SPRACHTECHNOLOGIEN FÜR DIGITALE ASSISTENTEN

Reinhold Haeb-Umbach

Fachgebiet Nachrichtentechnik, Universität Paderborn
haeb@nt.uni-paderborn.de

Kurzfassung: Nachdem digitale Assistenten, die sich über Sprache bedienen lassen, zunächst auf dem Smartphone ihren Siegeszug angetreten haben, sind sie mittlerweile auch in vielen Wohnzimmern allgegenwärtig. Die großen Internetfirmen bieten solche digitale Heimassistenten, auch "intelligente Lautsprecher" genannt, an, die berührungslos aus der Ferne per Sprache zu bedienen sind. Dabei ist der Übergang von Nahbesprechung auf eine Sprachbedienung aus der Ferne alles andere als ein gradueller Unterschied. In dem Vortrag werden die Herausforderungen und Lösungsansätze für Realisierung einer Sprachbedienung aus der Ferne dargestellt. Dabei zeigt sich, dass eine geschickte Kombination von Signalverarbeitung und tiefen neuronalen Netzen zu sehr effektiven Lösungen führt.

1 Einleitung

Digitale Heimassistenten, von denen wohl Amazons "Echo" das bekannteste ist, haben in den letzten Jahren einen ungeahnten kommerziellen Erfolg erfahren. Anfänglich unterstützten die Assistenten nur eine geringe Anzahl von Anwendungen, wie etwa das Abspielen von Musik oder das Führen von Einkaufslisten oder Terminkalendern. Mittlerweile gibt es jedoch eine Vielzahl sogenannter "Skills" von Drittanbietern, die das Gerät als Bedienschnittstelle für die unterschiedlichsten Anwendungen verwenden, wie beispielsweise für die Heimautomatisierung.

Dieser Erfolg war nur möglich, weil die Sprachbedienung so zuverlässig funktioniert. Dass dies wiederum so ist, hat entscheidend mit den Erfolgen tiefer neuronaler Netze für die Sprachverarbeitung zu tun. Insbesondere die Erkennungsgenauigkeit der automatischen Spracherkennung bei verrauschten und verhallten Sprachaufnahmen, wie sie für Anwendungen im Bereich der Heimassistenten typisch sind, hat sich in den letzten Jahren drastisch verbessert. Dabei zeigte sich, dass eine explizite Signalbereinigung, die dem eigentlichen Spracherkenner vorgeschaltet ist, die Spracherkennungsleistung deutlich steigern kann. Dies gilt insbesondere für Störungen, die für den Spracherkenner schwer zu beherrschen sind. Dazu gehört zum einen der Raumhall, der einen Laut über mehrere Analysefenster der Kurzzeit-Fouriertransformation in der Merkmalsextraktion des Erkenners verschmiert und somit schwierig mit dem akustischen Modell des Erkenners zu modellieren ist. Auch ist es für den Erkenner schwer zu entscheiden, welches Sprachsignal zu dekodieren ist, wenn mehrere Sprecher gleichzeitig aktiv sind. Auch hier ist eine vorgeschaltete Signalbereinigung sehr hilfreich.

Wenn mehrkanalige Signalaufnahmen vorhanden sind, d.h. wenn statt eines einzelnen Mikrofons eine Mikrofongruppe verwendet wird, kann die Signalentstörung besonders effektiv sein. Da Nutzsignal und Störung häufig aus unterschiedlichen Raumrichtungen auf die Mikrofongruppe eintreffen, kann sich eine akustische Strahlformung auf das Nutzsignal ausrichten und so den Störschall unterdrücken, wie in diesem Beitrag noch weiter ausgeführt wird.

Die Forschungsanstrengungen in Wissenschaft und Industrie zur robusten Erkennung wurden ergänzt durch durch wissenschaftliche Wettbewerbe zur Spracherkennung bei ungünstigen

akustischen Umgebungsbedingungen. Hier sind vor allem die "REVERB Challenge" zur Erkennung verhallter Sprache [1], und die Wettbewerbe CHiME-3 [2], CHiME-4 [3], und CHiME-5 [4] zu nennen (CHiME: Computational Hearing in Multisource Environments). Durch Definition von Referenzdatenbasen konnten Vergleichstests durchgeführt werden, die Aufschluss darüber gaben, welche Verfahren besonders effektiv sind, um in akustisch schwierigen Umgebungen eine gute Erkennungsgenauigkeit zu erzielen.

Dieser Beitrag ist wie folgt gegliedert. Im Kapitel 2 beschreiben wir zunächst die typischen Signalbeeinträchtigungen, die sich bei einem großen Abstand zwischen Sprecher und Mikrofon ergeben, um dann im nächsten Kapitel 3 die Sprachverarbeitungskette vorzustellen, wie sie in einem digitalen Heimassistenten realisiert ist, und Kapitel 4 liefert eine kurze Zusammenfassung.

2 Signalbeeinträchtigungen bei großem Abstand zwischen Sprecher und Mikrofon

In einem Anwendungsfall beträgt der Abstand zwischen dem Sprecher und dem intelligenten Lautsprecher mehrere Meter. Dieser große Abstand hat signifikanten Einfluss auf die Qualität des aufgenommenen Signals. Im Vergleich zu einer Nahbesprechungssituation, wie sie etwa bei der Bedienung eines Smartphones vorliegt, ergeben sich folgende Arten von Beeinträchtigungen:

- Das Signal erfährt eine Dämpfung aufgrund der Ausbreitung im Raum. Im Freiraum verringert sich die Signalleistung proportional zu dem Quadrat des Abstandes zwischen Signalquelle und Sensor. Das bedeutet, dass sich bei einem isotrop abstrahlenden Sender die Signalleistung bei einer Erhöhung des Abstandes zum Sensor von 2 cm auf 1 m um 34 dB reduziert. In der Realität ist die Dämpfung deutlich geringer, da der Mund eine gewisse Strahlformung bewirkt. Dennoch erfährt das Sprachsignal eine Reduktion der Leistung, die in der Regel mit einer Reduktion des Signal-zu-Rauschleistungsverhältnisses einhergeht.

- Weiterhin führt die Signalausbreitung in Räumen zu einem Halleffekt. Damit bezeichnet man die Mehrwegeausbreitung, denn das Sendesignal wird von Wänden und Gegenständen reflektiert und gelangt über unterschiedliche Ausbreitungspfade mit jeweils unterschiedlichen Laufzeiten und Dämpfungen zum Mikrofon. Dies führt dazu, dass die akustische Impulsantwort von der Quelle zum Sensor keineswegs nur aus einem zeitversetzten Impuls besteht. Die Impulsantwort ähnelt vielmehr einem weißen Rauschprozess mit einer exponentiell abklingenden Einhüllenden. Insbesondere der Nachhall, d.h. Mehrfachreflexionen, deren Laufzeiten um mehr als 50 ms größer sind als die der Signalkomponente, die auf direktem Weg vom Sprecher zum Mikrofon gelangt, führen zu Problemen in der Spracherkennung.

- Bei einem entfernten Mikrofon ist es wahrscheinlich, dass das Mikrofon neben dem Nutzsignal auch andere akustische Signale aufnimmt, wie etwa das des Fernsehers oder auch Gespräche anderer Personen im Raum.

- Schließlich haben digitale Heimassistenten eine Audioausgabe zum Abspielen von Musik oder Sprachantworten. Das Lautsprechersignal wird ebenfalls vom Mikrofon des Geräts aufgenommen und hat häufig eine deutlich höhere Leistung als die des entfernten Sprechers. Dieses akustische Echo kann die Spracherkennung völlig unmöglich machen.

Dies alles führt dazu, dass zuverlässige Spracherkennung bei entfernten Mikrofonen sehr viel schwieriger ist als bei einer Aufnahme mit einem Nahbesprechungsmikrofon. Dies hat Auswirkungen darauf, wie die Signalverarbeitungskette aufgebaut ist, wie im nächsten Kapitel erläutert wird.

3 Systemüberblick

Abb. 1 zeigt das Blockschaltbild einer typischen Sprachverarbeitungskette in einem digitalen Heimassistenten. Das Gerät hat einen oder mehrere Lautsprecherausgangskanäle zur Sprachausgabe und zum Abspielen von Musik. Auf der Eingabeseite gibt es eine Mikrofongruppe mit typischerweise 2 – 8 Mikrofonen. Die Signalverarbeitung erfolgt in der Regel im Frequenzbereich, d.h. die Mikrofonsignale werden einer Kurzzeit-Fouriertransformation unterzogen, bei der auf sich überlappenden gefensterten Signalabschnitten jeweils eine diskrete Fouriertransformation ausgeführt wird. Im Folgenden werden die anschließenden Signalverarbeitungsblöcke kurz beschrieben. Für eine etwas ausführlichere Darstellung sei auf [5] verwiesen.

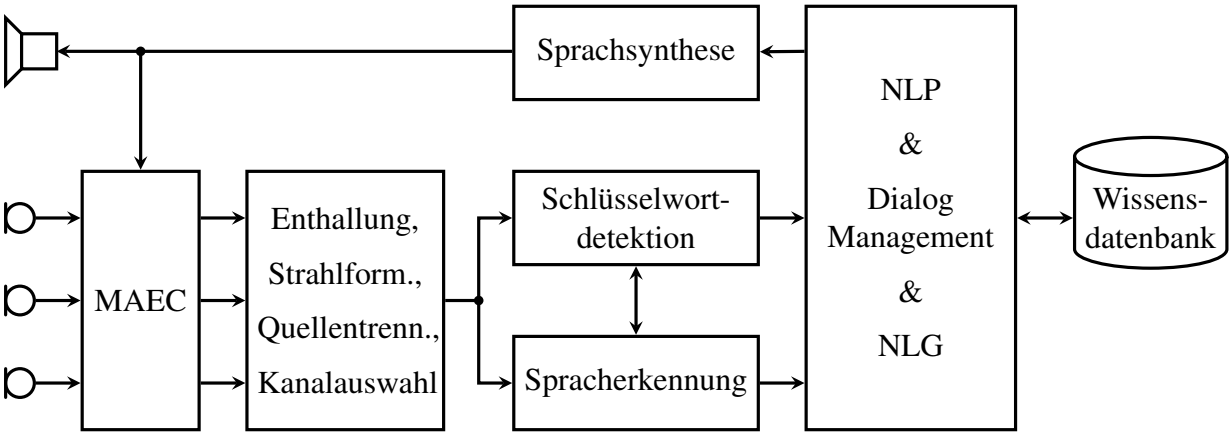

Abbildung 1 – Überblick über die Sprachverarbeitungsblöcke in einem typischen digitalen Heimassistenten (angepasst von [5]).

3.1 Mehrkanalige Echokompensation

Da der Lautsprecher in unmittelbarer Nähe der Mikrofone angeordnet ist, wird das Lautsprechersignal unerwünschterweise in die Mikrofone eingekoppelt. Dabei kann das Lautsprechersignal um mehrere 10 dB stärker sein als das eigentlich interessierende Sprachkommando des Sprechers. Es muss daher unterdrückt werden, damit das Sprachkommando erkennbar wird. Mehrkanalige akustische Echokompensation (“multi-channel acoustic echo cancellation” (MAEC)) ist eine etablierte Technik [6]. Die Filter schätzen den akustischen Übertragungsweg vom Lautsprecher zu den Mikrofonen, um dann das Lautsprechersignals aus dem Mikrofonsignal herauszurechnen. Lineare adaptive Filter erreichen in der Regel nur eine Echounterdrückung in der Größenordnung von 10 bis 20 dB, die jedoch für die Anwendung nicht ausreichend ist. Noch größere Unterdrückung ist nur mit sehr langen Filtern möglich, die aber nicht eingesetzt werden können, da sonst die Adaptionszeit zu lang wird. Außerdem können lineare Filter nicht die nichtlinearen Komponenten im Echosignal entfernen, die etwa durch Vibrationen oder Nichtlinearitäten im Lautsprecher hervorgerufen werden.

Um das Restecho zu unterdrücken, wird ein neuronales Netz eingesetzt: Das Netz wird so trainiert, dass es für jeden Zeitfrequenzpunkt eine Sprachpräsenswahrscheinlichkeit (SPW) schätzt. Von dieser SPW kann eine Maske berechnet werden, mit dessen Hilfe die Zeitfrequenz-

punkte, die von Sprache dominiert werden von denen getrennt werden können, die von Restecho dominert sind. Daraus wiederum können die Koeffizienten eines Filters zur Restechounterdrückung berechnet werden [7]. Zusammen mit den linearen Filtern erreicht man damit eine Unterdrückung von über 30 dB.

3.2 Sprachsignalverbesserung

Das mehrkanalige, vom akustischen Echo des Lautsprechersignals weitgehend befreite Signal wird einem Sprachsignalverbesserungsblock zugeführt, welcher aus den Komponenten Enthallung, akustische Strahlformung, Quellentrennung und Kanalauswahl besteht. Diese Verarbeitungsschritte werden nun besprochen.

3.2.1 Enthallung

Die Enthallungskomponente hat zum Ziel, den späten Nachhall aus dem Signal zu entfernen, d.h. die Signalkomponenten, die um mehr als 50 ms gegenüber der direkten Signalkomponente verzögert sind. Das für die Spracherkennung am häufigsten eingesetzte Enthallungsverfahren trägt den Namen "Weighted Prediction Error" (WPE) Methode [8], welche in [9] auf eine MIMO-Variante mit mehrkanaligem Eingang und mehrkanaligem Ausgang erweitert wurde. In diesem Verfahren wird das verhallte Sprachsignal als autoregressiver Prozess modelliert, um dann den nächsten Sprachsignalrahmen aus vergangenen Sprachsignalrahmen zu prädizieren. Anschließend wird das prädizierte Signal vom tatsächlichen Signal abgezogen. Damit damit nicht das Sprachsignal vollständig zerstört wird, wird eine Verzögerung eingebaut, so dass die nächste Wert nur aus weiter in der Vergangenheit liegenden Beobachtungen vorhergesagt wird. Die Korrelation, die diese Messungen mit dem momentanen Signalwert haben, können nur durch den Hall hervorgerufen worden sein, denn die dem Sprachsignal inhärente Korrelation beschränkt sich auf wenige 10 ms.

Die Schätzung eines linearen Prädiktors für einen autoregressiven Prozess ist eine gut etablierte Technik. Hier kommt jedoch die Schwierigkeit hinzu, dass der treibende Prozess (das unverhallte Sprachsignal) nichtstationär ist und ein unbekanntes, zeitvariantes Leistungsdichtespektrum besitzt. Aus diesem Grund haben die Autoren in [8] ein iteratives Verfahren vorgeschlagen, welches abwechselnd das Leistungsdichtespektrum des unverhallten Signals und die Prädiktorkoeffizienten schätzt.

Viele Autoren haben berichteten, dass eine WPE-basierte Enthallung die Wortfehlerrate eines Spracherkenners verbessert. In [10] wird eine Fehlerratenverbesserung um 5 bis 10% berichtet, die auf simulierten Daten erreicht wurde, die ein typisches Anwendungsszenario für digitale Heimassistenten nachbilden.

3.2.2 Akustische Strahlformung

Da die WPE-Komponente am Ausgang ein mehrkanaliges Signal zur Verfügung stellt, kann anschließend eine akustische Strahlformung erfolgen. Die Theorie statistisch optimaler mehrkanaliger Filterung ist aus Lehrbüchern bekannt. Für die Berechnung von Filterkoeffizienten, die beispielsweise das "Minimum Variance Distortionless Response" (MVDR) Kriterium optimieren, wird jedoch die Kenntnis der Kreuzleistungsdichtematrix des Rauschen und die Kenntnis der Übertragungsfunktion vom Sprecher zu den Mikrofonen benötigt, die natürlich in der Praxis unbekannt und sogar zeitvariant ist. In den vergangenen Jahren wurden jedoch Verfahren entwickelt, die diese Größen aus dem Eingangssignal schätzen können. Wiederum läuft dies darauf hinaus, dass zunächst für jeden Zeitfrequenzpunkt eine Sprachpräsenswahrscheinlichkeit geschätzt wird, die angibt, ob der Zeitfrequenzpunkt vornehmlich das Nutzsignal oder

die Störung enthält. Dies kann entweder mit Hilfe neuronaler Netze [11] oder unter Verwendung räumlicher Mischungsmodelle [12] erfolgen. Unter Verwendung der SPW können nun die Kreuzleistungsdichtespektren der Störung und des Nutzsignals geschätzt werden, indem jeweils nur die Zeitfrequenzpunkte herangezogen werden, die von dem jeweiligen Signalanteil dominert sind. Wenn erst einmal die Kreuzleistungsdichtespektren bekannt sind, lassen sich die akustische Übertragungsfunktion und schließlich die Strahlformerkoeffizienten daraus ermitteln [13].

Akustische Strahlformung hat sich als sehr effektiv zur Verbesserung der Spracherkennungsrate erwiesen. Auf der Datenbasis des CHiME-3 und CHiME-4 Wettbewerbs konnte die Fehlerrate fast halbiert werden. Auf Datensätzen, die für digitale Heimassistenten typisch sind, liegen die Fehlerratenreduktionen in der Größenordnung von 10 bis 30%.

3.2.3 Quellentrennung und Kanalauswahl

Mit akustischer Strahlformung lassen sich Störungen aus anderen Raumrichtungen als die des Nutzsignals gut unterdrücken. Wenn andere Sprecher gleichzeitig mit den Nutzsprecher aktiv sind, ist die Unterdrückung aber nicht stark genug, um eine Erhöhung der Wortfehlerrate des Spracherkenners zu vermeiden. Aus diesem Grund wird zusätzlich eine blinde Quellentrennungskomponente vorgesehen, die das Sprachgemisch am Eingang in die Signale der beteiligten Sprecher zerlegt. Dies erfolgt auf ähnliche Weise wie die akustische Strahlformung. Die SPW Schätzung wird lediglich ersetzt durch eine Schätzung, welcher Sprecher in jedem Zeitfrequenzpunkt jeweils dominant ist. Diese Information kann wiederum entweder mit einem neuronalen Netz oder mit einem räumlichen Mischungsmodell geschätzt werden. Ist einmal bekannt, welche Zeitfrequenzpunkte von welchem Sprecher dominiert werden, können Masken oder akustische Strahlformer für jeden Sprecher berechnet werden, um das Signal jedes einzelnen Sprechers aus dem Gemisch zu extrahieren.

Die Auswahl, welches dieser Signale denn nun an den Spracherkenner weitergereicht werden soll, welches also das Nutzsignal darstellt, erfolgt danach, dass man bestimmt, welches Signal das Schlüsselwort (z.b. “Alexa” oder “Hey Google”) enthält, welches anzeigt, dass der Heimassistent angesprochen wird.

3.3 Schlüsselwortdetektion

Für die Schlüsselwortdetektion werden dedizierte neuronale Netze verwendet, die speziell auf die Erkennung des Musters, z.B. “Hey Siri”, trainiert wurden. Eine zuverlässige Erkennung in einer verrauschten und verhallten Umgebung ist jedoch eine große Herausforderung. In einem Ansatz wird zunächst eine Sprachaktivitätsdetektion durchgeführt, so dass das neuronale Netz nur arbeiten muss, wenn Aktivität im Mikrofonsignal erkannt worden ist. Wenn Sprache detektiert worden ist, wird ein gleitendes Fenster über die Daten gelegt, dessen Länge der Länge des Schlüsselwortes entspricht. Für jede Lage des Fensters soll nun das neuronale Netz entscheiden, ob das Fenster das Schlüsselwort enthält oder nicht. Die Entscheidung des Netzes wird häufig noch einem zweiten Test unterzogen, bei dem etwa überprüft wird, ob die Länge des detektierten Schlüsselwortes passend ist und ob es die Phoneme des Schlüsselwortes enthält.

Neben dieser Schlüsselwortdetektion gibt es weitere Klassifikatoren, die entscheiden, wann denn die Aussage des Nutzers endet bzw. nicht mehr an das Gerät adressiert ist und ob eine zweite Aussage nach einer Pause ebenfalls an das Gerät gerichtet ist. Häufig wird auch eine Sprechererkennung durchgeführt, damit das System erkennt, welcher Sprecher es gerade bedient, um Anfragen, etwa zum Terminkalender, auch korrekt beantworten zu können.

3.4 Automatische Spracherkennung

In digitalen Heimassistenten werden meist Spracherkenner nach dem “Hybridansatz” eingesetzt. Sie verwenden ein als neuronales Netz realisiertes akustisches Modell, welches für alle möglichen Laute der Sprache die Wahrscheinlichkeit berechnet, dass dieser dem momentanen akustischen Signal zugrundeliegt. Außerdem verwenden sie ein Sprachmodell, das ebenfalls mit Hilfe eines neuronalen Netzes realisiert ist. Dieses gibt für alle möglichen Wörter und Wortfolgen deren Wahrscheinlichkeit an. Diese beiden Wissenquellen werden mit Hilfe eines “hidden Markov Modells” (HMMs) kombiniert, und es wird diejenige Wortsequenz ermittelt, die am wahrscheinlichsten dem akustischen Signal zugrundeliegt.

Die spezielle Herausforderung im Bereich digitaler Heimassistenten war, dass dies eine neuartige Anwendung der Spracherkennung ist, für die keine anwendungsspezifischen Trainingsdaten vorlagen. Daher wurden anfänglich große Anstrengungen unternommen, um realistische Trainingsdaten zu simulieren, etwa indem Raumimpulsantworten simuliert oder gemessen wurden und die ungestörten Sprachaufnahmen aus früheren Datenbasen dann damit gefaltet wurden, um verhallte Signale zu generieren. Neben diesem wurden vielfältige weitere Methoden zur künstlichen Vermehrung der Trainingsdaten entwickelt.

3.5 Dialogmanagement

Nachdem der Spracherkenner die Nutzereingabe in eine maschinenlesbare Form transkribiert hat, erfolgt nun die Verarbeitung dieser Eingabe (“natural language processing” (NLP)), um sie semantisch zu interpretieren und um eine Antwort vorzubereiten, indem a) die von Nutzer gewünschte Information durch Abfrage einer Wissensdatenbank bestimmt wird und b) ein Text zur Ausgabe vorbereitet wird (“natural language generation” (NLG). Auf diese Aspekte soll aber in diesem Beitrag nicht näher eingegangen werden.

3.6 Sprachsynthese

Schließlich muss der Antworttext noch zu einem Sprachsignal zur Ausgabe über den Lautsprecher synthetisiert werden. Auch hierfür wird heutzutage ein neuronales Netz eingesetzt. Hier hat sich das sogenannte Wavenet durchgesetzt [14], da es eine hohe Natürlichkeit der generierten Sprache erreicht. Allerdings ist das Wavenet aufgrund seiner autoregressiven Struktur schlecht zu parallelisieren und daher rechenaufwändig und langsam. Aber auch hierfür gibt es Lösungen, wie beispielsweise dass ein einfacher zu parallelisierendes “Feedforward” Netzwerk so trainiert wird, dass es die Ausgaben des Wavenet emuliert [15].

4 Zusammenfassung

Dieser Beitrag gibt ein Überblick über die Sprachverarbeitungskomponenten eines typischen digitalen Heimassistenten. Die einzelnen Blöcke wurde jeweils nur kurz beschrieben, ohne in die Details der zugrundeliegenden Theorie oder der Implementierung zu gehen. In praktisch allen Komponenten werden tiefe neuronale Netze eingesetzt, häufig in Kombination mit klassischer Signalverarbeitung, wie etwa bei der Echokompensation oder der akustischen Strahlformung. Auch wenn die Beschreibung in diesem Beitrag relativ oberflächlich ist, so wird doch deutlich, dass die Assistenten ein echtes Hightech- Produkt sind.

Literatur

[1] KINOSHITA, K., M. DELCROIX, S. GANNOT, E. HABETS, R. HAEB-UMBACH, W. KELLERMANN, V. LEUTNANT, R. MAAS, T. NAKATANI, B. RAJ, A. SEHR, und T. YOSHIOKA: *A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research*. *EURASIP Journal on Advances in Signal Processing*, 2016.

[2] BARKER, J., R. MARXER, E. VINCENT, und S. WATANABE: *The third "CHiMEßpeech separation and recognition challenge: Analysis and outcomes*. *Computer Speech and Language*, 46, S. 605–626, 2017.

[3] VINCENT, E., S. WATANABE, A. A. NUGRAHA, J. BARKER, und R. MARXER: *An analysis of environment, microphone and data simulation mismatches in robust speech recognition*. *Computer Speech & Language*, 46, S. 535–557, 2017.

[4] BARKER, J., S. WATANABE, E. VINCENT, und J. TRMAL: *The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines*. In *Proc. Interspeech*, S. 1561–1565. 2018.

[5] HAEB-UMBACH, R., S. WATANABE, T. NAKATANI, M. BACCHIANI, B. HOFFMEISTER, M. L. SELTZER, H. ZEN, und M. SOUDEN: *Speech processing for digital home assistants: Combining signal processing with deep-learning techniques*. *IEEE Signal Processing Magazine*, 36(6), S. 111–124, 2019. doi:10.1109/MSP.2019.2918706.

[6] BENESTY, J., T. GÄNSLER, D. MORGAN, M. SONDHI, und S. GAY: *Advances in network and acoustic echo cancellation*. Springer, 2001.

[7] AUDIO SOFTWARE ENGINEERING AND SIRI SPEECH TEAM: *Optimizing Siri on HomePod in far-field settings*. 2018. URL `https://machinelearning.apple.com/2018/12/03/optimizing-siri-on-homepod-in-far-field-settings.html`.

[8] NAKATANI, T., T. YOSHIOKA, K. KINOSHITA, M. MIYOSHI, und B.-H. JUANG: *Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation*. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2008.

[9] YOSHIOKA, T. und T. NAKATANI: *Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2012.

[10] CAROSELLI, J., I. SHAFRAN, A. NARAYANAN, und R. ROSE: *Adaptive multichannel dereverberation for automatic speech recognition*. In *Proc. Interspeech*. 2017.

[11] HEYMANN, J., L. DRUDE, und R. HAEB-UMBACH: *Neural network based spectral mask estimation for acoustic beamforming*. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016.

[12] YOSHIOKA, T., N. ITO, M. DELCROIX, A. OGAWA, K. KINOSHITA, M. FUJIMOTO, C. YU, W. J. FABIAN, M. ESPI, T. HIGUCHI, S. ARAKI, und T. NAKATANI: *The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices*. In *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, S. 436–443. 2015. doi:10.1109/ASRU.2015.7404828.

[13] WARSITZ, E. und R. HAEB-UMBACH: *Blind acoustic beamforming based on generalized eigenvalue decomposition*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(5), S. 1529–1539, 2007. doi:10.1109/TASL.2007.898454.

[14] VAN DEN OORD, A., S. DIELEMAN, H. ZEN, K. SIMONYAN, O. VINYALS, A. GRAVES, N. KALCHBRENNER, A. W. SENIOR, und K. KAVUKCUOGLU: *Wavenet: A generative model for raw audio*. *CoRR*, abs/1609.03499, 2016. URL `http://arxiv.org/abs/1609.03499`. `1609.03499`.

[15] VAN DEN OORD, A., Y. LI, I. BABUSCHKIN, K. SIMONYAN, O. VINYALS, K. KAVUKCUOGLU, G. VAN DEN DRIESSCHE, E. LOCKHART, L. COBO, F. STIMBERG, N. CASAGRANDE, D. GREWE, S. NOURY, S. DIELEMAN, E. ELSEN, N. KALCHBRENNER, H. ZEN, A. GRAVES, H. KING, T. WALTERS, D. BELOV, und D. HASSABIS: *Parallel WaveNet: Fast high-fidelity speech synthesis*. In *Proc. ICML*, S. 3918–3926. 2018.