ETH zürich Research Collection

Doctoral Thesis

Speech processing strategies based on the sinusoidal speech model for the profoundly hearing impaired

Author(s):
Timms, Olegs

Publication Date:
2003

Permanent Link:

https://doi.org/10.3929/ethz-a-004554764 →

Rights / License:

In Copyright - Non-Commercial Use Permitted →

This page was generated automatically upon download from the ETH Zurich Research Collection. For more information please consult the Terms of use.

ETH Library

Diss. ETH no 15167

Speech Processing Strategies Based on the Sinusoidal Speech Model for the Profoundly Hearing Impaired

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of

Doctor of Technical Sciences

presented by

Olegs Timms

Dipl. Phys. ETH

born October 19, 1973

citizen of Latvia

accepted on the recommendation of

Prof. Dr. Peter Niederer, examiner

PD Dr. Norbert Dillier, co-examiner

Dr. Stefan Launer, co-examiner

2003

Abstract

Average speech recognition for profoundly sensorineural hearing impaired subjects using conventional high-power hearing devices is very restricted. For most of these people, speech communication is very limited, and they use their hearing devices only as support for lip-reading, to make acoustical contacts in their environment and to hear warning sounds.

The reasons for the limited hearing capacity in case of profound sensorineural hearing impairment are the restricted audible frequency area, the strongly reduced dynamic range, and the limited temporal and spectral resolution. In spite of these handicaps it is theoretically possible to use the remaining narrowed information channels of the auditory system to provide the hearing impaired subjects with temporal and spectral information for improving their speech perception capacities. For this purpose, particular speech perception oriented signal processing strategies enabling spectral reduction and the transposition of the essential spectral components into the residual hearing area can be employed. The successful application of spectrally reduced and transformed speech signals in the cochlear implant was one of the basic motivations for the present investigation.

For the design of the signal processing strategies for the profoundly hearing impaired, different approaches for the identification of the essential speech cues and their acoustic presentation were studied based on the literature and on cochlear implant technology. In order to investigate and apply the different proposed spectral and temporal modifications of the acoustic signal, a signal processing system was implemented based on the "sinusoidal speech" algorithm. This system enables the choice of different signal processing parameters including different possibilities for signal reconstruction.

In order to investigate the perception of spectrally reduced speech, a study with normal hearing native German speaking adults was performed. For this purpose, the speech materials of the Oldenburg sentence test and the German C12 consonant and V08 vowel tests were processed with the signal processing system, using a limited number of spectral components (1 - 5) and three different temporal and spectral resolutions. Speech comprehension tests using the processed speech material were carried out to determine the minimally required number of spectral components required for near 100% recognition scores for each of the time/frequency resolutions. The results of this study showed that in order to achieve a satisfactory speech recognition score, different time respectively frequency resolutions require a different number of spectral components per time unit. It was therefore concluded that this spectral component per time ratio is very important for speech perception. The minimum number of spectral components per time unit for near 100% speech recognition was found to be approximately two to four spectral components per 1.5 ms. Hence, signal processing schemes using longer analysis/synthesis frames (*i.e.* high frequency resolution respectively low temporal resolution) require a larger number of spectral components for signal reconstruction or a greater overlap in the analysis/synthesis frames (increased temporal resolution). The study showed also that speech perception for normal hearing subjects is preserved even if a dramatical spectral reduction of the speech signal is applied. This

observation was an important step towards the implementation of different spectral manipulation schemes including different kinds of spectral compression.

In the following two studies, speech perception of the combined spectral reduction and linear spectral compression on both the FFT and the SPINC scale was investigated with normal hearing and moderately severe to profoundly hearing impaired subjects. The study with the normal hearing subjects showed significantly increased vowel (~+20%) and consonant (~+10%) identification scores for the spectrally reduced and compressed signals with respect to the reference signal lowpassed at 2 kHz (the lowpass filtering approximates the loss of high frequencies). It was also observed that even though linear spectral compression on the FFT scale with spectral compression ratios larger than 1.6 can improve consonant identification, it results in a considerably decreased sentence perception (~-40%) although the vowel identification did not change significantly. In addition, it was found that the linear spectral compression on the SPINC scale showed larger consonant identification score improvements than the linear spectral compression on the FFT scale. For vowel identification, linear spectral compression on the FFT scale with a compression ratio of 1.3 was better than linear spectral compression on the SPINC scale.

The study with hearing impaired subjects showed that the profoundly hearing impaired subjects with an average hearing loss for the low frequency tones (125, 250, and 500 Hz) close to 90 dB can benefit from spectrally compressed speech. Compared with the spectrally non-compressed reference signal, they achieved improvement in sentence (~+5%) and consonant (~+10%) identification. The identification of the unprocessed reference sentences and consonants for these subject class were close to 10% and 20% respectively. The hearing impaired subjects with moderately severe to profound steeply sloping hearing loss with average hearing thresholds between 20-60 dB at the low frequency tones (125, 250, and 500 Hz) did not profit from any spectral compression signal processing scheme but showed even significantly decreased identification scores for sentences, consonants and vowels. Based on this observations it was proposed to use the low frequency pure tone average in combination with poor sentence recognition scores as a criterion for the potential benefit of spectral compression for profoundly hearing impaired patients.

In addition, competitive speech perception experiments with temporally modified speech were carried out with normal hearing subjects using the Oldenburg sentence test material. It was observed that the prolongation of the whole speech signal or any of its segments improved the signal-to-noise ratio of the temporally modified speech with respect to the unprocessed reference. However, temporal shortening of any speech segment as well as the whole signal caused a decrease of the signal-to-noise ratio. The bidirectional temporal modification, *i.e.* the simultaneous prolongation and shortening of different speech segments applied in order to preserve the original duration of the signal, lead to a decreased signal-to-noise ratio. It was therefore concluded that the temporal modification strategy for processing speech for the profoundly hearing impaired does not make sense.

Zusammenfassung

Die mittlere Satzverständlichkeit von hochgradig Schwerhörigen ist stark limitiert. Die sprachliche Kommunikation ohne Nutzung von Lippenlesen ist für diese Personen meistens sehr begrenzt oder gar unmöglich. Die meisten hochgradig Hörbehinderten nutzen daher ihre konventionellen Hochleistungs-Hörgeräte oft nur zur einfachen akustischen Kommunikation mit ihrer Umgebung und für die Erkennung von Warnsignalen.

Die Gründe für die ungenügende akustische Wahrnehmung bei hochgradiger sensorineuraler Schwerhörigkeit sind der eingeschränkte hörbare Bereich, der stark reduzierte Dynamikbereich sowie das begrenzte zeitliche und spektrale Auflösungsvermögen. Trotzdem ist es theoretisch möglich, den verbleibenden stark reduzierten Informationskanal zur Übertragung der für Sprachverständlichkeit notwendigen spektralen und zeitlichen Information zu nutzen. Um die begrenzte Sprachwahrnehmung der Betroffenen zu verbessern können speziell für die Sprachverständlichkeit zugeschnittene Signalverarbeitungsstrategien verwendet werden, welche eine spektrale Reduktion sowie die Transposition der wesentlichen spektralen Komponenten in den Resthörbereich der hochgradig Schwerhörigen ermöglichen. Der Erfolg der Cochlea Implantate, welche ein sehr stark spektral reduziertes und transponiertes Signal verwenden und trotzdem eine gute Sprachverständlichkeit ermöglichen, war eine der Hauptmotivation für die vorliegende Doktorarbeit.

Für den Entwurf sinnvoller Signalverarbeitungsstrategien für hochgradig Schwerhörige wurden verschiedene aus der Literatur und von Cochlea Implantanten bekannte Ansätze zur Identifikation der wesentlichen Information des Sprachsignals und deren akustische Darbietung untersucht. Um die verschiedenen Signalverarbeitungsstrategien genauer zu untersuchen und testen zu können, wurde basierend auf dem „sinusoidal speech" Modell ein Signalverarbeitungssystem implementiert. Dieses System erlaubt eine grosse Auswahl verschiedener Signalverarbeitungsparameter inklusive der freien Wahl der Signalresynthese-Methode.

Um die Grenzen der spektralen Reduktion zu untersuchen wurde eine Studie mit normalhörenden Personen deutscher Muttersprache durchgeführt. Dazu wurde das Sprachmaterial des Oldenburger Satztests, des deutschen C12 Konsonantentests und des deutschen V08 Vokaltests mit einer begrenzten Anzahl spektraler Komponenten (maximal 5) und drei verschiedenen zeitlichen und spektralen Auflösungen mit Hilfe des Signalverarbeitungssystems verarbeitet. Diese Sprachmaterialien wurden zur Identifikation der minimal nötigen Anzahl spektraler Komponenten für jede der drei Zeit- und Frequenzauflösungen für nahezu 100% Spracherkennung verwendet. Die Resultate der Studie zeigen, dass für verschiedene Zeit- und Frequenzauflösungen eine unterschiedliche Anzahl spektraler Komponenten erforderlich ist, um eine befriedigende Sprachverständlichkeit zu erreichen. Daraus wurde gefolgert, dass die Anzahl der spektralen Komponenten per Zeiteinheit von entscheidender Bedeutung für die Sprachverständlichkeit ist. Der minimale Wert der benötigten Anzahl spektraler Komponenten per Zeiteinheit für 100% Sprachverständlichkeit konnte aus der Studie bestimmt werden und liegt bei zwei bis

vier spektralen Komponenten per 1.6 msec. Dies bedeutet, dass Signalverarbeitungsverfahren welche längere Analyse/Synthese Fenster verwenden (d.h. hohe Frequenzauflösung / niedrige Zeitauflösung) auch eine grössere Anzahl spektraler Komponenten oder einen grösseren Überlappungsbereich der Analyse/Synthesefenster (Erhöhung der Zeitauflösung) benötigen. Die Untersuchung der Verständlichkeit eines spektral reduzierten Sprachsignals hat gezeigt, dass ein stark reduziertes Sprachsignal noch immer verständlich sein kann und war eine wesentliche Voraussetzung für die Implementation von diversen spektralen Transpositionsverfahren inklusive lineare und nichtlineare spektrale Kompression.

In den zwei darauffolgenden Studien wurde die Sprachverständlichkeit der Kombination von spektraler Reduktion mit linearer Frequenzkompression auf der FFT Skala und auf der SPINC Skala mit normalhörenden und hochgradig sensorieneural schwerhörigen Probanden untersucht. Die Studie mit normalhörenden Probanden zeigte im Vergleich mit dem bei 2 kHz tiefpassgefilterten Referenzsignal (die Tiefpassfilterung approximiert den Verlust der Wahrnehmung hoher Frequenzanteile) eine wesentliche Verbesserung der Vokalidentifikation (~+20%) und Konsonantenidentifikation (~+10%) für spektral reduzierte und komprimierte Sprachsignale. Es hat sich gezeigt, dass die lineare spektrale Kompression auf der FFT Skala mit Kompressionsfaktoren grösser als 1.6 zwar eine Verbesserung der Konsonantenidentifikation bewirkt, gleichzeitig aber die Sprachverständlichkeit von Sätzen signifikant verschlechtert (~-40%), obwohl die Vokalidentifikation im Wesentlichen unverändert geblieben ist. Zusätzlich konnte gezeigt werden, dass die lineare spektrale Kompression auf der SPINC Skala im Vergleich zur linearen spektralen Kompression auf der FFT Skala eine grössere Verbesserung für die Konsonantenidentifikation bewirkt. Für die Vokalidentifikation erwies sich die lineare spektrale Kompression auf der FFT Skala mit einer Kompressionsrate von 1.3 als besser als die spektrale Kompression auf der SPINC Skala.

Die Studie mit hochgradig schwerhörigen Probanden, welche einen Hörverlust von etwa 90 dB für Tiefton-Frequenzen (125, 250, und 500 Hz) aufweisen, zeigte eine Verbesserung der Sprachverständlichkeit in Sätzen (~+5%) und eine Verbesserung der Konsonantenidentifikation (~+10%) für spektral komprimierte Sprachsignale im Vergleich zum spektral nicht komprimierten Referenzsignal. Die gemessene Spracherkennung für das Referenzsignal dieser Probanden betrug etwa 10% für die Satzidentifikation und etwa 20% für die Konsonantenidentifikation. Für die hochgradig schwerhörigen Probanden mit einem Hörverlust zwischen 20-60 dB für Tiefton-Frequenzen (125, 250, und 500 Hz) zeigten die verschiedenen spektralen Kompressionsschemen keinen Nutzen sondern führten zu einer signifikante Verschlechterung der Sprachverständlichkeit. Es wird daher vorgeschlagen, die Hörverluste der Tiefton-Frequenzen (125, 250, und 500 Hz) in Kombination mit der Satzverständlichkeit als Kriterium für den potentiellen Nutzen von spektralen Kompressionsschemen bei hochgradig Schwerhörigen zu verwenden.

Des weiteren wurden auch Experimente zur Sprachverständlichkeit von zeitlich modifizierten Sprachsignalen mit normalhörenden Probanden durchgeführt. Für diese Studie wurde das Sprachmaterial des Oldenburger Satztests verwendet. Generell zeigten die Sprachsignale mit einer Ganzsignal-Verlangsamung oder mit einer Verlangsamung bestimmter Sprachsegmente einen verbesserten Signal-Rausch-Abstand als die (zeitlich unmodifizierten) Referenzsignale. Dagegen zeigte die Verkürzung von Einzelsegmenten oder die Verkürzung des Gesamtsignals einen geringeren Signal-Rausch-Abstand als die Referenz. Die bidirektionalen zeitlichen Modifikationen, d.h. die zeitliche Verlängerung der einen und die zeitliche Verkürzung der anderen Sprachsegmente mit den Zweck der etwa

gleichbleibenden Länge des zeitlich modifizierten Signals im Vergleich zum Originalsignal, wies einen schlechteren Signal-Rausch-Abstand als die Referenz auf. Daraus wurde gefolgert, dass die Implementation zeitlicher Modifikationen zur Verbesserung der Sprachverständlichkeit bei hochgradig schwerhörigen Personen nicht sinnvoll ist.