Schätzung der idealen binären Maske mittels Bayes'scher Klassifikation unter Einfluss von Störgeräusch und Nachhall

Christoph Kowalski, Tobias May und Steven van de Par
Carl von Ossietzky Universität Oldenburg, Institut für Physik - Akustik, Email: christoph.kowalski@uni-oldenburg.de

Einleitung

Während Normalhörende die erstaunliche Fähigkeit besitzen Sprachinformationen aus einem von Störeinflüssen veränderten Signal zu nutzen, stellt dies für computergestützte Systeme eine Herausforderung dar. Wird ein Sprachsignal, das teilweise durch ein Störgeräusch maskiert ist, z.B. zur automatischen Sprechererkennung verwendet, so hat dies einen negativen Einfluss auf die Erkennungsraten zur Folge. Eine zusätzliche Verhallung des Signals bewirkt aufgrund der daraus resultierenden zeitlichen Verschmierung eine weitere Verschlechterung. Eine Möglichkeit mit diesem Problem umzugehen und somit die Erkennungsraten zu verbessern bietet der s.g. *Missing Data* Ansatz, hierfür muss jedoch das gestörte Signal in verlässliche und nicht-verlässliche Zeit-Frequenz Punkte aufgeteilt werden [1].

Ideale binäre Maske

Die Aufteilung einer Zeit-Frequenz Repräsentation zwischen verlässlichen und nicht-verlässlichen Punkten erfolgt durch die ideale binäre Maske (IBM). Eine Möglichkeit diese zu erhalten besteht darin, jeden Punkt einer Zeit-Frequenz-Darstellung S_{sp} eines Sprachsignals, welcher einen höheren Pegel aufweist als der zugehörige Zeit-Frequenz-Punkt des Störgeräusches S_n, mit 1 zu markieren, andernfalls mit 0. Dies bedeutet ein lokales SNR-Kriterium von 0 dB. Als Repräsentation werden im Folgenden s.g. Ratemaps verwendet (z.B. nach [4]).

$$IBM(i,j) = \begin{cases} 1, & \text{wenn } S_{sp}(i,j) > S_n(i,j) \\ 0, & \text{sonst} \end{cases} \quad (1)$$

Abbildung 1 zeigt beispielhaft eine Ratemap und die daraus resultierende IBM. Um die IBM 1 zu erhalten, wird jedoch vorausgesetzt, dass das Sprach bzw. das Störsignal bekannt sind. Da dies in der realen Anwendung jedoch selten gegeben ist, muss eine Schätzung der IBM erfolgen. Hierfür kann bspw. das *a priori* bekannte Störgeräusch S_n durch eine Schätzung des Störgeräusches ersetzt werden. Ein weiterer Ansatz, der in dieser Arbeit verfolgt wird, besteht darin die IBM unter Verwendung eines Klassifikationssystems zu schätzen. Dies hat den Vorteil, dass das lokale SNR-Kriterium nicht empirisch festgelegt werden muss, sondern anhand von verschiedenen Merkmalen statistisch modelliert wird. Desweiteren kann das im Folgenden vorgestellte Klassifikationssystem beliebig durch weitere Merkmale erweitert werden.

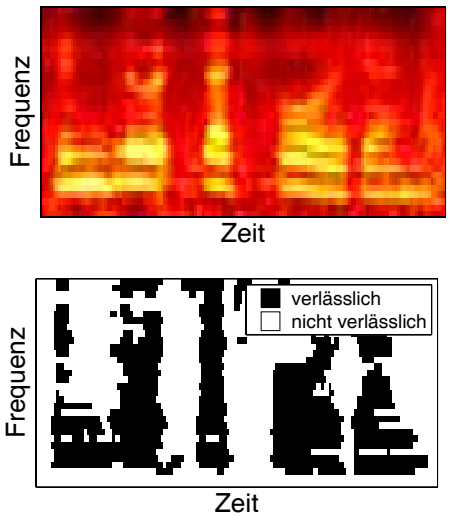

Abbildung 1: Darstellung einer Ratemap und der dazugehörigen IBM eines verrauschten Sprachsignals.

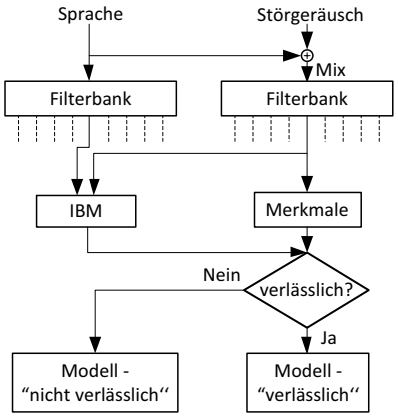

Abbildung 2: Trainingsphase des entwickelten Klassifikationssystems zur Schätzung der IBM.

Klassifikationssystem

Zu Beginn der Trainingsphase (Abb. 2) werden die Signale mit dem gewünschten SNR erstellt und anschließend über eine 32-kanälige Filterbank in Bänder aufgeteilt. Da während der Trainingsphase *a priori* Wissen über die Leistung von Sprache und Störgeräusch vorhanden ist, kann die IBM direkt berechnet werden. Parallel dazu werden bestimmte Merkmale aus dem verrauschten Signal extrahiert. Mittels der IBM kann anschließend für jeden Kanal der Filterbank je ein Modell für die verlässlichen und die nicht-verlässlichen Merkmale trainiert werden. Als Merkmal wird im Folgenden ein Störgeräuschschätzer nach [2] verwendet, welcher zur Schätzung der IBM in [3] angepasst wurde und somit im Vergleich zu anderen SNR-Schätzern gute Ergebnisse erzielt hat. Da ein Sprachsi-

gnal mehrere Frequenzbereiche anregt, ist es naheliegend, dass kanalübergreifende Informationen zu einer Verbesserung der Maskenschätzung beitragen können. Um diese Informationen zu nutzen, wurde die Möglichkeit integriert, dass als Merkmale für den aktuellen Kanal zusätzlich die aus den benachbarten Kanälen extrahierten Merkmale verwendet werden können. In der Testphase werden die so erhaltenen Modelle verwendet, um eine Klassifikation anhand von Merkmalen aus einem gemischten Sprachsignalen, ohne *a priori* Wissen, durchzuführen. Dabei wird jeder Zeit-Frequenz Punkt entweder der verlässlichen oder nicht-verlässlichen Klasse zugeordnet um somit die geschätzte IBM zu erhalten. Um mit der zeitlichen Verschmierung durch Nachhall umzugehen, kann die geschätzte IBM mit einer Onset-Maske [4] kombiniert werden. Da die Onsets den Direktschallanteil des Signal widerspiegeln, soll dies dazu führen, dass die durch Nachhall beeinflussten Anteile der IBM verringert werden.

Ergebnisse

Die Evaluation der verschiedenen Verfahren zur Maskenschätzung erfolgt anhand von Sprechererkennungsraten (Methode ähnlich [3]), wobei die Sprachsignale bei verschiedenen SNRs mit einem Fabrikgeräusch aus der NOISEX-92 Datenbank [5] gemischt wurden. Dabei war es das Ziel zwischen zehn verschiedenen Sprechern zu unterscheiden und dabei möglichst hohe Erkennungsraten zu erreichen. Getestet wurde sowohl unverhallt, als auch mit $T_{60} = 1.2$ s verhallten Signalen. Wird bei unverhallten Signalen die *IBM* verwendet um

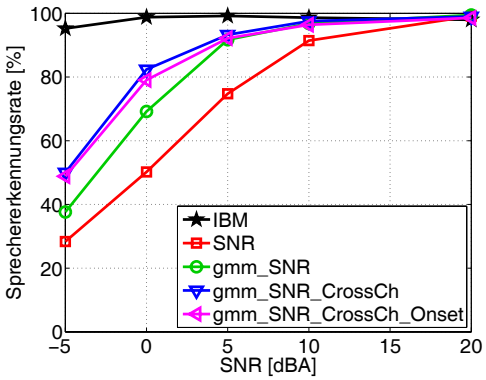

Abbildung 3: Sprechererkennungsraten in Abhängigkeit von Hindergrundrauschen.

die Sprechererkennung durchzuführen (Abb. 3), so werden über alle SNRs durchweg sehr hohe Erkennungsraten erzielt. Verwendet man hingegen jedoch den SNR-Schätzer (*SNR*), so sinken die Erkennungsraten zu niedrigen SNRs hin stark ab. Durch die Verwendung des Klassifikationssystems zur Maskenschätzung, mit dem SNR-Schätzer als Merkmal (*gmm_SNR*), lässt sich eine signifikante Steigerung der Erkennungsraten erreichen. Werden zudem die benachbarten Kanäle ebenfalls als Merkmale genutzt (*gmm_SNR_CrossCh*), so lässt sich eine weitere Steigerung beobachten. Eine Onset-Selektion (*gmm_SNR_CrossCh_Onset*) führt zu einem leichten Abfall der Erkennungsraten über alle SNRs. Grund hierfür ist, dass durch die Selektion eines unverhallten Signals mittels Onset-Maske, Zeit-Frequenz Punkte verworfen werden, welche nicht durch Nachhall beeinflusst wurden. Werden die Signale zusätzlich verhallt, so hat dies einen

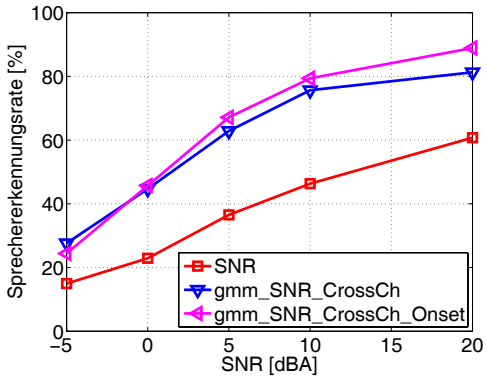

Abbildung 4: Sprecherkennungsraten von verhallten und verrauschten Signalen.

zusätzlich negativen Einfluss auf die Erkennungsraten (Abb. 4). Auch hier erreicht das Klassifikationssystem mit dem SNR-Schätzer und kanalübergreifenden Merkmalen deutlich höhere Erkennungsraten gegenüber der Methode ohne Klassifikationssystem. Durch die Onset-Selektion bei verhallten Signalen steigt hier die Sprechererkennungsrate zu hohen SNRs jedoch an.

Fazit und Ausblick

Die Schätzung der IBM mittels Klassifikationssystem zeigt einen deutlichen Zuwachs gegenüber der Methode ohne Klassifikationssystem. Auch die Onset-Selektion zeigt eine Verbesserung bei verhallten Signalen, wobei bei unverhallten Signalen nur eine leichte Verringerung der Erkennungsraten erfolgt. Durch eine Kombination mit weiteren Merkmalen und eine Optimierung verschiedener Parameter könnte eine zusätzliche Verbesserung der Masken erfolgen und somit eine weitere Annäherung an die Erkennungsraten, welche mit der IBM erreicht werden, erfolgen.

Literatur

[1] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001), „Robust automatic speech recognition with missing and unreliable data“, *Speech Commun.* 34, pp. 267-285.

[2] Lin, L., Holmes, W., and Ambihirajah, E. (2003), „Subband noise estimation for speech enhancement using a perceptual wiener filter“, *Proc. of ICASSP*, pp. 80–83.

[3] May, T., van de Par, S., and Kohlrausch, A. (2012), „Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling“, *IEEE Trans. Speech Audio Process.* 20, pp. 108-121.

[4] Palomäki, K., Brown, G., and Barker, J. (2006), „Recognition of reverberant speech using full cepstral features and spectral missing data“, *Proc. of ICASSP*, pp. 289–292.

[5] Varga, A. and Steeneken, H. (1993), „Assessment for automatic speech recognition: ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,“ *Speech Commun.* 12, pp. 247–251.