Entwicklung einer Simulationsumgebung zur Bewertung automatischer Mikrofonmischer

Natanael Nieland

TU Kaiserslautern, EIT, Digitale Signalverarbeitung, Email: nieland@eit.uni-kl.de

Einleitung

Zum automatischen Mischen von Mikrofonsignalen existiert eine Vielzahl unterschiedlicher Algorithmen, was sich auch in der Vielfalt der heute verfügbaren Geräte zeigt. Dem Anwender stellt sich die Frage, welches Verfahren für seine Beschallungsanwendung, z. B. eine Konferenz oder eine Kirchenbeschallung, am besten geeignet ist. Die Beurteilung der Verfahren beruht häufig auf individuellen Erfahrungen. Eine Methode zur objektiven Bewertung der Verfahren hat sich bisher nicht etabliert. Im folgenden Beitrag wird deshalb ein Gütekriterium definiert, das die Auswirkung der vom Mikrofonmischer berechneten Abschwächungsfaktoren auf die Qualität des Mischsignals erfasst. Die zur Berechnung des Gütekriteriums erforderlichen Signale werden während des Betriebs eines Mischverfahrens gewonnen. Um den Betrieb unter realistischen Umgebungsbedingungen zu simulieren, wurde eine Simulationsumgebung entwickelt, die Einflüsse von Nachhall und Störquellen, sowie die Eigenschaften der Quellen und Mikrofone berücksichtigt.

Zunächst werden sieben Verfahren kurz vorgestellt. Anschließend wird das Gütekriterium hergeleitet und ein Überblick über die Simulationsumgebung und die verwendeten Testszenarien gegeben. Die Verfahren werden für 96 unterschiedliche Beschallungsszenarien bewertet und die Ergebnisse vorgestellt. Außerdem wird am Beispiel des *Gainsharingmischers* gezeigt, dass ein Verfahren durch Parameteroptimierung, die mithilfe der Simulationsumgebung effizient durchgeführt wird, verbessert werden kann. Die Simulationsumgebung ermöglicht es, den Einfluss von Umgebungsbedingungen auf das Verhalten der Verfahren zu analysieren. Beispielhaft wird dazu der Einfluss des Abstands zwischen Sprecher und Mikrofon auf das Verhalten der Verfahren untersucht.

Verfahren

Im Folgenden werden einige der heute verwendeten Verfahren vorgestellt.

Verfahren mit fester Schwelle: Beim Verfahren mit fester Schwelle (F) wird ein Kanal aktiviert, das heißt die Abschwächung beträgt 0 dB, wenn der Signalpegel eine voreingestellte Schwelle überschreitet. Inaktive Kanäle werden um einen Abschwächungsfaktor g_d, der typischerweise -15 dB entspricht, abgeschwächt. Der sogenannte NOM-Abschwächer schwächt das Mischsignal zusätzlich um 3 dB pro Verdopplung der Anzahl aktiver Kanäle ab. Ein einmal aktivierter Kanal wird für die Haltezeit T_H, die typischerweise 1 s beträgt, aktiv gehalten, um eine ständige Aktivierung und Deaktivierung zu verhindern.

Verfahren mit variabler Schwelle: Beim Gatingmischer mit variabler Schwelle wird ein Kanal aktiviert, wenn die Differenz des Mikrofonpegels und des Pegels eines Referenzsignals einen einstellbaren Schwellwert überschreitet [1]. Das Referenzsignal entspricht der Summe aller Mikrofonsignale (Verfahren S) oder dem Signal eines Raummikrofons (Verfahren R). Zusätzlich wird ein Halteglied und ein NOM-Abschwächer eingesetzt.

Mischer mit *Max Bus* und *NAT*: Das *Maxbusverfahren* (MB) [2] berechnet für alle Kanäle ein gefiltertes Gleichrichtersignal und eine *NAT* (Noise Adaptive Threshold), die den Pegel stationärer Störgeräusche annähert. Überschreitet das Gleichrichtersignal die *NAT* um 6 dB, ist die erste Bedingung zur Aktivierung eines Kanals erfüllt. Zentrales Prinzip dieses Verfahrens ist jedoch der sogenannte *Max Bus*. Das Gleichrichtersignal eines Kanals muss das Maximale unter allen sein, um den *Max Bus* zu treiben und damit die zweite Bedingung zur Aktivierung eines Kanals zu erfüllen. Bereits aktive Kanäle erhalten bei der Ermittlung des maximalen Gleichrichtersignals einen Vorteil von 6 dB. Eine verbesserte Version dieses Verfahrens (MBRI) verwendet den sogenannten *Reverb Inhibit Bus*, durch den die *NAT* bei aktiv sein eines Sprechers angehoben wird, um die Aktivierung weiterer Mikrofone durch Nachhall zu verhindern.

Gainsharing: Der *Gainsharingmischer* kann als Pendant zum Gatingmischer mit Summensignal angesehen werden und unterscheidet sich primär in den kontinuierlichen Abschwächungsfaktoren [3]. Mit dem Operator $p\{\cdot\}$ eines ballistischen Spitzenwertgleichrichters wird der zeitabhängige Abschwächungsfaktor eines Kanals k

$$g_k(n) = \frac{p\{x_k(n)^e\}}{p\{\sum_{i=1}^{N} x_i(n)^e\}}, e \in \mathbb{N} \tag{1}$$

aus dem Verhältnis des Spitzenwertes des Mikrofonsignals $x_k(n)$ des Kanals k zum Spitzenwert des Summensignals errechnet, wobei N die Anzahl der Kanäle ist. Die Attack-Zeit des Detektors beträgt typischerweise 4 ms und die Release-Zeit 1 s. Die NOM-Abschwächung ist dem Verfahren inhärent. Für den *Gainsharingmischer* (GS) gilt $e = 1$, während eine modifizierte Version des Verfahrens (MGS) die Mikrofonsignale mit dem Exponenten $e = 4$ potenziert, um eine kontrastreichere Verstärkungsverteilung zu erzielen.

Gütekriterium

Ziel der folgenden Betrachtung ist die Herleitung eines Gütekriteriums, das für Gating- und Gainsharingverfahren gleichermaßen geeignet ist. Zur Bewertung automatischer Mikrofonmischer kann das Mischsignal herangezogen werden und Maße für die Sprachqualität wie z.B. *PESQ* (Perceptual Evaluation of Speech Quality) berechnet werden [4]. Bei automatischen Mischern ist jedoch auch der Fall zu betrachten, dass mehrere Sprecher mehrere Mikrofone aktivieren. Die Berechnung eines Qualitätsmaßes aus dem Mischsignal wird dann dadurch erschwert, dass sich das Mischsignal aus Sprachsignalen mehrerer Sprecher zusammensetzt. Das optimale Verhalten eines Verfahrens lässt sich einfacher mit den Abschwächungsfaktoren beschreiben: Unbenutzte Kanäle sollen vollständig abgeschwächt werden, während benutzte Kanäle lediglich eine Abschwächung durch die NOM-Abschwächung erhalten sollen. Die Bewertung erfasst dann das Vermögen eines Verfahrens unter den gegebenen Umgebungsbedingungen die gewünschten Abschwächungsfaktoren zu berechnen. Um eine Bewertungsfunktion zu ermitteln, wird die Beschallungssituation in Abbildung 1 betrachtet. Für den Hörer ist sowohl das SNR des Mischsignals als auch der Pegel des Mischsignals für die Sprachverständlichkeit von Bedeutung, weshalb hier nicht ausschließlich das SNR des Mischsignals, sondern das SNR am Hörerort betrachtet wird.

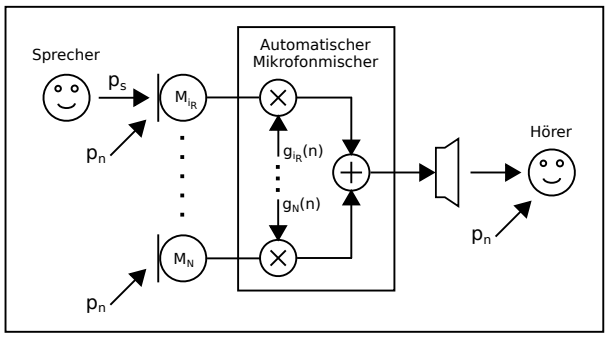

Abbildung 1: Betrachtung einer Beschallungssituation zur Herleitung eines Gütekriteriums

Es wird angenommen, dass im gesamten Raum der gleiche Störschall mit dem Effektivwert p_n herrscht. Ein Sprecher ruft zusätzlich einen Nutzschall mit dem Effektivwert p_s am Mikrofon M_{i_R} des Kanals i_R hervor. Die durch den automatischen Mischer verursachten zeitabhängigen Kanalabschwächungen seien $g_i(n), i = 1, ..., N$, wobei N die Anzahl der Mikrofonkanäle ist. Der Nutzanteil des Mischsignals beträgt $g_{i_R}(n)p_s$ und der Störanteil $\sqrt{\sum_{i=1}^{N} g_i(n)^2}\, p_n$. Es wird nun der Hörer betrachtet, dem das Mischsignal zugespielt wird und der zusätzlich den Störanteil p_n empfängt. Es wird angenommen, dass die Anlage so eingestellt ist, dass der Schallpegel am Ort des Mikrofons M_{i_R} einen gleichgroßen Schallpegel am Ort des Hörers hervorruft, wenn der entsprechende Kanal nicht abgeschwächt wird. Der Hörer befindet sich dann "virtuell" am Ort des Mikrofons. Am Hörerort beträgt das zeitabhängige SNR:

$$snr_{i_R}(n) = \frac{p_s}{p_n}\sqrt{\frac{g_{i_R}(n)^2}{1+\sum_{i=1}^{N} g_i(n)^2}}. \tag{2}$$

Sind A Sprecher aktiv und erhalten Kanäle, denen ein aktiver Sprecher zugeordnet ist, lediglich eine gewünschte NOM-Abschwächung von $1/\sqrt{A}$ und werden weitere Kanäle vollständig abgeschwächt, ergibt sich das im Optimalfall erzielte SNR:

$$snr_{opt} = \frac{p_s}{p_n}\sqrt{\frac{(\frac{1}{\sqrt{A}})^2}{1+A(\frac{1}{\sqrt{A}})^2}} = \frac{p_s}{p_n}\sqrt{\frac{1}{2A}}. \tag{3}$$

Das Verhältnis von $snr_{i_R}(n)$ zu snr_{opt}

$$D^A_{i_R}(n) = \sqrt{\frac{2Ag_{i_R}(n)^2}{1+\sum_{i=1}^{N} g_i(n)^2}} \tag{4}$$

nimmt im Optimalfall den Wert eins an. Der automatische Mikrofonmischer liefert dann das Mischsignal mit bestmöglichem SNR. Ohne den Einsatz eines automatischen Mikrofonmischers müssten die Abschwächungsfaktoren zur Beibehaltung der Gesamtverstärkung in allen Kanälen $1/\sqrt{N}$ betragen. In diesem Fall ergibt sich der Wert $D^A_{i_R}(n) = \sqrt{A/N}$. Ein Verfahren sollte Werte erzielen, die nahe bei eins liegen, auf jeden Fall aber größer als $\sqrt{A/N}$ sind.

Das Verhalten eines Verfahrens ist von der Anzahl aktiver Sprecher abhängig. Deshalb soll das Gütekriterium sowohl für den Fall, dass genau ein Sprecher aktiv ist, als auch für die Fälle, dass zwei oder mehr Sprecher aktiv sind, bestimmt werden. Dazu wird das logarithmierte Mittel

$$\overline{D^A_{i_R}} = 20\log_{10}\left(\frac{1}{|T_A|}\sum_{n\in T_A}\min(D^A_{i_R}(n),1)\right) \tag{5}$$

von $D^A_{i_R}(n)$ über die Menge der Abtastzeitpunkte T_A, zu denen genau A Sprecher aktiv sind, berechnet. Erzielt ein Kanal Werte $D^A_{i_R}(n) > 1$, geschieht dies stets auf Kosten eines anderen Kanals, beispielsweise wenn ein zweiter Kanal fälschlicherweise nicht aktiv ist. In Gleichung (5) werden deshalb nur Werte kleiner eins berücksichtigt.

Das Gütekriterium $\overline{D^A_{i_R}}$ gibt den mittleren Abstand zu dem im Optimalfall erzielten SNR für den Fall, dass A Sprecher aktiv sind, an. Insbesondere die Fälle $A = 1$ und $A = 2$ sind interessant, da in praktischen Fällen nur in Ausnahmesituationen mehr als zwei Sprecher gleichzeitig verstanden werden müssen. Da ein lauter Sprecher für schlechte Ergebnisse in einem anderen Kanal sorgen kann, sollten in den Fällen $A > 1$ die Werte aller Kanäle betrachtet werden.

Simulationsumgebung

Zur Berechnung des hergeleiteten Gütekriteriums werden die Abschwächungsfaktoren und die Anzahl aktiver Sprecher zu allen Abtastzeitpunkten benötigt. Zur Beschaffung dieser Signale wurde eine Simulationsumgebung entwickelt, die in Abbildung 2 schematisch dargestellt ist. Die Abschwächungsfaktoren werden vom zu untersuchenden Mischverfahren aus den Eingangssignalen berechnet. Diese weisen zwei für die Mischverfahren relevante Eigenschaften auf. Erstens handelt es sich um Sprachsignale, die eine transiente Charakteristik und für Sprache typische spektrale Eigenschaften aufweisen, zweitens sind die Eingangssignale nicht voneinander unabhängig, da eine Schallquelle stets mehrere Mikrofone beschallt. Um diese Eigenschaften zu berücksichtigen, werden Sprachaufnahmen verwendet und die Übertragungsstrecken zwischen allen Quellen-Mikrofon-Pärchen berücksichtigt.

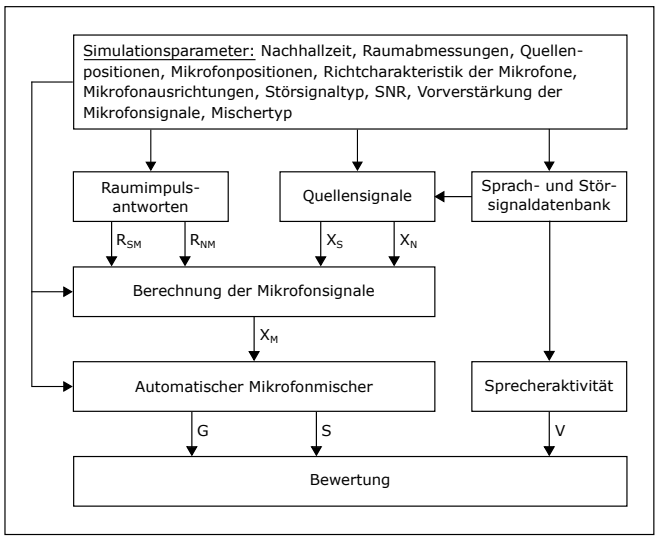

Abbildung 2: Ablaufdiagramm der Simulation

Zunächst werden die Raumabmessungen, die Nachhallzeit, die Positionen der Schallquellen und die Positionen, Ausrichtungen und Richtcharakteristiken der Mikrofone festgelegt. Für die Schallquellen werden kugelförmige Richtcharakteristiken angenommen. Die Raumimpulsantworten zwischen allen Quellen-Mikrofon-Pärchen werden mit der Spiegelquellenmethode [5] ermittelt. Die Impulsantworten zwischen Sprechern und Mikrofonen werden in der Matrix R_{SM} zusammengefasst, die Impulsantworten zwischen Störquellen und Mikrofonen in der Matrix R_{NM}. Es werden Aufnahmen aus der Sprachsignaldatenbank TSP Speech [6] verwendet, die in einem reflexionsfreien Raum gewonnene Aufnahmen einzelner gesprochener Sätze von 24 Sprechern in englischer Sprache enthält. Für jeden zu simulierenden Sprecher werden einzelne Aufnahmen bei Bedarf mit Pausen aneinandergereiht und in der Matrix X_S zusammengefasst. Die Sprecheraktivität wird in der Matrix V zusammengefasst. Zur Simulation der Störquellen werden weißes Rauschen und Aufnahmen von Hintergrundgesprächen, einer Lüftungsanlage und „Papierkritzeln“ aus der Datenbank MUSAN [7] verwendet.

Zur Berechnung der Mikrofonsignale werden die Quellensignale mit den Impulsantworten gefaltet und überlagert. Anschließend werden die Mikrofonsignale mit Verstärkungsfaktoren gewichtet, die sich aus der Sensitivität der Mikrofone und den Verstärkungen der Mikrofonvorverstärker ergeben. Die resultierenden Signale werden in der Matrix X_M zusammengefasst und dienen dem Mischverfahren als Eingangssignale. Das zu untersuchende Verfahren berechnet aus den Mikrofonsignalen die Abschwächungsfaktoren G und das Mischsignal S. Aus den Abschwächungsfaktoren und der Aktivität der Sprecher wird das Gütekriterium berechnet. Das Mischsignal S kann zusätzlich zur subjektiven Bewertung eines Verfahrens herangezogen werden.

Ergebnisse

Das Gütekriterium wird für 96 Testszenarien ermittelt. Dabei werden die Raumgröße, die Nachhallzeit, der Abstand zwischen Sprechern und Mikrofonen, die Störquellensignale und der Signalrauschabstand variiert. Es werden 8 Sprachmikrofone, ein Raummikrofon, zwei Sprecher und eine Störquelle positioniert. Für eine erste Sequenz, in der eine dem Kanal $i = 1$ zugeordnete Schallquelle zwanzig einzelne gesprochene Sätze eines Sprechers ausgibt, wird der Wert $\overline{D_1^1}$ ermittelt. In einer zweiten Sequenz gibt eine zweite Quelle, die dem Kanal $i = 2$ zugeordnet ist, zusätzlich dauerhaft gesprochene Sätze eines zweiten Sprechers aus. Die Werte $\overline{D_1^2}$ und $\overline{D_2^2}$ beider Kanäle werden im Wert

$$\overline{D_{1,2}^2} = 20 \log_{10} \left(\frac{10^{\overline{D_1^2}/20} + 10^{\overline{D_2^2}/20}}{2} \right) \qquad (6)$$

zusammengefasst. Für Verfahren, bei denen eine Schwelle einstellbar ist (F, R und S), werden in einem Vorlauf die Schwellwerte ermittelt, die im Mittel zum höchsten Wert $\overline{D_1^1}$ führen. Diese werden anschließend in der Simulation verwendet.

Mithilfe der Simulationsumgebung und des Gütekriteriums lassen sich die Parameter eines Verfahrens optimieren. Aus dem Verfahren MGS wurde das Verfahren MGS-OPT gewonnen, indem der Exponent e und die Release-Zeit τ_R des Spitzenwertgleichrichters des *Gainsharingmischers* bezüglich der Kostenfunktion

$$K(e, \tau_R) = 10^{\overline{D_1^1}/20} + (10^{\overline{D_1^2}/20} + 10^{\overline{D_2^2}/20})/2 \qquad (7)$$

optimiert wurden. Dazu wurde die Simulation für mehrere Variationen dieser Parameter durchgeführt und das Parameterpärchen ermittelt, für das die Kostenfunktion im Mittel über alle Testszenarien den höchsten Wert liefert ($\tau_{R_{opt}} = 4$ s, $e_{opt} = 3,3$). Um auch Dezimalzahlen als Exponenten zu erlauben, wurde die Potenzierung in Gleichung (1) auf die Beträge der Mikrofonsignale angewandt.

Abbildung 3 zeigt die Ergebnisse der Simulation, aus denen sich folgende Schlussfolgerungen ziehen lassen: Alle Verfahren erzielen im Mittel bessere Ergebnisse, als sie sich ohne Verwendung eines automatischen Mischers (grüne Linie) ergeben würden. Die Verfahren unterscheiden sich deutlich in der Streuung der Ergebnisse. Eine

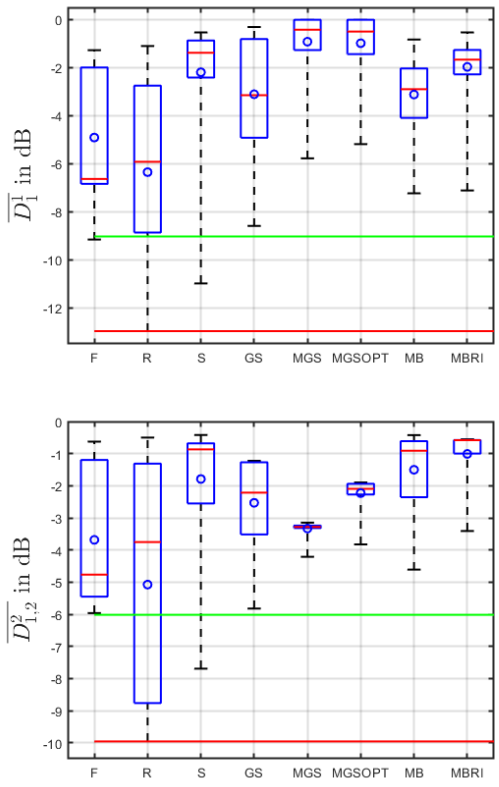

Abbildung 3: Oben: Werte $\overline{D_1^1}$, unten: Werte $\overline{D_{1,2}^2}$. Die Box zeigt den Bereich vom unteren zum oberen Quartil an und die schwarzen Linien die Extremwerte. Die roten Linien markieren den Medianwert, der Kreis den Mittelwert. Die grüne Linie zeigt das Ergebnis, das sich ohne Verwendung eines Mikrofonmischers ergibt, die rote Linie das Ergebnis, wenn bei Verwendung eines Mischers alle Kanäle deaktiviert sind.

geringe Streuung kann als Robustheit eines Verfahrens gegenüber Änderungen der Umgebungsbedingungen aufgefasst werden. Die Verfahren F und R weisen eine geringe Robustheit auf, was im Mittel zu schlechten Ergebnissen führt. Beim Verfahren R wirkt sich zudem die in der Simulation ungünstig gewählte Positionierung des Raummikrofons nahe der Störquelle negativ aus. Das Verfahren S hingegen erzielt bessere Ergebnisse bei geringerer Streuung. Beim Verfahren MGS wird das bessere Ergebnis gegenüber dem Verfahren GS bei einem aktiven Sprecher durch schlechtere Ergebnisse bei zwei aktiven Sprechern erkauft. Das Verfahren MGS konnte durch die Optimierung der Release-Zeit und des Exponenten deutlich verbessert werden. MGSOPT erzielt höhere Mittelwerte als das Verfahren GS und weist dabei eine geringere Streuung der Ergebnisse auf. Gegenüber dem Verfahren MB erzielt das Verfahren MBRI bessere Ergebnisse bei einer etwas geringeren Streuung. Von den Gainsharingverfahren liefert das Verfahren MGSOPT die besten Ergebnisse, von den Gatingverfahren das Verfahren MBRI.

Abbildung 4 zeigt für die Verfahren F, S, GS, MGSOPT und MBRI die Werte $\overline{D_1^1}$, die sich aus einer Simulation ergeben, bei der der Abstand zwischen Sprecher und Mikrofon zwischen 0,2 m und 2,5 m variiert wurde. Alle Verfahren, insbesondere das Verfahren F, erzielen mit

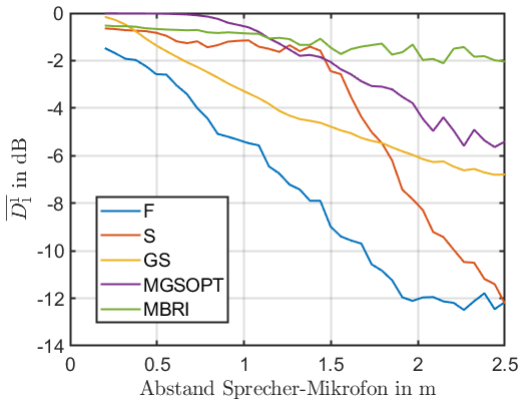

Abbildung 4: Einfluss des Abstands Sprecher-Mikrofon auf das Verhalten der Verfahren

zunehmenden Abstand schlechtere Ergebnisse. Das Verfahren MBRI erzielt jedoch auch bei großen Abständen noch gute Ergebnisse. Gegenüber dem Verfahren GS verschlechtern sich die Ergebnisse des Verfahrens MGSOPT erst ab einem etwas größeren Abstand ($\approx 0{,}75$ m), da die durch einen höheren Abstand bedingten, kleineren Pegeldifferenzen der Signale, durch die Potenzierung der Eingangssignale beim Verfahren MGSOPT, vergrößert werden.

Fazit

Die entwickelte Simulationsumgebung und das Gütekriterium ermöglichen die detailreiche Analyse automatischer Mikrofonmischer. Verfahren können objektiv bewertet und verglichen werden. Einflüsse der Umgebungsbedingungen können untersucht werden. Am Beispiel des Gainsharingverfahrens wurde gezeigt, dass die Parameter eines Verfahrens mithilfe der Simulationsumgebung und des Gütekriteriums effizient optimiert werden können. Ein informeller Hörtest zeigte eine grundsätzliche Übereinstimmung der Werte des Gütekriteriums mit dem subjektiven Qualitätseindruck des Mischsignals.

Literatur

[1] Dugan, D.: Control Apparatus for Sound Reinforcement Systems, US Patent 3,814,856, 1974

[2] Julstrom, S.: Microphone Actuation Control System, US Patent 5,297,210, 1994

[3] Dugan, D.: Automatic Microphone Mixer, US Patent 3,992,584, 1976

[4] Johansson, D.: Automatic Microphone Mixing for a Daisy Chain Connected Multi-Microphone Speakerphone Setup, Umeå University, 2016

[5] Allen, J.B. und Berkley, D.A.: Image Method for Efficiently Simulating Small-Room Acoustics, Journal Acoustic Society of America, 65(4), April 1979, p 943

[6] Kabal, P.: TSP Speech Database, McGill University, 2002

[7] Snyder, D., Chen, G., Povey, D.: Musan: A Music, Speech, and Noise Corpus, arXiv:1510.08484, 2015