Ein rückwärtskompatibles räumliches Telefonkonferenzsystem mit automatischer Sprechergruppierung

Jens Ahrens, Alexander Raake, Sascha Spors, Jitendra Ajmera
Deutsche Telekom Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Deutschland
Email: {jens.ahrens, alexander.raake, sascha.spors, jitendra.ajmera}@telekom.de

Einleitung

Die Verwendung von herkömmlicher Telefontechnik in Kommunikationsszenarien wie Telefonkonferenzen führt zu verminderter Verständlichkeit der Teilnehmer und vermindertem Komfort. Die Hauptursachen dafür liegen in dem Verlust der räumlichen Merkmale und der reduzierten Bandbreite der Signale. Die Übertragung mehrerer paralleler Sprachsignalströme zu einem Empfänger ermöglicht hingegen eine räumliche Wiedergabe eines solchen Szenarios. Die bedeutendsten Vorteile liegen dann in der vereinfachten Identifikation sowie der vereinfachten Unterscheidbarkeit der einzelnen Sprecher durch den Hörer [1].
Wir schlagen ein System vor, das automatische Sprecheridentifizierung mit anschließender räumlicher Darbietung des entsprechend segmentierten Signals kombiniert. Dadurch wird Rückwärtskompatibilität zu bestehenden Übertragungstechniken wie dem klassischen Festnetz gewährleistet, die die parallele Übertragung mehrerer Sprachkanäle nicht erlauben. Im Endgerät werden Sprecherwechsel detektiert, Sprecher identifiziert, und das Signal wird entsprechend segmentiert. Die einzelnen Sprecher werden dann in einer virtuellen auditiven Umgebung räumlich verteilt wiedergegeben. Dieses kombinierte System wurde implementiert mit dem Ziel, die Identifikation der Sprecher für den Hörer zu erleichtern. Um dieses Ziel zu evaluieren, wurde die Fähigkeit der Hörer verglichen, die einzelnen Stimmen bei diotischer Wiedergabe, sowie räumlicher Wiedergabe mit automatischer bzw. fehlerfreier Segmentierung zu identifizieren.

Signalsegmentierung und Sprechergruppierung

Zur Erkennung von Sprecherwechseln und zur Gruppierung der Sprecher verwenden wir das Bayesian Information Criterion (BIC), wie in [3, 4] vorgeschlagen. Aus dem Gesamtsignal werden alle 10 ms Merkmalsvektoren mit relativ kleiner Fensterbreite extrahiert. Das Problem der Detektion der Sprecherwechsel ist als Hypothesentest formuliert. Die Nullhypothese stellt die Annahme dar, dass sich zwischen zwei aufeinander folgenden Merkmalsvektoren kein Sprecherwechsel befindet. Die Hypothese wird mittels der Betrachtung von Sprechermodellen, welche auf diesen Merkmalsvektoren trainiert wurden, angenommen oder verworfen.
Die Sprechergruppierung verläuft ebenso, wobei aber längere Fenster von Merkmalsvektoren und Sprechermodelle mit mehr Parametern betrachtet werden.

Binaurale Wiedergabe

Eine einfache Möglichkeit der räumlichen Darbietung stellt die binaurale Wiedergabe über Kopfhörer dar, die hier verwendet wurde. Dabei werden dem Signal Merkmale aufgeprägt, die das menschliche Gehör zur Lokalisation benutzt. Zu diesen Merkmalen gehören u.a. Laufzeitunterschiede zwischen den Ohren sowie spezifische spektrale Merkmale [2]. Der Hörer nimmt dann eine virtuelle Schallquelle wahr, deren Position über die Wahl der Lokalisationsmerkmale gesteuert werden kann. Im vorliegenden Fall wurden die einzelnen Signalsegmente mit den entsprechenden kopfbezogenen Raumimpulsanworten (Binaural Room Impulse Response, BRIR) für die gewünschte Position gefaltet. Die hierbei verwendeten Impulsantworten wurden in einem Aufnahmestudio mittels eines Kunstkopfes gemessen.

Evaluierung

Das System wurde sowohl objektiv als auch subjektiv evaluiert, wobei sich die objektive Evaluierung auf den Sprechersegmentierungs-/-gruppierungsalgorithmus beschränkte. Als Testsignale wurden zufällige Aufzählungen von Ziffern in deutscher Sprache von verschiedenen Sprechern aus der VeriDat Datenbasis [5] verwendet. Es wurde ein Beispiel mit zwei Sprechern und jeweils zwei Beispiele mit drei und vier Sprechern vorbereitet. Die Bandbreite der Signale betrug 4 kHz, die Länge der Zusammenstellungen jeweils ungefähr eine Minute.

Objektive Evaluierung der Sprecheridentifikation/-gruppierung

Die fünf Testbeispiele enthalten insgesamt 48 Sprecherwechsel, welche alle vom System korrekt detektiert wurden. Darüber hinaus meldete das System zwölf zusätzliche Sprecherwechsel, von welchen die meisten durch den Gruppierungsalgorithmus ignoriert wurden. Nach der Gruppierung blieben 46 Sprecherwechsel, wobei zwei davon inkorrekt waren. Zwei Sprecherwechsel wurden also nicht erkannt. Insgesamt beträgt die Performanz des Gruppierungsalgorithmus 88,20%, d.h. 88,20% der Verarbeitungsblöcke wurden korrekt zugeordnet.

Perzeptive Evaluierung

Die fünf Testbeispiele wurden in zufälliger Reihenfolge auf drei verschiedene Arten dargeboten: (1) diotisch

(„mono“) und jeweils eine binaurale Aufarbeitung des (2) automatisch segmentierten Datenstroms („auto“) und des (3) fehlerfrei segmentierten Datenstroms („ideal“). Um die Probanden an das System zu gewöhnen, begannen die Sitzungen immer mit den drei Beispielen mit zwei Sprechern.
Bei den räumlichen Beispielen wurden die Sprecher symmetrisch in der Horizontalebene aus Richtungen aus der Menge $\{60°, 30°, 0°, -30°, -60°\}$ angeordnet. 0° ist „geradeaus“. Alle Beispiele wurden über einen AKG K240 DF Kopfhörer wiedergeben, wobei den Probanden keinerlei Auskunft über die momentane Darbietungsweise gegeben wurde. 16 deutsche Muttersprachler (9 Männer, 7 Frauen) im Alter von 25 bis 49 Jahren wurden getestet.

Ergebnisse

Abbildung 1 zeigt die Performanz der Probanden bezüglich der Identifikation der Sprecher, aufgeschlüsselt nach Darbietungsart und Anzahl der Sprecher. Es ist zu erkennen, dass die Sprecheridentifikation bei der Darbietung von mehr als zwei Sprechern bei räumlicher Darbietung am zuverlässigsten ist. Dabei ist die Performanz bei idealer Segmentierung am größten, gefolgt von der automatischen Segmentierung. Die Performanz bei automatischer Segmentierung liegt dazwischen. Eine Varianzanalyse (ANOVA) gibt die Signifikanz der Faktoren „Anzahl der Sprecher“ und „Darbietungsweise“ mit $p = 0,000$ bzw. $p = 0,022$ an, was auf einen signifikanten Einfluss hinweist.

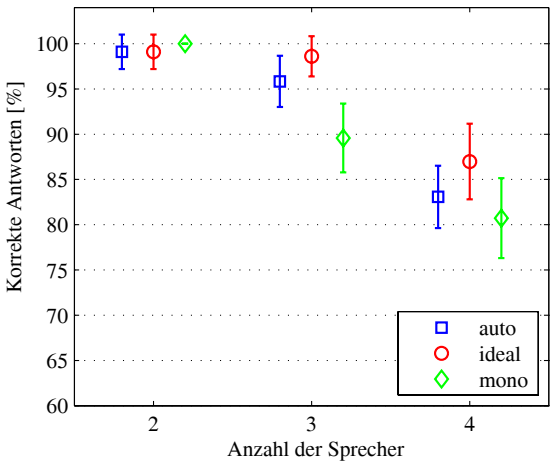

Abbildung 1: Identifikation der Sprecher; Die Balken geben das 95% -Konfidenzintervall an;

Die gleichen Tendenzen sind bei der Betrachtung der Anzahl der Sprecherwechsel, die fälschlich hinzugefügt wurden, sowie der verpassten (nicht erkannten) Sprecherwechsel zu erkennen, weshalb diesbezüglich keine expliziten Ergebnisse aufgeführt werden.
Abbildung 2 zeigt die Antworten der Probanden auf die Frage nach der Annehmlichkeit der Darbietung auf einer Skala von 0 (unangenehm) bis 100 (angenehm). Hier wird die ideale Segmentierung mit anschließender räumlicher Wiedergabe am angenehmsten empfunden. Die automatische Segmentierung mit räumlicher Wiedergabe wird am unangenehmsten bzw. als kaum angenehmer als diotische Darbietung empfunden. Offenbar werden also Segmentierungsfehler als sehr störend empfunden.

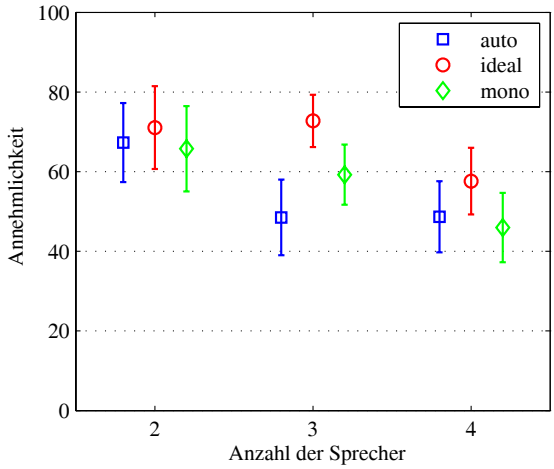

Abbildung 2: Wahrgenommene Annehmlichkeit auf einer Skala von 0 bis 100;

Zusammenfassung

Dieser Beitrag stellt ein System zur räumlichen Darbietung von Telekommunikationsszenarien mit mehreren Sprecher vor. Das System basiert auf automatischer Sprechersegmentierung und -gruppierung und binauraler Wiedergabe. Es ist rückwärtskompatibel zu bestehenden Übertragungstechniken wie dem klassischen Festnetz, da nur ein einziger Empfangskanal benötigt wird.
Erste Evaluierungsergebnisse zeigen, dass die Segmentierung der Sprecher mit anschließender räumlicher Darbietung die Identifikation der Sprecher erleichtert. Allerdings erscheinen Verbesserungen bzgl. der Zuverlässigkeit der automatischen Segmentierung und Gruppierung notwendig.

Literatur

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

[2] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, MIT Press, 1996.

[3] I. McCowan, J. Ajmera and H. Bouarlard: Robust speaker change detection. IEEE Signal Processing Letters, vol. 11, no. 8, pp. 649–652, 2004.

[4] J. Ajmera and C. Wooters: A robust speaker clustering algorithm. IEEE Automatic Speech Recognition and Understanding workshop (ASRU), pp. 357–366, 2004.

[5] U. Turk and F. Schiel: Speaker verification based on the German VeriDat database. Eurospeech, pp. 3025–3028, 2003.