum_{n=0}^{N-1} \sum_{i=1}^{P} \left(l_i^O(n) - l_i^T(n) \right)^2 \tag{2}$$

zwischen Original-LSF $l_i^O(n)$ und transformierten als Optimierungskriterium verwandt, um so die Auswirkungen der Transformation in konsistenter Weise berücksichtigen zu können. Durch Minimierung von ε gemäß

$$\frac{1}{2} \cdot \frac{\partial \varepsilon}{\partial \alpha_{i,i+j}} \stackrel{!}{=} 0 \tag{3}$$

folgt ein lineares Gleichungssystem für die Elemente der Tridiagonalmatrix. Dieses kann in Teilgleichungssysteme

$$\sum_{k=-1}^{1} \left(\sum_{n=0}^{N-1} l_{i+j}^D(n) \cdot l_{i+k}^D(n) \right) \cdot \alpha_{i,i+k} = \sum_{n=0}^{N-1} l_i^O(n) \cdot l_{i+j}^D(n) \tag{4}$$

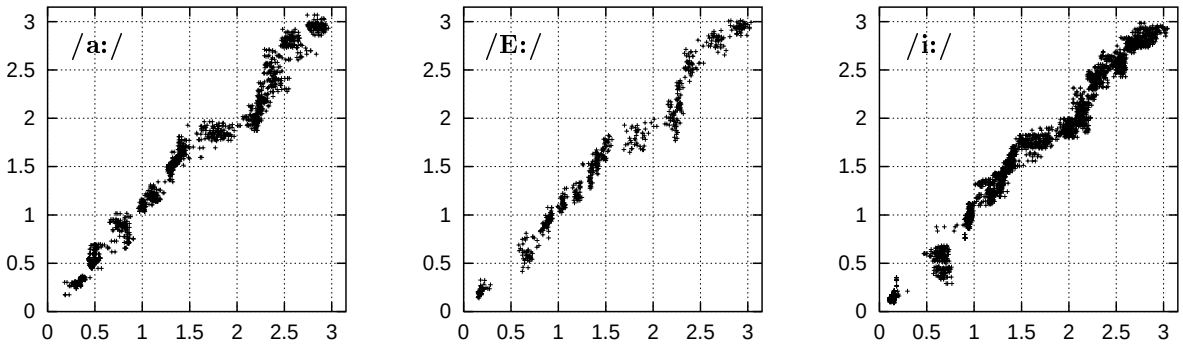

Bild 1: Lautspezifischer Vergleich der LSF von Datenbasissprecher (Abszisse) und Originalsprecher (Ordinate)

der Dimension 3 aufgespalten werden, aus denen sich jeweils die Elemente eines Zeilenvektors berechnen lassen. Somit ist eine sehr effiziente Lösung des Gesamtgleichungssystems realisierbar.

4 Training der Transformationsparameter

In weiteren Experimenten wurde untersucht, inwieweit die anhand einzelner Trainingssätze bestimmten Transformationsmatrizen für deren Sprecher repräsentativ sind. Indem die Matrizen mit der im folgenden dargestellten Methode quantisiert werden, erhält man ein sprecherspezifisches Codebuch, anhand dessen für eine disjunkte Menge von Testsätzen des gleichen Sprechers die Genauigkeit der Transformation für verschiedene Codebuchgrößen kontrolliert werden kann. Ausgehend von einem Initialisierungscodebuch wird in einem iterativen Algorithmus erst eine Zuordnung der Lautsegmente der Trainingsdaten zu den bezüglich des Fehlermaßes (2) bestpassenden Codebuchmatrizen vorgenommen, um in einem zweiten Schritt die Matrizen anhand der Paare von Original- und Datenbasis-LSF der ihnen jeweils zugeordneten Trainingsdaten nachzuoptimieren. Die durch die Tridiagonalstruktur der Matrizen begrenzte Zahl von Freiheitsgraden wirkt dabei einer Überadaption der Transformationsvorschrift an die verwandten Trainingssätze entgegen.

Eine Auswertung der resynthetisierten Testsätze zeigt, daß bereits mit einer einzigen Codebuchmatrix eine nennenswerte Annäherung der Stimmcharakteristik an die der Originalsprache erreicht werden kann. Durch die Verwendung von mehr Matrizen läßt sich dies noch steigern, wobei deutliche Korrelationen zwischen der Auswahl der quantisierten Matrizen und dem Lautinhalt beobachtet werden können. Folglich wurden im weiteren lautspezifische Codebücher berechnet. Hier zeigt sich, daß deren Sprachqualität mit der Datenrate kaum noch zunimmt und somit eine Matrix pro Laut als ausreichend betrachtet werden kann.

5 Vergleichende Hörtests

Im subjektiven Hörvergleich wurde, bei Verwendung einer weiblichen Synthesedatenbasis, das hier vorgestellte Verfahren dem in [4] vorgeschlagenen VQ-basierten gegenübergestellt. Dabei zeigt sich für die Transformation auf andere weibliche Stimmen eine für beide Methoden vergleichbare, gute Wiedergabe der sprecherspezifischen Klangeigenschaften. Bei der matrixbasierten Methode sind darüberhinaus Störgeräusche in der transformierten Synthesesprache deutlich geringer. Die noch verbleibenden Störgeräusche in einzelnen Sprachsignalen treten in ähnlicher Form – wenn auch zum Teil an anderer Stelle – ebenfalls in untransformierten Synthesesprachsignalen auf, so daß deren Ursache eher innerhalb des implementierten konkatenativen Sprachsynthesesystems vermutet werden können.

6 Zusammenfassung

In diesem Beitrag wurde eine neuartige Methode zur Sprechertransformation vorgestellt. Die Abbildung auf den Zielsprecher erfolgt hierbei durch Multiplikation von – als Vektoren aufgefaßten – in grundfrequenzsynchronen Analyserahmen bestimmten LSF-Tupeln mit einer jeweils lautspezifisch optimierten Tridiagonalmatrix. Durch die begrenzte Zahl von Freiheitsgraden wird eine Überadaption der Transformationsvorschrift an die verwandten Trainingssätze vermieden. In subjektiven Hörtests wurde das vorgestellte Verfahren mit einem VQ-basierten verglichen. Dabei zeigt sich eine für beide Methoden vergleichbare, gute Wiedergabe der sprecherspezifischen Klangeigenschaften. Bei der matrixbasierten Methode sind darüberhinaus Störgeräusche in der transformierten Synthesesprache deutlich geringer.

Literatur

[1] Balss, U., Englert, F., Reininger, H., Schlothauer, M.: „Automatische Extraktion von Intonations-Parametern aus Sprachdaten“, Fortschritte der Akustik – DAGA 97, DEGA, Universität Kiel 1997, S. 549–550.

[2] Moulines, E., Charpentier, F.: „Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones“, Speech Communication 9 (1990), S. 453–467.

[3] Paliwal, K.K.: „Interpolation Properties of Linear Prediction Parametric Representations“, Proc. Eurospeech, Madrid 1995, S. 1029–1032.

[4] Geravanchizadeh, M.: „Spektrale Transformation von Stimmen“, Studientexte zur Sprachkommunikation, „Elektronische Sprachverarbeitung“, Heft 20, Cottbus 2000, S. 70–77.