Vorhersage und Kontrolle der Sprachverständlichkeit in räumlich dargebotenen Audio-Konferenzschaltungen

Alexander Raake[1], Brian FG Katz[2], Gabriel Perez de la Sota[1]

[1] *Deutsche Telekom Laboratories, Technische Universität Berlin, 10587 Berlin, Deutschland, Email: alexander.raake@telekom.de*

[2] *LIMSI-CNRS, 91403 Orsay, Frankreich, Email: brian.katz@limsi.fr*

Einleitung

Heutige Telekonferenzsysteme sind gekennzeichnet durch schmalbandige Übertragung (300–3400 Hz) und das Zusammenmischen der Sprachsignale. Werden die unterschiedlichen Sprachkanäle stattdessen mittels paketbasierter Übertragung breitbandig (≥50–7000 Hz) und separat übermittelt und mit räumlichen Audio-Widergabemethoden dargeboten, so werden Telekonferenzen möglich, die zunehmend Merkmale einer realen Konferenzsituation aufweisen. Dabei kann die räumliche Darbietung je nach Anwendung mittels Binauraltechnik oder Mehrkanal-Lautsprecherwiedergabe erfolgen. Bei einer räumlichen Audio-Konferenz hängt die Sprachverständlichkeit von unterschiedlichsten Faktoren ab. Der Beitrag beschreibt einen Modellansatz zur Vorhersage der Sprachverständlichkeit für ein bestimmtes Zielsprecher-Hörer-Paar und eine gegebene Quellenanordnung, inklusive zur Übertragung eingesetzter Sprachkodierung. Das Modell basiert auf der Schätzung des so genannten Speech Reception Thresholds (SRT), d.h. der 50% Verständlichkeitsschwelle bei festem Störschallpegel. Modell und Vorhersagegenauigkeit werden anhand mehrerer Sprachverständlichkeitstests erörtert.

Virtuelle Audio-Konferenzen und Sprachverständlichkeit

Bei räumlicher Darbietung können zukünftige Anwendungen wie sprach-basierte virtuelle "Chat-Rooms" stark von einer gesteigerten Trennbarkeit der Quellen, besseren Sprachverständlichkeit und geringerem Höraufwand profitieren. Dabei wird insbesondere der Cocktail-Party-Effekt´´ unterstützt, d.h. die Fähigkeit der menschlichen auditiven Wahrnehmung, unterschiedliche Quellen in einer Situation mit mehreren Kommunikationspartnern zu trennen und zu verstehen [1].

In einem virtuellen Chat-Room gibt es eine Vielzahl von Einflussfaktoren für die Sprachverständlichkeit, wie beispielsweise die Anzahl, die relativen Pegel und die Eigenschaften der Quellen (Stationarität, Grundfrequenz, Spektrum, Intonation, etc.), ihre Position relativ zum jeweiligen Hörer, die dargebotene Raumakustik, Beeinflussungen durch den Übertragungskanal (Kodierung, evtl. Bandbegrenzung, Übertragungsfehler, etc.), usw. Die gemessene Sprachverständlichkeit in % korrekt (auf Silben-, Wort- oder Satzebene) hängt u.a. auch von der verwendeten Messmethode ab. Zur Steigerung der Sensibilität beim Vergleich zwischen Sprecher-Maskierer-Konfigurationen können adaptive Methoden eingesetzt werden. Ein prominentes Beispiel ist der "speech reception threshold" (SRT) für die 50% Verständlichkeitsschwelle. Die Empfindlichkeit der Messung basiert auf der Steigung der psychometrischen Funktion der Sprachverständlichkeit in Rauschen am SRT, der typischerweise zwischen 10 und 20% pro dB Signal-zu-Rausch-Abstand liegt (z.B. [2]).

In dieser Studie wird folgende Messgröße als Ausgabemaß des Vorhersagemodells verwendet: Der Unterschied in dB SRT zwischen der jeweils betrachteten Konfiguration und einer festen Referenzsituation, bei der der Zielsprecher und ein stationäres Sprach-Rauschen als Referenz-Maskierer jeweils von vorne dargeboten werden (Azimuth $\varphi = 0°$).

$$\Delta SRT_i = SRT_i - SRT_{ref} \quad (1)$$

SRT Tests

Im Rahmen dieser Arbeit wurden zwei Reihen von SRT-Tests durchgeführt, jeweils in drei Teilen à 17 Versuchsbedingungen, d.h. insgesamt pro Testreihe $3 \cdot 17 = 54$. Jeder Test-Teil wurde mit zehn Versuchspersonen durchgeführt.

Die erste Versuchsreihe (Test 1) diente der Entwicklung des Modells. Sie umfasste Konfigurationen mit unterschiedlicher Anzahl und Art der Quellen, Quellen-Eigenschaften (Grundfrequenz, Spektrum, etc.), Quellen-Kodierung und räumlicher Quellen-Anordnung (nur Horizontalebene). Die zweite Versuchsreihe (Test 2) diente zur Evaluierung und Verbesserung des Modells, mit Kombinationen unterschiedlicher Maskierer bei unterschiedlichen Pegeln. Die Konfigurationen waren der Art $AixBjy$. A, B steht dabei für die Art der Maskierer (stationäres Sprachrauschen; gleicher Sprecher; anderer, weiblicher oder männlicher Sprecher; Mehr-Sprecher-Rauschen – "Babble"). Die Indizes i und j stehen für den Azimuth der jeweiligen Quelle, und x und y für den jeweiligen Pegel.

Für beide SRT Test-Reihen wurde die in [3] beschriebene Methode eingesetzt. Als Referenz-Maskierer dient ein 60 Sekunden langes stationäres Sprachrauschen mit gleichem Langzeit-Spektrum wie das Sprachsignal des Ziel-Sprechers [3]. Die Testsignale aller Versuchsbedingungen wurden durch Faltung mit im reflexionsarmen Raum im Abstand von 1.9 m aufgezeichneten Außenohr-Impulsantworten ("Head-related impulse responses", HRIRs) realisiert. Sämtliche Pegel wurden digital hinsichtlich des Quell-RMS-Wertes ("Root Mean Square", d.h. des Effektivwertes) eingestellt.

Modell

Ausgehend von den Test-Ergebnissen und der Literatur wurde ein erster Ansatz für ein Vorhersagemodell entwickelt. Es geht von einer parametrischen Szenenbeschreibung und einer groben Klassifikation der Quellen aus, die in Sprache, Rauschen und Babble unterschieden werden. Das Herzstück des Modells ist eine leicht abgewandelte Form des Modells aus [4]:

$$\Delta SRT = C\left[\alpha\left(1-\frac{1}{n}\sum_k \cos(\theta_i-\varphi)\right) + \beta\frac{1}{n}\left|\sum_k \sin(\theta_i-\varphi)\right|\right] \quad (2)$$

Dabei ist ΔSRT der Vorteil in dB SRT einer bestimmten Konfiguration von Zielsprecher und n stationären Sprach-Rauschquellen im Vergleich zur Situation, dass alle Quellen von vorne präsentiert werden. C, α und β sind Konstanten, n und Θ_i sind die Gesamtzahl und die jeweiligen Winkel der Maskierer in der Horizontalebene, und φ ist der Azimuth des Zielsprechers (alle Winkelangaben relativ zum Hörer). Der cos-Term beschreibt den Zugewinn durch zunehmendem Abstand zwischen Zielsprecher und Maskierer; der sin-Term beschreibt den Zugewinn durch die Asymmetrie der Anordnung und die dadurch ermöglichte binaurale Verarbeitung.

In der vorliegenden Arbeit wurde der Ansatz zudem anhand der Ergebnisse von Test 1 und der Literatur (z.B. [5]) erweitert auf andere Maskierer als stationäres Sprachrauschen und den Einfluss von Sprachkodierung. Darüber hinaus berücksichtigt das Modell unterschiedliche Maskierer-Pegel sowie den Fall, dass die Maskierer für eine gegebene Konfiguration unterschiedlicher Art sind. Mehr Einzelheiten über das Modell finden sich in [6].

Das Modell geht von einer zumindest teilweise gegebenen Additivität der Effekte hinsichtlich des SRT-Einflusses aus. Unterschiedliche Module wurden dazu realisiert, mit folgenden Eigenschaften:

- Transformation der zunächst absolut vorliegenden Pegel- und Positions-Informationen der Quellen in Werte relativ zum Hörer.
- Quellen, die nicht zur Verständlichkeit beitragen werden ignoriert.
- Berechnung des Unterschieds in dB SRT zwischen der Situation dass alle vorhandenen Quellen von vorne kommen und der Referenz-Situation mit nur einem Sprach-Rauschen als Maskierer (Zielsprecher und Maskierer beide von vorne).
- Berechnung des relativen Unterschieds zwischen einer Konfiguration mit räumlich verteilten Sprach-Maskierern und einer identischen Konfiguration bei der alle Maskierer stationäres Sprachrauschen sind.
- Modell gemäß Gleichung (2) [4].
- Schätzund des SRT-Einflusses durch Sprachkodierung.

Diese Funktionen und die Berechnung eines Gesamt-RMS werden angewendet auf (a) alle Maskierer (unter der Annahme dass alle die gleiche Signalleistung wie die stärkste Quelle haben), und (b) nur auf die Quelle mit dem höchsten Einzel-RMS. Eine erste Schätzung von ΔSRT wird durch lineare Interpolation zwischen den nach (a) und (b) bestimmten ΔSRT-Werten berechnet, ausgehend vom tatsächlichen Gesamt-RMS der Konfiguration an der Hörerposition.

Im Vergleich zu Test 1 zeigt das Modell eine lineare Korrelation von $\rho^2 = 0.98$ und einen mittleren quadratischen Fehler von $RMSE = 1.39\ dB$. Die Vorhersagegenauigkeit für Test 2 (unbekannte Daten) ist mit $\rho^2 = 0.86$ und $RMSE = 3.37$ erwartungsgemäß geringer.

Eine genauere Analyse der Vorhersagen und Test-Ergebnisse ergab folgende Schwachstellen des Modells:

(1) "Informational Masking" wird nicht betrachtet; sind die Sprachsignale konkurrierender Sprecher perzeptiv ähnlich, kann es statt zum erwarteten (energetischen) Vorteil gegenüber stationären Maskierern zu einer schlechteren Sprachverständlichkeit kommen (z.B. [4, 5]).

(2) Die Interpolationsfunktion zwischen dem signalstärksten Maskierer und dem kompletten Satz aller Maskierern bei angenommener gleicher Signalleistung ist nur im Falle stationärer Maskierer als linear anzunehmen.

Beide Aspekte werden hinsichtlich einer gemischt Szenen- und Signal-basierten, verbesserten Modellierung derzeit näher untersucht.

Literatur

[1] Cherry, E.C.: Some experiments on the recognition of speech with one and with two ears. J. Acoust. Soc. Am. 25 (1953), 975-979.

[2] Brand, T., Kollmeier, B.: Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J. Acoust. Soc. Am. 111 (2002), 2801-2810.

[3] Raake, A., Katz, B.F.G.: SUS-based method for speech reception threshold measurement in French. In: Proc. LREC (Language Resources and Evaluation Conference, 2006), 2028-2033.

[4] Bronkhorst, A.: The Cocktail Party phenomenon: A review of research on speech intelligibility in multitalker conditions. Acta Acustica utd w. Acustica 86 (2000), 117–128.

[5] Hawley, M., Litovsky, R., Culling, J.: The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. J. Acoust. Soc. Am. 115 (2004), 833-843.

[6] Raake, A., Katz, B.F.G.: Measurement and Prediction of Speech Intelligibility in a Virtual Chat Room. In: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems (2006), 40-43.