Auswirkungen spektraler Eigenschaften auf die Klangfarbe und Qualität übertragener Sprache

Marcel Wältermann[1], Alexander Raake[2], Sebastian Möller[2]

[1] *Institut für Kommunikationsakustik, Ruhr-Universität Bochum, 44780 Bochum, Deutschland*

[2] *Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Deutschland*

Einleitung

Die Klangfarbe eines Hörereignisses wird hauptsächlich durch spektrale Eigenschaften des korrespondierenden Schallereignisses bestimmt. Sie hat dabei häufig einen mehrdimensionalen Charakter und kann durch eine oder mehrere perzeptive Dimensionen beschrieben werden. Im Kontext der Sprachübertragung können lineare Verzerrungen des Übertragungskanals (Leitungsfilter, Endgeräte) zu unterschiedlichen Klangverfärbungen führen und zudem die Sprachqualität beeinträchtigen. Die sendeseitige Raumakustik kann Einfluss auf die *Direktheit* der Sprache haben, die sich widerum im Frequenzgang niederschlägt (Kammfilter-Effekt) [1].
Um den multidimensionalen Wahrnehmungsraum der Klangfarbe und die Auswirkung auf die Sprachübertragungs-Qualität näher zu untersuchen, wurden zwei auditive Experimente durchgeführt. Mit Hilfe einer sogenannten *Sortieraufgabe* als eine Methode der Ähnlichkeitsskalierung wurden von den Versuchspersonen (VPn) ähnlich klingende Sprachproben für jeweils unterschiedliche Sprecher gruppiert. Hierin wurde eine Vielzahl von Sprachproben unterschiedlicher spektraler Eigenschaften untersucht. Mit Hilfe einer Multidimensionalen Skalierung konnte der Wahrnehmungsraum quantitativ bestimmt werden. Durch Abbildung der Raumkonfiguration auf Gesamtqualitätsurteile können so Aussagen über den Einfluss spektraler Eigenschaften auf die Sprachqualität formuliert werden.

Sprachproben und Versuchsbedingungen

Für das Experiment wurde eine Vielzahl von Frequenzcharakteristika einer Telefon-Übertragungstrecke generiert. Im Einzelnen sind dies:

- 3 Sende-/Empfangscharakteristika konventioneller Handapparate
- 9 Sende-/Empfangscharakteristika mobiler Endgeräte
- 7 Aufnahmen in jeweils zwei unterschiedlichen Büroräumen/KFZ-Innenräumen mit Kunstkopf und Freisprecher

Darin enthalten sind 4 Bedingungen, die als relevant für die Qualitätsdimension *Direktheit/Frequenzgang* identifiziert wurden [1]. Weitere 38 Sprachproben wurden mit einem in [2] entwickelten System erzeugt, das es durch die Variation von 5 Parametern erlaubt, eine gezielte spektrale Beschaffenheit zu erzeugen: Die *equivalent rectangular bandwidth* $ERB \in \{5, 8, 11, 14\}$ Bark, der Schwerpunkt des Frequenzganges $\theta_G \in \{4, 6, 8, 10, 12, 14\}$ Bark, die Steigung $\beta \in \{-2, 0, 2\}$ dB/Bark, die Tiefe γ_D der Welligkeit des Frequenzganges ($\gamma_D \in \{0, 8\}$ dB) und die Häufigkeit γ_R der Welligkeit ($\gamma_R \in \{0, 0.5, 3\}$ ripple/Bark). Zusätzlich wurden Parameterwerte für die erstgenannten (realen) Sprachproben mit einer erweiterten Version des in [3] vorgestellten Verfahrens geschätzt und mittels [2] entsprechend idealisierte Sprachproben hinzugefügt.
Insgesamt umfasst das Experiment 80 spektral unterschiedliche Bedingungen. Als Quellmaterial wurden vier verschiedene Sprecher/Satzkombinationen verwendet (2 m, 2 w), für die das Experiment getrennt durchgeführt wurde. Sämtliche Sprachproben wurden auf einen Frequenzbereich von 300-3400 Hz begrenzt. In den Experimenten kam ein Standard-Handapparat mit neutraler Übertragungsfunktion zum Einsatz. Es nahmen $K = 20$ VPn (14 m, 6 w) im Alter von 19 bis 37 Jahren teil ($\varnothing = 27$ Jahre).

Multidimensionale Skalierung

Die Multidimensionale Skalierung (MDS) ist eine statistische Technik, die es erlaubt, Unähnlichkeiten δ_{ij} zwischen zwei Objekten i und j in Distanzen d_{ij} zu transformieren. Sind alle Unähnlichkeiten zwischen einer Anzahl von I Objekten bekannt, können die Objekte in einem R-dimensionalen Raum ($R < I$) dargestellt werden (s.a. [4]). So können Beziehungen zwischen den Objekten identifiziert und interpretiert werden.
Üblicherweise erhält man die Unähnlichkeiten δ_{ij} durch vollständigen Paarvergleich aller Objekte, d.h. durch $I(I-1)$ Beurteilungen. Da die vorliegenden $I = 80$ Sprachproben einen vollständigen Paarvergleich aus Aufwandsgründen nicht zulassen, kommt in der hier beschriebenen Untersuchung eine alternative Datenerhebungsmethode zur Anwendung, deren Ergebnisse in vielen vergleichenden Untersuchungen diejenigen eines Paarvergleichs zufriedenstellend annähern (vgl. [5]). Die Aufgabe der VPn besteht darin, unter Zuhilfenahme einer geeigneten Software ähnliche Proben in jeweils eine gemeinsame Gruppe einzusortieren. Daraus läßt sich für jede VP und Sprecher/Satzkombination eine $I \times I$ Inzidenzmatrix mit einem binären $(0, 1)$-Eintrag pro Objektpaar ableiten, der anzeigt, ob die entsprechenden Objekte in eine gemeinsame Gruppe einsortiert wurden. Anschließend werden diese Matrizen über die K VPn zu einer Ähnlichkeitsmatrix $\mathbf{\Sigma}$ summiert, die dann die Anzahl der Vorkommnisse jeweils eines Objektpaares in einer gemeinsamen Gruppe enthält (die Diagonalelemente sind mit K belegt). Diese wird durch $\mathbf{\Delta} = K - \mathbf{\Sigma}$ in eine Unähnlichkeitsmatrix

$\mathbf{\Delta}$ transformiert, deren Elemente die Unähnlichkeiten δ_{ij} sind. Die Matrix $\mathbf{\Delta}$ wurde nicht-metrisch multidimensional skaliert (vgl. [4]). Das Ergebnis ist jeweils eine Konfiguration $\mathbf{X}$ ($I \times R$-Matrix) für jeden Sprecher, die die Koordinaten der die Sprachproben repräsentierenden Punkte im R-dimensionalen Raum enthält.

Analyseergebnis

Eine Dimensionalität von $R = 2$ erwies sich für alle vier Sprecher sowohl bezüglich statistischer Parameter wie Anpassungsgüte zwischen δ_{ij} und d_{ij} ($Stress \leq 0.10$) und der Varianzabdeckung ($R^2 \geq 95.2\%$) als auch im Sinne der Interpretierbarkeit der Konfiguration als adäquat. Die Konfigurationen für die vier Sprecher wurden durch orthogonale Rotation, Translation und Skalierung zu einer gemeinsamen (Konsens-)Konfiguration $\mathbf{X}_C$ transformiert (*Generalized Procrustean Analysis*, vgl. [4]). Die Summe der Fehlerquadrate, normiert auf die quadrierten Elemente von $\mathbf{X}_C$, beträgt 7.4..13.8% und deutet eine hohe Ähnlichkeit der Sprecher/Satz-Konfigurationen an. Die Punktkonfiguration zeigt Abb. 1.

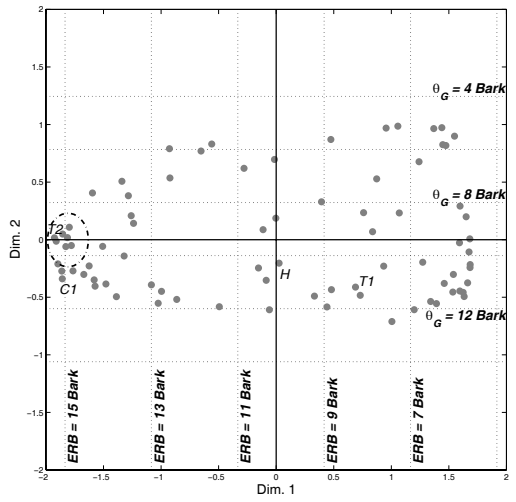

Abbildung 1: 2-dim. Konsens-Konfiguration.

Zur Interpretation können a-priori bekannte Eigenschaften der Sprachproben mit der Konfiguration in Verbindung gebracht werden. Z.B. lassen sich axiale Partitionierungen durch die Parameter ERB und θ_G erkennen, angedeutet durch die „Iso-ERB- und Iso-θ_G-Konturen“(gestrichelte Linien) in Abb. 1 (Korrelation von ERB und Dim. 1: $r = -0.92$, $RMSE = 0.51$, Korrelation von θ_G und Dim. 2: $r = -0.80$, $RMSE = 0.29$). Weiterhin liegen Sprachproben mit Raumeffekten eher mittig oder in der rechten Halbebene der Konfiguration (z.B. H in Abb. 1, Kunstkopf/Freisprecher-Aufnahme). Insgesamt verliert der Klang der Sprachproben in positiver Richtung von Dim. 1 an Fülle, Nähe, oder *Direktheit*. Dim. 2 bestimmt durch die Korrelation mit dem Schwerpunkt θ_G den *Frequenzgehalt* bei fester Bandbreite (z.B. die Dominanz von tiefen oder hohen Frequenzen). Die Positionen derjenigen Stimuli sind eingezeichnet, die in [1] zur Dimensionsbezeichung *Direktheit/Frequenzgehalt* führten.

Auswirkung auf die Sprachqualität

In einem weiteren Hörversuch wurden Gesamtqualitätsurteile zu etwa der Hälfte der Sprachproben und zwei der Sprecher gesammelt. Hierin erhielten die in Abb. 1 umkreisten Sprachproben die besten Bewertungen auf einer 5-Punkte-Skala (u.a. die spektral neutrale Sprachprobe *T2*, vgl. [1]). Die Qualität fällt mit steigenden Werten auf Dim. 1, während es bzgl. Dim. 2 offensichtlich einen idealen Punkt gibt, von dem aus gesehen die Qualität in beide Richtungen abfällt. Modelliert man diese Zusammenhänge mathematisch, so gelingt es, die Qualität anhand der Raumstruktur mit einer Varianzabdeckung von $R^2 = 98.4\%$ vorherzusagen. Der Großteil der vorliegenden (reduzierten) Qualitätsurteile wird allerdings bereits durch Dim. 1 erklärt.

Zusammenfassung

Mit Hilfe einer MDS eines großen Satzes von Sprachproben konnten zwei Dimensionen der Klangfarbe von übertragener Sprache identifiziert werden: *Direktheit* und *Frequenzgehalt*. Die hohe Korrelation mit den Parametern ERB und θ_G bildet die Grundlage für die Modellierung und instrumentelle Schätzung des Wahrnehmungsraumes [3], und dient letztendlich der Vorhersage der Qualität von spektral veränderter übertragener Sprache. Die gewonnenen Erkenntnisse lassen z.B. Rückschlüsse auf optimale Übertragungseigenschaften von Übertragungskomponenten wie z.B. Endgeräten zu.

Die Autoren danken der DFG (MO 1038), den DT Laboratories, Harman/Becker AS und HEAD acoustics für die Unterstützung dieser Arbeit.

Literatur

[1] Wältermann, M.; Scholz, K.; Raake, A.; Heute, U.; Möller, S.: Qualitätsdimensionen bei der Sprachübertragung in Modernen Telekommunikationsnetzen. DAGA 2006, Braunschweig, 2006.

[2] Huo, L.; Scholz, K.; Heute, U.: Idealized System for Studying the Speech-quality Dimension "Directness/Frequency content". In: *2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Berlin, 2006.

[3] Scholz, K.; Wältermann, M.; Huo, L; Raake, A.; Möller, S.; Heute, U.: Messtechnische Erfassung der Qualitätsdimension 'Direktheit/Frequenzgehalt' zur Instrumentellen Analyse und Beurteilung von Sprachqualität. DAGA 2006, Braunschweig, 2006.

[4] Borg, I.; Groenen, P.: *Modern Multidimensional Scaling – Theory and Applications*. Springer Series in Statistics, New York NY, 2005.

[5] Tsogo, L.; Masson, M.H.; Bardot, A.: Multidimensional Scaling Methods for Many-object Sets: A Review. *Multivariate Behavioral Research*, 35(3), 2000.