Erstellung eines Testsignals mit Sprachcharakteristik

Inga Holube[1], Stefan Fredelake[1], Jörg Bitzer[1], Marcel Vlaming[2]

[1] *Institut für Hörtechnik und Audiologie, 26121 Oldenburg, Deutschland, Email: Inga.Holube@fh-oow.de*
[2] *VU University Medical Center, 1081HV Amsterdam, Niederlande, Email: m.vlaming@vumc.nl*

Einleitung

Zur Analyse und Charakterisierung der Übertragung von Sprache durch ein System (z.B. ein modernes nichtlineares Hörgerät oder ein Telekommunikationsgerät) wird ein Testsignal benötigt, das reproduzierbare Messbedingungen ermöglicht und möglichst alle Eigenschaften von Sprache aufweist. Zu diesen Eigenschaften zählen neben dem Spektrum z. B. das Modulationsspektrum und der Grundfrequenzverlauf sowie dessen Harmonische. Künstlich erzeugte Signale erfüllen diese Anforderungen nur unzureichend. Aufnahmen mit realen Sprechern repräsentieren dagegen nur eine Sprache und sind deshalb u. U. nicht international einsetzbar. In diesem Beitrag wird ein internationales Sprach-Testsignal (International Speech Test Signal, ISTS) vorgestellt, das in Zusammenarbeit mit der ISMADHA-Arbeitsgruppe der European Hearing Instrument Manufacturing Association (EHIMA) entworfen wurde. Das Testsignal beruht auf natürlicher Sprache, ist aber trotzdem im Wesentlichen unverständlich. Zur Erzeugung wurden Aufnahmen mit weiblichen Sprechern in sechs verschiedenen Sprachen erstellt. Diese Aufnahmen wurden in Segmente zerlegt und in zufälliger Reihenfolge wieder aneinander gehängt. Dadurch blieben alle relevanten Eigenschaften von Sprache erhalten. Eine mögliche Anwendung des neuen Testsignals ist die Bestimmung der Verstärkung von Sprache in Hörgeräten. Viele weitere Anwendungen, bei denen Sprache übertragen, verarbeitet oder als Störer verwendet wird, sind denkbar.

Sprachaufnahmen

21 Sprecherinnen mit sechs verschiedenen Muttersprachen (Arabisch, Chinesisch, Deutsch, Englisch, Französisch und Spanisch) lasen mehrmals die Geschichte „Der Nordwind und die Sonne“ [1] in ihrer Muttersprache mit möglichst natürlicher Artikulation vor. Die Aufnahmen wurden mit einem Neumann KM184 Richtmikrofon mit einer Abtastrate von 44,1 kHz und einer Auflösung von 24 bit in einem modifizierten Büroraum (Nachhallzeit 0,5 s bei 500 Hz) erstellt. Aus diesen Aufnahmen wurde jeweils eine Aufnahme einer Sprecherin für jede Sprache ausgewählt. Dabei wurde die regionale Herkunft der Sprecherinnen, die Stimmqualität (z. B. keine Heiserkeit) und die mittlere Grundfrequenz der Sprecherinnen berücksichtigt. Die Dauer der Sprachpausen innerhalb der Aufnahmen wurden auf 650 ms begrenzt (nur in wenigen Fällen notwendig) und auf das mittlere Langzeitspektrum weiblicher Sprecherinnen zwischen 100 Hz und 16 kHz nach [2] gefiltert. Ziel war die Optimierung der Homogenität des Testsignals. Außerdem wurde eine Statistik der Dauer von Sprachabschnitten zwischen längeren Pausen (mehr als 100 ms) erstellt und eine Wahrscheinlichkeitsfunktion angepasst, die bei der Mischung der Aufnahmen benötigt wurde.

Zerlegung der Aufnahmen

Die Aufnahmen wurden mit einem automatischen Programm in Segmente zerlegt. Dazu wurde ein Abschnitt von 500 ms aus der Aufnahme entnommen und die letzten 400 ms davon betrachtet. Das 10ms-Intervall mit der geringsten Leistung innerhalb dieses 400ms-Abschnitts wurde ausgewählt und der niedrigste Absolutwert innerhalb dieses 10ms-Intervalls bestimmt. Ein Segment umfasste dann die Aufnahme vom Beginn des 500ms-Abschnitts bis zu diesem niedrigsten Absolutwert. Der nächste 500ms-Abschnitt startete direkt nach dem niedrigsten Absolutwert. Diese automatische Zerlegung wurde per Hand modifiziert, um Schnitte innerhalb von Vokalen und zusammenhängenden Phonemfolgen möglichst zu vermeiden. Daraus resultierten Sprachsegmente mit einer Dauer zwischen 100 und 600 ms. Sprachpausen, die länger als 100 ms waren, wurden im Anschluss an den vorhergehenden Sprachabschnitt im gleichen Segment belassen, damit sie in ihrer natürliche Position verbleiben. Diese Segmente mit langen Pausendauern und die nachfolgenden „Beginn-Segmente“ wurden markiert.

Mischung der Segmente

Die erstellten Segmente wurden in einer zufälligen Reihenfolge zu Abschnitten von 10 s und 15 s wieder aneinander gehängt. Die Segmente wurden dazu mit einem Hann-Fenster mit einer Flankendauer von 1 ms gefenstert, um Artefakte zu vermeiden. Darüber hinaus wurde berücksichtigt, dass sich die Sprache von Segment zu Segment ändert und jede Sprache innerhalb von sechs aufeinander folgenden Segmenten genau einmal vorkommt. Außerdem konnte jedes einzelne Segment nur einmal innerhalb eines 10- oder 15s-Abschnitts verwendet werden. Um Sprünge der Grundfrequenz zu minimieren, wurde die Grundfrequenz innerhalb der ersten und der letzten 50 ms jedes Segmentes bestimmt. Wenn bei dem Mischungsvorgang zwei stimmhafte Artikulationen aufeinander trafen, wurde nur ein Grundfrequenzunterschied von 10 Hz zugelassen. Die Kombinationen von einer stimmhaften und einer stimmlosen sowie von zwei stimmlosen Artikulationen waren jedoch immer möglich. Diejenigen Segmente, die Pausen mit einer Dauer von mehr als 100 ms beinhalteten, wurden dann ausgewählt, wenn die Sprachdauer einen Wert überstieg, der mit Hilfe der oben beschriebenen Wahrscheinlichkeitsfunktion bestimmt wurde. Dadurch wurde gewährleistet, dass die Sprachpausen in einem natürlichen Abstand voneinander positioniert wurden. Nach einer Sprachpause folgte ein „Beginn-Segment“ einer

anderen Sprache. Am Ende eines jeden 10s- und 15s-Abschnitts wurde ein Segment mit Sprachpause eingefügt und auf die notwendige Gesamtlänge des jeweiligen Abschnitts begrenzt. Alle generierten Abschnitte wurden wiederum auf das internationale Spektrum nach [2] gefiltert. Aus den Abschnitten wurde ein Testsignal mit einer Dauer von 60 s erstellt, je nach Anforderung sind jedoch auch andere Dauern möglich. Für die Bestimmung der Übertragungseigenschaften von Hörgeräten soll zunächst ein Abschnitt von 15 s zum Einschwingen der Signalverarbeitungsalgorithmen und danach ein Messzeitraum von 45 s verwendet werden. Zur groben Abschätzung der Messergebnisse soll es möglich sein, diesen Messzeitraum auf 10 s zu begrenzen.

Analyse des Testsignals

Das erstellte internationale Testsignal wurde auf seine Eigenschaften hin analysiert und mit den ursprünglichen Aufnahmen verglichen. Damit konnte nachgewiesen werden, dass es in allen relevanten Kriterien mit natürlicher Sprache überein stimmt. Im Folgenden werden die wichtigsten Ergebnisse für das Testsignal mit einer Dauer von 45 s zusammengefasst.

Langzeitspektrum

Das Langzeitspektrum des Testsignals wie auch die 10s- und 15s-Abschnitte weichen um maximal 1 dB vom internationalen Langzeitspektrum weiblicher Sprache aus [2] ab.

Kurzzeitspektrogramm

Das Kurzzeitspektrogramm des Testsignals weist Sprünge in der Grundfrequenz auf, die jedoch auch in den ursprünglichen Aufnahmen, je nach Sprache unterschiedlich stark, vorhanden sind.

Grundfrequenz

Der Median der Grundfrequenz des Testsignals beträgt 196 Hz, während die Sprecherinnen eine mediane Grundfrequenz von 203 Hz aufweisen. Dies kann als hinreichend ähnlich angesehen werden. Die Standardabweichung liegt sowohl beim Testsignal als auch bei den ursprünglichen Aufnahmen bei ca. 44 Hz.

Modulationsspektren

Die Modulationsspektren des terzband-gefilterten Testsignals und der ebenfalls gefilterten ursprünglichen Aufnahmen weisen ein Maximum im Bereich von ca. 2-8 Hz auf und zeigen keine systematischen Abweichungen.

Komodulationsanalyse

Zur Analyse der Komodulationen wurden die Einhüllenden der terzband-gefilterten Signale berechnet und miteinander korreliert. Mit zunehmender Entfernung der Terzbänder reduziert sich die Stärke der Kreuzkorrelationen. Dies ist sowohl beim Testsignal als auch bei den ursprünglichen Aufnahmen beobachtbar.

Pausendauer

Die Verteilung der Pausen und ihre Dauer im Testsignal entsprechen im Wesentlichen denjenigen in den ursprünglichen Aufnahmen, wobei jedoch berücksichtigt werden muss, dass das Testsignal kürzer und deshalb die Verteilung weniger gleichmäßig ist. Der Anteil der Pausen an der Gesamtdauer des Signals liegt bei ca. 1/6.

Perzentilverteilung

Die Signale wurden terzband-gefiltert und die Pegel in 125ms-Fenstern (50%-Überlapp) bestimmt. Aus der Pegelverteilung wurde die Differenzen zwischen den 99%- und den 30%-Perzentilen berechnet. Diese liegen sowohl für das Testsignal als auch die ursprünglichen Aufnahmen zwischen 20 und 30 dB (siehe Abb. 1).

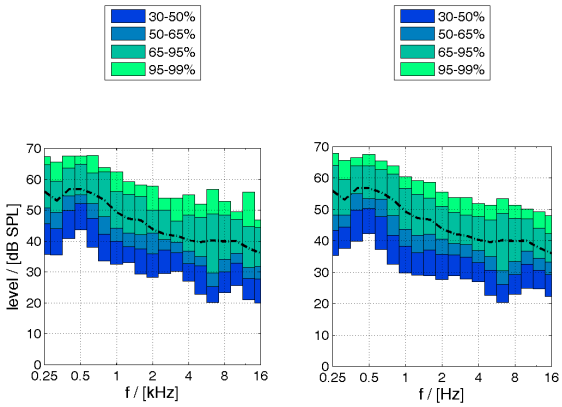

Abbildung 1: Perzentilverteilung der Pegel in 125ms-Fenstern für einen 10s-Abschnitt (links) und den 45s-Abschnitt (rechts) des Testsignals gemeinsam mit dem internationalen Sprachspektrum aus [2] (gestrichelt).

Anteil stimmloser Abschnitte

Der Anteil der stimmlosen Abschnitte der Signale liegt bei dem Testsignal mit 44% geringfügig oberhalb des mittleren Wertes für die ursprünglichen Sprachaufnahmen (35%).

Instantane Amplitudenverteilung

Die Verteilung der instantanen Amplituden des Testsignals ist derjenigen der ursprünglichen Sprachaufnahmen sehr ähnlich.

Crest-Faktor

Der CREST-Faktor des Testsignals ist mit einem Wert von 17 sehr ähnlich zu dem Wert von 18 der ursprünglichen Sprachaufnahmen.

Literatur

[1] Handbook of the International Phonetic Association, Cambridge University Press.

[2] Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wibraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M., Nasser, N., El Kholy, W., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavatkiladze, G., Frolenkov, G., Westerman, S. und Ludvigsen, C.: An international comparison of long-term average speech spectra. J. Acoust. Soc. Am. 96 (1994), 2108-2120.