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

Visiem tiem kuri manī ticējuši…

For all who believed in me…

… and of course for all of those who know
what the meaning of 42 is

Acknowledgments

This dissertation describes the work effectuated at the Laboratory of Experimental Audiology of the Department of Otorhinolaryngology, Head and Neck Surgery of the University Hospital of Zurich in cooperation with Phonak AG, Stäfa, Switzerland from summer 1999 until summer 2003. During this time I had the pleasure of meeting and working with many people from whom I learned a lot and who helped me to complete the present work.

I am especially indebted and very grateful to my advisor, Dr. Silvia Allegro for the ardent and selfless support she has given me throughout my whole work period, even during sad times for her. I do appreciate this very much and know that my own work would have been much harder without her competent advice, enthusiasm and confidence.

I am also very grateful to PD Dr. Norbert Dillier and Dr. Stefan Launer for their professional advice, personal effort and very interesting suggestions during and especially in the beginning and concluding phases of my thesis, as well as for acting as co-examiners.

Furthermore, I would like to thank Prof. Dr. Peter Niederer, head of the Institute of Biomechanical Engineering at the ETH Zurich, for giving me the opportunity to complete a doctoral dissertation.

I am further obliged to the Phonak AG which financially supported this dissertation.

I would like to thank also PD Dr. Wai Kong Lai, Dr. Volker Kuehnel, PD Dr. Hugh McDermot, Maurice Boonen, Dr. Herbie Bächler, Dr. Hansueli Roeck, and Dr. Andreas von Buol for the interesting discussions and suggestions on the topics of speech perception, signal processing, statistic analysis of the results of psycho-acoustical speech perception tests or/and helping me out of difficult MATLAB programming problems.

I am thankful to Christian Aebi whose diploma thesis materials were essential for chapter 8 of the present PhD thesis.

I would also like to thank the vice dean of the Faculty of Physics and Mathematics at the University of Latvia, Dr. Aivars Simanovskis†, rector of the Latvian Maritime Academy Dr. Jānis Bērziņš, Prof. Dr. Mārcis Auziņš, Martina Bächli, Dr. J. Brunner, and Prof. Dr. Peter Rüegsegger who all enabled my physics studies at the ETH Zurich.

Very special thanks go to all my LEA colleagues who spend many hours of their time in the sound proof room being exposed to all kind of strange sounds trying to recognize human speech in it. Dr. Michael Büchler many times helped me out of various unpleasant situations and was always “quick with a joke or to light up a smoke” (beside that he was able to fold thousands of paper butterflies and perform excellent scientific work). My office neighbor Mattia Ferrazzini always found a topic for interesting scientific or political discussions, which normally took place in our small but cozy office leaving strangely colored trails on the whiteboard. Urban Willi was always ready to explain various zoological, biological, biochemical, photo-logical, and quantum mechanical problems and secrets (of course mainly

hearing related). Simone Volpert was for a long time the spirit of our small group, simultaneously being a sister, a mother and sometimes also a nurse until she was stolen by a nice looking Oldenburg lad, with as she told us, beautiful eyes. My special thanks go also to our group secretaries Belja Dillier-Breginc and Connie El-Mokdad-Braccini. Belja was reading and correcting my English and German written letters and emails as well as posting daily email-jokes and cat care information, watering my poor office plants and making sure that daily coffee and milk dose would always be available (she was also stolen, however, within the group). Connie was always very responsive and from time to time informed us about news and problems from the Middle East. Herbert Jakits with his characteristic Bavarian accent introduced me into the life of a good husband and father, allowing me to babysit his little son Nicki. René Holzreuter gave me a super short but effective crash course in data basis. Felix Beerli was always ready to help out with different software and hardware problems and Markus Schmid knew which capacitor, transistor or diode is dead in various electronic devices. Hubert Hauschild showed me once where angels and starlets live (in a snow dome of course!). I would also like to mention Pierrette Bachofen-Dufrène, Evelyn Leitner, Astrid Som and Erika Witschi-Haldimann from the neighboring Audiometry and Christoph Wille, Franziska Conod, Béatrice Wyss, Christiane Kühn, Juliane Schweizer, and Peter Seligman. It was a great pleasure to work and to have coffee together.

I am also thankful to my friends Dainus Perednis who was there almost from the very beginning to the end (and of course his *alter ego* Maximus with his "strength and honor"), Raquel Orellana for her confidence, Valerie Mäder for introducing me in nuances of cat sailing, Ivan Vjunitsky for being so very Russian and not necessarily making my PhD writing easier, Louise Siemens and Regina Weber for giving me the idea of Mark Twain's "Adams Diary", Wanda Stelmachowicz for correcting my English texts, Saulius Vosylius, Josef and Aida Kurmanavicius, Katerina Grabovska and Sergiy Katrych as well as and many others.

Finally, I would like to thank Susan Botosh for her diplomacy, her caring and her confidence in me during moments when someone's mind was becoming "too beautiful" (fortunately it did not) especially in the very last period of my PhD writing, and my mother Mārīte Timma for her strength and patience during the passed nine years with me being so far away from home.

Oļegs Timms

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.

Contents

Chapter 1

Monday

"This new creature with a long hair is a good deal in the way. It is always hanging around and following me about. I don't like this; I am not used to company."

M. Twain

General introduction

1.1 Historical background

During the seventh week in the development of a human embryo, cells of the cochlear duct differentiate to form the spiral organ of Corti (the structure that bears the hair cell receptors responsible for transducing sound vibrations into electrical impulses) [B74]. Approximately on the 60th day of the development a human fetus starts to hear and in the third trimester of pregnancy it already is able to respond to audio frequency stimuli [B7]. It would not be surprising if the very first sound experienced had been the rhythmic beat of the mother's heart.

Human babies are learning about the speech a long time before they begin to talk. Newborn babies go beyond the actual physical sounds they hear, dividing them into more abstract categories. According to Patricia Kuhl, "babies love the sounds almost as much as they like milk" [B54]. At about seven or eight months, babies begin to babble and at the age of approximately one year the child starts to form its first words. Later on they put them together in two word combinations and simple three word sentences.

The vocabulary of an average English adult consists of approximately 75 000 words. The number of various different sentences that can be formed out of this vocabulary is indeed astronomical. The key for such rapid speech development is the ability to hear and a daily listening and speech training especially during the first years of life. Child's language prototypes for example are assumed to be formed between six and twelve months of age [B54]. After this the chaotic world of sounds becomes organized into complicated but coherent structures which are unique to their particular language, making hearing some of the distinctions of other languages almost impossible. If the hearing of a newborn is damaged or restricted the natural speech development becomes more complicated, and in the case of profound deafness without any particular training it can be nearly impossible.

However, there is always such a thing as "expectation" of sound in terms of pronunciation, loudness, quality, or other unique features. This expectation of how these characteristics of sound have to be, play an important role in our behavior, especially when considering the processing of new information or sound not conforming to our expectations or experience. In most of these cases this can contribute to confusions. Unfulfilled expectations can lead to great disappointment, an experience each has surely experienced at least once in our lives.

Humans are social beings. Sign or motion language and written language can be used as alternatives to the acoustical speech communication. Sign and motion communication is probably the oldest way of communicating. However, before humans learned to write, they spoke to each other. Even today, only some 1000 of the world's approximately 7300 languages and dialects have writing systems, *i.e.* for most of those besides sign language, there is no possible alternative communication available.

The German philosopher Immanuel Kant has written: "blindness separates people from things, deafness separates people from people" (it was rephrased in English by H. Keller, a woman who lost both her eye sight and hearing [B53]). Social life is very important for humans. Without society or collective organizations our civilization would never have developed like it has in the last five thousand years. Without communication, such organizations would be unthinkable. Without language, such communication would be impossible.

If for some reasons speech communication for an individual becomes impaired or limited, this can lead to a considerable discomfort or even loss of social connections and can result in an extensive deterioration in the quality of life. Hearing loss (especially profound hearing loss) may even lead to a complete lack of speech communication. To compensate for this handicap, people have developed sophisticated lip-reading skills and/or the use of sign or written language. These improved skills can help, but they never completely restore acoustical speech communication.

In addition to developed communication skills (lip reading and/or sign language), people have attempted to develop technical devices or discover medicinal remedies to cure hearing loss. A brief historical summary on methods to treat hearing impairment is given by Clark [B1] an additional information is also available on the webpage of the John Q. Adams Centre for the History of Otolaryngology, Head and Neck Surgery (http://www.entnet.org/museum/exhibits/earsandhearing_page1.cfm).

According to these materials, Archigenes[1], a physician of Rome in the first century BC, treated hearing disorders by sending loud, irritating noises to the ear through a tuba. However, the first references to mechanical hearing aids (hearing tubes) were documented in the 16th century. Prior to this, various remedies as the juice of cockroaches, sea baths, the use of leeches, or the fat of the green tree frog were used to attempt to cure hearing loss.

[1] Archigenes was a Roman physician from Syria. He is also reputed to have established the practice of dentistry by using drills to treat dental infections.

In 1902, Miller Reese Hutchinson developed the world's first practical electrical hearing aid. This hearing device used the transmitting potential of carbon. It caused a lot of static noise and distortions; however, the amplification of the acoustic signal was greater than that of the mechanical hearing devices of this period.

In the 1920's, the first vacuum tube hearing devices were introduced (for these systems an extra separate battery pack was necessary in order to heat the vacuum tubes). In the year 1947, the first transistor based hearing devices were made. At this point, hearing devices started to become not only more powerful but also increasingly smaller.

One of the first recorded electrical stimulations[2] of the auditory nerve was performed by Lundberg in 1950 [B51]. He used sinusoidal current stimuli during a neurosurgical operation. The patient could hear only noise. In the year 1957 A. Djourno and C. Eyries [B21] during an operation for cholesteatoma performed electrical stimulation of the acoustical nerve by placing electrodes directly on it. The patient was able to detect differences in pitch and to distinguish certain words such as "papa", "maman", and "allo".

Doyle *et al* [B28] were the first who in 1964 reported inserting an array of electrodes into the cochlea of a patient with total perceptive deafness [B1]. The totally implantable cochlear implant systems was produced initially by Michelson (1971) [B91], Clark (1977) [B14], and Hochmair (1980) [B59]. Since then, the hearing of profoundly deaf people can be partially restored. However, according to accounts from people who lost hearing subsequently to having a normal organ function, the electric hearing is not the same as natural hearing. Subjects with cochlear implants (CI) have to learn how to interpret the new kind of stimulation. They normally need rehabilitation to learn to interpret the CI signal and to recognise speech in it. In addition, cochlear implantation is expensive and requires relatively complicated surgery. However, for pre-lingual deaf children, CIs are one of the few opportunities for acquiring speech. If deaf children are not provided with cochlear implants by their fifth year of age, their chances of learning to speak normally and acoustically understanding speech are dramatically reduced.

[2] The first who electrically stimulated the auditory system was Italian scientist Alesandro Volta. He connected a battery of 30 or 40 "couples" to two metal rods which were inserted into his ears. When the circuits was completed he received "une secousse dans la tète" and a few moments later a noise like the boiling of thick soup [B1].

1.2 Motivation

In highly developed industrial countries, the percentage of profoundly hearing impaired individuals (pure-tone average hearing loss ≥90 dB) is approximately 0.05% (accounting to approximately 3500 persons in Switzerland). However, the number of annual performed cochlear implantations in Switzerland is not much larger than 100.

On the other hand, for profoundly hearing impaired people, which still have some residual hearing, high power conventional hearing devices can be used. However, in many cases, the auditory capacities of the profoundly hearing impaired are so limited, that they can hardly understand spoken speech. The use of their hearing device remains limited to hearing warning sounds or making acoustical contact within their environment.

Acoustical amplification in cochlear regions with no intact hair cells left is practically meaningless [B143]. For the group of subjects with large intra-cochlear damaged regions, conventional high-power hearing devices can be used only as support for lip-reading. Without the use of lip-reading, speech communication for this population group is very limited or even impossible. Average speech recognition of sentences for the profoundly sensorineural hearing impaired remains about 30% or less [B44].

Possible improvement of the given situation could be the development of special acoustical hearing devices, which would consider the specifically restricted residual hearing area of the profoundly hearing impaired subjects and provide special speech processing strategies with the purpose of improving speech perception. The motivation for the present study was to investigate the possibilities for the development of speech processing strategies which would implement transformations of the acoustical information necessary for speech perception into the residual hearing area of the profoundly hearing impaired subject. An important prerequisite for the present investigations is the development of a speech processing system based on the sinusoidal speech (SiSp) algorithm proposed by McAulay and Quatieri [B84]. The SiSp algorithm was chosen because it enables implementation of a broad range of spectral and temporal modifications of the acoustic signals. The background for the possible speech processing strategies is derived from the various existing speech processing schemes presently in use for cochlear implants in conjunction with studies with acoustical hearing devices on spectrally-reduced and transformed speech within the normal hearing and the hearing impaired subjects. Many speech processing algorithms for the profoundly hearing impaired subjects simplifies speech to only a few components such as the fundamental frequency or the frequencies of the first and the second formants, which did not yield the expected positive results. This is particularly the case for the sinusoidal voice algorithm mentioned in the study by Rosen *et al.* [B114], in which the speech signal was simplified to the fundamental frequency. Within the context of the present study, the meaningful spectral reduction of the speech signal was investigated by carrying out speech perception studies with spectrally reduced speech on normal-hearing adults.

Various possibilities were considered for the transposition of the inaudible spectral speech component into the residual hearing area of profoundly hearing impaired subjects. The spectral compression is assumed to be one of the most promising spectral processing methods which can be successfully used for speech transformation. In the development of the spectral compression schemes, psychoacoustic and auditory phenomena, including different

auditory frequency scaling, were considered. A strong decrease of frequency resolution capacities which normally occurs within the sensorineural hearing impaired is also taken into account. Within the present study, linear spectral compression on the auditory frequency scale is proposed. In comparison with conventional linear spectral compression, the proposed new compression scheme results in limited distortion of the processed speech signal, yet delivers a greater amount of spectral information within the residual hearing area of the profoundly sensorineural hearing impaired subjects.

Most spectral compression schemes described in the literature do not achieve large improvements in speech perception of profoundly hearing impaired subjects (see the overview in chapter 3). Most studies implemented only linear spectral compression with different compression ratios. In the present study, the potential use of the additional spectral information delivered by the various different spectral compression schemes was investigated on normal-hearing and sensorineural profoundly hearing impaired subjects.

Considering the wide range of profound sensorineural hearing impairments [B8], it is obvious that different subject groups will react differently on the same signal processing scheme. During the study with the hearing impaired subjects, some propositions for selection criteria were made, which should assist identifying potential subjects which might profit from the implemented signal processing algorithms. However, it is also to be expected that there will be no general signal processing scheme for acoustic hearing devices which will completely compensate for the handicaps of the profoundly sensorineural hearing impaired subjects.

Different unconventional signal processing schemes may even be implemented in hearing devices in combination with cochlear implants stimulating overlapping or non-overlapping regions of the cochlea. Several studies are currently investigating the simultaneous use of conventional hearing aids and cochlea implants.

To summarize, the main motivation for the present thesis is to make speech communication for people with profound sensorineural hearing loss easier, and thus improving their quality of life. The development of new speech processing strategies is a step towards better speech comprehension resulting in improved possibilities for speech communication of these people. Such improvements may provide for profoundly hearing impaired people increased possibilities for integration within the society, both socially and professionally.

1.3 Overview

1.3.1 Objectives and approach

The objectives of the present study are the following:

- Investigation of the possibilities of using signal processing schemes presently implemented in cochlear implants within future acoustical hearing devices.
- Investigation and implementation of different possibilities for spectral reduction. Examination of the limits of spectral reduction by means of speech recognition tests.

- Investigation and implementation of different spectral compression schemes enabling the compression of the original speech spectrum into the residual hearing area of profoundly hearing impaired subjects. Quantification of the most promising spectral compression schemes by means of speech perception tests with profound sensorineural hearing impaired subjects.
- Investigation and implementation of different temporal modifications of various speech segments involving temporal prolongation and temporal shortening of the acoustic signal while preserving its absolute duration. Quantification of the effects of temporal modification by means of speech perception tests.
- Quantification of the combined effects of the spectral reduction, the spectral compression and the temporal modifications on speech perception.

The various experimental signal processing schemes are implemented based on the sinusoidal speech algorithm of McAulay and Quatieri [B84].

1.3.2 Contributions

One contribution of the present work is the implementation and modification of the signal processing system based on the sinusoidal speech algorithm proposed by McAulay and Quatieri [B84]. The system was used for implementing various promising schemes of speech signal processing for profoundly hearing impaired subjects.

The modifications of the baseline SiSp algorithm involved implementation of various spectral component selection and spectral manipulation algorithms, and different signal reconstruction methods enabling significant simplification of the original reconstruction algorithm. Four different methods can be used for the reconstruction of the processed signal reconstruction: the original polynomial phase interpolation, the inverse fast Fourier transformation IFFT, the "Teilton" (partial tone) algorithm [B95], and the sinusoidal noise speech (SiNoSp). The SiNoSp was introduced as an alternative signal generation mechanism, and it basically implements a switch between a conventional sinus-wave generator for the generation of low frequency tones and the narrow frequency band noise generator for the generation of high frequency spectral components. In addition to that, a frequency interpolation method within the FFT spectrum was also developed, allowing an increased precision of the spectral analysis of the signal processing system.

The signal processing strategies proposed for profoundly hearing impaired are partially based on ideas from cochlear implant signal processing. In particular, various approaches of spectral compression, spectral reduction, and temporal modification are proposed and implemented.

A major contribution of the present work is the introduction of linear spectral compression performed on the auditory frequency scale. In addition, some propositions were made for the selection of suitable candidates which might profit from the new spectral compression scheme.

All proposed signal processing schemes were thoroughly evaluated in psychoacoustic speech tests with normal hearing and profoundly hearing impaired subjects.

The studies carried out in the context of this thesis demonstrated that the sinusoidal speech algorithm can be successfully used for the different speech signal manipulations including spectral reduction, linear and non-linear spectral compressions, and speech segment duration manipulations. The developed signal processing system can also be used as a signal processing tool for other studies which involve special spectral and temporal manipulations on the audio signals.

1.3.3 Thesis Outline

Different possibilities for processing the speech signal for the profoundly hearing impaired and their possible impacts on the speech signal are discussed in chapter 2. This chapter also describes the different auditory frequency scales.

In chapter 3, a literature review on the state of the art in spectral compression and spectral reduction is given, involving a description of different spectral compression and spectral reduction hearing devices along with the results of studies using hearing-impaired subjects. Prior work on temporal modifications in speech processing is described as well.

In chapter 4, the developed signal processing system based on the sinusoidal speech analysis-synthesis technique is described in detail.

In chapter 5, a study on spectral reduction and speech perception with normal hearing subjects is described. This study involves the determination of the necessary spectral components per time unit ratio for nearly 100% speech perception in sentences.

In chapter 6, a comparative study on linear and non-linear spectral compression with normal hearing subjects is described. Spectrally reduced and compressed low-pass filtered speech comprehension is also compared with low-pass filtered speech comprehension.

In chapter 7, a comparative study on the linear and non-linear spectral compression with severe to profound hearing impaired subjects is described. Influences of the linear and the non-linear spectral compression on the different speech segments for the hearing impaired subjects are analyzed in detail.

Chapter 8 describes a study on various temporal modifications of different speech segments, which is performed with normal hearing subjects.

Finally, chapter 9 summarizes the contributions of the present thesis and gives some recommendations for the future improvements and development.

Chapter 2

Tuesday

"Been examining the great waterfall. It is the finest thing on the estate, I think. The new creature calls it Niagara Falls – why, I am sure I do not know."

M. Twain

Different aspects of speech processing for profound hearing impairment

2.1 Overview

In this chapter different possibilities for speech processing for profoundly hearing impaired subjects are discussed. Taking into account different non-linear spectral compressions including linear spectral compressions on an auditory scale, various auditory frequency scales are briefly described.

Different speech perception aspects which are used in the following chapters are briefly explained, and a description of the consonant and vowel characteristics of the German language is given.

Finally, changes in vowel and consonant spectra caused by linear spectral compression along with spectral compression on auditory frequency scales are also discussed.

2.2 Hearing and hearing impairment

The auditory system delivers environmental acoustic information to the brain. The physical parameters of sound interpreted by the auditory system are the sound pressure level and the frequency. These two physical parameters as well as their combination and their temporal changes are carriers of information. The frequency versus sound pressure level diagram with the average threshold of hearing and threshold of pain functions for the intact auditory system are given in Fig 2.1. The sound pressure difference between the threshold and the upper comfortable limit of hearing is called the dynamic range. The spectral region between 16 Hz and 22 kHz is called the audible frequency range.

Different impairments or damages to the auditory system can cause significant restrictions within the auditory area. Any auditory range restrictions decrease the potential

amount of acoustical information which would be provided to the brain are called hearing impairments. Two general types of hearing impairment can be classified dependent upon the processing structure where the impairment occurs. A conductive hearing impairment affects only the outer and the middle ear. A sensorineural hearing impairment is caused by damage to the cochlear and the higher stages of the auditory pathway. The conductive hearing impairment can be compensated for by medical or surgical treatment, or by the frequency-specific increase of sound energy (linear amplification). The sensorineural hearing impairment involves more complex mechanisms which can affect sound processing in more complex ways.

The perception of speech for individuals with profoundly limited auditory capabilities is so restricted that speech conversation is not possible without lip-reading. A profound hearing impairment is defined as an average hearing loss of more than 90 dB HL for the pure-tone average frequencies (PTA: 0.5; 1.0; and 2.0 kHz) [B50]. In many cases, hearing loss is stronger for the higher frequencies than for the lower frequencies, and often involves also a reduced frequency resolution.

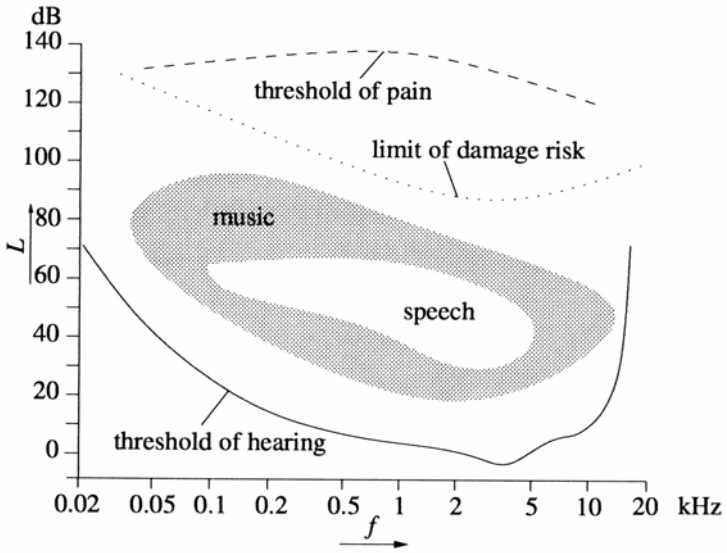


Fig. 2.1 Frequency versus sound pressure level diagram (auditory area) for speech and music [B148] [B147].

The present study investigates speech signal processing for the moderately-severe to profoundly sensorineural hearing-impaired subjects. Sensorineural hearing impairment is a combination of restricted dynamic and audible frequency range. The present work concentrates mainly on the effects caused by the restricted audible frequency span, considering also the possible effects of the limited frequency resolution. It has been assumed that sounds in a frequency region with a dynamic range smaller than 10 dB are practically inaudible. A similar situation occurs within subjects with severe to profound steeply sloping hearing loss in the high frequency area and normal to mild hearing impairment for the frequencies below 0.5 kHz. However, spectral information is very important for proper speech comprehension [B33]. Assuming that the reason for the limited speech perception is

insufficient in-coming spectral information, one main goal of the present investigation is to provide the individuals with additional spectral information to improve their speech perception capabilities [B138].

The following strongly simplified model of the sensorineural hearing-impaired subject is assumed:

1. The auditory frequency range of the subject is limited by a frequency threshold of 2000 Hz (best case) or 1000 Hz (worst case). All frequencies above this threshold are inaudible (i.e. dynamic smaller than 10 dB). However, the frequencies, which below this threshold frequency, are assumed to be audible, and their dynamic ranges sufficient for perceiving the speech signal.

2. The frequency resolution within the audible frequency range is limited and therefore, much worse than for cases of normal hearing [B134]. The broadening of the auditory filters is the reason for the decreased frequency resolution which can be observed within the profound sensorineural hearing impairment population [B93;B94]. The total ratio of the auditory filter broadening is assumed to be unknown.

3. Additional masking effects are to be expected, considering the phenomenon of the loudness recruitment and the fact, that all of the potential hearing impaired subjects are hearing device users, which normally present the sound at the higher than normal loudness levels.

As mentioned above, this hearing impairment model does not consider the limited dynamic range of the residual hearing area or any worsening of the temporal resolution [B29], [B30], [B42], [B136], [B52], [B137]. The temporal resolution can, however, significantly correlate with speech intelligibility [B136] and normally occurs along with the profound sensorineural hearing impairment.

This simplified model of the sensorineural impairment can be compared with a physical system of a fluid flowing through a funnel (Fig. 2.2), where the fluid represents acoustic information which is sent to the brain in case of the non-disturbed hearing, and the funnel represents the sensorineural hearing impairment which restricts the amount of the conducted information per time unit (to simulate the broadening of the auditory filters, the fluid must expand). So if the incoming fluid quantity per time unit is larger than the amount of the fluid, which can be conducted through the funnel, then it leads to the overflow of this physical system and causes the loss of the overflowed liquid. Approximately the same happens with the spectral information of an acoustic signal which is conducted into the impaired inner ear – it simply gets lost because it can not be transferred.

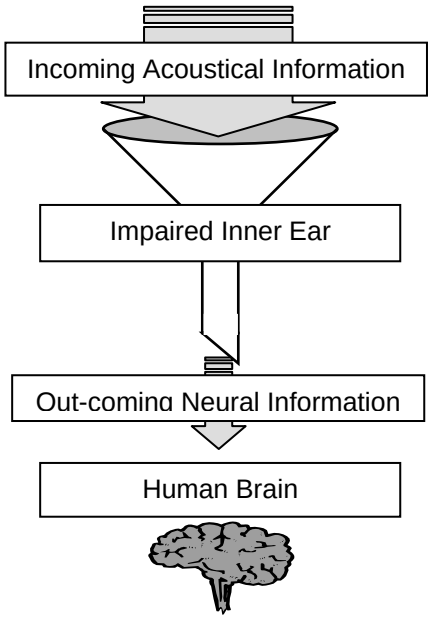


Fig. 2.2 Simplified scheme of problems caused by sensorineural hearing impairment.

The only manner possibly allowing the funnel system to work normally would be to reduce the amount of incoming liquid. A possible way to improve speech perception of the profoundly hearing-impaired subject is to reduce the amount of incoming information intelligently and provide for only the speech perception. The question remains which part of the incoming acoustic information can be reduced.

2.3 Discussion of different signal processing aspects

2.3.1 Spectral reduction

One possibility for reducing the acoustic information is also presently used in the CI signal processing schemes. It is the division of the original spectral range into the given number of the spectral channels, and the use of the selective stimulation of different parts of the cochlea, which corresponds to the desired channel number. The number of channels and their physical characteristics such as place and width can differ from system to system. In the case of CI the number of channels and the electrical stimulation rate can be chosen. The electrical stimulation sequence plays a very important role as soon as only one electrode can be stimulated simultaneously. In cases of acoustic stimulation of the cochlea, the stimulation of all channels can and normally is performed simultaneously. However, in the case of impaired hearing the number of provided spectral channels per time unit can be critical due to the limited spectral resolution. A reduction of the used spectral component[3] per time unit ratio will result in decreasing overlapping masking effects. Nevertheless, up to certain spectral

[3] The terminology differences between spectral components, spectral lines and spectral channels are discussed in Chapter 3. In the further text, the term “spectral component” is used as a general notion.

reduction stages, further reductions can cause a significant decrease of speech signal intelligibility. If temporal resolution of the hearing-impaired subjects could be preserved, the alternative method, which would possibly obey the loss of the frequency resolution capacities, would be the high-ratio acoustical stimulation of only one spectral component per time unit. Unfortunately, usually the temporal resolution of the hearing impaired subjects is also strongly reduced.

Chapter 5 will describe studies using different spectral component per time unit ratios, which correspond to differing acoustical stimulation ratios and synthesis frame lengths, and an increasing number of spectral components per synthesis frame. The purpose of these studies was to investigate the minimal spectral component per time unit ratio which is required for near to 100% speech perception for normal-hearing adults, respectively the maximally allowed amount of spectral reduction for sufficient speech perception. As soon as the sinusoidal speech algorithm [B84] was used for speech signal modification, it was assumed that the carriers of the spectral information are spectral maxima (see chapter 4 and 5).

2.3.2 Spectral transposition

Considering the limited residual auditory area of the profound sensorineural hearing impaired subject, the selected spectral components, which are assumed to be important for speech perception, have to be transposed into the residual hearing area in order to make them audible again. However, spectral information transposition into other spectral regions can cause serious confusions leading to the total loss of intelligibility of the processed speech signal. Too large or non-linear spectral transpositions are believed to destroy speech perception [B86], [B87]. There are many possibilities of the spectral component transposition. However, basically they all involve a re-location of spectral components consisting of two mathematical operations: spectral shifting and spectral compression. These two operations can have linear or non-linear character on the physical (FFT) frequency scale. The spectral manipulation parameters can be dependent on time or specific speech segment characteristics. An overview on the existing spectral transposition systems is given in chapter 3. The most investigated and practiced method of the spectral transpositions is the linear spectral compression on the linear physical frequency scale. The obvious reason for this is its simplicity of implementation, requiring a linear slowing down of the acoustic signal record. Unfortunately, the spectral compression ratios which are necessary to compress the speech spectrum up to 5 kHz within the spectral range of 2 or 1 kHz, remain much too large. After applying such large linear spectral compressions, the speech signal totally loses its intelligibility. Acceptable spectral compression ratio values, however, do not significantly increase the amount of the provided spectral information of speech.

One possible alternative would be the use of time- or speech segment-dependent spectral compression ratios in order to transpose the high frequency speech components into the lower frequency areas only if they are present and important for the actual speech segment perception [B89], [B56], [B116]. This strategy necessarily requires the use of a speech segment detector as it involves different spectral compression ratios for the voiced and unvoiced speech parts. The positive aspect of this spectral compression scheme is the possible low spectral compression ratio values for the voiced speech segments, which are

highly sensible on the spectral compression. However, the switching between different spectral compression ratios and the use of the same spectral regions for principally different kinds of spectral information can be very confusing for the potential user [B5]. Another possible aspect, which can negatively effect this signal-processing scheme, is the indefinite acoustical environment state, from the speech segment detector point of view, which occurs within multi-talker environments or in situations with great amount of background noise.

A plausible alternative for the afore mentioned linear or temporary linear spectral compressions could be the implementation of non-linear spectral compression on the physical frequency scale, or a combination of spectral compression and spectral shifting [B117], [B61], [B116].

The present study proposes the use of auditory frequency scales to provide for spectral compression. The various auditory scales that can be used for the rescaling of the FFT frequency scale according to human sound perception, along with the impact of spectral compression on speech signals will be discussed later in the present chapter. All spectral compressions on the auditory frequency scale, including those with constant compression ratios, are non-linear on the physical (FFT) frequency scale. However, it will be shown that such non-linear spectral compressions preserve the naturalness and intelligibility of the speech signal up to certain compression ratio values. The positive aspects of spectral compressions on the auditory frequency scales are their relatively small effective spectral compression for the low frequency area and the increasing effective spectral compression for the high frequency areas. This kind of spectral compression enables transposition of the whole speech frequency range into the 2 kHz frequency range using relatively small spectral compression ratios on the auditory frequency scales. The main disadvantage of these spectral compression schemes is the partial non-linear spectral transposition of the voiced speech segments.

The question of the ability of the profoundly hearing impaired subjects to make use of the additional high frequency information transposed into the lower frequency areas remains open [B55], [B60], [B10], [B121], [B47] [B132]. In chapters 6 and 7, studies using linear spectral compressions on the FFT scales and linear spectral compression on the auditory frequency scale with normal hearing and hearing-impaired adults are described.

As mentioned above, the use of the transposed spectral information by the brain is one of the most important aspects in the applying of different speech transposing signal processing schemes. The main requirement of all spectral manipulations on the speech signal is the preservation of speech comprehension. This remains arduous as in most cases spectral compression or spectral shifting can cause the speech signal to sound unnatural or to totally lose intelligibility. In this case, all endeavors to provide better speech perception fail, even if, theoretically, sufficient information about spectral components is delivered to the brain.

This situation is mostly similar to that of foreign language learning. In the case of the normal hearing person, listening to an unknown language, the individual is obviously provided with sufficient acoustical information to understand the speech, but still cannot interpret it due to the lack of knowledge and training. In the beginning, the individual probably does not understand anything and may become frustrated with the confluence of unintelligible sounds. An excellent example of this is given by Mark Twain in “A Trump Abroad” [B135], wherein he describes the German language. As well, the social situation within the new language milieu can be comparable to the experiences of hearing-impaired

persons. Additionally, without practicing and using the newly acquired language, it would be nearly impossible to maintain it. Even languages one speaks fluently can be forgotten if not used for an extended period. The brain has to learn how to interpret the new sound patterns and to use them continuously.

For the similar situation applied to hearing-impaired listeners the example of cochlear implants can be given. Most of the CI implanted subjects have to learn to interpret the new sound of the speech signal as a language. This is a process which often requires effort from the cochlear implanted persons. However, motivation for this effort is the negative alternative of using only lip-reading skills to enable speech communication.

The fact that the hearing impaired individuals have an intact brain, which is capable of dealing with and interpreting different information, is very important. In many cases, hearing devices or signal processing systems are considered separately, and the fact that the human brain is still present and requires a certain learning time to interpret incoming information is simply overlooked. In fact, it means higher requirements for the signal processing system due to the additional condition of attempting to preserve the high acoustical naturalness of the speech. In this way, nowadays speech processing systems attempt to choose the simplest way for the brain and take over most of the signal interpretation difficulties. However, the human brain is a highly developed biological system capable of dealing with and adapting to more complex and different information presentation schemes as any presently developed signal-processing system. One example of this is the Morse code. Persons, who are specially trained in this code, can achieve remarkable code interpretation and conversation tempo. This, however, does not mean that signal processing quality can be worse or that we should work with the Morse code system. The profoundly hearing-impaired subjects' adaptation to the different signal-processing strategies or rehabilitation should be relevant considerations for the development of the special signal processing systems for profoundly hearing impaired subjects. There are also various studies on different spectral manipulations for the hearing impaired subjects (and the present one is not an exclusion), where speculations about possible learning effects and possible improvement of speech perception are made [B26], [B114], [B47], [B17], [B86]. However, in the world of commercial hearing devices, it is difficult to develop a complete system and hope that potential benefits for customers will appear after a year or even later.

2.3.3 Temporal modifications

The temporal modifications of the speech signal such as temporal compression or temporal expansion are often considered together with linear spectral compression due to similar production techniques [B15]. Our overall experience shows that slower speaking is more easily understood than faster speaking. It is to be expected that the slowing down of the speaking tempo could improve speech perception also for hearing-impaired subjects. However, the main difficulty with the implementation of a slowing-down technique in a hearing device is caused by requirements of the simultaneous speech processing and transmitting, which requires the same input-output signal durations. If this is not fulfilled, enormous delays may occur in the transmission of the acoustic signal. This requirement automatically sets the limitations on all time stretching or compression constants which can be applied. The slowing down of one of the speech segments requires the speeding up of

other speech segments in order to keep the input-output delay time sufficiently small. A second difficulty is caused by the continuous change of the speech rhythm itself, which requires the continuous adaptation of the time expanding and compression ratio values. This, however, requires a special speech rhythm detector, for which the appropriate accompanying mechanism and technique is still unclear.

2.4 Auditory (psychophysical) frequency scaling

Besides the logarithmic frequency scale, there are three other auditory-based frequency scales: the Spectral-Pitch Increment (SPINC) scale; the Critical-Band Rate (BARK) scale; and the Equivalent Rectangular Bandwidth (ERB) scale.

The logarithmic frequency scale is rather from traditional usage in acoustics. It is based on the human sensation of the doubled fundamental frequency as a harmonic structure [B127].

The SPINC scale is based on spectral pitch increment measurements for two sinusoidal tones with the duration of at least 200 ms [B126], [B127]. According to experimental data collected from various studies, the empirical equation for a minimal frequency resolution interval is given by the following:

$$\Delta f_D(f) = 1 + \left(\frac{f[Hz]}{\sqrt{2} * 1000} \right)^2, \tag{2.1}$$

where Δf_D is the minimal frequency resolution and f the measured frequency in *Hertz*. One *spinc* is defined as the interval of minimal frequency resolution. The SPINC function itself is given by the equation:

$$\Phi(f) = const * \arctan\left(\frac{f[Hz]}{const}\right), \tag{2.2}$$

where $\Phi(f)$ is the SPINC function in *[Spinc]* units, f is the physical frequency in *[Hertz]*, and $const = \sqrt{2} * 1000$.

The BARK and the ERB scale are based on the fact that auditory filter width or critical-bandwidth corresponds to a certain length along the Basilar membrane [B148]. The differences between these two scales are the differences in assumed interval length and so called “cut off” frequency. For frequencies greater than the “cut off” frequency the auditory filter bandwidth increases with increasing center frequency. The Basilar membrane interval length corresponding to the auditory filter lengths for BARK- and ERB- scale is 1.3 mm and

0.86 mm respectively. The corresponding "cut-off" frequencies are 500 Hz for the BARK and 100 Hz for the ERB-scale, respectively [B94], [B93], [B75].

For the BARK scale the frequency resolution interval is given by the equation:

$$\Delta f_D(f) = 25 + 75 * \left(1 + 1.4 * f[kHz]^2\right)^{0.69}, \tag{2.3}$$

where Δf_D is the minimal frequency resolution and f the measured frequency in *[Hertz]* [B94]. An empirical equation describing the BARK – scale is given by:

$$z(f) = 13 * \arctan\left(0.76 * f[kHz]\right) + 3.5 * \arctan(f[kHz]/7.5)^2, \tag{2.4}$$

where *z(f)* is the BARK function in *[Bark]* units and *f* is the frequency in *[Hertz]* [B94].

The frequency resolution interval for the ERB scale is given by the equation:

$$\Delta f_D(f) = 24.7 * \left(4.73 * f[Hz] + 1\right), \tag{2.5}$$

where Δf_D is the minimal frequency resolution and f the measured frequency in *[Hertz]* [B94]. The ERB rate according to Moore and Glasberg is given by the equation:

$$r_{ERB}(f) = 21.4 * \log_{10}\left(4.73 * f[kHz] + 1\right), \tag{2.6}$$

where $r_{ERB}(f)$ is the ERB function and *f* is the physical frequency in *[Hertz]* [B94].

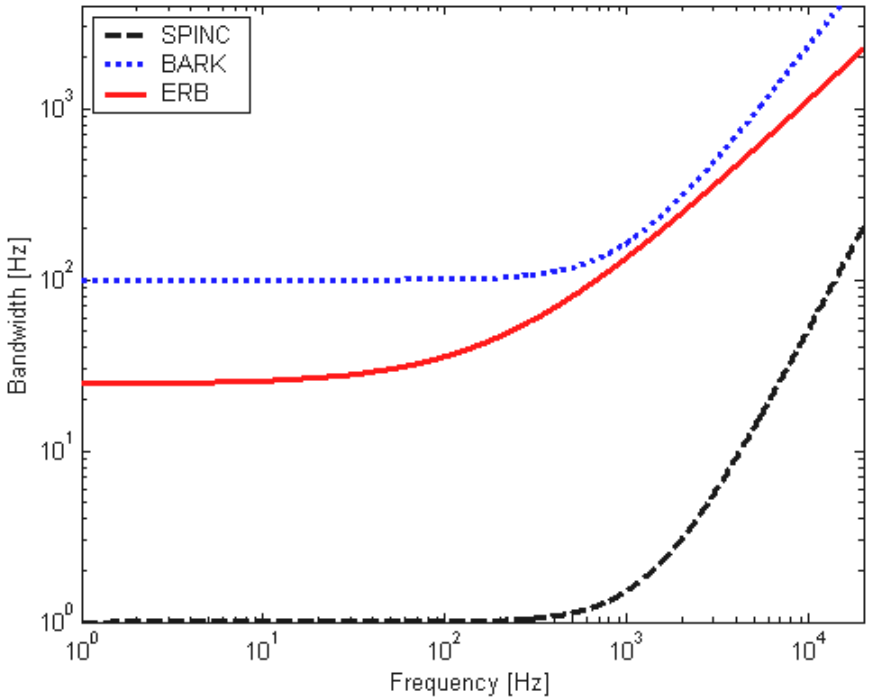


Fig. 2.3 Comparison of calculated frequency resolution intervals (auditory filter bandwidth) plotted versus center frequency of the respective interval (auditory filter) obtained using SPINC-, BARK- and ERB-scales.

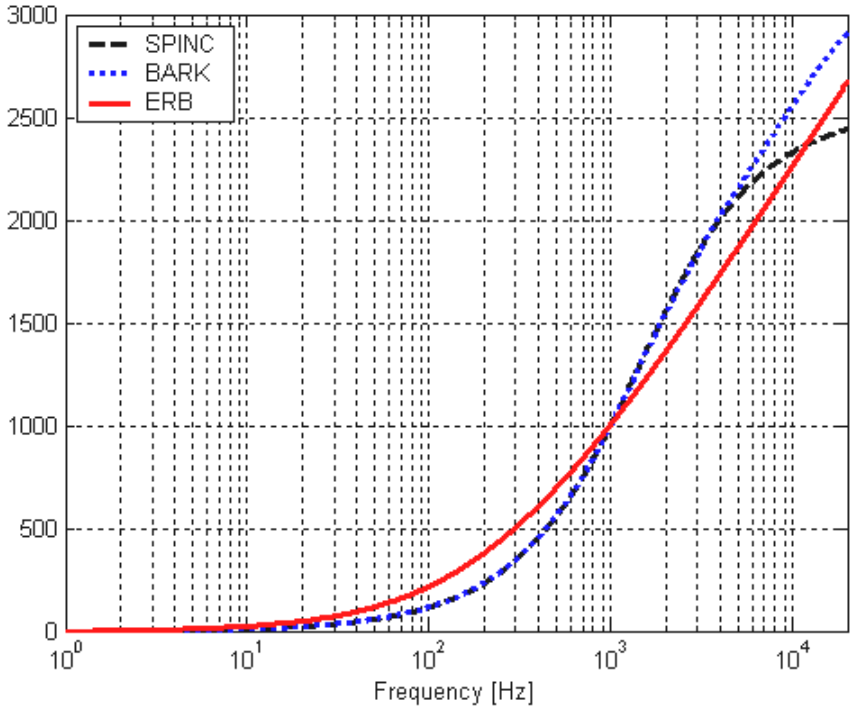


Fig. 2.4 Comparison of three psychophysically based frequency scales (SPINC-, BARK- and ERB). All scales are rescaled so that their values of 1 kHz is equal to 1000 rescaled units.

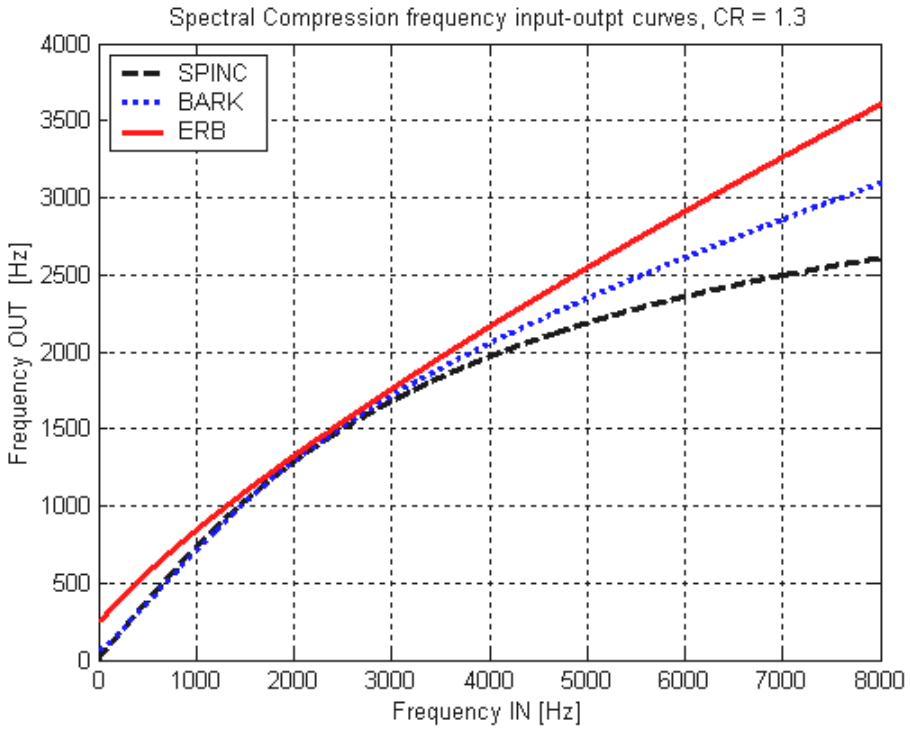


Fig. 2.5 Comparison of calculated spectral compression frequency input-output curves based on three different auditory frequency scales (SPINC-, BARK- and ERB). Spectral compression was calculated using a compression ratio of CR=1.3.

For all three psychophysically-based frequency scales, the frequency resolution intervals or auditory filter bandwidths are shown in Fig. 2.3. Note that the calculated frequency resolution intervals for the SPINC-scale are significantly smaller than those of the ERB and BARK scales below 300 Hz where the SPINC resolution is about 1 Hz.

In Fig. 2.4 all three psychophysically-based frequency scale functions are rescaled so that their function values (bandwidth) at 1 kHz is equal to 1000 units. The rescaled SPINC and BARK scale bandwidth differ from each other significantly only for frequencies higher than 4 kHz. ERB- function grows steeper for low frequencies than SPINC- and BARK-functions. However, for frequencies over 300 Hz the BARK and SPINC functions become steeper than the ERB-function and cross it by a frequency of 1000 Hz.

A comparison plot between the spectral compression frequency input-output curves for three different psychophysical frequency scales using a spectral compression ratio of CR=1.3 is given in Fig. 2.5. It can be observed that spectral compression on the SPINC scale is stronger especially in the higher frequency area. The spectral compression on the ERB scale, on the other hand, is more intense for low frequencies. Differences between spectral compression on the BARK scale and the SPINC scale become significant for input frequencies higher than 3000 Hz.

Spectral compression on the SPINC scale provides the strongest spectral compression in the higher frequency area, do to this the SPINC scale was chosen for further spectral compression experiments on the auditory frequency scale.

2.5 Essential speech cues and their susceptibility to spectral manipulations

2.5.1 Vowels

Vowels are usually produced by voicing (If the vocal folds are brought together during exhalation, the escaping air causes them to vibrate, alternately opening and closing the glottis. The air is thus released in brief, repetitive bursts, generating a complex tone called voicing [B11]). Exceptions are whispered vowels in which the sound source is glottal frication. The sole resonator for voicing is the oral cavity. The frequency spectrum of vowels is characterized by peaks at certain frequencies. These peaks are called formants and are numbered starting from the lowest frequency peak in growing order. The first three formants are the most important ones for vowel identification [B103], [B66], [B22], [B32], [B22], [B67], [B57], [B62].

There are 17 vowels in English, and 15 German vowels. Besides classification according to the first and second formant frequency place, vowels can be classified depending on the tong position. In the present study German vowel classification according to their formant frequency placing was used. The first formant (F1) and the second formant (F2) frequency place classification is given in table 2.1. The German vowel classification based on formants is given in table 2.2.

F1	F2	Group
<320 Hz	<1100 Hz	1
320-500 Hz	1100-1400 Hz	2
500-650 Hz	1400-1700 Hz	3
>650 Hz	>1700 Hz	4

Tab. 2.1 First and second formant location for German vowels [B20].

Phonemes	Nr	**F1**	**F2**
I (Gr***ie***s)	1	1	4
1 (g***i***lt)	2	1	4
E (s***eh***t)	3	2	4
@ (B***ä***r)	4	3	4
4 (G***e***ld)	5	3	4
A (P***aa***r)	6	4	2
# (g***a***lt)	7	4	2
O (Br***o***t)	8	2	1
3 (G***o***ld)	9	2	1
0 (b***ö***se)	10	2	3
: (St***ö***cke)	11	2	3
U (Br***u***t)	12	1	1
6 (L***u***st)	13	1	1
Y (k***üh***l)	14	1	4
! (St***ü***cke)	15	1	4

Tab 2.2 German vowel classification according to their features (vowel nomenclature according Dillier [B19]).

The approximate locations of German vowels as defined in table 2.1 on the F1 versus F2 plot are given in Fig. 2.3. They form the so called vowel triangle plots. According to their duration, vowels can be classified into long and short vowels. The short vowels (SV) span a smaller F1 versus F2 triangle inside the larger long vowel triangle (LV).

Vowel identification by subjects with rather moderate cochlear hearing loss is often rather good [B94], [B44]. Possible reasons for this are the usually large spectral differences between individual vowels and their temporal cues which might be used to compensate for the effects of reduced frequency selectivity.

However, if the residual hearing area of a hearing impaired subject is very limited then it can happen that the only information that can be received is the first formant. For example in case if the residual hearing area is limited to ~1500 Hz, then the vowels in the groups {/:/ (St**ö**cke), /4/ (G**e**ld) and /@/ (B**ä**r)}, {/Y/ (k**üh**l) and /I/ (Gr**ie**s)}, and {/!/ (St**ü**cke) and /1/ (G**i**lt)} are very difficult to distinguish because their first formant frequencies are situated closely together, with the second formant already inaudible. The situation becomes even more dramatic if the residual hearing area is limited to lower frequencies. In this case, only very few of the vowels can be separated and identified.

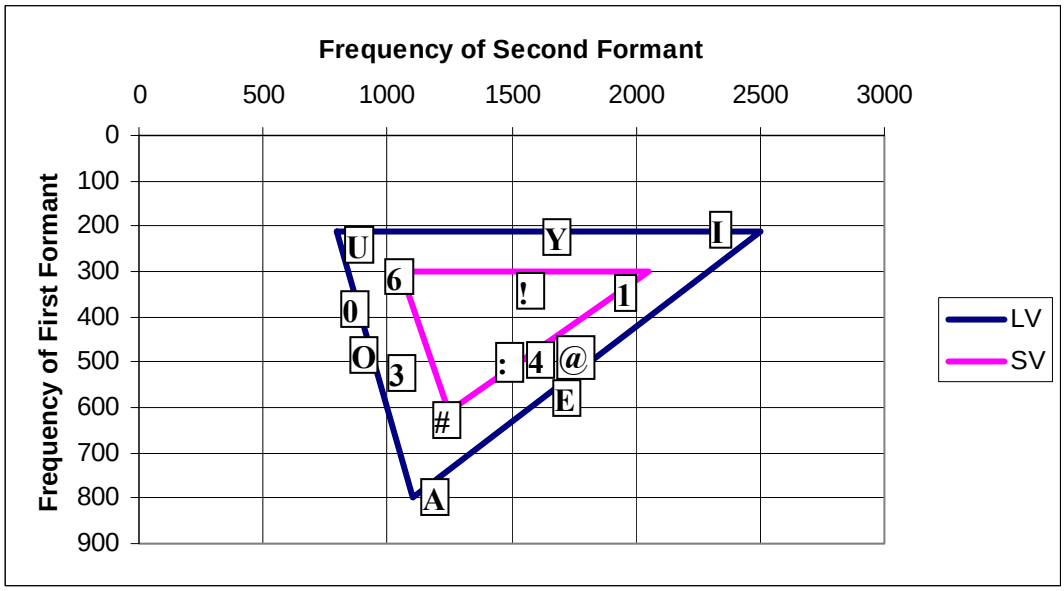


Fig. 2.6 Location of German vowels on F1 versus F2 plot according to nomenclature given in table 2.2. The short vowels (SV) are spanning a smaller F1 versus F2 triangle inside the larger long vowel (LV) triangle.

2.5.2 Impact of frequency compression on vowel perception

By using any kind of spectral compression, vowel triangles are shifted to the lower frequency area and additionally compressed. This circumstance causes overlapping of original and compressed vowel triangles. At the points of the intersections and overlapping of the original and compressed vowel triangles, different vowel identification confusions can be expected.

An example of vowel triangle transformation using linear spectral compression with CR=1.3 is shown in Fig. 2.7. (Linear spectral compression with CR=1.3 is used in studies described in chapter 6 and 7.) The compressed vowel triangles lie completely inside the 2000 Hz frequency border. Hence, for subjects with residual hearing up to the 2000 Hz region, linear compression (with CR=1.3) can theoretically improve vowel identification scores.

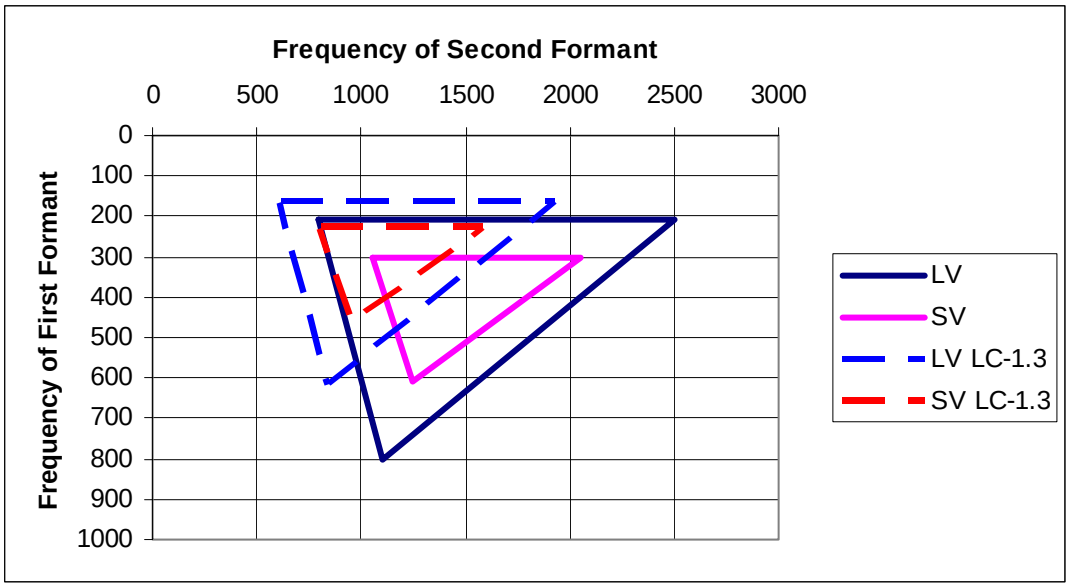


Fig. 2.7 Non-compressed long and short vowel triangles (LV and SV) and linearly spectrally compressed (CR=1.3) vowel triangles (LV LC-1.3 and SV LC-1.3) on the F1 versus F2 chart.

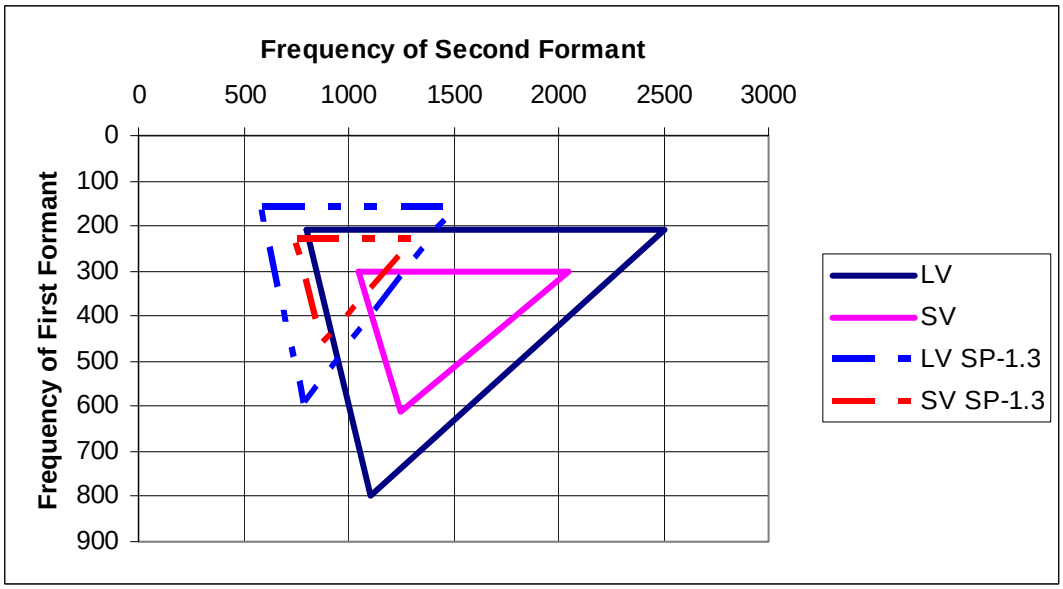


Fig. 2.8 Non-compressed long and short vowel triangles (LV and SV) and on auditory frequency scale spectrally compressed (CR=1.3) vowel triangles (LV SP-1.3 and SV SP-1.3) on the F1 versus F2 chart.

However, identification confusions can occur for example between /!/ (St**ü**cke) and spectrally compressed /1/ (g**i**lt); /3/ (G**o**ld) and spectrally compressed /A/ (P**aa**r), and /U/ (Br**u**t) and spectrally compressed /6/ (L**u**st). Theoretically, however, after certain acclimatization and training periods, these confusions should disappear. On the other hand, due to narrowing of formant frequency placing, other vowel consonant group confusions might simultaneously occur. Hence, confusions inside the /0/ (b***ö***se), /O/ (Br***o***t), /3/ (G***o***ld) and /U/ (Br**u**t) group, and the /:/ (St***ö***cke), /4/ (Ge**l**d), /@/ (B**ä**r) and /E/ (s***eh***t) group for example are possible.

Spectral compression on the auditory frequency scale causes non-proportional vowel triangle transformations (see Fig 2.8). If the lower frequencies are compressed approximately by the same amount as if under linear spectral compression conditions, then higher frequencies are compressed more. A comparison between vowel triangle displacements caused by linear spectral compression and spectral compression on the auditory frequency scale is shown in Fig. 2.9.

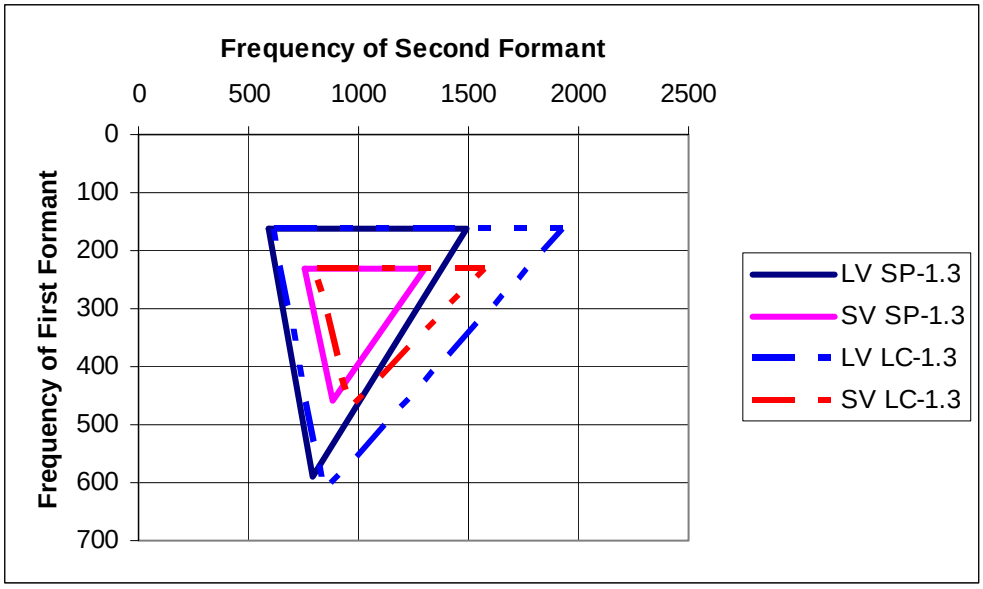


Fig. 2.9 Linearly spectrally compressed (CR=1.3) long and short vowel triangles (LV LC-1.3 and SV LC-1.3)and on auditory frequency scale spectrally compressed (CR=1.3) vowel triangles (LV SP-1.3 and SV SP-1.3) on the F1 versus F2 chart.

The theoretical advantage of spectral compression on the auditory frequency scale is that even if the hearing impaired subject's residual hearing area is limited to say approximately 1500 Hz, information on the second formant is still provided. This circumstance, however, is not only beneficial. The additional narrowing of the vowel triangle results also in narrowing of certain vowel groups, especially in the spectral area close to 1500 Hz. Therefore the same possible vowel confusions which can occur when using linear spectral compression, the confusion group of /0/ (b**ö**se), /O/ (Br**o**t), /3/ (G**o**ld) and /U/ (Br**u**t) e.g. can now include also /Y/ (k**üh**l), /I/ (Gr**ie**s), /!/ (St**ü**cke) and /1/ (g**i**lt). This potential problem is particularly acute if the broadening of auditory filters by sensorineural hearing loss is taken into account.

The highest original second formant position is approximately 2500 Hz. For transforming this spectral range into, e.g. 1000 Hz, linear spectral compression with a compression ratio of 2.5 would be necessary. For transposition of the same range into the 2000 Hz spectral area, linear spectral compression with CR=1.25 is required. Using spectral compression on the auditory frequency scale e.g. on the SPINC scale, the same conditions would be fulfilled with CR= 1.72 and 1.11 respectively. From the studies described in the literature, it is known that linear spectral compressions with compression ratios larger than 2 destroy speech intelligibility completely. It is possible that this reaction is caused by the destruction of vowel comprehension, in particular through the enormous lowering of the first formant frequency. The advantage of spectral compression on the auditory frequency scale is that it affects frequencies in the lower spectral area less than higher frequencies.

Hence, when applying spectral compression schemes, a compromise between providing the subject with spectral information and at the same time loosing this information due to spectral smearing must be found.

2.5.3 Consonants

Consonants can be divided into five categories: vowel-like consonants, nasals, fricatives, stop-plosives, and affricates. Consonants contain generally higher frequencies (1000-5000 Hz) and are normally weaker in loudness level than vowels. There are 25 consonants in the English language and 19 in German. English consonant [B19] classification according to their features was performed by Miller and Nicely [B92]. German consonant classification according to their features such as voicing (VOI), nasality (NAS), sonorance (SON), frication (FRI), sibilance (SIB), placing (PLC) and manner (MAN) is given in table 2.3. In the following passage the consonant classification groups are described in detail.

The source of vowel-like consonants is voicing. The oral cavity is normally involved as sole resonator. This kind of consonants is normally weaker than vowels and they are short lived. However, the spectrum of vowel-like consonants shows a formant structure which depends on the oral cavity parameters. These formants have a rapidly changing character.

For the nasals, the resonator cavity is the nasal cavity with the mouth acting as a closed side cavity. As for vowel like consonants, the source of nasal consonants is voicing. In the spectrum of nasals, low formants of fixed frequency are predominating. The differences between nasals are defined in the formant changes.

Fricatives can be divided into voiceless and voiced. Frication is the sound source for the voiceless fricatives. However, for voiced fricatives the sound source is both frication and voicing. In both cases, the oral cavity serves as a sole resonator. The addition of voicing by voiced fricatives results in a periodicity in a random noise. This results in a mix between random noise and complex tones.

Stop-plosives have also two variants – voiceless and voiced. The source of the voiced stop-plosives is voicing and stop-plosion. The oral cavity is the sole resonator. The main acoustic features of the stop-plosives are a period of silence following by the sudden appearance of energy across a wide range of frequencies (voiceless stop-plosives), and a brief period of frication with clear formant structure and rapidly increasing amplitude (voiced stop-plosives).

There are also voiced and voiceless affricates. The voiceless affricates have two sound sources, frication and stop-plosion. The sole resonator is the oral cavity.

Phoneme	Nr	VOI	NAS	SON	FRI	SIB	PLC	MAN
P (***P***ass)	1	-	-	-	-	-	1	1
T (***T***al)	2	-	-	-	-	-	2	1
K (***k***ahl)	3	-	-	-	-	-	3	1
B (***b***and)	4	+	-	-	-	-	1	2
D (***D***ank)	5	+	-	-	-	-	2	2
G (***g***ebt)	6	+	-	-	-	-	3	2
M (***M***und)	7	+	+	+	-	-	1	3
N (***n***ass)	8	+	+	+	-	-	2	3
9 (Fa***ng***)	9	+	+	+	-	-	3	3
L (***l***acht)	10	+	-	+	-	-	2	4
R (***r***und)	11	+	-	+		-	3	4
F (***F***ass)	12	-	-	-	+	-	2	5
S (***S***and)	13	-	-	-	+	+	2	5
$ (***Sch***acht)	14	-	-	-	+	+	3	5
V (***W***ind)	15	+	-	-	+	-	2	6
Z (***S***aal)	16	+	-	-	+	+	2	6
X (Da***ch***)	17	-	-	-	+	-	3	7
J (***J***ahr)	18	+	-	-	+	-	3	7
H (***H***ass)	19	-	-	-	+	-	4	7

Tab 2.3 German consonant classification according to specific features (cosnonant nomenclature according Dillier [B19]).

2.5.4 Impact of frequency compression on consonant perception

Considering the circumstance that the consonant energy is concentrated in the higher frequency area, it can be expected that spectral compression would generally improve consonant recognition for subjects with residual hearing in the lower frequency area.

Assuming that a lot of consonants also have formant like structures then it is possible that spectral compression affects these structures in a similar way as the vowels, meaning that

the formant-like structures become spectrally closer. As for the vowels this circumstance can lead to reduced ability of consonant identification.

It is very difficult to predict how linear spectral compression and spectral compression on the auditory frequency scale would affect consonant identification. However, subjects will certainly be confronted with additional spectral information, following the fact that the important information for consonant recognition is concentrated in the higher frequency regions [B27], [B12], in different places of the spectrum. As for the vowel recognition, the patients would require time for adaptation and for learning to use the new information for consonant identification.

2.6 Summary

Based on a simple model of profound hearing impairment and signal processing approaches employed in cochlear implants, three strategies for signal processing for the profound hearing impaired are proposed: information reduction, transposition of information essential for speech into the residual hearing area, and emphasis of particular speech segments by temporal prolongation. Spectral transposition can be achieved by means of spectral compression on either the physical frequency scale or on any auditory frequency scale.

In order to allow an understanding of the effects of the proposed signal processing strategies on the speech signal, the main characteristics of vowels and consonants have been reviewed. The impact of the processing on these speech cues, in particular frequency compression, is discussed in detail.

Chapter 3

Wednesday

"Built me a shelter against the rain, but could not have it to myself in peace. The new creature intruded."

M. Twain

Prior work

3.1 Overview

This chapter provides a brief overview of existing literature on spectral reduction, different types of spectral compression and various temporal modifications. Some suggestions on further signal processing studies using spectral reduction, spectral compression, and temporal modification are also given.

3.2 Existing work on spectral reduction

3.2.1 Introduction and terminology

A number of studies have shown that high levels of speech comprehension could be achieved when the speech spectrum was dramatically reduced to only a few spectral components or spectral channels. However, there remains no definitive answer to the question of how many spectral components or spectral channels are necessarily required to reach 100% speech perception. In order to understand the previous work, the terminology differences between the spectral channel, spectral lines, and spectral components should be explained first.

Studies working with *spectral channels* or *spectral lines* normally divide the speech spectrum into a certain number (normally smaller than twenty two) of channels or regions. The output signal is then reconstructed as a sum of the same number of noise bands with fixed center frequencies and bandwidths, or sine waves with fixed frequencies equal to the center frequencies of the spectral channels. Thus, the difference between the spectral channels and the spectral lines lies only within the signal reconstruction. In the spectral line method, single sinusoids are used for signal reconstruction, whereas in the spectral channels method, noise bands are used. Some studies demonstrate small or negligible differences between perception

of speech signals reconstructed with spectral lines and those using the spectral channel method [B119], [B26]. For this reason the difference between spectral lines and channels is neglected. The general notation used in the present work for both methods is taken to be the spectral channel.

Signal processing schemes which are based on *spectral component* reconstruction do not divide the spectrum of the incoming signal into a number of predefined spectral channels. Instead, they analyze the whole spectrum, but use only a limited number of non-fixed spectral components for output signal reconstruction. The frequencies of the spectral components used in signal reconstruction are determined by the signal analysis and may therefore change from frame to frame. The spectral components employed for signal reconstruction are usually sinusoids or amplitude modulated narrow-band noises.

Hence, the main difference between the spectral channel and the spectral component signal reconstruction is that when spectral channels are employed, the resulting output frequencies or noise bands are fixed, and when using spectral components, the output frequencies are variable and depend on the spectral analysis.

Note that the spectral component method passes into the spectral channel method when the resolution of the employed spectral analysis is not fine enough. Likewise, if the number of employed spectral channels increases and only a limited number of channels become activated in the output signal reconstruction, this method divergates to the spectral component method. Note also that if the spectral component method is observed during the smallest time period (one analysis frame duration), then there is no difference between the two methods. In other words, the spectral component method is equivalent to the spectral channel method with periodically changing spectral channel dislocation. An overview of all three spectral reconstruction methods is given in table 3.1.

Signal processing	Type of acoustic stimulation	Location mode of stimulation
Spectral Channels	Band limited noises	Fixed
Spectral Lines	Sinusoids	Fixed
Spectral Components	Sinusoids or band limited noises	Variable

Table 3.1 Comparison between three different spectral reconstruction methods.

Examples of spectral channel stimulation are cochlear implants. Therefore the effects of using a limited number of spectral channels have been investigated in detail for electric hearing applications.

In the spectral channel method, it is very important to choose the locations of the spectral channels in a meaningful way. Speech perception depends very much on where the spectral channels are positioned. Pavlovic and Studebaker [B124], [B101] reported about articulation index and band importance functions to describe the auditory channel positions and bandwidth importance for speech perception. Different spectral channels have different

importance for speech perception. The critical spectral bands are located between 315 and 5000 Hz. However, this is only an analysis problem which depends primarily on the distribution of critical speech pattern information in frequency and time [B119]. When using only one modulated noise channel, it is even possible to remove all spectral cues leaving only temporal information of the speech [B118;B119].

Otherwise, in all cases both temporal and spectral information is present for acoustic stimulation using spectral component methods. For this reason, the spectral component method might be more suitable for acoustic-only stimulation. That means that the number of required spectral components for 100% speech perception could be smaller than when using the spectral channel method.

Taking into account the above discussion, it is not surprising that there are discrepancies between different studies investigating the number of channels or spectral components required to understand speech. In the following paragraphs the most relevant work is summarized.

3.2.2 Studies on spectral reduction

The number of spectral channels or spectral components needed to understand spectrally reduced speech has been investigated and discussed intensively by many authors. Some studies consider the performance of normal hearing people on CI simulations as a function of the number of channels, type of signal synthesis (sinusoids or narrow-band noises), amount of temporal information left in the signal (channel envelopes), dynamic range, location of the frequency bands (channel spacing), and altered spectral distributions (shifts and warping) [B82], [B23], [B26], [B22], [B27], [B25], [B32], [B118], [B47], [B119], [B36], [B43], [B114], [B113], [B112]. Furthermore, the speech understanding performance of CI patients (fixed number of channels) has been investigated both in quiet and noise and compared to the performance of normal hearing subjects presented with CI processor simulations [B23], [B24], [B48], [B64]. It is generally agreed that speech understanding does not require the fine spectral detail present in naturally-produced utterances [B82]. However, it remains unclear how much spectral detail, and thus, how many frequency channels are really necessary for understanding speech.

In order to simulate speech with a reduced number of channels for the normal hearing, the signal was usually band-passed and the envelopes in each analysis band were computed. The RMS value of the envelopes was then used to modulate sinusoidal or band-limited carriers which were combined to form the simulated speech signal. The synthesis bands were either the same as the analysis bands, or they were shifted or warped to investigate effects of frequency transposition. Tests were reported with as few as 1 to as many as 16 channels. The temporal resolution was between 4 and 9.6 ms.

Warren *et al.* [B146] tested normal hearing subjects on speech perception abilities using sentences which were limited to narrow spectral channels. The experimental sentence set was prepared using nine 1/3 and 1/20 octaves width band-pass filters at the center frequencies of 370, 530, 750, 1100, 1500, 2100, 3000, 4200, and 6000 Hz (used sampling frequency was 20 kHz). Only one or two out of nine selected spectral channels were used for speech signal reconstruction. The speech perception tests were performed with 420 normal hearing adults.

Very high sentence recognition scores (95%) were observed when 1/3 octave width band pass filters at center frequencies of 1100, 1500, or 2100 Hz were used. The recognition scores were significantly lower (only 23%) when the pass-band filter centre frequencies were taken at 370 or 6000 Hz. However, the sentence identification scores increased significantly (from 23% to 78%) when both spectral channels were used simultaneously. The overall sentence recognition was lower if the pass-band filters with 1/20 octave width were used. The best observed sentence recognition (77%) for the single channel case was observed by pass-band filter with a center frequency of 1500 Hz. Given the results of this study, even one spectral channel with sufficient width and the appropriate placing ([B124], [B101]) can provide near 100% speech perception. However, other authors reported requiring a greater number of required spectral channels for near to 100% speech perception. The importance of the spectral channel bandwidth was also reported by Van Tassel *et al.* [B139]. Increasing consonant logotome recognition was observed by increasing spectral channel bandwidth from 20 to 200 Hz. To produce spectrally reduced signals, three band-limited noise channels were amplitude modulated using low-pass filtered speech signal envelopes with cutoff frequencies by 20, 200, and 2 kHz.

Shannon *et al.* report 90% correct identification with as few as only 3 band-limited white noise channels [B118]. For the band-limited noise amplitude modulation, different low-pass filtered (cutoff frequencies at 16, 50, 160 and 500 Hz) speech signal envelopes were used. No significant differences between spectrally reduced speech perceptions were observed for signals for which amplitude modulation was performed using 50, 160, and 500 Hz envelopes. However, reduced sentence and consonant identification was observed when at least one of the band-limited noise channels of the spectrally reduced signal was stimulated with 16 Hz envelope modulation. Lower consonant and vowel identification scores were observed using only one or two spectral channels. They explain the fewer number of channels required with the use of band-limited noise as carriers instead of sinusoids. Still, Dorman *et al.* [B26] reported that the nature of the carrier (band-limited noise or sinusoids) is not important and makes no difference in the intelligibility of spectrally reduced speech. Note that, the result of Shannon can also be explained with the extensive training that was applied during the test performance.

Shannon *et al.* [B119] repeated their tests with four spectral channels but different filter crossover frequencies and observed good speech perception independent from crossover frequency placing. They tested eight native English speaking normal-hearing subjects.

Loizou and Dorman [B26], [B81], [B82] found that 5 channels are required for 90-100% correct recognition of sentences, and 8 channels are required for 100% recognition of vowels and consonants. Increasing the number of channels beyond 8 improved the sound quality but not the intelligibility. This minimum number of 5 channels required for understanding speech is interpreted as being due to the fact that a minimum number of three channels are required to code the F2 formant and / or high frequency information, and an additional two channels are required to code the F1 formant. The temporal resolution in this and the previous study was 4 ms, the carriers were sinusoids.

Friesen *et al.* [B46] showed that normal hearing subjects (tests were performed on five normal hearing subjects) require 6 spectral channels for 100% sentence recognition. However, with the same number of used spectral channels, only 85% “aCa” consonant logatome and 70% “hVd” vowel logatome recognition could be achieved. To increase the consonant and vowel recognition scores to the value of 95%, approximately 20 spectral

channels required. They also performed tests with an increasing number of spectral channels on cochlear implanted subjects and found that most of them reach their saturation of speech perception using eight spectral channels.

Hill *et al.* [B58] reported 70% consonants and vowels recognition scores using 6 to 8 spectral channels. They tested 3 normal hearing adults with considerable experience in listening with spectrally reduced speech. To reconstruct audio signals spoken by a single male speaker they used a band-pass filter-oscillator system. Slight improvement of the recognition sores was also observed using oscillator amplitude quantitization.

Kates [B64] observed 73%, 83%, and 92% consonant identification scores by five normal-hearing subjects using 8, 16, and 32 spectral component for reconstruction. To produce a spectrally reduced speech signal, the sinusoidal speech algorithm was employed. All speech signal materials used were low-pass filtered at 7.5 kHz and digitalized using 20 kHz sampling frequency. Spectral analysis of the signal was performed using a 512 (25.6 msec) sample-long FFT calculation with a 62.5% overlap.

The following studies investigated speech perception using a constant number of spectral channels with normal-hearing and hearing-impaired subjects:

Erber [B38] observed improvement of word recognition (+10%) in visual-acoustic presentation over visual-only presentations by 6 normal hearing adults, 6 normal hearing children, and 6 profoundly hearing impaired children using only one spectral channel. The spectrally-reduced speech signal was produced by using one octave limited-band noise with centre frequency of 500 Hz. The limited-band noise amplitude was modulated with 20 Hz low-pass envelope of the original speech signal. The overall word recognition scores for the normal hearing adults improved from 40 to 50%.

Breeuwer and Plomp [B13] reported 33.2% correctly reproduced vowel syllable scores in auditory experiments on 18 normal hearing subjects. The speech signal was reduced to only two spectral lines which represented the two first formants (F1 and F2).

Remez *et al.* [B111] reported good sentence recognition on 18 normal-hearing subjects using 1 to 3 spectral components, which followed the first formant frequencies. However, the sentence material in the performed test was limited.

Additionally, it should be also mentioned that learning can significantly improve the performance on spectrally altered and reduced speech signals. In their investigations on frequency-shifted spectrally reduced speech, Fu and Shannon [B47] have shown that CI patients become accustomed to the tonotopic patterns of their individual speech processors (a 1:1 frequency mapping of the analysis bands to the stimulated nerves is not guaranteed in cochlear implants). When they were presented with frequency allocations different to their own speech processor, they showed similar performance decreases as normal hearing subjects presented with mismatched analysis and synthesis bands. The performance of CI users proves that at least upward frequency shifts can be learned.

Rosen and colleagues [B113] have investigated the learning effects of normal hearing subjects in spectrally shifted 4-channel speech by explicitly providing training through audio-visual connected discourse tracking over a large number of sessions. The spectral shifts caused enormous performance decrease at the beginning. Yet, all listeners learned to compensate spectral shifts (at least partially) within a reasonable time. The performance increased across sessions, although to a diminishing degree towards later sessions (*i.e.* the

steepest performance increase occurred at the beginning). Thus, practice has a significant learning effect, and therefore short-term experiments without sufficient training exaggerate the long-term consequences of spectral shifts considerably.

3.2.3 Summary of important issues in spectral reductions

From the previously discussed literature, the following important issues in spectrally reduced speech have been identified:

- number of frequency channels
- frequency representation (band-limited noise or sinusoids)
- filter spacing
- temporal resolution
- dynamic range of amplitudes
- amplitude resolution
- nature of the carriers for synthesis
- type of signal processing
- training [B17]

Of course, other very important issues are the testing and the evaluation of the results. In testing, the nature of the tests (sentences, vowels, consonants, naturally or synthetically produced, context, phonetic richness, etc.), the number of speakers used, the sequence of the tests (allowing more or less training), testing in quiet or in noise, and at which SNR may all significantly influence the results of the tests.

In the evaluation of the results, the employed measure of performance (e.g. %, asymptotic performance) plays a role. In order to compare and interpret different results, the above mentioned issues and the type of testing must be considered.

3.2.4 Conclusions on spectral reduction

To conclude, it can be stated that the auditory system uses the information “how much energy (amplitude) is where (frequency) and when (temporal information) in the frequency space” for speech recognition. Clearly, the reduction of the spectral information decreases the performance both in quiet and noise. It also seems that the recognition of vowels is more sensitive to spectral distortions than the recognition of consonants.

Nevertheless, speech recognition is possible from spectrally reduced speech. Depending on the nature of the tests carried out, the number of channels required for asymptotic performance in quiet is between 5 and 8. In noise and with decreasing SNR, the number of

channels to achieve a given level of performance must be increased. Therefore, it appears as if the fine spectral information is more important for speech in noise than for speech in quiet, or that these situations require more spectral information than clean speech.

In reducing the frequency content of speech, the details of the spectral shapes are lost. Significant spectral peaks and zeros are difficult to perceive, just as it is the case for the shape of broad spectral distributions. For vowels this means that the formant structures are severely degraded (depending on the number of channels, formant frequency transitions are either completely lost or just presented as temporal change in the relative level of adjacent bands), and for consonants the width and broadness of the spectral shape cannot be properly represented. This is why a minimum of 5 frequency channels is proposed to allow the coding of F1 and F2 and / or high-frequency information. In this way, the listeners can rely on relative amplitude differences across channels to infer frequency information (distinguish formants), as long as the dynamic range of the amplitudes is wide enough to allow the identification of spectral peaks and valleys.

It has also been stated that the temporal information is more crucial for speech understanding than high frequency resolution and the exact representation of the spectral information. Thus, even though the reduced number of channels severely degrades the distribution of the spectral energy, 100% speech recognition is possible as long as temporal information is available with sufficient resolution[4]. Indeed, preservation of the temporal information is possible in reducing the frequency content when the involved signal processing is fast enough. In trading spectral resolution versus temporal resolution, temporal cues should therefore be given preference.

Furthermore, it has been reported that altering the frequency representations of the original speech signal may severely degrade the performance in speech recognition. However, it has been proven that practice implies a significant learning effect. Therefore, short-term experiments, without sufficient training, exaggerate the long-term consequences of spectral shifts or other frequency transpositions considerably[5]. Since CI patients are able to accommodate to the tonotopic patterns of their particular speech processor and electrode configuration, it is expected that with the appropriate training also profoundly hearing impaired will learn to understand speech subjected to spectral reduction and frequency alterations even though the sound quality of spectrally reduced speech will always remain inferior.

For the present investigations it was decided to perform a study on the minimal number of spectral components per time unit used for spectral compression on normal hearing subjects (see chapter 5). The goals of the study were to find the limitation on maximal spectral reduction which can be applied for speech processing for profound sensorineural

---

[4] Based on the finding that speech recognition can be achieved with primarily temporal cues and that hearing impairment often implies impaired or absent spectral resolution, Shannon [B118] suggested already in 1995 alternative signal-processing strategies for auditory prostheses!

[5] Note however, that only upward spectral shifts as occurring in CI signal processing due to the shallow insertion of the electrodes were investigated. It is expected that the same results apply also to downward spectral shifts, but this still needs to be proven!

hearing-impaired subjects, assuming that their spectral resolution can be significantly limited by the broadening of the auditory filters.

3.3 Prior work in frequency transposition

3.3.1 Overview

The method of transposing spectral information essential for understanding speech into a hearing area where residual hearing is still available has become known and has been investigated since the beginning of the last century. The various methods can be summarized as follows:

- Reduced playback speed (no real-time playback)
- Reduced playback speed and time compression
- Linear frequency shift
- Linear frequency compression
- Various non-linear frequency transposition schemes.

Reviews on the earlier work on frequency transposition are e.g. presented in Ling [B80], Erber [B37], Mazor [B83], and Lafon [B73]. More recent reviews are given by McDermott [B89] and Gravel [B56]. A chronological list of the work on frequency transposition is presented below.

- 1925: frequency transposition into high-frequency range, Perwitzschky [B102].
- 1952: speech recording and playback at half speed, Tato [B125].
- 1954: speech recording at decreased record speed, König and Eichler [B70], 1955: Springer [B122], 1963: Oeken [B100], and 1976: Nagafuchi [B96].
- 1962: VENUS-Vocoder by Pimonow, 6 to 8 channel signal synthesis at lower frequencies than original (several different implementations) [B105;B106].
- 1963: „amplification by 2 compensation tones“, low-frequency information is transmitted unmodified, together with the information in two high-frequency bands which modulate the amplitude of two tones in the residual hearing range (Lafon & Isaac, mentioned in [B73], commercialized by Philips).
- 1965 / 1966: „Transposer“, low signal frequencies are transmitted unmodified, high-frequent components modulate a carrier frequency, the difference between the original and the carrier frequency is transmitted to the listener (Risberg 1965, Johansson 1966 [B63], mentioned in [B142]). A similar device was investigated by Oticon (TP72, see [B142].
- 1967: frequency transposition by partial vocoding described by Ling and Druz [B80], [B78], [B79],

- 1969 / 1977: PARME vocoder [B105], similar as „amplification by 2 compensation tones“, but with a modulation of band limited noise instead of tones.
- 1971: DIFA device, 10 ms intervals are recorded, only every second interval is played back at half sampling rate [B144].
- 1972: presentation of the Oticon TP72, a device similar to the “transposer” (Nielsen 1972, in [B142]).
- 1973: FRED system, linear downshifting of high-frequent components by 4 kHz (Velmans & Marcuson [B141], [B142]).
- 1976: Formant-Vocoder, spectral peaks modulate white noise within 12 low-frequency bands (Carrat *et al*., 1978, mentioned in [B73]).
- 1980: Galaxie system, signal squaring and lowpass filtering in different frequency bands [B73]
- 1983: Mini-Fonator, tactile stimulation (*i.e.* vibration) with a carrier frequency modulated by high-frequency components (Siemens), mentioned in [B73].
- 1986: SiVo device, presentation of single sine waveform representing the maximum spectral component as a support for lip reading (EPI group, UK, [B114], [B145]).
- ca. 1986: Emily / TDL, addition of F2/2 and 2*F2 to the spectrum [B34], [B35].
- 1989: first description of frequency transposing prototype by AVR [B115].
- 1999: linear spectral compression by means of a phase vocoder by McDermott [B86], [B87].
- 1999: proportional spectral compression by Turner and Hurtig [B133].
- 2000: spectral compression hearing aid by RION [B116].

3.3.2 Description of frequency transposition approaches

3.3.2.1 AVR devices

AVR Ltd. (Israel / USA) is the only company, which successfully sells frequency transposing hearing aids (http://www.avrsono.com). The first device sold was the TranSonic; so far, the ImpaCt, and the Logicom-20 device are on the market. All of their devices provide two different, constant compression factors for low- and high-frequency spectral components. The cut-off frequency is set to 2.5 kHz. In addition, the transposition hearing aids provide a dynamic consonant boost (amplification of high-frequent components by up to 16 dB). It has been reported from various sources that the fitting of the AVR devices is not satisfactory as there are no objective and clear fitting rules, [B89], [B56].

Linear spectral compression is achieved by writing the input into a memory and reading it out at a slower speed [B4]. Not utilized portions of the input signal are discarded which leads to audible distortions of the signal. The amount of compression is programmable and can be adapted to the individual patient’s hearing abilities.

While the TranSonic was a body-worn device [B115], the more recent ImpaCt is available as a BTE. The Logicom-20 combines an ImpaCt with FM technology. It also includes a dynamic speech recoding system [B16]; however, it is not clear whether this is a new or further development of the spectral compression scheme or the same principle as applied already in the TranSonic and ImpaCt.

Gravel and Chute in their investigations on the use of transposition for children [B56] report on two studies with the TranSonic device. Plant & Franklin (1994, mentioned in [B56]) have performed systematic training and testing of a subject with acquired profound hearing loss during a three-month period (2 two-hour trainings with the TranSonic per week, additional use of TranSonic during at least one hour per day). Although no benefit on vCv identification could be shown for both auditory-alone and visual-auditory conditions, a significant benefit on the detection of consonants (s,n,t,d), on closed-set identification of consonant-initial contrasts in words, and on the recognition of connected speech was found. Nevertheless the subject preferred the conventional hearing instrument. The second study (uncompleted) by Chute *et al.* (1995, mentioned in [B56]) was on the combination of a cochlear implant on one ear and the TranSonic on the other ear during a 6 month trial (the use of a TranSonic device alone during two hours daily; for the remainder of time, concurrent use of a TranSonic and a CI). No formal training was given. Evaluated were phoneme detection, discrimination of speech features, closed-and open-set word identification, and sentence identification. At the time of publication, no benefits could be shown; however, 2 of 5 subjects felt their understanding to be clearer and easier when using the CI together with the TranSonic device.

Gravel and Chute also evaluated the fitting of the TranSonic to young children and found that the fitting of the TranSonic as proposed by the manufacturer was too subjective. They proposed an objective fitting method based on DSL together with guidelines for the study of transposition hearing aids [B56]. The fact that the TranSonic device, although proposed for moderately severe to profound hearing losses by the manufacturer, had been propagated as an alternative to CIs is considered very unfortunate. On one hand, this has lead to unreasonable expectations of performance, and on the other hand, it has eliminated some potentially good candidates for frequency transposition (due to “too much” hearing capabilities).

The TranSonic as well as the ImpaCt device were also evaluated by McDermott at the CRC in Melbourne, Australia. For the evaluation of the TranSonic [B89], 5 experienced hearing aid users with no prior exposure to frequency transposition used the device for a total of 12 weeks. The performance of the subjects with their conventional aids and with the TranSonic was compared. A performance increase with the TranSonic was found for 4 of the 5 patients; however, it was questioned whether the benefits achieved in comparison to the conventional aid are really a result of the spectral compression or due to the good electro-acoustic characteristics and better amplification at low frequencies of the TranSonic (input AGC, frequency response, maximum possible output, no amplification to uncomfortable level and thus higher overall gains). Evidence for a benefit from transposition was found for only 2 of the 5 subjects, these two also kept the TranSonic in spite of cosmetic and practical disadvantages. It was stated that the fitting strategies of the TranSonic are not satisfying (absence of precise fitting rules). Furthermore it was not clear which patients might benefit from such a spectral compression scheme. Also unknown was the effect of longer usage of the TranSonic (learning effects).

For the evaluation of the ImpaCt hearing instrument [B90], 3 experienced hearing aid users with no prior exposure to frequency transposition were employed. All of them had rather flat moderate to profound hearing losses. No benefits of spectral compression were found in comparison to the own hearing aid. In particular, no benefit could be shown for recognition of monosyllables and medial consonants, and sentence understanding in noise was even poorer with spectral compression than without. Based on these results the question was raised as to whether a longer training period with the new signal-processing scheme is required (the trial was carried out during a few weeks), and which subjects are suitable candidates for frequency compression.

Davis [B16] presented two case studies for the measurement of benefit of the AVR Logicom-20 device. A child and a young adult had both an AVR ImpaCt device and an own instrument. They were evaluated on the Ling Five Sound Test plus t (/a/, /u/, /i/, /sh/, /s/, /t/). Both performed much better with the ImpaCt than with their conventional hearing aid. It was mentioned that the use of the Ling Test is suitable to predict speech discrimination ability (correlation with aided thresholds, WIPI, and NU-6 test), since it can serve as a quick and accurate test to assess hearing instrument benefit.

3.3.2.2 Proportional spectral compression by McDermott

McDermott at the CRC in Melbourne, Australia, has implemented a real-time spectral compression scheme in a body-worn processor. The system is based on a phase vocoder and compresses the complete frequency range by a constant factor [B86], [B87].

The effect of this frequency transposition scheme was evaluated on 6 hearing impaired and 5 normal hearing subjects [B88]. The hearing impaired subjects had all a steeply sloping hearing loss and near normal hearing at low frequencies. None of them had used hearing instruments. For the normal hearing subjects the signals were low-pass filtered at 1.2 kHz. It was first shown that the hearing impaired subjects received the same amount of information as the normal-hearing subjects with simulated loss (low-pass filtering) on CNC words (consonant – vowel nucleus - consonant) [B104] in speech shaped noise. Proportional spectral compression with a compression factor of 0.6 was then applied to both the hearing impaired and normal hearing with simulated hearing loss. The same CNC stimuli as for the previous test (but without noise) were used for evaluation. Training was provided in 10 sessions for 1 hour in weekly intervals. No evidence for any benefit was found for both the hearing impaired and the normal hearing subjects with simulated hearing loss. As possible reasons for this result, insufficient training and the unnatural pitch of frequency-lowered speech were named.

3.3.2.3 Proportional spectral compression by Turner & Hurtig

Proportional frequency transposition, preserving ratios between formant peaks of speech has also been implemented and investigated by Turner and Hurtig at the University of Iowa [B133]. Their real-time algorithm is based on FFT, multiplication of all frequency bins by a constant factor < 1, and back-transformation with IFFT. Data loss in compressing the spectrum is minimized by linear frequency interpolation. The compression rate is selected depending on the hearing loss. Upwards frequency shifting is provided to shift the speech spectrum compressed to lower frequencies into the region of usable hearing. An optional time domain trimming after compression / shifting is provided so that the output signal has the

same duration as the input signal. The FFT-IFFT implementation was patented in 1999 [B61].

A study on normal hearing and hearing impaired subjects showed that frequency compressed speech is well understandable with compression factors of up to 60%, and that spectral compression may provide a benefit for hearing impaired with residual hearing better than 60 dB below 2 kHz. In particular female voices seemed to have greater intelligibility. It also seemed that patients with more severe hearing loss in high frequencies benefited most from frequency compression. Some fricatives however sounded unnatural. It was concluded that, even though spectral compression might have a benefit, it does not solve all problems, and thus traditional amplification is still required.

3.3.2.4 RION spectral compression hearing aid

In a recent publication, Sakamoto *et al.* [B116] describe a new portable (probably body-worn) hearing aid by RION (HD11). The implementation is based on PARCOR analysis-synthesis (extraction of LPC, pitch, excitation power, and voiced/unvoiced analysis, non-linear transformation of frequency scale combined with linear compression) and allows the separate adjustment of the compression ratio for unvoiced and voiced speech, the separate adjustment of fundamental frequency and spectral envelope, and the adjustment of the frequency response (digital filter). The compression ratio for both voiced (FCR-V) and unvoiced speech (FCR-U) is adjustable from 10% to 90% in 10% steps. The fundamental frequency can either be left unmodified or compressed by FCR-V or (FCR-V+100)/2. Another mode of the instrument is without compression at all. For adjusting the frequency response, 10 basic frequency response curves are available (5 amounts of bass reduction, and 5 amounts of speech component enhancement) which can be chosen individually. The application of such frequency response shaping is proposed to decrease the disturbing effects of frequency transposition (information overloading of remaining narrow hearing range).

The device was tested on 11 subjects with severe-to-profound hearing losses (> 100 dB for frequencies of 2 kHz and higher), which were all dissatisfied with conventional hearing aids. Fittings were carried out in a rather heuristic way (trial and error) until the optimum seting was found. Performance and qualitative judgment in everyday situations was reported individually, and speech recognition tests (mono- and disyllabic Japanese words) with the spectral compression hearing aid in both optimum settings and no-compression mode, and with the own aid were carried out. In addition, sentence recognition was tested in audio-visual, audio only, and visual only situations (again with the spectral compression hearing aid in optimum settings and no-compression mode, and with the individual's own aid). All in all, no improvements in speech recognition scores could be shown, and no obvious benefits could be demonstrated. It was suspected that spectral compression is mainly of benefit for the hearing-impaired with higher lip-reading ability (one subject showed a significant improvement of performance in audio-visual tasks when using the spectral compression device). It was found that 7 out of the 11 subjects liked the frequency compression, 5 wanted to continue using it. (Note that three of these 5 subjects who liked the compression needed to communicate by writing or sign-language before using the new device.) Two subjects did not like the compression because they perceived it as loud and noisy, another two because they found it sounded uncomfortable and strange. One found no difference between the conventional hearing aid and frequency compression. Clinical trials with the new hearing instruments have just started.

3.3.2.5 Thomson-CFS signal processing method for hearing correction of hearing impaired

The goal of the Thomson-CFS [B130] approach is to provide a solution for intermediate hearing losses having frequency zones with no hearing, for which no hearing aids are available on the market right now.

Proposed is a parametric speech model based on the vocoder technique. It extracts pitch, voicing, energy, and spectrum, modifies these features and re-synthesizes the signal based on the modified features. The approach provides wide possibilities of control (any parameter can be tuned individually).

3.3.2.6 Improvement of hearing instruments by Lafon

The invention refers to an older patent by Lafon [B72], where high-frequency information form two frequency bands within the speech spectrum is presented within 2 tones of lower frequency modulated by the amplitude envelope of the two high-frequency bands.

The actual patent describes the transposition and amplification of different bands of the spectrum into lower-frequency bands within the audible range. The number of "image" bands equals the number of original bands. The spectral compression ratio might be different for each band but is constant within each band. The image bands are arranged stepwise together (*i.e.* one band joining the next band without interruption). Frequencies below 500 Hz of the original signal are amplified with a constant gain. For frequencies between 500 and 1000 Hz the amplification decreases gradually with increasing frequency in order to free this part of the spectrum for the image bands which are actually transposed to frequencies above 500 Hz.

3.3.2.7 Signal processing apparatus by Adelman

In this invention from Adelman [B2], the input is subject to adaptive noise canceling for suppression of signal portions not related to the strongest dominant frequency in a predetermined range. Then an AGC followed by a filter bank is applied. In each frequency band, the following processing is done: non-linear amplification, harmonic frequency transposition, band filtering, and gain shaping. Finally, all frequency bands processed in this way are summed together and subject to further amplification before being output. Harmonic frequency transposition denotes the selective multiplication or division of all frequencies by a constant value, retaining the harmonic relationships within each band. The entire processing takes place in the time domain.

Claimed to be effective is the application of adaptive noise canceling to the input, the application of a filter bank with at least two bands to this signal, the selective transposition of the frequencies in at least one frequency band, the temporal matching of individual bands, and the summation of the temporally matched bands for creating an acoustic output. The transitional discontinuities between two succeeding signal segments are minimized by the amplitude and the rate of change matching.

3.3.2.8 Speech coding hearing aid system utilizing formant frequency transformation by Strong & Palmer

This approach is very sophisticated for its time of publication [B123]. The most important parameters of speech are extracted from the input, namely formant frequencies and amplitudes, fundamental frequency, and voiced/unvoiced information. Formant frequency estimation is carried out based on the spectral envelope generated via LPC or FFT. The output signal is synthesized by means of pure tones and noises based on modified formant frequencies and voiced/unvoiced information.

Claimed is a hearing aid system with the extraction of the spectral envelope and the formant frequencies, the modification of the formant frequencies, and sound generation based on the modified formant frequencies for forming the output signal. Formant frequency modification is achieved by division by predetermined values from 2 to 6 and the addition of predetermined values to these divided frequencies. In addition, the fundamental frequency is determined and modified by division. The formant frequency division does not need to match the fundamental frequency division (pitch lowering can be less than the lowering of the complete spectrum). Synthesis of the output signal is achieved by means of tone and noise generators driven by the modified formant and pitch frequencies. A voiced / unvoiced detection allows synthesizing with tones in case of voiced signals and narrow-band noises in case of unvoiced signals. The synthesized signal is subject to amplitude compression and gain. The gain is determined in function of the RMS of the input amplitude.

3.3.2.9 Speech transformer by Pimonow

The patent of Pimonow [B105] describes the VENUS vocoder mentioned in the introduction in Sections 3.1. A filter bank is applied to the input signal, and the output is synthesized within the audible range by generating tones and / or noises which are modulated by the output of the analysis frequency bands.

Claimed are a method and device for transferring useful speech information into the reduced hearing range of hearing impaired by dividing the input spectrum into a number of bands, and modulating a tone or noise within the residual hearing range based on the output of each analysis band (number of analysis bands = number of sound generators). The output sound is the concurrent generation of all tones and noises.

3.3.2.10 Improving the intelligibility of a speech signal by Ericsson

The additive combination of the original sound with one or more transposed versions of the original within cellular phones and hearing instruments is reported [B39]. Speech signals modified in this way sound as if they were pronounced by several fictive voices simultaneously (chorus in the background of the original); the resulting signal contains speech information in frequencies where the user has better perception (presence of noise / hearing loss). The additional signals can be generated with either lower or higher frequencies than the original, they can be switched on and off according to the actual demand, and their amplitudes can be regulated individually to adapt the device to specific hearing deficiencies and to specific noise situations. Frequency transposition is achieved by means of pitch modification with a possible implementation being time-domain harmonic scaling. To increase the

intelligibility but preserve the naturalness of the original sounds/speaker (the speaker should not sound like a different person), the amplitude of the original is always kept higher than the additional signals.

3.3.2.11 Other spectral compression schemes

Besides the above cited important frequency transposition schemes, there exist various other approaches (see chronological list and review papers cited in Section 3.1). Here only a few systems shall be described briefly.

The EMILY device, sometimes also denoted by TDL processing, was developed by Dupret and Lefèvre [B34], [B35]. It determines the second formant F2 be extracting the dominant frequency within an band pass of 1 – 2 kHz. F2 is added together with F2/2 and F2*2 to the original signal, whereby each F2 component can by amplified individually based on the audiogram of the patient. (An earlier version added a compressed and expanded band pass determined via the function of the hearing loss to the original spectrum.) The resulting system is implemented in a body-worn case and used together with Phonak hearing instruments. Although the EMILY inventors state that the proposed concept increases speech intelligibility in noise and quiet [B34], there exist no published reports on the usefulness of this scheme [B56]. The non-linear signal processing also introduces additional harmonics into the spectrum. The EMILY system is currently worn by ca. 500 children in France and 20 in Texas [B56].

Another French system is Galaxie developed by Lafon & Gharbi [B73]. The goal of this processing was to present information in the low-frequency range and preserve the prosodic structure of the signal. For this purpose, the high-frequency envelope is projected into the low-frequency range by means of signal squaring and low-pass filtering in different frequency bands. In addition, impulse-like signals are lengthened and presented in the low-frequency range, the amplitude of these signals corresponds to the original duration of the impulse. Although several prototypes of Galaxie were built (table systems) the scheme was never commercialized due to funding lacks for miniaturization. The authors also report a benefit, but no clear numbers can be stated.

A linear frequency-shifting device is the FRED system described by Velmans [B141], [B142], [B140]. Linear frequency shifting of frequencies greater than 4 kHz is achieved by subtracting 4 kHz from the 4 to 8 kHz frequency range. These shifted frequencies are added to the original frequency range, and the combined spectrum is amplified. It was reported that this spectrum preserving processing scheme sounds natural to normal hearing subjects.

Another frequency transposition scheme described by Velmans is the Oticon TP72 [B142] which transposes the spectrum in a non-linear way. The frequency range > 4 kHz is converted into broadband noise by means of a non-linear operation (similar to squaring) and then low-passed at 1.5 kHz. The resulting spectrum is added to the original signal and amplified. This spectrum-destroying scheme was reported to sound unnatural. A comparison of the FRED and the Oticon TP72 showed that frequency transposition gives a benefit on speech perception, but that the spectrum preserving transposition is preferred over the spectrum destroying approach [B142].

The Johansson transposer [B63] modulates a 5 kHz carrier with the energy above 4 kHz, and the difference components (noise below 1.5 kHz) are mixed with the original

signal. This processing is only effective when high-frequency components are present. No statistical benefit of this approach could be proven for auditory tasks, but a significant better performance than with conventional amplification was achieved in combined visual-plus-auditory conditions. This is interesting since a similar effect (no benefit in auditory tasks alone but improvement of visual-auditory tasks) was also mentioned by Sakamoto *et al.* [B116]. In addition, significant learning effects were reported.

Additionally to the described studies and patents there are also some from Shigeru [B120], Vigneron & Lamotte [B144], and Biondi [B9].

3.3.3 Summary and conclusions on frequency transposition

The study of frequency transposition for the presentation of audio signals in residual hearing areas has continued through decades. Technological limitations, distortions introduced into the signals by means of the employed processing schemes, the absence of suitable candidate identification means and objective fitting rules, as well as missing clinical studies, convincingly proving the benefits of frequency transposition have so far impeded this technique to be widely employed for hearing impaired patients. Today, only one company (AVR) sells frequency transposing hearing instruments.

The many techniques allowing for frequency transposition can be coarsely subdivided into frequency shifting, frequency compression, and reducing the playback speed while discarding parts of the input signal to preserve the original duration. Among frequency compression, many linear and non-linear techniques including FFT/IFFT processing, vocoding, high-frequency envelope transposition and mixing with unmodified low-frequency components have been investigated. Since harmonic patterns and formant relations are very important in the perception of speech, it seems meaningful to distinguish spectrum preserving and spectrum destroying techniques.

The most serious and promising activities in frequency transposition today are:

- Intelligent reduction of playback speed, separate compression of voiced/unvoiced components (AVR, see http://www.avrsono.com)
- Linear compression by means of phase vocoding (McDermott, Melbourne, [B86])
- Linear compression by means of FFT/IFFT processing (Turner & Hurtig, Iowa, [B133])
- Feature extraction and signal resynthesis based on the vocoder technique [B130]
- Separate compression of voiced/unvoiced components and fundamental frequency (RION, Japan [B116]

Depending on the approach used for frequency transposition, the resulting speech signals may be strongly distorted and sound unnatural. In addition, a particular problem in this context seems to be the modification of the user's own voice. Furthermore the mixing of original low-frequency components with transposed high-frequencies may mask important

signal components and lead to unnatural sound as well. The following recommendations have been collected from the literature:

- Keep compression > 60 .. 70%
- Preserve frequency ratios as an invariant cue for speech recognition
- Transpose the complete spectrum if at all (alter the global acoustic structure and not only parts of it)
- Do not introduce spectral distortions / avoid non-linear frequency distortions
- Free the frequency range into which to transpose from the original sound
- Avoid mixture of low-frequency components with transposed high-frequency components
- Preserve the temporal structure of sound (both temporal characteristics and duration)
- Transmit intonation patterns
- Reduce information content (e.g. by removing F1 and other low-frequency components).

In the reported experiments, very few patients received benefits from frequency transposition. This may on the one hand be due to the above-mentioned degradation of sound quality. On the other hand, however, no criteria exist today to identify potential candidates which might receive a benefit from frequency transposition. The choice of subjects who participated in the various studies is, hence, rather arbitrary. Furthermore, the lack of objective fitting rules and insufficient training with the new processing schemes within most of the studies performed might also have negatively influenced the results of frequency transposition studies.

The influence of learning effects have been mentioned by both McDermott [B89], [B90], [B18], [B88], [B86] and Gravel [B56]. According to Rosen *et al.* [B113], at the beginning of the experiments with frequency transposition an enormous performance decrease is usually experienced. However, practice on frequency-transposed signals shows a significant learning effect. Listeners can learn to compensate for spectral shifts (at least partially), and therefore, short-term experiments are not appropriate to seriously predict the long-term consequences of spectral transformations. This is also confirmed by Fu and Shannon [B47] who showed that CI patients are able to adapt to altered spectral representations, and that they have similar performance decrease when sounds are presented in tonotopic patterns different from what they are known, their CI device as having normal-hearing listeners which are presented with shifted spectral information.

To conclude, the following guidelines for the study of transposition hearing aids proposed by Gravel and Chute [B56] shall be mentioned:

- Clear definition of candidacy criteria
- Definition of objective fitting methods
- Definition of test materials used

- Quantification of benefit of transposition over conventional amplification: examination of user's perceptual ability in quiet and noise
- Study of short- and long-term efficacy
- Collection of data from various locations using the same fitting protocols and test procedures
- Provision of specific auditory training.

It is clear that without suitable criteria for the identification of potential benefitters and without the availability of suitable fitting and training rules, frequency transposition will continue to remain in a rather experimental stadium.

Assuming the guidelines and suggestions mentioned in the literature and the different spectral processing possibilities discussed in chapter 2, it was decided to concentrate the present investigations on the linear spectral compression on the FFT scale and the linear spectral compression on the SPINC scale. In this order, different spectral compression ratios will be tested with and without spectral reduction on normal hearing and profoundly hearing impaired subjects (see chapters 6 and 7).

3.4 Existing work on temporal manipulation

3.4.1 Studies on temporal manipulation

Most literature findings on temporal manipulations of speech signals are performed using the slowing down of replay velocity technique which results in additional spectral compression of the processed signal [B70], [B99], [B77], [B76], [B125], [B68], [B131], [B122], [B100], [B15], [B96]. However, some authors used spectral processing schemes which enabled temporal modifications without significant changes in the signal spectrum (these approaches were based on the technique introduced by Fairbanks [B40] and Kurtzrock [B71]). They reported that the temporal modifications of speech signals are less destructive for speech intelligibility than the linear spectral compression of speech signals [B15], [B96].

Daniloff *et al.* [B15] performed vowel identification tests in a /h-d/ context with 20 normal hearing subjects using temporally shortened signals. They observed a significant decrease in vowel identification scores only when the signal was 80% time compressed (~50% correct identifications, similar results were reported also by Klumpp and Webster [B68], Beasley [B6], Fairbanks and Kodman [B41], and Garvey [B49]). However, using 70% temporally shortened signals, subjects still achieved 90% correct vowel identifications.

Nagafuchi [B96] performed speech intelligibility tests with 160 normal hearing, normal IQs children between the ages of 4 and 11 years. Test materials consisted of 20 phonetically balanced meaningless monosyllables in Japanese speech audiometry, which were temporally shortened between 75% and 30% or temporally prolonged between 150% and 300%. The achieved recognition scores for temporally shortened speech signals decreased significantly only when the 75% shortening was applied (~55% correct identifications). For 50% temporally shortened signals, 90% monosyllable recognition was achieved. In the prolongation of the monosyllables, no significant decrease in recognition was observed, even

if the test material was 300% prolonged (~90% correctly recognized monosyllable). For 200% temporally prolonged signal, 100% recognition was achieved.

An overview of the linguistic uses of different segmental durations in the English language is given by Klatt [B65]. He concludes that, in English, duration often serves as a primary perceptual cue in the distinctions between inherently long versus short vowels, voiced versus voiceless fricatives, phase-final versus non-final syllables, voiced versus voiceless postvocalic consonants, stressed versus unstressed or reduced vowels, and the presence or absence of emphasis.

Liberman *et al* [B76] reported that the voiced stops in initial position could be made to sound like their voiceless counterparts by cutting back the beginning of the first-formant transition.

Liberman *et al* [B77] showed that it is possible to convert different spectral patterns of consonant-vowel syllables into different other speech sounds by varying only the speed of the transitions. So for example, as the transitions are progressively slowed, the patterns for the stop consonant "*b*" plus the vowel "*a*" (as in *bottle*) begins to be heard, as the semivowels "*w*" plus "*a*" (as in *wobble*).

To our knowledge, no previous work on bidirectional temporal modifications of different speech segments was reported.

3.4.2 Summary and conclusions on temporal manipulation

The following conclusions can be drawn from the previous work on temporal modifications of speech signals:

- Speech monosyllables do not lose intelligibility when temporally prolonged even to a considerable amount and can still be correctly recognized even when they are 300% temporally expanded.
- The speech signal loses its intelligibility with increasing temporal shortening. However, the monosyllables can still be correctly recognized when they are 60-70% shortened.
- The general slowing down of the speech signal spoken in sentences may increase it's intelligibility.
- The speech segment duration is an important feature of the English language. Changing of segment or segment transition durations can result in recognition confusions.

Assuming these four points, it was decided to perform speech perception experiments with temporally modified speech including the prolongation and shortening of the whole speech signal or some specific speech segments on normal hearing subjects (see chapter 8).

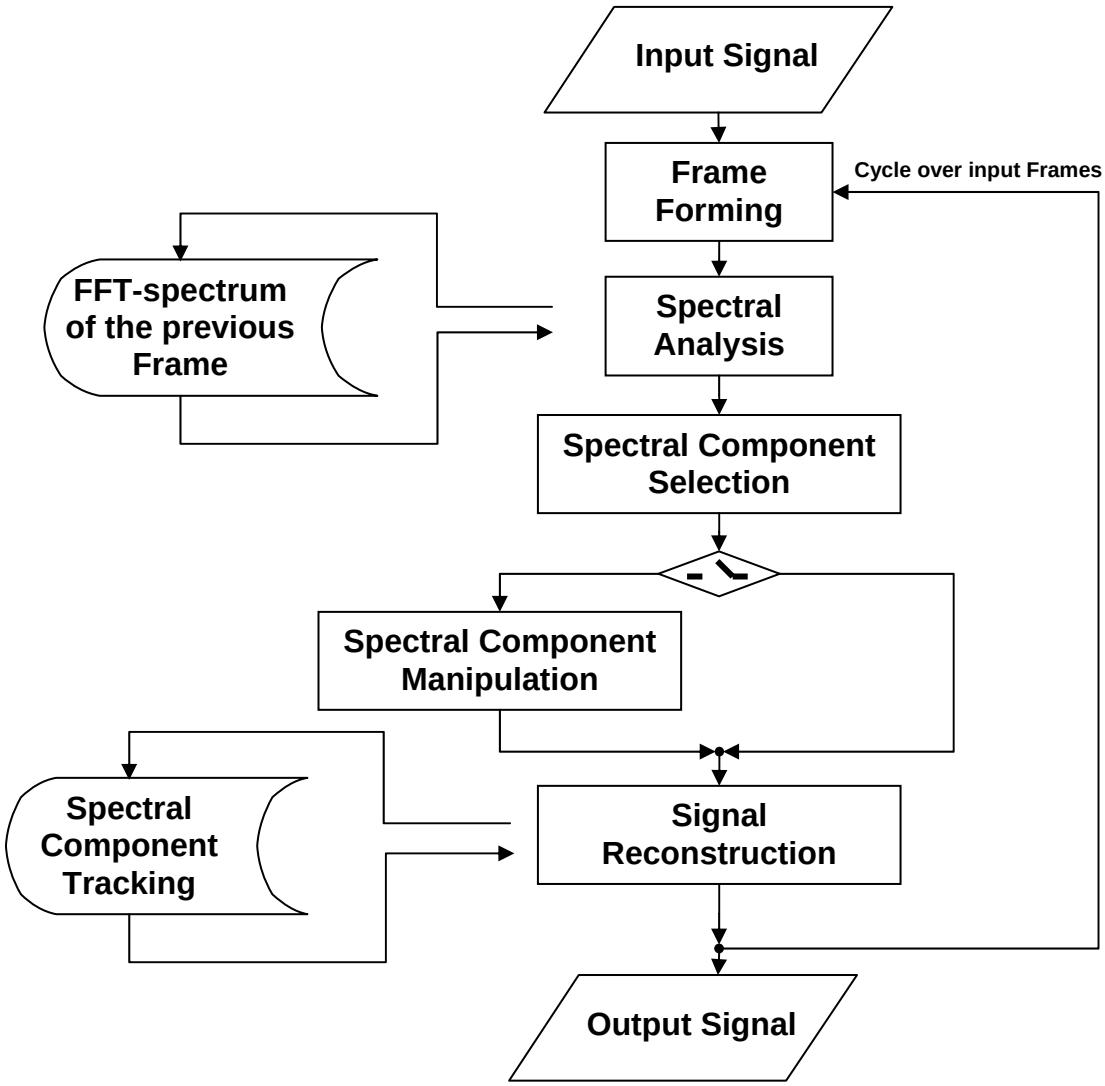
Chapter 4

Baseline sinusoidal speech analysis/synthesis system

4.1 Motivation

The present signal processing system was developed with the purpose of realising a working tool for providing for different manipulations on an audio signal. The most important manipulations motivated and set by the goals of the present study are: spectral reduction, spectral transposition in meaning of spectral compression, and temporal stretching and shrinking of different audio file segments. The sinusoidal speech (SiSp) algorithm developed by McAulay and Quatieri [B84], [B109], [B108], [B110], [B85] was taken as the basis for the developed signal processing system because it enables the implementation of all the required spectral operations and temporal modifications (spectral reduction and various linear and non-linear spectral compressions prolongations and shortenings of different speech segments without changing of signal's pitch). Some parts of the SiSp were simplified or modified according to the special needs defined by the requirement that the present system was designed to perform speech processing for profoundly hearing-impaired subjects. This circumstance did set requirements for the maximal simplicity and even choice of some signal processing parameters of the developed system to allow implementation of some parts of it in a potential hearing device. The spectral reduction of the audio signal, for example, was larger than for the original SiSp algorithm. To produce appropriate selection of spectral components used for signal reconstruction, different algorithms were implemented. One of the implemented spectral selection criterions is described by Kates [B64]. Besides the original signal reconstruction described by McAulay and Quatieri, a broad range of various additional reconstruction methods was also implemented. One of these was based on the partial tone ("Teilton") signal reconstruction algorithm developed and described by Mummert [B95]. The developed system was later used to process materials of German auditory tests for testing of normal-hearing and hearing impaired subjects (see chapters 5, 6 and 7).

The present system was developed using MATLAB and enabled the processing and saving of audio files in the *.wav format. System operation was completely digital. The baseline sinusoidal speech system described by McAulay and Quatieri consists of the frame forming block, the spectral analysis block, the spectral component selection block, the spectral component manipulation block, and the signal reconstruction block. The present study introduced the signal processing system consisting of the same blocks. In the following paragraphs, each of these will be described in detail. In Fig. 4.1 a block diagram of the implemented signal processing system is illustrated.

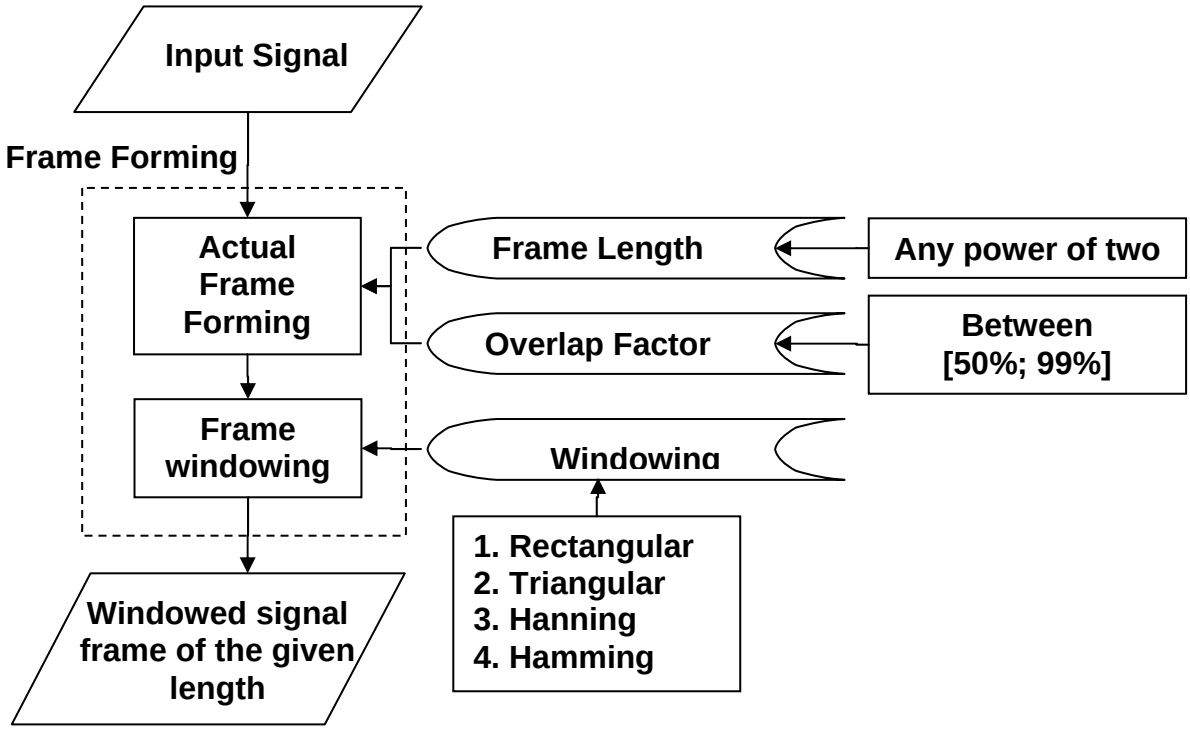


4.1 Block diagram of the implemented signal processing system.

4.2 Frame formation

The detailed block diagram of the frame formation is given in Fig. 4.2. The task of this signal processing block is the formation of the small signal chain packages out of the given input signal chain. These packages, consisting of a given number of samples, are then passed to the spectral analysis block. The parameters necessary for this operation are: frame length, overlap factor, and multiplicative window shape. All of these parameters influence the further signal processing. The frame length determines the maximum number of spectral peaks which can be selected in the spectral component selection block. The frame length, together with the overlap factor, determines the length of the reconstruction frame. All three parameters, together, influence temporal and spectral resolution of the signal processing.

The valid frames have a length of the power of two (2^n), due to the optimal calculation length of the Fast Fourier transformation (FFT). The optimal frame lengths, used in the present study, were 128, 256, and 512 samples. Note that the frame lengths presently used in commercially-available hearing devices are normally not longer than 128 samples.



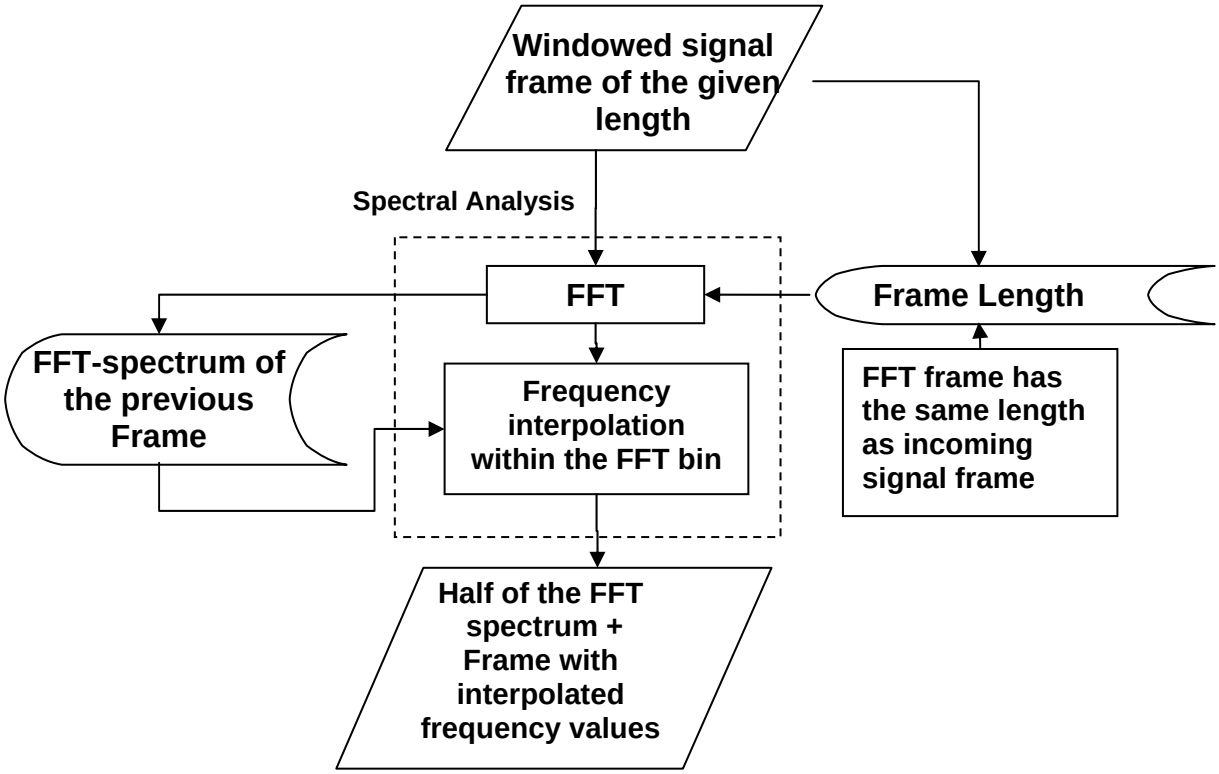
4.2 Detailed block diagram of the signal frame formation.

The overlap factor can be set between 50% and 99% of the frame length. Together with the sampling frequency of the input audio file, which in the present study was always equal to 22050 Hz, predefines the time resolution of the signal processing. The smallest value of the processing frame overlap 50%. This value is predefined by reconstruction algorithms, which are using overlap of the reconstructed frames. With increasing overlap factors increases also the time resolution of the signal processing. On the other hand, this increases the calculation time (depending on the operational resources of a system). In the present study, two different overlap factors were used: 50% and 75%.

Four different user selectable windowing shapes were offered: rectangular, triangular, Hanning, and Hamming. The form of the windowing influences the spectral shape of the calculated FFT power spectrum and the spectral resolution of the signal processing. In the present study the triangular and the Hamming windowing of sampled frames were used.

4.3 Spectral analysis

The detailed block diagram of the spectral analysis block is given in Fig. 4.4. For the spectral analysis of an audio signal the calculations of the FFT were performed. The lengths of the FFT analysis frames were the same as those formed in the frame forming block (as mentioned earlier. In the present study three length of the analysis frame were used: 128, 256, 512 points).



4.4 Detailed block diagram of the spectral analysis block.

The FFT frame-length and the sampling frequency of the audio file predefine the resolution of the spectral analysis. According to the sampling theorem, half of the sampling frequency value defines the highest frequency which can possibly be detected (at least two samples per oscillation period are required). All audio files used in the present study had sampling frequencies equal to 22050 Hz. This sets the range of the spectral analysis between zero and 11025 Hz. The half length of the FFT frame defines the maximal number of spectral divisions. The calculation of the FFT spectrum, as well as the length of the spectral intervals corresponding to the spectral division, is isotopic. Therefore it follows that the theoretically[6] maximal spectral resolution of the FFT analysis is given by the relation:

[6] Spectral resolution depends not only on the sampling frequency and the FFT frame length but also on the used windowing function.

$$\Delta Fr = \frac{Fr_{Sampl.}}{L_{FFT}}, \tag{4.3}$$

where ΔFr is the resolution of the FFT analysis, $Fr_{Sampl.}$ is the sampling frequency, and L_{FFT} is the FFT frame-length. The spectral resolution corresponds to 172.2 Hz, 86.1 Hz, and 43.0 Hz for the 128, 256 and 512 sample long frames, respectively.

4.3.1 Frequency interpolation for increasing the accuracy of spectral analysis

If the procedures of spectral manipulation do not require improved spectral resolution and the Inverse Fast Fourier Transformation (IFFT) is used for signal reconstruction, then the given spectral resolution is sufficient. If pure tones or narrow-band noise generators are used for signal reconstruction, then the frequency resolution provided by the FFT calculation is insufficient especially for the low frequencies.

A frequency approximation inside the FFT bin was applied to improve the frequency resolution of the spectral analysis block (more detailed description is given by Nelson [B98]). These calculations were based on information from two temporally subsequent signal frames respective to their FFT magnitude spectra. For each sample of the FFT spectrum the frequency interpolation was performed based on the following conditional equation:

$$\sin(2\pi * Fr * T + \theta_1) = \sin(\theta_2 + 2\pi * s), \tag{4.4}$$

where Fr is the sought after (interpolated or true) frequency, T the time period between two subsequent analysis signal frames respectively phase measurements, θ_1 the corresponding phase angle value from the first signal frame, θ_2 the corresponding phase angle value from the second signal frame, s the FFT bin factor. Both phase values are calculated from the FFT spectrum. The general equation for the phase is given by:

$$\theta = \arctan\left(\frac{\mathrm{Im}(V_{FFT}^n)}{\mathrm{Re}(V_{FFT}^n)}\right), \tag{4.5}$$

where θ is the phase angle, and V_{FFT}^n is the value of the n-th bin of the FFT spectrum. Following from equation 4.4, the equation for frequency estimation results to:

$$Fr = \frac{\theta_2 + 2\pi s - \theta_1}{2\pi T}. \tag{4.6}$$

The only unknown factor in this equation is the FFT frequency bin factor *s*. This factor can be estimated from the approximate frequency value, which can be calculated from the FFT spectrum. The following equation can be used for estimation of the *s*-factor:

$$s = \mathrm{int}\left(Fr_{appr} * T + \frac{\theta_1 - \theta_2}{2\pi} \right) \tag{4.7}$$

where Fr_{appr} is the approximate frequency as calculated from the FFT spectra (*i.e.* center frequency of FFT bins). An equation which can be used for estimation of the approximate frequency is given by:

$$Fr_{appr} = \frac{Fr_{Sampl}}{L_{FFT}}\left(n - \frac{1}{2} \right), \tag{4.8}$$

where $Fr_{Sampl.}$ is the sampling frequency, L_{FFT} is the FFT spectrum length, and *n* is the number of the corresponding bin of the FFT spectrum. Combining together 4.6, 4.7 and 4.8, the equation for the interpolated frequency is given as:

$$Fr = \frac{\theta_2 - \theta_1}{2\pi T} - \frac{1}{T}\mathrm{int}\left(\frac{T * Fr_{Sampl}}{L_{FFT}}(n - 1/2) + \frac{\theta_1 - \theta_2}{2\pi} \right). \tag{4.9}$$

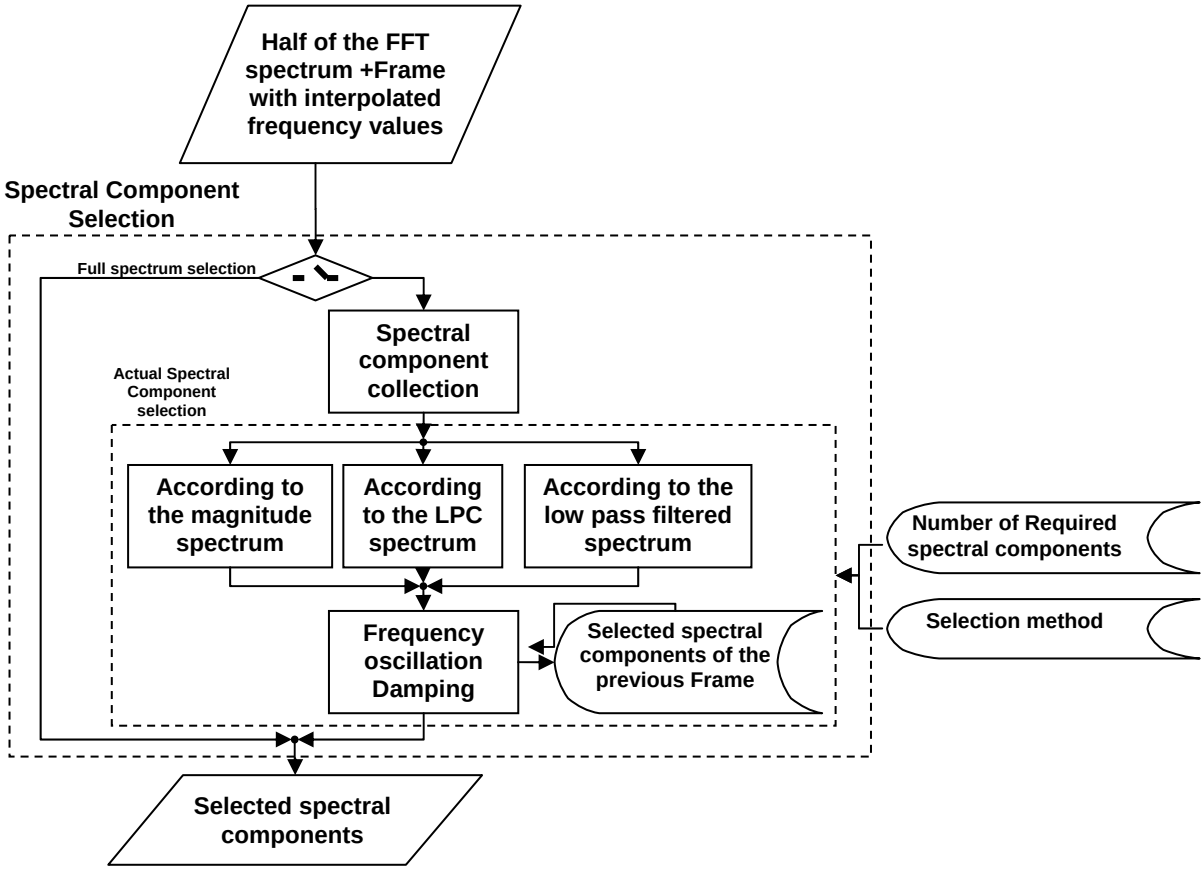
The precision of the interpolated frequency calculation depends on the phase difference and time period between two subsequent overlapping signal frames (*T*). The necessary condition for this frequency interpolation is that the frame formation frequency ($Fr_{Frame}=1/T$) has to be larger than the precision of the FFT analysis given by equation 4.3 ($Fr_{Frame}>\Delta Fr$). If this condition is not fulfilled then confusions in the calculation of the s-factor can occur. These confusions can cause mistakes in the interpolated frequency values equal to the value of the frame formation frequency ($\pm Fr_{Frame}$).

Two alternative frequency interpolation methods are the parabolic-approximation and the Feldtkeller-method [B95]. The frequency calculations for both of these methods make use of the directly neighboring FFT bin values. It is assumed, that all the frequency interpolation methods can increase spectral resolution of the conventional FFT analysis up to 20 times [B95].

If the polynomial phase interpolation is used for signal reconstruction [B84] then additional frequency interpolation is not required.

4.4 Spectral component selection

A detailed diagram of the spectral component selection block is given in Fig. 4.5.



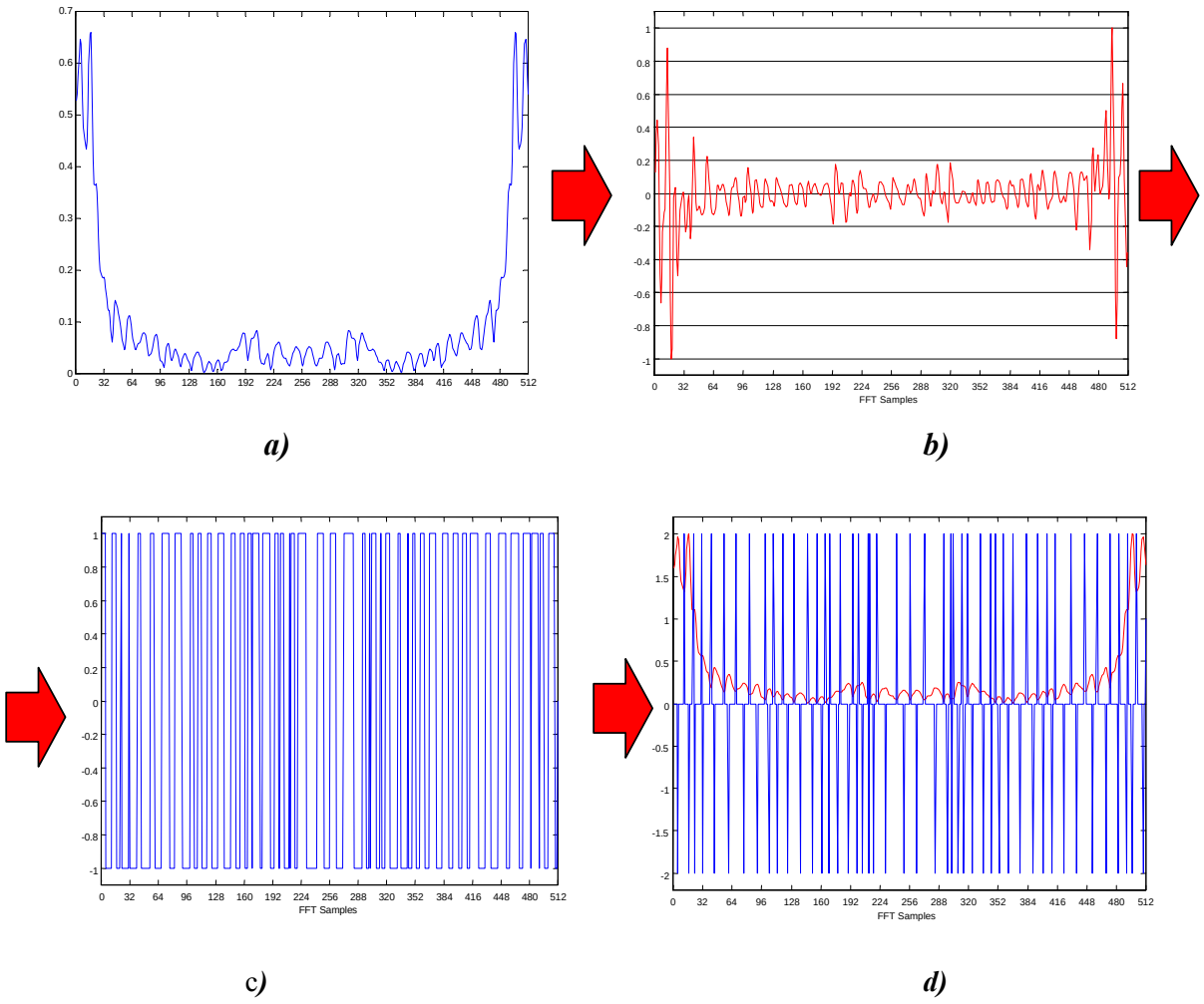
4.5 Detailed block diagram of the spectral component selection block.

Spectral component selection is performed in two steps. During the fist step all maxima of the FFT magnitude spectrum are selected. In the following second step a given number of the selected spectral maxima is chosen for further signal processing. If the full spectrum is required for further signal processing, then every sample of the FFT magnitude spectrum is considered as a selected spectral component. The maximal number of spectral components which can be used for further processing is limited to the half-length of the analysis frame. However, the maximal number of spectral maxima respectively the number of spectral components, which can be selected during spectral reduction, is limited to a quarter of the analysis frame-length.

4.4.1 Collection of spectral maxima

For the selection of the spectral maxima the FFT magnitude spectrum is used. Each sample of the FFT magnitude spectrum which is larger than its two direct neighbors is classified as spectral maximum. To collect all samples, the following mathematical operations are performed:

- calculation of the forward difference [B107],
- mathematical *signum* function, which for each element of X, SIGN(X) returns 1 if the element is greater than zero, 0 if it equals zero and -1 if it is less than zero [B128],
- repeated calculation of the forward difference,
- the selection of the samples whose values are equal to minus two (-2).



a) ***b)*** c***)*** ***d)***

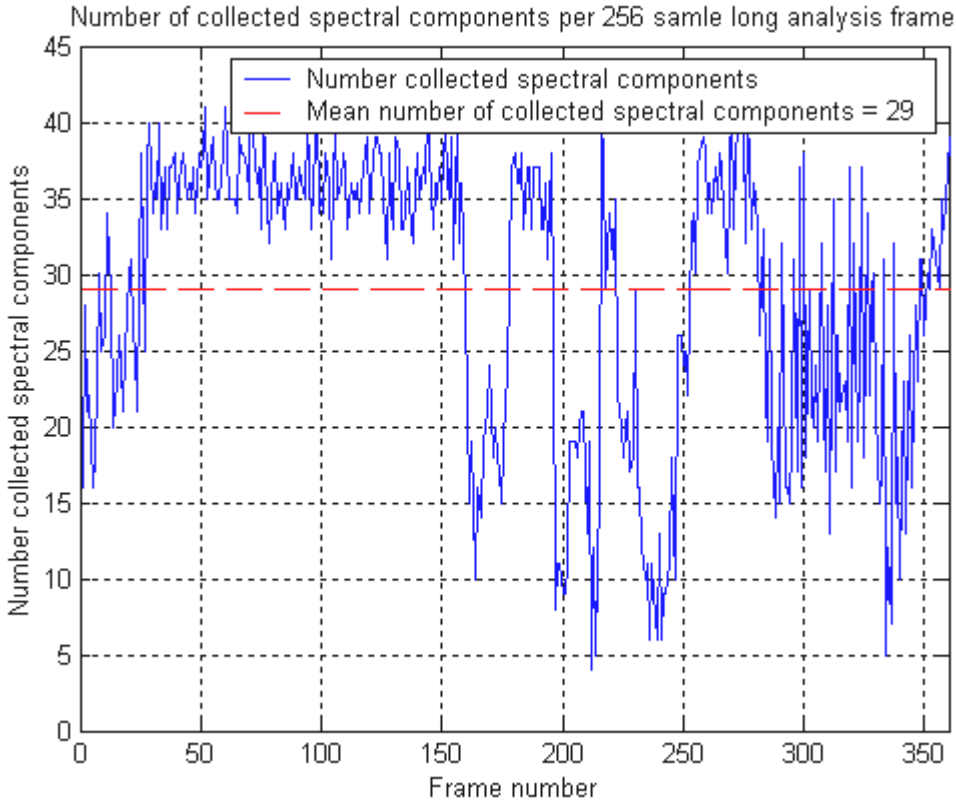
4.6 Spectral maximum collection. a) Original FFT spectrum, b) first forward difference of the FFT spectrum, c) signum function of the forward difference FFT spectrum, d) the second forward difference and original FFT spectrum. Each spectral maximum corresponds to a negative spike with a value of minus two in the two times forward differenced function.

The advantage of this calculation method is that all operations can be performed by vector algebra, which means that a loop over the sample number of the FFT spectrum is not required. The impact of these calculations on the FFT magnitude spectrum is shown on an example in Fig. 4.6.

Making use of the property that the input signal is always real (spectrum becomes symmetrical), the spectral maxima are selected only in the first half of the spectrum (positive

frequency part) and then complexly conjugated and flipped, *i.e.* symmetrically depicted in the half of the FFT spectrum.

It is obvious that the maximal number of the theoretically possible spectral maxima cannot be larger than a quarter of the analysis signal frame (FFT length). This follows from the condition that only the half of the discrete FFT magnitude spectrum corresponds to the real frequencies, and only every second sample of the discrete FFT spectrum can theoretically have a larger value than its directly neighboring samples. The number of collected spectral components per analysis frame of a speech signal, which were calculated using 256 sample FFT frame length, is given in Fig. 4.7.



4.7 The total number of collected spectral components per 256 samples long analysis frame of a speech signal; mean number of collected spectral components is equal to 29.

4.4.2 Choice of the spectral components

Speech perception depends very much on the selection of the right spectral components if only a few of them are used for signal reconstruction. Different criteria can be applied for the selection of a few spectral components out of all spectral maxima. Most of today's approaches are based on the spectral bin magnitudes, their absolute displacement within the FFT magnitude spectrum, or their displacement relative to each other.

Three different algorithms of the spectral component selection were applied in the present signal processing system: spectral component selection depending only on the magnitude, spectral component selection according to the method described by Kates [B64], which implements low pass filtering of the FFT spectrum, and a similar method which instead of using low pass filtering uses calculation of the linear prediction coefficients (LPC) of the FFT spectrum. Kates grouped the magnitude spectrum into overlapping auditory critical bands from 100 Hz to 5 kHz and made use of the smoothed spectrum in order to find the potential formant areas.

The spectral component selection based on magnitude only is the simplest possible way to select the required number of spectral components. The algorithm basically chooses the largest spectral maxima out of all identified spectral maxima. This selection mechanism is very robust. However, if it is applied to a speech signal, it tends to prefer spectral components from the lower spectral regions because these magnitudes are predominant in the speech spectrum (fundamental and formant frequencies), see Fig. 4.8.

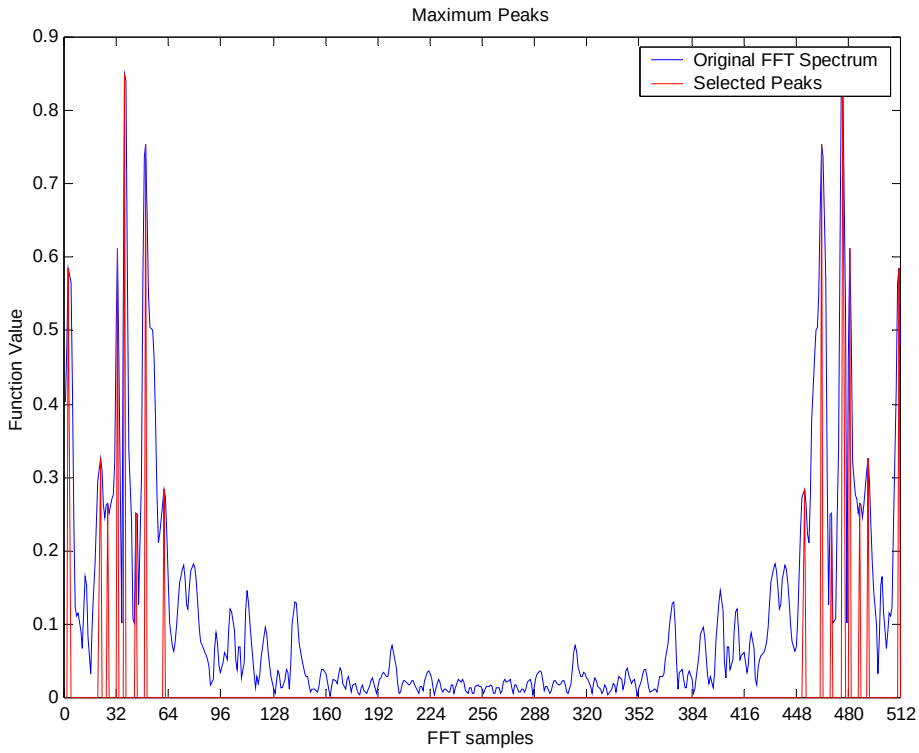


Fig. 4.8 Full 512 sample FFT magnitude spectrum of a speech signal with 8 selected spectral components in the positive and symmetrical (negative) part of spectrum.

The spectral component selection method described by Kates [B64] is based on the analysis of the low-pass filtered (LPF) FFT spectrum. According to Kates, all maxima of the LPF spectrum are calculated in the same way as described earlier (see section 4.4.1). Then the collected maxima of the LPF spectrum are sorted according to their magnitudes and considered for the choice of spectral components in the FFT spectrum beginning with the largest one. The collected maxima of the LPF spectrum predefine the spectral areas of the original FFT spectrum, from which the spectral components are selected. The center of the spectral component seeking area in the FFT spectrum is defined by the location of the spectral

maximum selected in the LPF spectrum; the width of the seeking area is set equal to the smallest distance between any two selected maxima in the LPF spectrum. The algorithm selects the largest spectral maximum in the seeking area within the original FFT spectrum. In addition, the two largest side-lobes in the seeking area are selected if their magnitudes are within a 6 dB difference to the main spectral maximum. The spectral component selection out of the collected spectral maxima continues until the required number of spectral components is chosen, or there are no remaining collected spectral maxima which have not already been selected. The same spectral maxima selection procedure is also applied if the linear prediction coefficient (LPC) spectrum is used instead of the LPF spectrum.

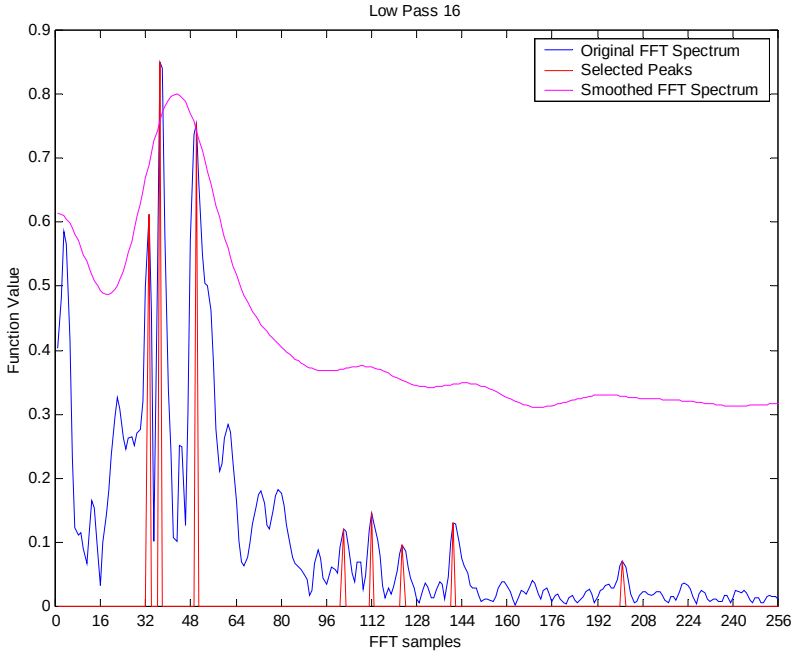


Fig. 4.9 FFT magnitude spectrum with 8 selected spectral components. Spectral component selection was performed using a modified algorithm of Kates. The low pass filtered and rescaled function of the original FFT spectrum is plotted over the original FFT magnitude spectrum.

The form of the LPF or LPC spectra depends on the choice of the low pass or linear prediction coefficient parameter settings. According to the form of the LPF or LPC spectra, different spectral components are selected (see Fig. 4.9-4.11). During the present study, experiments with normal hearing subjects using 1-4 components and with normal hearing and hearing impaired subjects using 8 selected spectral components were performed. Using this small number of spectral components, no audible differences of the processed speech signals to the original signal were observed. Note that all described spectral component selection methods chose the same spectral components if only one or two components are to be chosen. For this reason, the most simple spectral component selection method, which selects spectral components only according to their magnitudes, was emploed for speech signal processing in studies described in chapters 5-8.

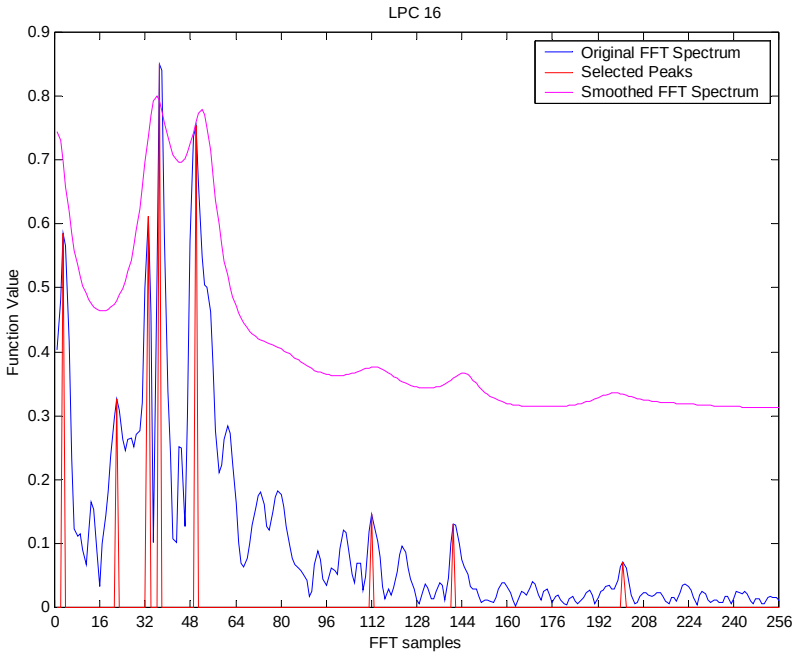


Fig. 4.10 FFT magnitude spectrum with 8 selected spectral components. Spectral component selection was performed using the LPC spectrum with 16 prediction coefficients. The rescaled LPC function of the original FFT spectrum is plotted over the original FFT magnitude spectrum.

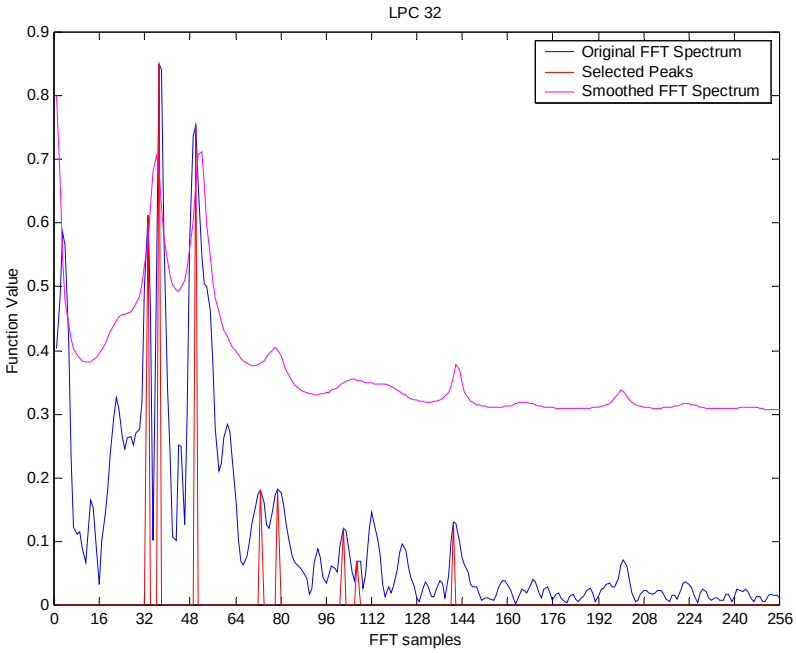


Fig. 4.11 FFT magnitude spectrum with 8 selected spectral components. Spectral component selection was performed using an LPC spectrum calculation with 32 prediction coefficients. The rescaled LPC function of the original FFT spectrum is plotted over the original FFT magnitude spectrum.

One problem deserves special attention: if a spectral component of the audio signal falls right between two FFT analysis bins, then its energy is also divided between these two frequency bins. This may cause the spectral maxima collecting algorithm to select different FFT bins for the same audio frequency in two subsequent FFT analysis frames. When sinusoids are employed for signal reconstruction, two subsequent pure tones which theoretically correspond to the same signal frequency might be generated with a frequency difference equal to the frequency resolution of the FFT analysis given by equation 4.3. Hence, undesired frequency oscillation between two neighboring FFT bins can occur.

In order to prevent these audible frequency oscillations, a damping operation is introduced. All selected peaks from each of two subsequent spectral frames are checked and their locations in the FFT spectrum are compared. In case that there are two spectral maxima, which are allocated in subsequent FFT frames, and whose position difference is on the order of only one FFT bin width, then their absolute magnitudes are compared. If their magnitudes are within 10% difference of each other then the frequency of the spectral maximum from the second FFT analysis frame is replaced with the frequency of the FFT bin whose position is the same as that of the spectral maximum in the first FFT frame. The spectral maximum in the second frame is hence shifted into the corresponding neighbouring FFT bin, and is assumed to be a regular spectral maximum in the following signal processing.

4.5 Spectral component manipulation

The developed signal processing toolbox enables different spectral manipulations. The detailed diagram of the spectral manipulation block is given in Fig. 4.12. The most important blocks, which are used in the studies described later, are the spectral compression and the spectral clipping. In addition to these two spectral operations, spectral transposition and spectral flipping were implemented. However, they were not used in the following studies.

The techniques of the spectral manipulations depend on the signal reconstruction method. If the inverse fast Fourier transformation (IFFT) is used for the reconstruction of the processed signal, then the discrete character of the FFT bin scale has to be considered. For some spectral operations, in particular for spectral compression, the discrete character of the FFT scale limits the resolution of the reconstructed frequencies (in particular at low frequencies) and can lead to ambiguous assignment between input-output and output-input frequencies. For all other spectral reconstruction methods which are operating with interpolated frequency values and reconstruct each spectral component independently, this restriction is irrelevant.

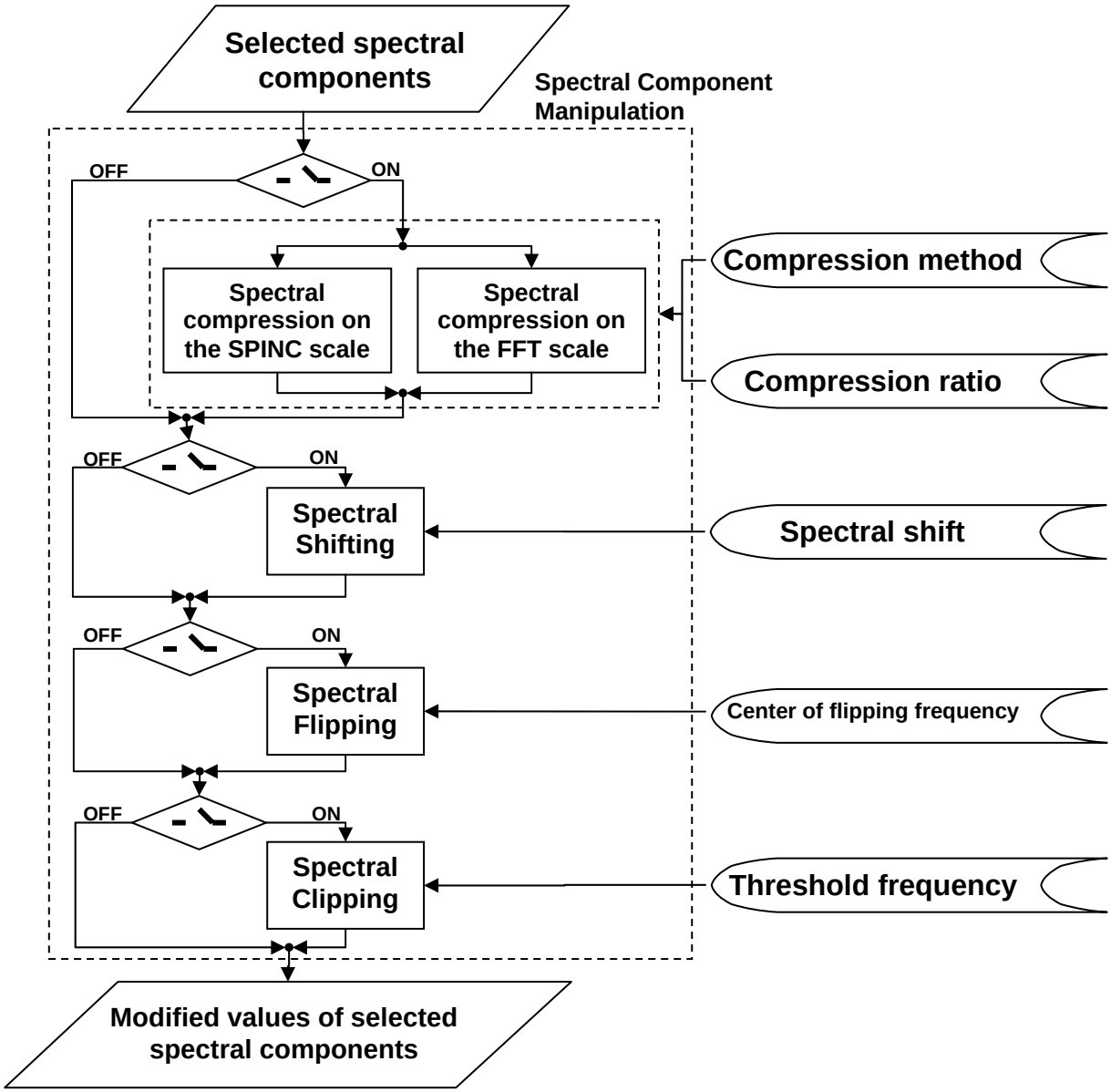
Note that in addition to the already implemented spectral manipulation methods, all spectral manipulations which can be described by any relation of the form:

$$Fr_{OUT} = f(Fr_{IN}) \tag{4.10}$$

where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency, and f is any functional dependence or law, can be easily implemented as well.

The spectral processing chain as depicted in Fig. 4.12 applies spectral manipulations in the following queue: spectral compression, spectral shifting, spectral mirroring, and spectral clipping. Any of the mentioned spectral operations can be removed from the spectral processing queue.

In the following paragraphs, each of the mentioned spectral manipulations are described in detail.



4.12 Detailed block diagram of the spectral component manipulation block.

4.5.1 Spectral compression

Generally, each spectral compression is defined by the spectral compression ratio. The spectral compression ratio is a value which quantitatively describes the amount of spectral compression and is defined by the relation:

$$CR = \frac{Fr_{original}}{Fr_{compressed}}, \tag{4.11}$$

where CR is the spectral compression ratio, $Fr_{original}$ is the original frequency, and $Fr_{compressed}$ is the spectrally compressed frequency. For linear spectral compression, the CR is a constant. For non-linear spectral compressions, the compression ratio can be a function of frequency and time:

$$CR_{NonLin} = f(Fr,t), \tag{4.12}$$

where CR_{NonLin} is the time and frequency dependent spectral compression ratio, f is any functional dependence, Fr is frequency, and t is time. Relation 4.11 also describes the particular momentaneous spectral compression ratio for non-linear spectral compressions. In the case of linear spectral compression, spectral compression ratios can also be defined by the relation between the spectral range of the audio signal before and after performing the spectral compression:

$$CR_{Lin} = \frac{SR_{original}}{SR_{compressed}}, \tag{4.13}$$

where CR_{Lin} is the constant spectral compression ratio, $SR_{original}$ is the original range of frequencies, and $SR_{compressed}$ is the spectral range after performing the spectral compression (see Fig. 4.13). Any linear and non-linear spectral compression can be defined by the equation:

$$Fr_{OUT} = \frac{Fr_{IN}}{CR(Fr_{IN},t)}, \tag{4.14}$$

where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency, and $CR(Fr_{IN},t)$ is a frequency and time dependent spectral compression ratio. Note that time-dependent spectral compression ratios are not considered during the present study.

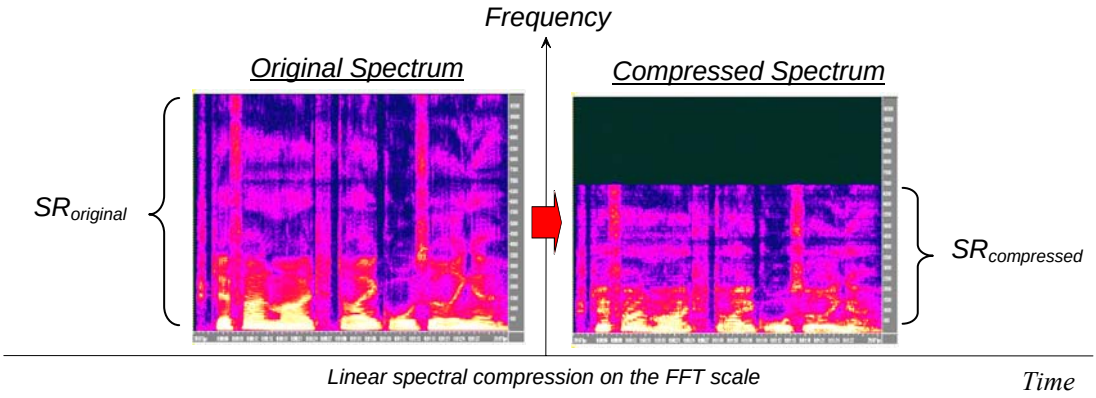


Fig. 4.13 Time-frequency plot of the linearly spectrally compressed speech signal on the FFT scale with CR = 1.6.

Two spectral compression equations were implemented in the present signal processing system: the linear spectral compression on the FFT scale, and the linear spectral compression on the SPINC scale (see capters 2.3 for the definition of the SPINC scale).

The equation describing the relation between input and output frequencies for the linear compression on the FFT scale is given by:

$$Fr_{OUT} = \frac{Fr_{IN}}{CR}, \tag{4.15}$$

where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency and CR is the spectral compression ratio (this is equal to equation 4.11).

For the spectral compression on the SPINC scale, the FFT frequency scale is transformed to the SPINC scale according to the equation:

$$\Phi(Fr) = const * \arctan\left(\frac{Fr}{const}\right), \tag{4.16}$$

where *Φ(Fr)* is the *spinc* scale *[Spinc]*, *Fr* is a frequency value on the FFT scale *[Hz]*, and *const* is an empirical constant equal to $const = \sqrt{2} * 1000$ [B126], [B127] (see also equation 2.2). After the transformation, a linear spectral compression on the SPINC scale is performed:

$$\Phi_{OUT} = \frac{\Phi_{IN}}{CR_{SPINC}}, \tag{4.17}$$

where Φ_{OUT} is the output spinc value, Φ_{IN} is the input spinc value, and CR_{SPINC} is the value of the spectral compression ratio on the SPINC scale. After the spectral compression on the SPINC scale, the inverse transformation to the FFT frequency scale is performed according to:

$$Fr_{OUT} = const * \tan\left(\frac{\Phi_{OUT}}{const}\right), \tag{4.18}$$

where Fr_{OUT} is the output frequency. Combining equations 4.16 to 4.18, the input-output frequency relation can be written as

$$Fr_{OUT} = const * \tan\left(\frac{\arctan\left(Fr_{IN}/const\right)}{CR_{SPINC}}\right). \tag{4.19}$$

Note that as already showed in Fig. 2.5, the linear spectral compression on the SPINC scale has non-linear characteristics on the FFT frequency scale[7].

By combining equations 4.15 and 4.19, the frequency-dependent spectral compression ratio on the SPINC scale according to equation 4.12 can be expressed in terms of linear FFT frequency as follows:

$$CR(Fr) = \frac{Fr}{const * \tan\left(\frac{\arctan\left(Fr/const\right)}{CR_{SPINC}}\right)}, \tag{4.20}$$

where $CR(Fr)$ is the frequency dependent spectral compression ratio on the FFT scale, Fr is the FFT frequency, and CR_{SPINC} is the constant compression ratio on the SPINC scale. This functional spectral compression ratio described by equation 4.20 is shown in Fig. 4.14. The character of the spectral compression is linear for FFT frequencies up to approximately 500 Hz and rises then exponentially (respectively according to the tangents function).

[7] The criterion of linearity is given by relation $f(Fr_1+Fr_2)=f(Fr_1)+f(Fr_2)$

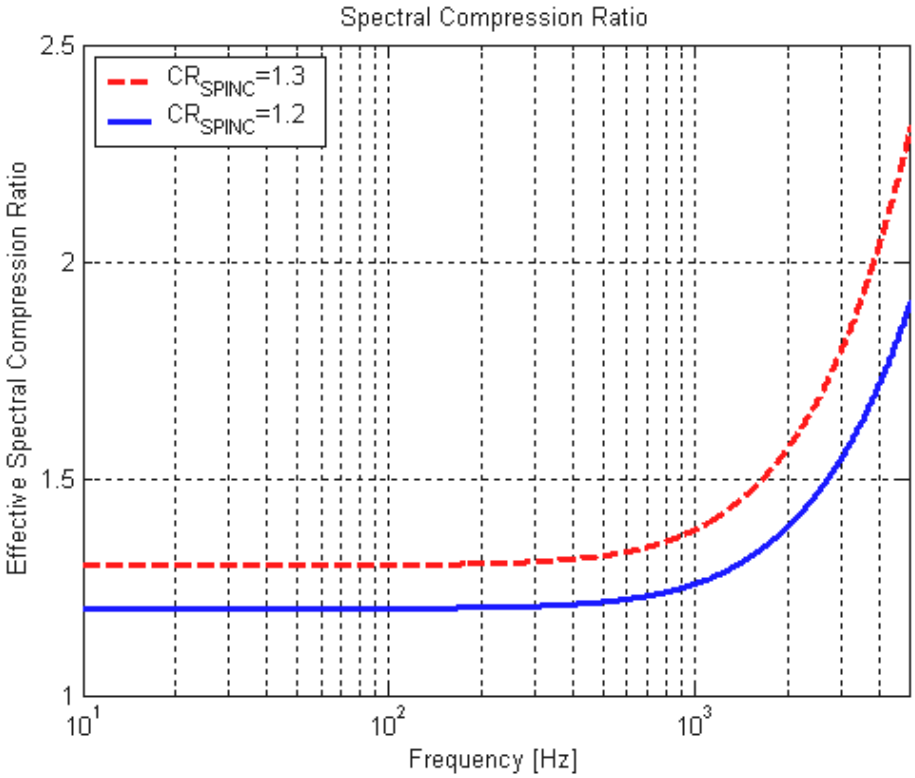


Fig. 4.14 Spectral compression ratio as a function of FFT frequency for linear spectral compression on the SPINC scale with CR=1.2 and 1.3 calculated according to equation 4.21.

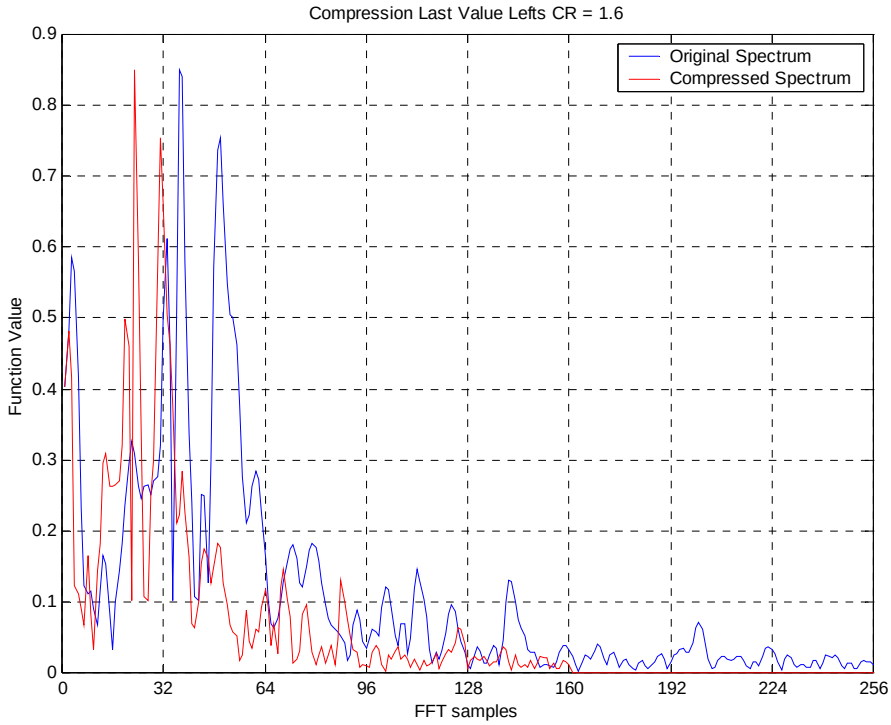


Fig. 4.15 Example of full spectrum linear spectral compression on the FFT scale with CR=1.6, original and compressed spectrum.

Both the linear spectral compression on the FFT scale and the linear spectral compression on the SPINC scale were used in studies with normal hearing and hearing impaired subjects described in chapters 6 and 7. The CR settings for the linear spectral compression on the FFT scale were CR = 1.2, 1.3, 1.6, and 1.7. The CR_{SPINC} settings used for the linear spectral compression on the SPINC scale were CR_{SPINC} = 1.2 and 1.7.

An example of full spectrum linear spectral compression on the FFT scale is given in Fig. 4.15.

4.5.1.1 Aspects of the spectral compression on the discrete FFT scale

If the IFFT is used for reconstruction of the spectrally compressed signal, then the discrete character of the FFT spectrum has to be considered. As soon as the FFT spectrum is coded in the bin number, then the linear spectral compression on the FFT scale given by equation 4.15 has to be modified to the form:

$$BN_{FFT}^{OUT} = round\left(\frac{Fr_{IN}}{CR * \Delta Fr}\right) + 1, \tag{4.21}$$

where BN_{FFT}^{OUT} is the new FFT bin number of the spectrally compressed spectral component, Fr_{IN} is the input frequency, CR is the spectral compression ratio, and ΔFr is the frequency resolution given by the FFT analysis (see equation 4.3). The addition of one is prescribed by the definition of the FFT calculations in MATLAB.

For the linear spectral compression on the SPINC scale, which is described by equation 4.19, the form of equation which is suitable for the discrete spectral compression is given by the following relation:

$$BN_{FFT}^{OUT} = round\left[\frac{const}{\Delta Fr} * \tan\left(\frac{\arctan\left(Fr_{IN}/const\right)}{CR_{SPINC}}\right)\right] + 1, \tag{4.22}$$

where Fr_{IN} is the input frequency in the linear FFT scale, and CR_{SPINC} is the spectral compression ratio in the SPINC scale.

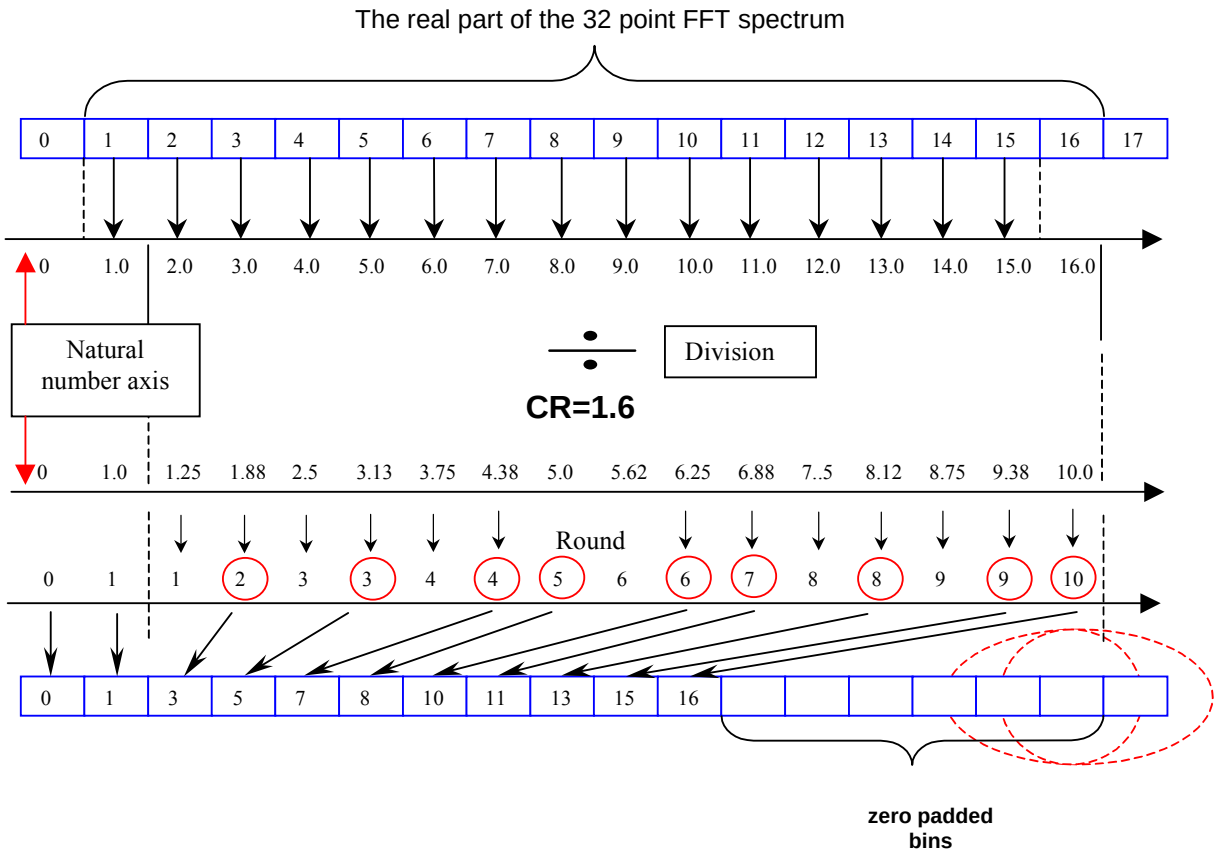


Fig. 4.16 Non-linearity caused by the non-integer spectral compression ratio. Example of the linear spectral compression on the FFT scale of the 32 point FFT spectrum with CR=1.6.

The rounding of the spectrally-compressed frequencies causes non-isomorphic spectral compression for all frequencies which belong to a particular compressed FFT bin. This effect becomes more intense especially if compression ratios with non-integer values are used (see Fig. 4.16). An additional problem is caused by the choice of the dominant spectral component after performing the spectral compression within the new FFT bin. In any case, because the IFFT reconstruction accepts only one spectral component per FFT bin, each possible choice leads to loss of spectral components if more than one component belongs to a single FFT bin to be reconstructed. In the present signal processing system, if more than one spectral component belongs to the same spectrally-compressed FFT bin, the spectral component with the highest frequency value is employed. If only one of the spectral components which belong to the same spectrally-compressed FFT bin, has non-zero magnitude then the nonzero spectral component is chosen. The compressed FFT bins whose values have no assigned spectral components are zero-padded.

The possible loss of some spectral components and also the non-isomorphic spectral compressions are proportional to the value of the spectral compression ratio (see Fig. 4.16 and 4.17). Therefore, especially for the linear spectral compression on the SPINC scale, discrete spectral compression can lead to the loss of high frequency spectral components. Considering the main goal of the spectral compression (to transpose high frequency information), the

discrete spectrum spectral compression has a rather destructive character if the linear spectral compression on the SPINC scale is applied.

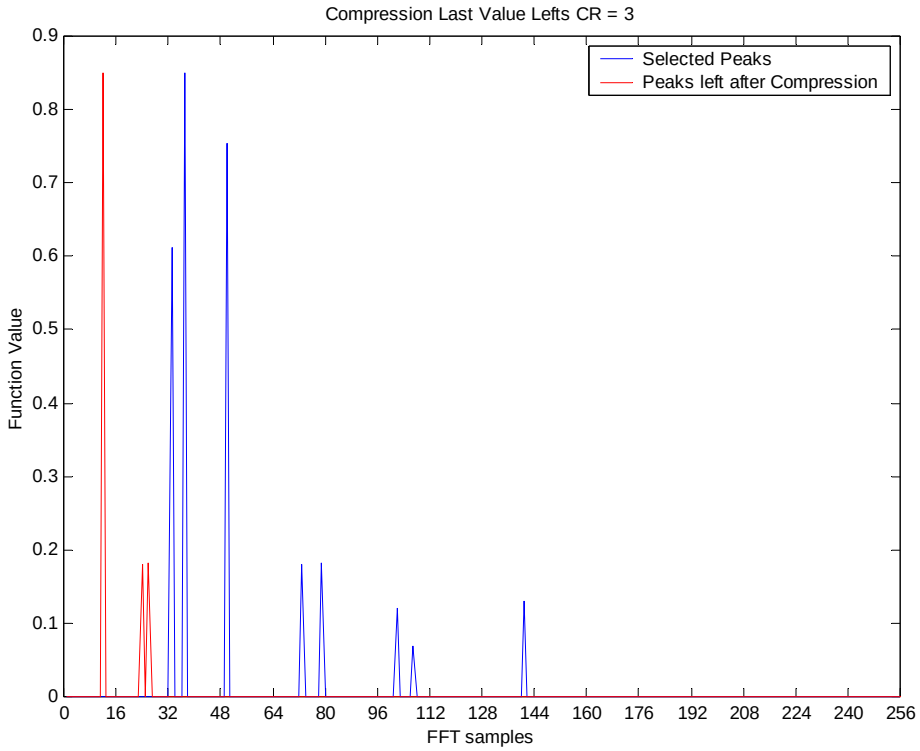


Fig. 4.17 Loss of selected spectral components caused by spectral compression on the discrete FFT scale. After spectral compression with CR=3 only three out of eight spectral components are reconstructed.

In spite of the mentioned disadvantages, there is one big practical advantage of the discrete spectral compression. When signal reconstruction after spectral manipulation is carried out by means of the IFFT, then the signal reconstruction part and therefore the entire signal-processing algorithm become simpler and the corresponding signal processing calculations become computationally more efficient and quicker than any other reconstruction approach described in section 4.6.

4.5.2 Spectral shifting

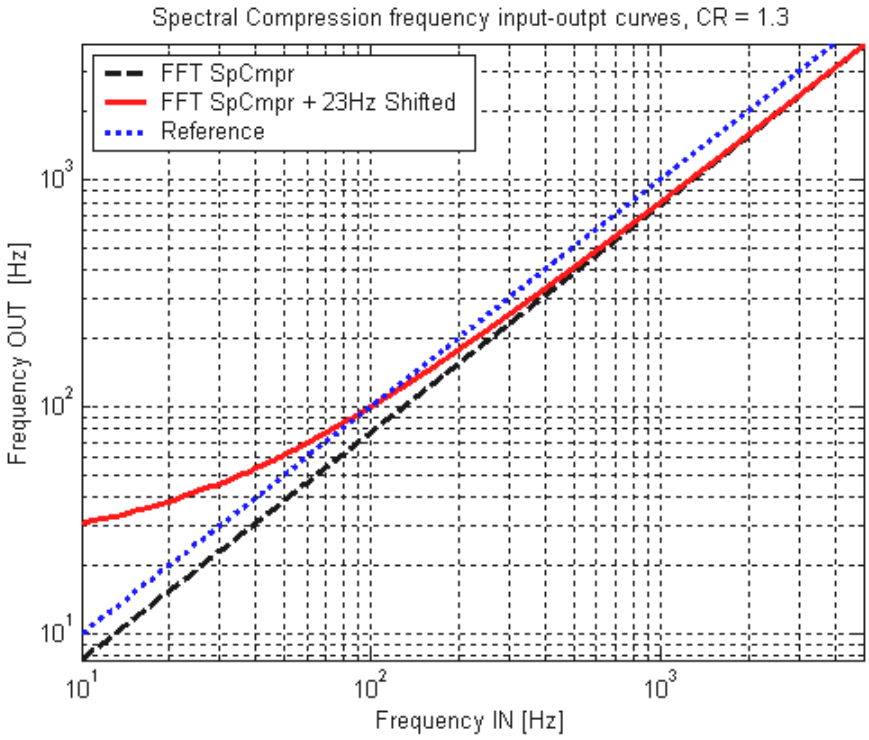
The spectral shifting can be applied separately and together with the spectral compression. It basically adds a constant spectral shift value given in Hz to the original frequency value. The frequency shift operation is described by equation:

$$Fr_{OUT} = Fr_{IN} + Fr_{Shift}, \tag{4.23}$$

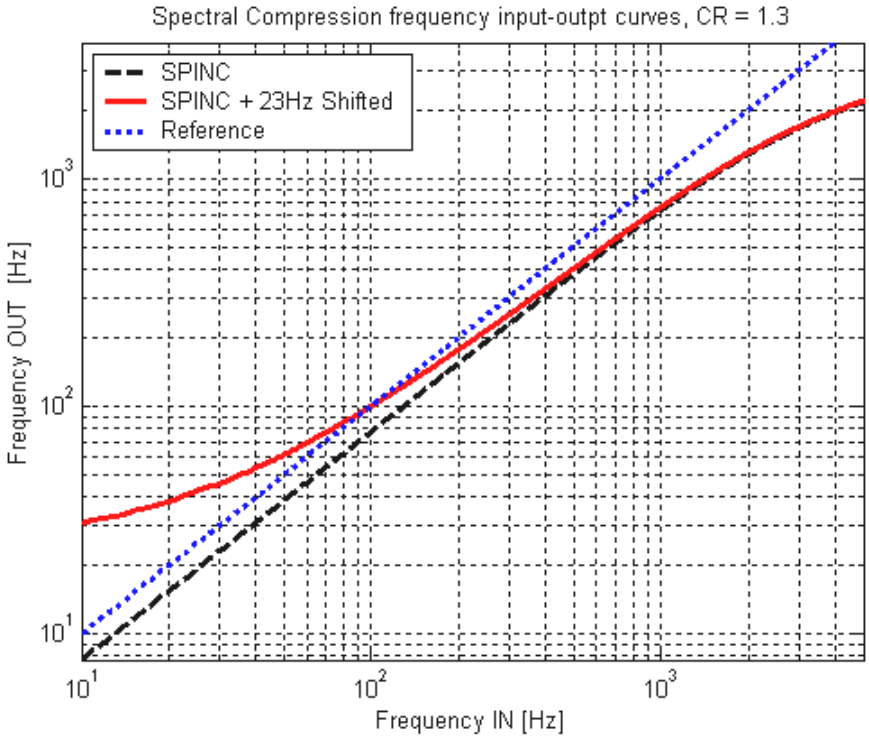
where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency and Fr_{Shift} is the spectral shift value. The restrictions caused by the discrete FFT spectrum shift are given by the frequency resolution of the FFT analysis described by equation 4.3.

Upward spectral shifting can be of interest for the effect of the low frequency lowering of both linear and non-linear spectral compressions. An example of a spectral shift of 23 Hz combined with linear spectral compression on the SPINC scale and linear compression on the FFT scale with CR = 1.3 is shown in Fig. 4.18. In this example, the spectral shift of 23 Hz results in an output frequency of 100 Hz after spectral compression and spectral shifting of the 100 Hz input frequency. This particular output frequency has hence the same value as prior to performing spectral manipulations for both the linear spectral compression on the SPINC scale and the linear compression on the FFT scale. The upward frequency shifting relative to the original frequencies which occurs for the low frequencies of up to 70 Hz, remain irrelevant because it does not cause any change for the potential speech signal (the lowest average fundamental frequency value [B11]). In the spectral area between 70 and 100 Hz, the upward frequency shift decreases from 10 to 0 Hz. The relative difference between the shifted-frequencies and non-shifted frequencies becomes less relevant, with increasing input frequency respectively for the differences between the shifted and non-shifted frequency input-output curves as they become smaller on the logarithmic scale.

It would be possible to use simple spectral shifting (without additional spectral compression usage) for improvement of vowel recognition. Example of a downward spectral shift is shown in Fig. 4.19.



a)



b)

4.18 Input-output frequency curves for spectral compression only and for spectral compression in combination with spectral shift of 23 Hz: a) linear spectral compression on the FFT scale; b) linear spectral compression on the SPINC scale.

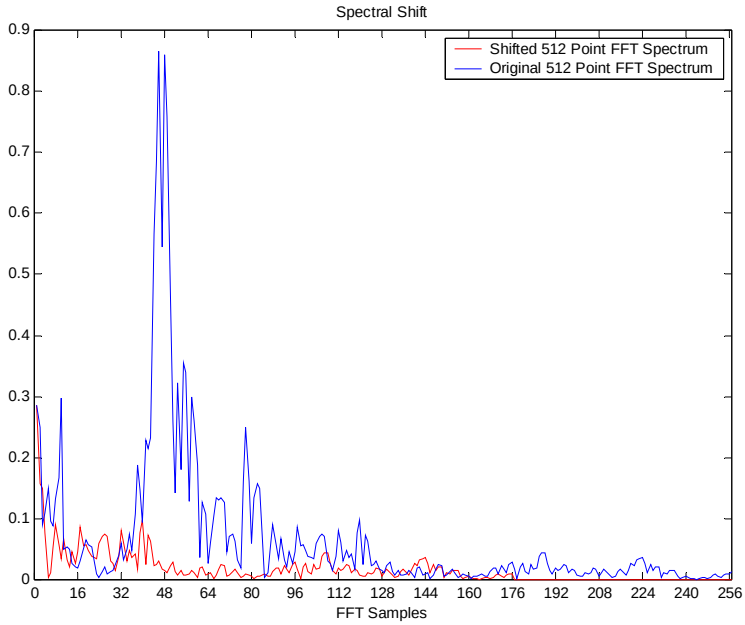


Fig. 4.19 An example of the downward spectrum shift of the 512-point FFT spectrum by 80 FFT bins corresponding to a spectral shift of 3440 Hz.

4.5.3 Spectral flipping

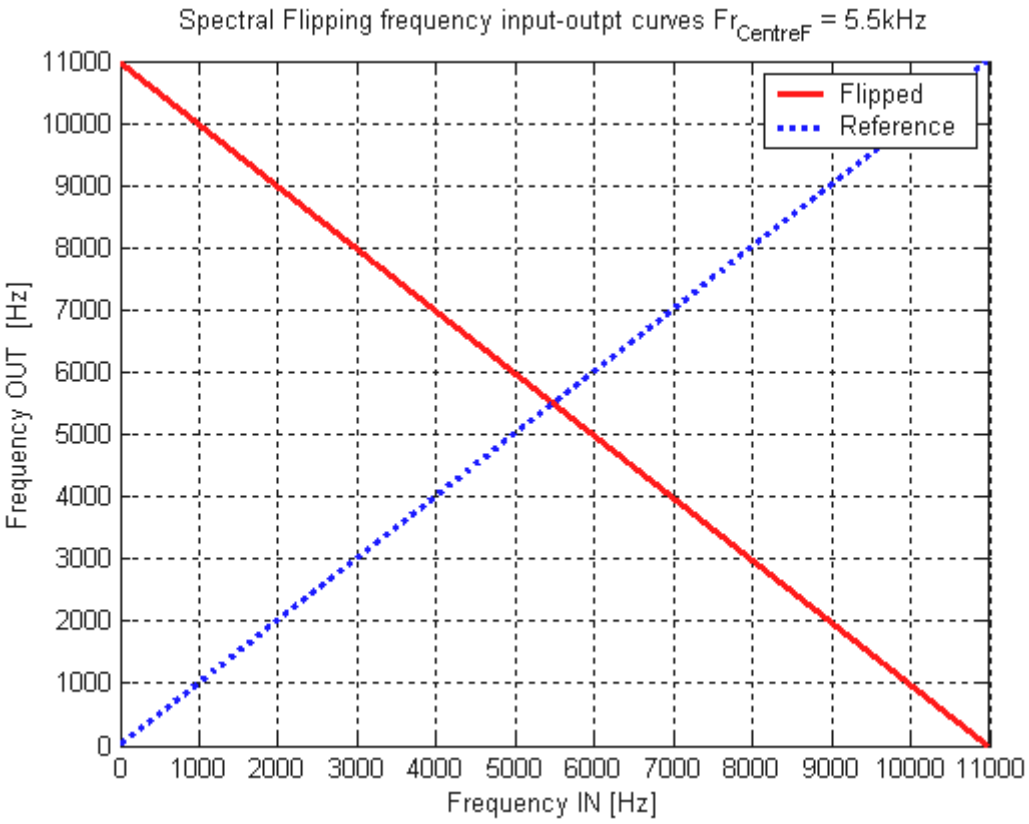
From an audiologist point of view the spectral mirroring is a spectral manipulation without any possible clinical relevance because it remains rather improbable that the higher frequency regions are less affected by hearing impairment than the low frequency areas. However, this spectral manipulation was motivated by a study of Blesser [B10] and included in the possible spectral manipulation option list.

The main parameter, describing spectral flipping, is the frequency of the spectral flipping center. This frequency is the only one, which after applying the spectral flipping operation, remains at the original value. The general equation describing spectral flipping is given as follows:

$$Fr_{OUT} = 2 * Fr_{CenterF} - Fr_{IN}, \tag{4.24}$$

where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency and $Fr_{CenterF}$ is the center of flipping frequency. Normally, the center of the analyzed spectrum is also the center frequency flipping. If the center of frequency flipping is placed below the center of the analyzed spectrum, then the frequencies which become higher than the double center of flipping frequency are not considered for further spectral operations and the reconstruction of the audio signal. If the center of frequency flipping is placed above the center of the analyzed spectrum, then only the frequency range between $Fr_{Sampl}/2$-$Fr_{CenterFv}$ and $Fr_{Sampl}/2$ is flipped,

and the remaining spectrum is left unchanged. This combination can be interesting also from a clinical point of view.



4.20 Input-output frequency curves for spectral flipping with center frequency $F_{centreF}$=5500 Hz.

Using spectral flipping, only preliminary tests, similar to those which are described by Blesser [B10], were performed. In the studies described in chapters 5, 6, and 7, spectral flipping was not applied. The input-output frequency curves for spectral flipping with center of flipping frequency equal to 5500 Hz is given in Fig. 4.20.

4.5.4 Spectral clipping

Spectral clipping is equivalent to digital low pass filtering. The main parameter of spectral clipping is the cutoff frequency. Spectral clipping replaces all spectral component values with zeros, if their frequencies are higher than the spectral clipping cut-off frequency. The equation, which describes spectral clipping, is given as follows:

$$Fr_{OUT} = \begin{cases} Fr_{IN}; Fr_{IN} < Fr_{CutOff} \\ 0; Fr_{IN} \geq Fr_{CutOff} \end{cases}, \tag{4.25}$$

where Fr_{OUT} is the output frequency, Fr_{IN} is the input frequency and Fr_{CutOff} is the cutoff frequency. An example of spectral clipping is given in Fig. 4.21.

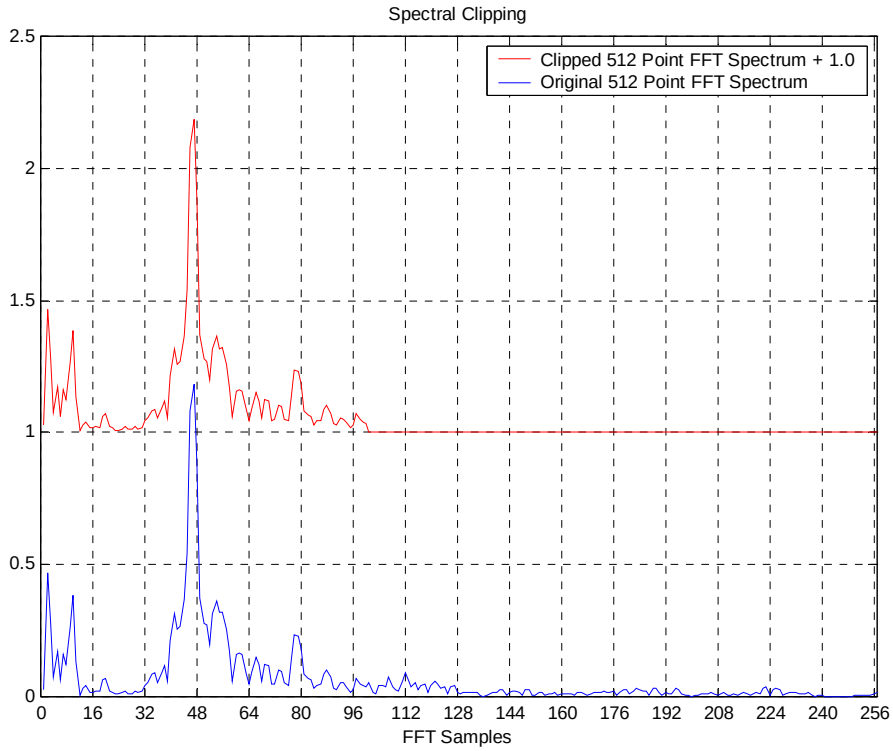
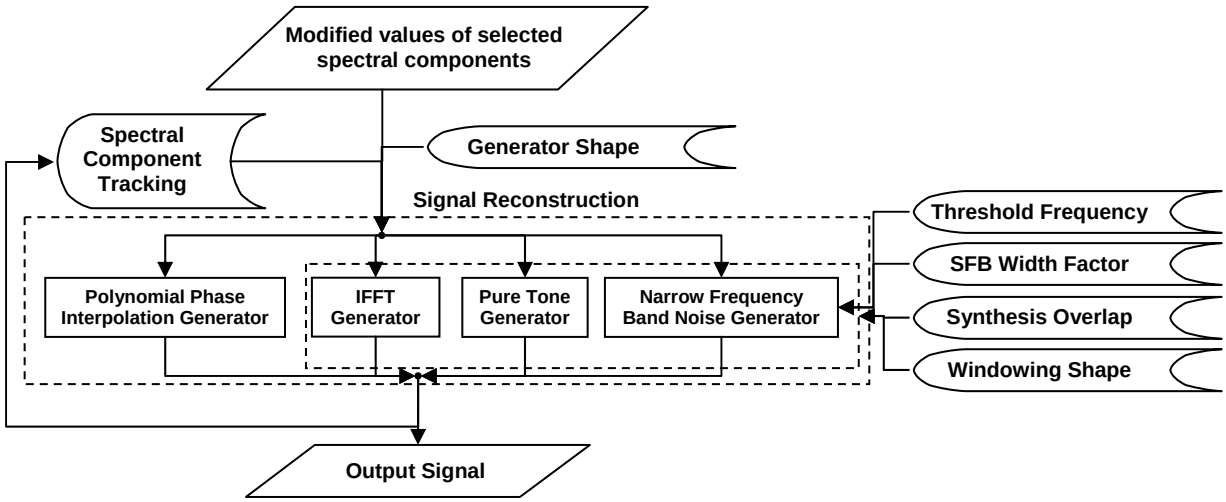


Fig. 4.21 Example of spectral clipping of the 512 point FFT spectrum with an applied threshold of 100 FFT bins which corresponds to a cut-off frequency of 4300 Hz.

Spectral clipping leads to the loss of the spectral components already selected by the spectral maxima selection algorithm. During the present study, spectral clipping with cutoff frequency Fr_{CutOff} = 2000 Hz was performed for simulating profound hearing impairment on normal hearing test subjects (see chapters 6 and 7).

4.6 Reconstruction of the processed signal

A detailed diagram of the signal reconstruction module is given in Fig. 4.22. Four different options for the reconstruction of the processed audio signal were implemented in the present signal processing system: the inverse-fast Fourier transformation (IFFT), the pure tone generator (PTG), the narrow frequency band noise generator (NFBNG) and the polynomial phase interpolation generator (PPIG) according to the baseline SiSp system described by McAulay and Quatieri [B84;B97]. The IFFT generator, the PTG and the NFBNG signal reconstruction methods use overlap of the reconstructed signal frames. Only the baseline PPIG uses a special phase interpolation method for a smooth coupling of two subsequent signal frames. In the following paragraphs, each of the algorithms for signal reconstruction will be described in detail.



4.22 A detailed block diagram of the spectral reconstruction block.

4.6.1 IFFT generator

Before the IFFT can be computed the negative part of the FFT spectrum must be reconstructed. The IFFT generator is the only signal reconstruction algorithm which reconstructs the whole spectrum in one operation. It uses the same overlap factor which is used by the signal frame building. The smallest value of the synthesis overlap factor is 50%. An output signal frame windowing is necessarily required in order to perform signal frame overlap. An optional choice between Hanning, Hamming, and triangular windowing is implemented.

As mentioned earlier, the advantage of the IFFT generator is its simplicity and the relatively quick and cheap calculation. The main disadvantage, however, is the fact that it operates with a discrete FFT spectrum which is insufficient for isotropic spectral compression operations, especially if non-linear spectral compressions have to be applied (see section 4.5.1.1). These disadvantages can partially be neglected if the value of the spectral compression ratio is an integer, or if the spectral compression is not used.

Note that the overlap of the synthesis frame causes differences between the effective number of spectral components in the reconstructed signal and the given number of spectral components used per signal reconstruction frame. In Fig. 4.23, this effect is illustrated on an example of a 128 point non overlapping FFT analysis of a signal, which was reconstructed using 50% overlap of 128 sample long analysis-synthesis frame using only one spectral component per reconstruction frame. In most of the FFT frames plotted in Fig. 4.23, more than one spectral component can be observed.

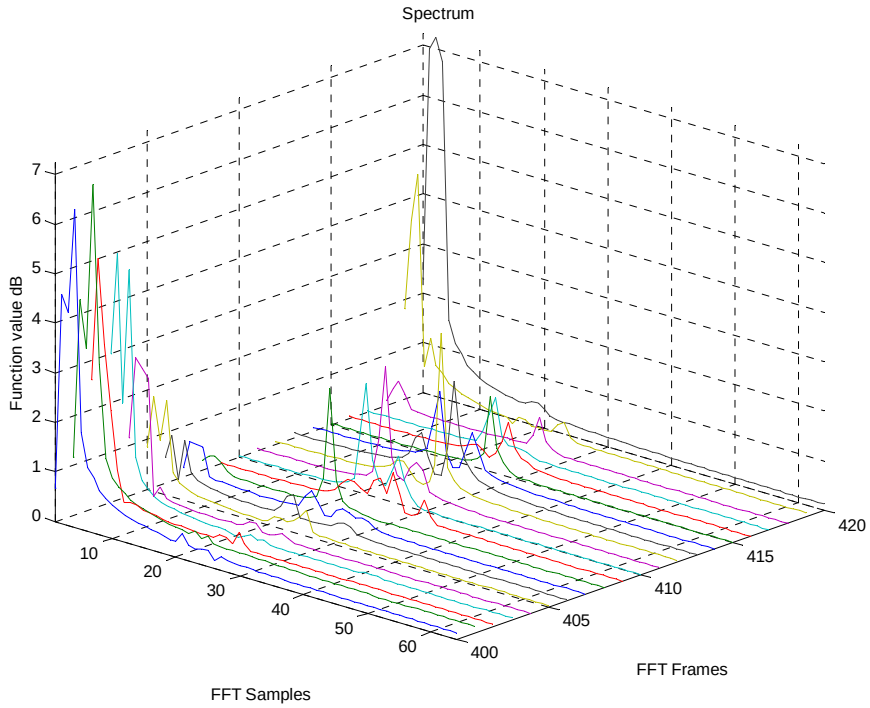


Fig. 4.23 Twenty FFT spectral frames calculated using a non-overlapping 128 point frame-length speech signal reconstructed using an IFFT generator, 128 point long analysis-synthesis frames with 50% overlap and only one spectral component per reconstruction frame.

The IFFT generator was used for the study which investigated the number of spectral components required for near to 100% speech perception using normal hearing adults.

4.6.2 Sinewave generator

The signal reconstruction using the pure tone generator is based on the assumption that every audio signal can be approximately reconstructed using a finite number of pure tones. This assumption follows directly from the theory of Fourier transformations.

The input parameters necessary for the signal reconstruction are the pure tone frequencies, and their amplitude and phase values. The frequency values are given by the frequencies of the selected spectral components assuming all previously applied spectral manipulations. The amplitudes of the pure tones are given by the absolute values of the corresponding FFT bins. In order to provide the phase information necessary for the signal reconstruction, the spectral component tracking was introduced. For this purpose, the phase, amplitude, and frequency values of each reconstructed spectral component were stored in an intermediate spectral component-tracking frame, whose content is updated after the synthesis of the reconstructed signal frame. All phase values for the first processed signal frame in the time sequence are set to zero. The tracking of the spectral components is performed by searching for the nearest spectral component within the intermediately stored spectral component tracking frame. The center of the track seeking area within the spectral component tracking frame is given by the frequency of the actually reconstructed spectral component. The width of the seeking area is equal to the width of six FFT bins given also by

six times the frequency analysis resolution (equation 4.3). If more then one spectral component is within the defined seeking area, then the one closest to the actual reconstructed spectral component frequency is taken. To calculate the phase of the reconstructed spectral component, the following equation is used:

$$\Theta_k = \frac{2\pi * Fr^{Tr} * L_{SF} * (1 - OF_{Synth}/100)}{Fr_{Sampl.}} + \Theta_k^{Tr}, \tag{4.26}$$

where Θ_k is the phase of the actually reconstructed spectral component, Fr^{Tr} is the frequency value which is taken from the tracked spectral component tracking frame corresponding to the frequency of the actually reconstructed spectral component, Fr_{Sampl} is the sampling frequency, L_{SF} the length of the synthesis frame, OF_{Synth} is the overlap factor of the synthesis frame formation, and $\Theta^{Tr}{}_k$ is the phase of the tracked spectral component taken from the spectral component tracking frame. If the corresponding spectral component cannot be found within the spectral component tracking frame, then the phase of the actual reconstructed spectral component is set to zero. Note that the phase values given by the FFT calculations are not required in the signal reconstruction.

The length of the synthesis frame, L_{SF}, is predicted by the choice of the synthesis frame overlap factor, analysis frame overlap factor, and the length of the signal analysis frame. The synthesis frame overlap factor value can be set between 50% and the maximal overlap factor value, which depends on the analysis overlap factor value and is calculated according to the following equation:

$$OF_{Synth}^{\max} = \left(50\% + \frac{OF_{Analysis}}{2} \right) [\%], \tag{4.27}$$

where $OF^{max}{}_{Synth}$ is the maximal overlap factor of the synthesis frame and the $OF_{Analysis}$ is the overlap factor of the signal analysis frame. The length of the synthesis frame is then given by:

$$L_{SF} = floor\left(L_{FFT} \frac{100 - OF_{Analysis}}{100 - OF_{Synth}} \right). \tag{4.28}$$

The synthesis equation of the signal reconstruction is finally given by:

$$A(t_m) = W(t_m) * \sum_{k=1}^{N} A_k \sin\left(\frac{2\pi * Fr_k * t_m}{Fr_{Sampl}} + \Theta_k \right), \quad (4.29)$$

where $A(t_m)$ represents the reconstructed signal at a discrete time sample $t_m = 1,...,L_{SF}$, k is the number of the actual reconstructed spectral components ($k=1,...,N$; N is the total number of spectral components used for signal reconstruction), A_k is the amplitude of the k-th reconstructed spectral component given by the magnitude of the corresponding FFT bin, Fr_k is the frequency of the k-th reconstructed spectral component Fr_{Sampl} is the sampling frequency ($Fr_{Sampl}=22050Hz$), Θ_k is the phase value of the k-th reconstructed spectral component given by equation 4.26, and $W(t_m)$ is the value of the windowing function at the corresponding discrete time. The windowing function is calculated according to the shape of the chosen synthesis window and has the same length as the synthesis frame. The three forms offered for the synthesis windowing function are the triangular, the Hanning and the Hamming window. Assuming that during signal reconstruction, spectral components are generated in a sequence, the signal generation is not as fast as the IFFT signal generation.

The difference between the implemented pure tone generator and the method described by Mummert [B95] is that the synthesis overlap factor value and the windowing shape can be chosen according to the user's requirement. The signal processing system developed by Mummert uses 50% synthesis overlap and a triangular shape of the windowing function. For the studies described in chapters 6 and 7, a 75% synthesis frame overlap and a Hanning windowing shape were used.

4.6.3 Narrow frequency band noise generator

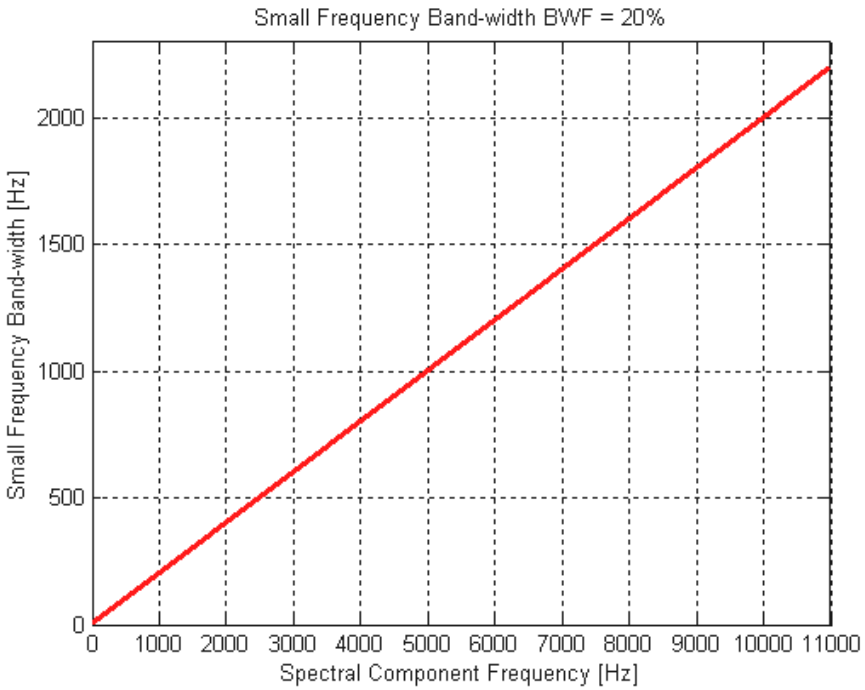
The narrow frequency band noise generator was implemented to suppress the generation of musical noise which occurs if noise-like signals have to be reconstructed. The narrow frequency band noise (NFBN) generator uses basically the same input parameters and frame construction characteristics as the sinwave generator described above. The additional parameters describing the NFBN generator are the threshold frequency and the narrow frequency bandwidth factor.

The main difference between the pure tone generator and the NFBN generator consists in the fact that if the frequency of the actually reconstructed spectral component is larger than a given threshold frequency, a narrow frequency band noise is generated instead of a sinusoid. All other spectral components, which are lower in frequency than the threshold frequency, are generated according to equation 4.29. The value of the threshold frequency can be freely set by the user of the signal processing system. The default frequency threshold value is 1000 Hz. The value of the threshold frequency remains the same during the whole signal processing.

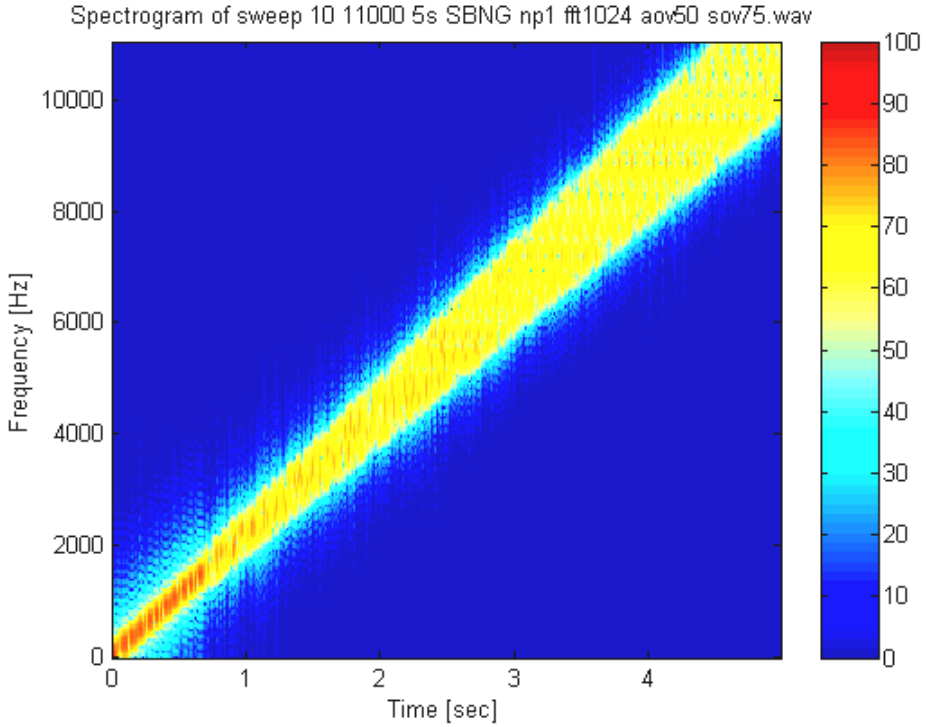
The width of the generated noise band is obtained by the narrow frequency bandwidth factor (BWF) multiplication with the spectral component frequency. The narrow frequency bandwidth factor is given in percents. The default value of the narrow frequency bandwidth factor is 20%. However, it can be changed according to the user's requirement. The narrow band noise bandwidth for the default BWF=20% depending on spectral component frequency is given in Fig. 4.24. If a large number of spectral components is used for signal

reconstruction, then using large values of the BWF leads to spectral overlapping of the reconstructed spectral components. This happens especially in the high frequency area.

For narrow band frequency noise generation, the FFT and the IFFT calculations of random number sequences are used. The length of the random number sequences and the FFT and IFFT frames is equal to 22050 samples. After the generation of the random number sequence, the FFT spectrum is calculated, followed by digital band-passing to the FFT spectrum portion which corresponds to the narrow frequency band with the center frequency equal to the frequency of the actually reconstructed spectral component. The part of the calculated FFT spectrum outside the narrow-band bandwidth is zero-padded. This operation is similar to the spectral clipping described earlier (section 4.5.4). After the spectral band clipping, the IFFT of the clipped spectrum is calculated and a frame with a length equal to the synthesis frame-length is taken from the reconstructed narrow band noise signal. Summing and windowing of all the reconstructed spectral components is performed in the same way as for the pure tone generator. Note that the spectral component tracking is relevant only for frequencies below the given threshold frequency. An example of narrow band noise generated with the described approach is shown in Fig. 4.25. The width of the noise band as a function of center frequency is clearly visible.



4.24 Narrow frequency band bandwidth used for the NFBN generator, with a narrow frequency band bandwidth factor equal 20%.



4.25 Sweep signal (10Hz-11kHz in 5 sec) which was generated using a NFBN generator, with a narrow frequency band bandwidth factor equal to 20% (analysis overlap 50%, synthesis overlap 75%).

4.6.4 Polynomial phase interpolating generator

The polynomial phase interpolating generator (PPIG) which is implemented in the present signal processing system is equivalent to the one described by McAulay and Quatieri [B84]. In contrast to the other signal reconstruction methods, this algorithm does require neither overlap of the synthesis frames nor their windowing. It uses instead a special phase interpolation mechanism enabling direct matching and junction of the sequential spectral components. The length of the synthesis frame, L_{SF}, is given by the equation:

$$L_{SF} = floor\left[L_{FFT} * \left(1 - \frac{OF_{Analysis}}{100\%} \right) \right], \tag{4.30}$$

where L_{FFT} is the analysis frame length, and $OF_{Analysis}$ is the analysis frame overlap factor. (Note that equation 4.30 equals equation 4.28 where the synthesis overlap has been set to zero.)

The spectral component tracking, which was described already for the pure tone generator (section 4.6.2), is essential for the signal reconstruction using the PPIG. However, there is a difference in using and saving the spectral components phase values. In contrast to the pure tone generator, the PPIG operates with the measured spectral component phases. In

addition to the phase interpolation, the PPIG performs the amplitude interpolation between the two subsequent spectral components. The amplitude interpolation is performed according to the following equation:

$$\tilde{A}(t_m) = \hat{A}^i + \frac{\left(\hat{A}^{i+1} - \hat{A}^i\right)}{L_{SF}} t_m , \tag{4.31}$$

where $\tilde{A}(t_m)$ is the amplitude at the t_m-th time sample in the synthesis frame $t_m = 0,1,...,L_{SF}-1$, $\hat{A}^i$ is a measured amplitude in the previous frame corresponding to the set of parameters in the current frame taken from the spectral component tracking frame, $\hat{A}^{i+1}$ is a measured amplitude in the current frame, L_{SF} is the synthesis frame length given in samples. If there is no correlated spectral component within the tracking area, then the $\hat{A}^i$ is set to zero.

For the unwrapped phase interpolation, the following cubical polynomial equation is used according to McAulay and Quatieri [B84]:

$$\tilde{\Theta}(t) = \hat{\Theta}^i + \hat{\omega}^i t + \alpha(M^*)t^2 + \beta(M^*)t^3 , \tag{4.32}$$

where i is the frame sequence coefficient corresponding to the previous frame, $\Theta(t)$ is the continuous phase function depending on the continuous time t, with $t = 0$ corresponding to the beginning of the i-th frame and $t=T$ corresponding to the beginning of frame i+1; T is the time interval between the two subsequent analysis frames measured in ms, Θ^i is the phase value taken from the spectral tracking frame corresponding to the actual reconstructed spectral component, $\acute{\omega}^i$ is multiplied by the value 2π of the corresponding spectral component frequency taken from the tracking frame, and $\alpha(M^*)$ and $\beta(M^*)$ are special phase interpolation coefficients. They can be calculated according to the following equation:

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \frac{3}{T^2} & \frac{-1}{T} \\ \frac{-2}{T^3} & \frac{1}{T^2} \end{bmatrix} \begin{bmatrix} \hat{\Theta}^{i+1} - \hat{\Theta}^i - \hat{\omega}^i T + 2\pi M \\ \hat{\omega}^{i+1} - \hat{\omega}^i \end{bmatrix}, \tag{4.33}$$

where M is an integer which has to be chosen as close to the value of the x-factor as possible. The x factor can be calculated according to the equation:

$$x = \frac{1}{2\pi}\left[\left(\hat{\Theta}^i + \hat{\omega}^i T - \hat{\Theta}^{i+1}\right) + \left(\hat{\omega}^{i+1} - \hat{\omega}^i\right)\frac{T}{2}\right]. \tag{4.34}$$

If there is no correlated spectral component within the tracking frame, then the $\acute{\omega}^i$ is taken equal to $\acute{\omega}^{i+1}$ and Θ^i is calculated backwards as if the $\acute{\omega}^i=\acute{\omega}^{i+1}$. The backward phase calculation is performed as follows:

$$\hat{\Theta}^i = \hat{\Theta}^{i+1} - \hat{\omega}^{i+1}T \,. \tag{4.35}$$

In this case, if $x=0$, the values of $\alpha(M)$ and $\beta(M)$ also equal zero, and the phase interpolation equation 4.33 can be written in the following form:

$$\tilde{\Theta}(t) = \hat{\Theta}^{i+l} - \hat{\omega}^{i+1}T + \hat{\omega}^{i+1}t \,. \tag{4.36}$$

Finally, the general equation for the synthesis frame construction is:

$$\tilde{s}(t_m) = \sum_{k=1}^{N} \tilde{A}_k(t_m)\cos\left[\tilde{\Theta}_k(t_m)\right], \tag{4.37}$$

where $\tilde{s}(t_m)$ is the value of the reconstructed frame at the sample t_m, $\tilde{A}_k(t_m)$ is the interpolated amplitude, $\tilde{\Theta}_k(t_m)$ is the interpolated phase value, and k is the number of the actual reconstructed spectral component. For more details see McAulay and Quatieri [B84]

4.7 Temporal modification scheme

To enable various temporal modifications of an audio signal, the implemented SiSp baseline system described in previous sections was modified. These modifications were necessary because on the one hand, many of the simplifications introduced in the baseline system can not be applied in the temporal-modification scheme, and on the other hand, the temporal modification scheme requires additional operations such as speech segment detection, pitch estimation, and time-scale modification. The block diagram of the temporal modification module is given in Fig. 4.26.

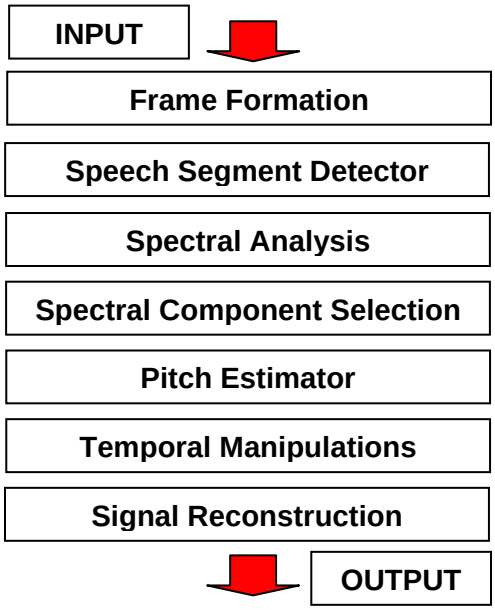


Fig. 4.26 Block diagram of the implemented temporal modification signal processing system.

The main components of the temporal modification module are: the frame formation block, the speech segment detector, the spectral analysis block, the spectral component selection block, the pitch estimator, the temporal manipulation block, and the signal reconstruction block. In the following, each of the components will be shortly described. For a more detailed description of the temporal modification system see [B3].

4.7.1 Frame formation, speech segment detector, spectral analysis, and spectral component selection

The input frame formation is provided in exactly the same way as described in section 4.2. The frame lengths were always equal to 128 samples, corresponding to 5.8 ms at the employed sampling frequency of 22.05 kHz. The formed frames are delivered to the speech segment detector.

The speech segment detector provides spectral center of gravity (CG) calculations and, depending on the estimated CG value, classifies the incoming frames into voiced or voiceless. To determine the centre of gravity, the FFT of the input signal frames is computed and the CG values are estimated according to the following equation:

$$CG = \frac{\sum_{f_i=f_{\min}}^{f_{\max}} f_i \cdot |X(f_i)|}{\sum_{f_i=f_{\min}}^{f_{\max}} |X(f_i)|}, \tag{4.38}$$

where $f_i = 0,\ldots,Fr_{sampl}$ is frequency bin i of the FFT, and $|X(f_i)|$ is the magnitude FFT spectrum at bin i.

If the estimated *CG* value is smaller than 2.2 kHz then the frame is classified as voiced, otherwise it is assumed to be voiceless. If the current frame is classified as voiceless, then it is delivered to the spectral analysis block without any changes. In case of a voiced frame, the previously analyzed frame is concatenated to the current frame, but only if it is also classified as voiced. This operation can concatenate up to four subsequent voiced frames and improves resolution of the spectral analysis block as a result of the increased analysis frame length.

The spectral analysis and spectral component selection blocks are similar to the blocks already described in section 4.3 and 4.4.

4.7.2 Pitch estimator

The signal's pitch period, T_{pitch}, is used for the onset time calculations, which are necessary for the signal reconstruction scheme. In every frame, the pitch of the signal is estimated by means of the auto-correlation function (ACF) of the time sequence. The peak value of the calculated ACF corresponds to the dominant period and is assumed to be the pitch period T_{pitch}, respectively the pitch frequency (f_{pitch}). The detailed pitch estimation technique used in the present system is described in [B69].

4.7.3 Temporal modification block and signal reconstruction

4.7.3.1 Onset time

The onset time, $t_0(m)$, of the current frame m is calculated from the onset time of the previous frame $t_0(m-1)$ and the pitch period T_{pitch} according to the following relation:

$$t_0(m) = t_0(m-1) + k \cdot T_{Pitch}, \tag{4.39}$$

where k = 0,1,2,… is an integer factor and is chosen such that the onset time is set as close as possible to the center of the current analysis frame.

4.7.3.2 Temporal modification factor

A multiplicative temporal modification factor, ρ, is introduced to apply various time-modifications of different speech segments. Values of the temporal modification factor smaller than one correspond to a temporal shortening, and values larger than one correspond to a temporal prolongation. For different speech segments which are classified according to the CG calculations (see equation 4.38), different temporal modification factor values can be applied.

4.7.3.3 Synthesis of temporally modified frames

The reconstruction of the processed signal frames is similar to the polynomial phase interpolating generator described earlier (see section 4.6.4) with the differences that the length

of the reconstructed frame and the interpolated phase values also have to be corrected according to the temporal modification factor.

4.7.3.3.1 Amplitude interpolation

For the reconstructed spectral component amplitude calculation, $\widetilde{A}(t_n)$, the following relation is used:

$$\widetilde{A}(t_n) = \hat{A}^i + \frac{\left(\hat{A}^{i+1} - \hat{A}^i\right)}{N_{\text{mod}}} t_n, \tag{4.40}$$

where t_n = 0,1,…(N_{mod}-1) is the sample time, $\hat{A}^i$ is the end amplitude of the spectral component used in previous frame, $\hat{A}^{i+1}$ is the amplitude of the spectral component in the current frame, and N_{mod} is the modified frame length. The modified frame length is estimated based on the temporal modification factor ρ:

$$N_{\text{mod}} = \rho \cdot T \cdot Fr_{Sampl}, \tag{4.41}$$

where T is the original signal frame period, and Fr_{Sampl} is the sampling frequency.

4.7.3.3.2 Phase interpolation

A modification of equation 4.33 is used for the estimation of the polynomial phase interpolation indices $\alpha(M)$ and $\beta(M)$:

$$\begin{bmatrix} \alpha(M) \\ \beta(M) \end{bmatrix} = \begin{bmatrix} \frac{3}{(\rho T)^2} & \frac{-1}{\rho T} \\ \frac{-2}{(\rho T)^3} & \frac{1}{(\rho T)^2} \end{bmatrix} \begin{bmatrix} \hat{\Theta}^{i+1} - \hat{\Theta}^i - \rho\hat{\omega}^i T + 2\pi M \\ \hat{\omega}^{i+1} - \hat{\omega}^i \end{bmatrix}. \tag{4.42}$$

where the modified factor M can be calculated according to:

$$M = round\left(\frac{1}{2\pi} \left[\left(\hat{\Theta}^i + \rho\hat{\omega}^i T - \hat{\Theta}^{i+1}\right) + \left(\hat{\omega}^{i+1} - \hat{\omega}^i\right) \frac{\rho T}{2} \right] \right). \tag{4.43}$$

The final phase interpolation is calculated according to equation 4.32, and the frame reconstruction is performed according to equation 4.37 (see section 4.6.4).

4.8 Summary

The employed signal processing system implements frame forming, spectral analysis, spectral component selection, spectral component manipulation, and signal reconstruction.

For frame forming, different frame lengths, frame overlaps and windows are available. Frame analysis is carried out by means of the FFT. Spectral component selection implies identification of the spectral maxima and different means for the selection of particular spectral components for further processing based on their magnitudes.

Spectral component manipulation comprises spectral compression, spectral shifting, spectral flipping, and spectral clipping. Signal reconstruction, finally, is implemented by means of IFFT, pure tone generation, narrow band noise generation, and polynomial phase interpolation.

The employed temporal modification signal processing system implements temporal frame forming, speech segment detection, spectral analysis, spectral component selection, modification, and signal reconstruction. For the temporal modifications, a multiplicative factor is introduced according to which phase frame lengths are modified. Different time modification factor values can be applied to different speech segments. The polynomial phase interpolation is used for the reconstruction of the temporally-modified signal.

All signal processing employed in the tests with experimental subjects described in the following is carried out by means of the signal processing toolbox described in this chapter of the thesis. The loudness of the individual processing schemes has not been investigated.

Signal processing system modulus	Parameter setting options		Used parameter settings
Spectral analysis	Frame length	2^n	128, 256, 512
	Windowing shape	Rectangular, triangular, Hamming, Hanning	Hanning
	Overlap	Optional	75%
Spectral component selection	Selection method	Max, low-pass, LPC	Max
	Number of spectral components	Optional	1-5, 8, full spectrum
Spectral component manipulation	Spectral compression method	LIN, SPINC	LIN – 1.3, 1.6, 1.7 SPINC- 1.2, 1.3
	Spectral shifting	Optional	
	Spectral flipping	Optional	
	Spectral clipping	Optional	2000 Hz
Signal reconstruction	Reconstruction method	IFFT, Sinus wave generator, NFBNG, PPIG	IFFT, Sinus wave generator

Tab. 4.1 Summary on implemented signal processing options and possible parameter settings and parameter settings used in further studies.

Chapter 5

Friday

"The naming goes recklessly on, in spite of anything I can do. I had a very good name for the estate, and it was musical and pretty – Garden of Eden."

M. Twain

Experiments with spectral reduction

5.1 Overview

In an investigation of a sinusoidal speech signal processing scheme, the number of spectral components as well as the temporal and spectral resolution required for 100% speech intelligibility were examined. German vowels, consonants and sentences (C12 and Vo8 tests and Oldenburg sentence tests) spoken by a male speaker were processed with an algorithm based on the sinusoidal speech model of Quatieri and McAulay. This algorithm uses a limited number of spectral components and provides different temporal and spectral resolutions. Speech intelligibility tests in quiet were performed in a sound proof room with normal hearing native German speaking adults.

The test results indicate that different temporal and spectral resolutions require different numbers of spectral components for 100% speech recognition in sentences. A surprising observation was that for the highest temporal resolution only one spectral component used in the reconstruction already provided 100% identification. For sufficient consonant and vowel identification (~90%), at least two spectral components were required. Vowels seem to be more affected by spectral reduction than consonants.

The second conclusion is that there is possibly a minimal spectral component per time ratio which is required for good speech perception. This minimum spectral component per time ratio lies between two and four spectral components per 1.5 ms. It was not possible to determine the minimum spectral component per time ratio more precisely with the actual signal processing strategy.

A tendency for increased speech intelligibility was observed by either increasing temporal resolution respectively decreasing spectral resolution, with an increasing number of spectral components. In addition, significant learning effects were observed and measured within the consonant identification test.

5.2 Motivation

This study on the minimal number of spectral components required for sufficient speech perception (close to 100%) was motivated by the inconsistent results of the various similar

studies (see chapter 3). The main purpose of this study was to adjust the presently implemented signal processing system in terms of the required minimal number of spectral components and the required spectral and temporal resolution for further investigation of different spectral manipulations (see chapters 6 and 7). Since the amount of auditory filter broadening by the potentially profoundly-impaired subjects was unknown, this study also investigated the upper limits of spectral reduction, which could possibly be used to compensate for the limited frequency resolution of the experimental subjects.

5.3 Method

5.3.1 Signal processing and parameter settings

In order to accomplish a flexible spectral reduction of audio signals the signal processing system based on the Sinusoidal Speech model [B84] described in chapter 4 was used. Fig. 5.1 shows the main components of this system.

In the frame forming block the incoming signal is divided into frames overlapping by 75 %. Three different frame lengths for processing the incoming signal were used. After frame formation, the frames are provided separately to the spectral analysis block.

In the spectral analysis block, each frame is Hanning windowed and subjected to Fast Fourier Transformation (FFT) of the processed frame is calculated. The lengths of the FFT windows are equal to the incoming frame lengths. The calculated FFT magnitude spectra are then conveyed to the spectral reduction block.

Fig. 5.1 Block diagram of sinusoidal analysis/synthesis system.

The spectral reduction of the incoming signal is divided into two separate operations: spectral peak picking and peak selection. The peak picking operation is performed by identifying all the points in the FFT magnitude spectrum that are larger than their two closest neighbors and by replacing all other points of the spectrum by zero. This operation causes a spectral reduction of at least a factor of two and emulates something similar to a spectral enhancement or spectral sharpening operation [B45]. Following the peak picking operation, peak selection is performed. For each analysis frame length the one to five largest spectral peaks are used for signal reconstruction. The only criterion required for peak selection is therefore the magnitude of the spectral peak.

By selecting only a limited number of the spectral components for further processing in the signal re-synthesis block, a strong spectral reduction is achieved. Note that the frequencies of the selected spectral peaks vary from frame to frame.

In the re-synthesis, an inverse fast Fourier transformation calculation is performed on the reduced FFT magnitude spectra. The IFFT window length is always equal to the analysis

window length. The re-synthesized signal frame is then multiplied with a Hamming window of the same length, and the output signal is reconstructed in a 75% overlap-add scheme.

The three different frame lengths in the analysis/synthesis system correspond to three different temporal and spectral resolutions. Based on a sampling frequency of 22.05 kHz and window length of 128, 256, and 512 samples, the achieved temporal and spectral resolutions amount to 5.8, 11.6, and 23.2 ms, and 172, 86, and 43 Hz respectively. These relationships are summarized in table 5.1.

Window length [samples]	FFT [points]	Temporal resolution [ms]	Spectral resolution [Hz]
128	128	5.8	172
256	256	11.6	86
512	512	23.2	43

Table 5.1 Spectral and temporal resolution achieved with the employed analysis/synthesis system at 22.05 kHz sampling frequency.

5.3.2 Performed speech tests

German sentences from the Oldenburg sentence test (each sentence includes five words), twelve German "a-Consonant-a" (C12) syllables, and eight German "d-Vowel" (Vo8) syllables spoken by a native German male speaker were used for testing. The standard "WAV"–files were processed using analysis and synthesis window lengths of 128, 256, and 512 samples, and between 1 to 5 selected spectral peaks (spectral components). All together, this results in fifteen different settings of processing parameters (3 window lengths * 5 different spectral component numbers = 15). All the processed *.wav files are mono and have a sampling frequency of 22050 Hz and a 16-bit amplitude resolution. Due to the limited number of employed spectral components, all the processed speech-files sounded more or less unnatural. In addition, the unvoiced or noise-like components of the speech signal got a tonal or music-like timbre.

All tests were carried out in a sound proof room. The stimuli were presented at 65 dB RMS without background noise from a distance of 1.5 m. A Westra type LAB-1001 audiometer box was employed. Playback was performed by a 16 bit PC sound card.

The Oldenburg sentence tests were carried out with 9 native German speaking adults. The subjects were asked to identify German words in 5-word-sentences presented only once and to repeat as many words as they could. The sentence test was provided only with the 128 and 512 point analysis-synthesis window length. Increasing the number of employed spectral components was stopped when the correct answer scores reached values close to 100%. For all parameter settings, two different lists including 10 sentences each were tested. Before each test sentence sequence, a 20 sentence training sequence was presented with the same parameters sets. Most of the tested subjects were inexperienced listeners.

The C12 consonant aCa-logatome tests were carried out with 15 native German speaking adults and the Vo8 vowel dV-logatome tests were carried out with 14 native German speaking adults. The subjects were asked to identify the correct syllable from a choice of syllables on a touch screen (MACarena test surface [B20]). Test items were not repeated. Both consonant and vowel tests were performed with 1 to 5 spectral components at each of the three window lengths, resulting in 15 different parameter settings.

The C-12 test consisted of 12 logatome identifications, recorded in three different utterances and presented in a random order. The Vo-8 test consisted of 8 logatome identifications, recorded in two different utterances and presented in a random order.

The different parameter settings were tested in a random sequence. Before each of the test series, a short training session comprising one or two testing lists was performed. In addition, learning saturation curves for two of the tested subjects were measured to estimate the amount of the "warm up" effect.

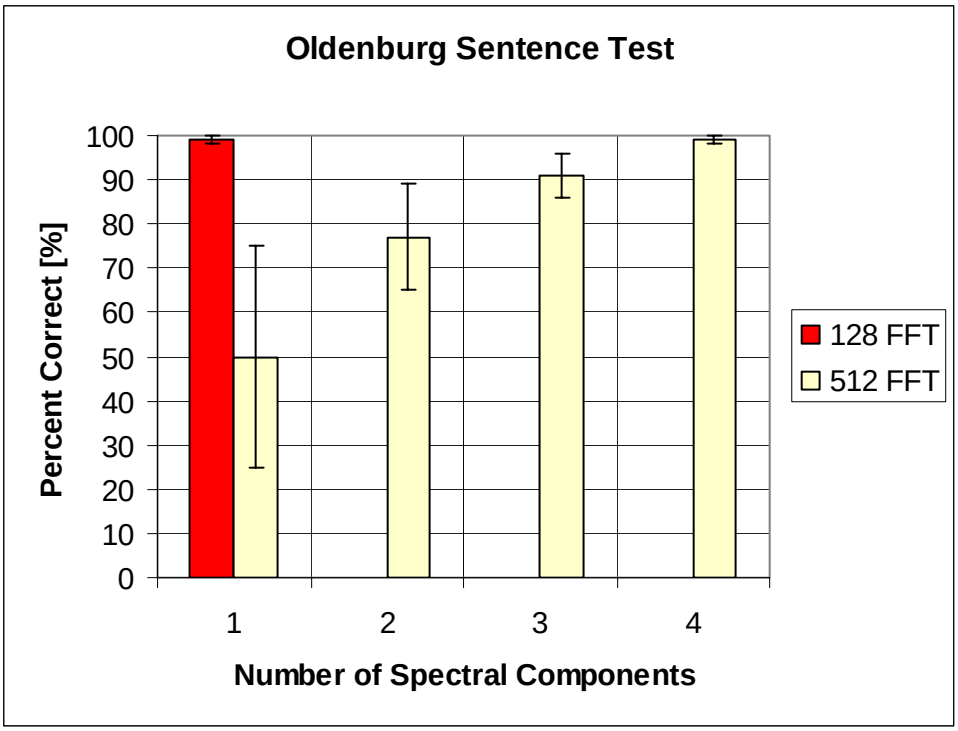
5.4 Results

5.4.1 Oldenburg sentence test

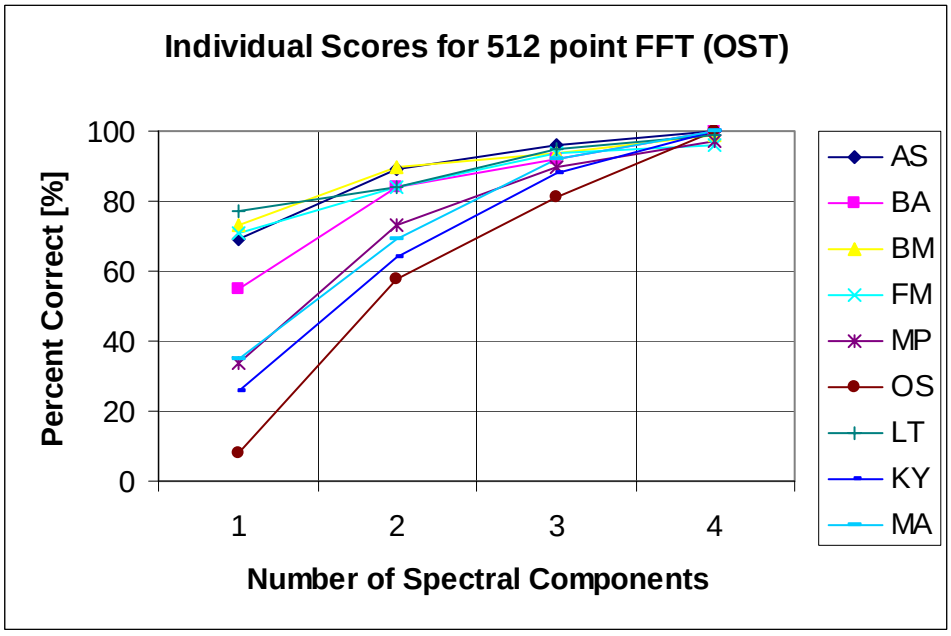
The results of the Oldenburg sentence test (Fig. 5.2) showed that nearly 100% German sentence recognition could be achieved with only one spectral component using a 128 point analysis-synthesis window (5.8 ms/172 Hz resolution). Four spectral components were required for the same sentence recognition scores with a 512 point analysis-synthesis window (23.2 ms/43 Hz resolution).

However, all tested subjects reported better speech quality with the 512 point analysis-synthesis window using four spectral components than with the 128 point analysis-synthesis window using only one spectral component. The percentage of correct-answer scores showed greater individual value variability using only one spectral component with the 512 point analysis-synthesis window than in case of more than one used spectral component (Fig. 5.2b). Some subjects could recognize nearly 80 percent of the words in sentences (subject LT in Fig. 5.2b). However, there were also subjects recognizing only 10 percent of the spoken words in sentences (subject OS in Fig. 5.2b). With the growing number of employed spectral components the range of correct answers score percentage strongly decreases from 70% to about 1%. Using two spectral components, the minimal recognition score is 60%, and using three spectral components, the minimal correct-answer percentage is 80%.

The sentences with the 256 point FFT analysis-synthesis window were tested with a few subjects and showed 100% speech recognition using one spectral component, similar to the results achieved with the 128 point analysis-synthesis window.



a)



b)

Fig. 5.2 Results of the Oldenburg sentence test a) mean percent correct answer scores with STD bars for 128 and 512 point FFT, b) individual percent correct answer scores for 512 point FFT.

5.4.2 Learning effects

Learning effects could not be completely excluded in this experiment. This is because the Oldenburg sentence test is a closed sentence set with a limited number of sentence lists with repeated words and is normally used in combination with the Oldenburg noise (in this case learning saturation occurs after 2-3 sentences lists) [B20]. Dorman *et al.* [B26] also reported about so called "warm up" effects for listeners exposed to altered speech signals of any kind. They managed to control this effect by the use of sequential instead of randomized test orders and by using familiarization procedures before each test condition. The subjects in their studies were allowed to use a "repeat" key during the consonant and vowel tests as many times as they wished.

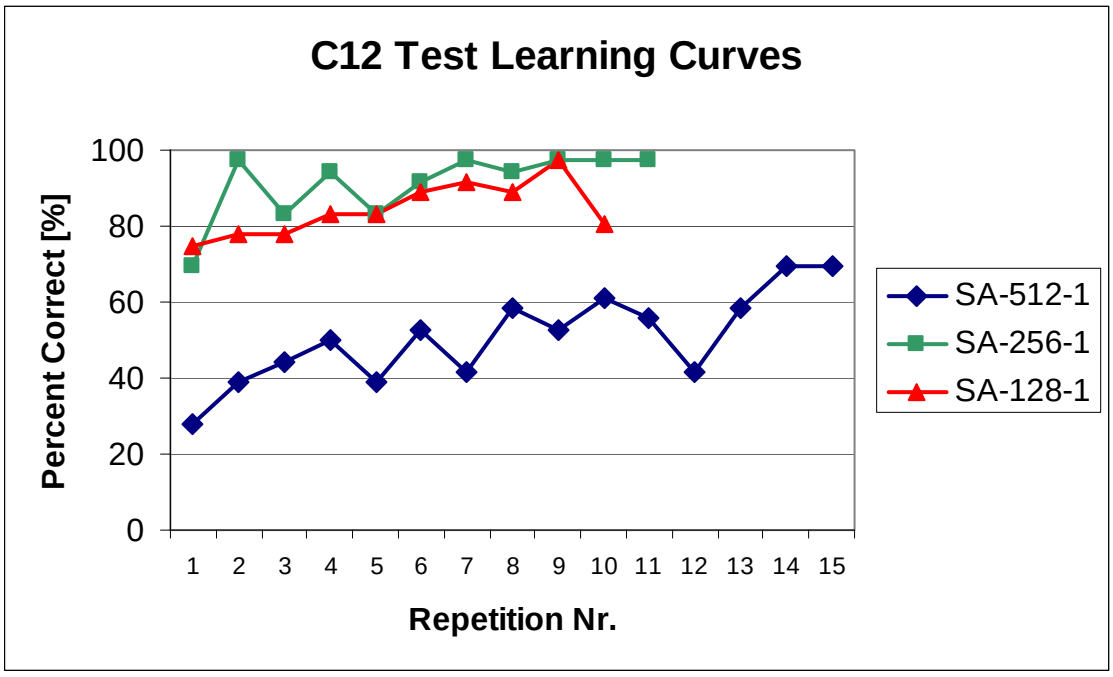


Fig. 5.3 Learning curves of one subject for the consonant test with three different analysis-synthesis window lengths and one spectral component used in signal reconstruction.

To estimate any possible influence of the "warm up" on the recognition score values, learning curves of consonant recognition scores were determined for two of the tested subjects. In this context, the C12 test was repeated without correct answer response until saturation was achieved. The saturation conditions were either 100% achieved consonant recognition or the sequence of two equivalent recognition answer scores following each other. The learning curves for one of the subjects are shown in Fig. 5.3.

For signal processing with the 512 point analysis-synthesis window length, 15 repetitions were necessary to achieve saturation conditions. For the 256 and the 128 point window case, the saturation condition was achieved after 10 and 11 repetitions. In parallel to the learning effects, a remarkable decrease in correct-answer scores during the test sequence could sometimes be observed, indicating tiring of the tested subject (see Fig. 5.3). When this occurred, the running test session was interrupted for one or two days.

Due to the "warm up" effect, an increase in correct consonant identification from 30% to 70% (40% difference!) could be observed for the 512 point analysis-synthesis window length with one spectral component. With the two other window lengths (one spectral component), the observed "warm up" effect was not as large (approximately 25%). When more than one spectral component was used in signal reconstruction, the observed learning effects were slightly smaller.

In order to decrease the influence of learning during the consonant and vowel tests, the most difficult test conditions (*i.e.* those with a smaller number of employed spectral components) were repeated four times in random sequence over all test conditions. This testing scheme enabled an additional training of the subjects on the most difficult signal-processing parameters, the learning curves of which indicate a larger saturation period. It was assumed that each test which employed smaller number of spectral components provides training for all other signal processing settings using a larger number of spectral components, but not *vice-versa*. This method does not exclude learning effects, but at least makes signal processing with different parameter choices comparable.

5.4.3 Consonant tests

The results of the C12 test over different analysis-synthesis window lengths and a different number of spectral components used in the reconstruction is shown in Fig. 5.4. Increases of the correct-answer scores are observable with each of the three different analysis-synthesis window lengths with an increasing number of used spectral components.

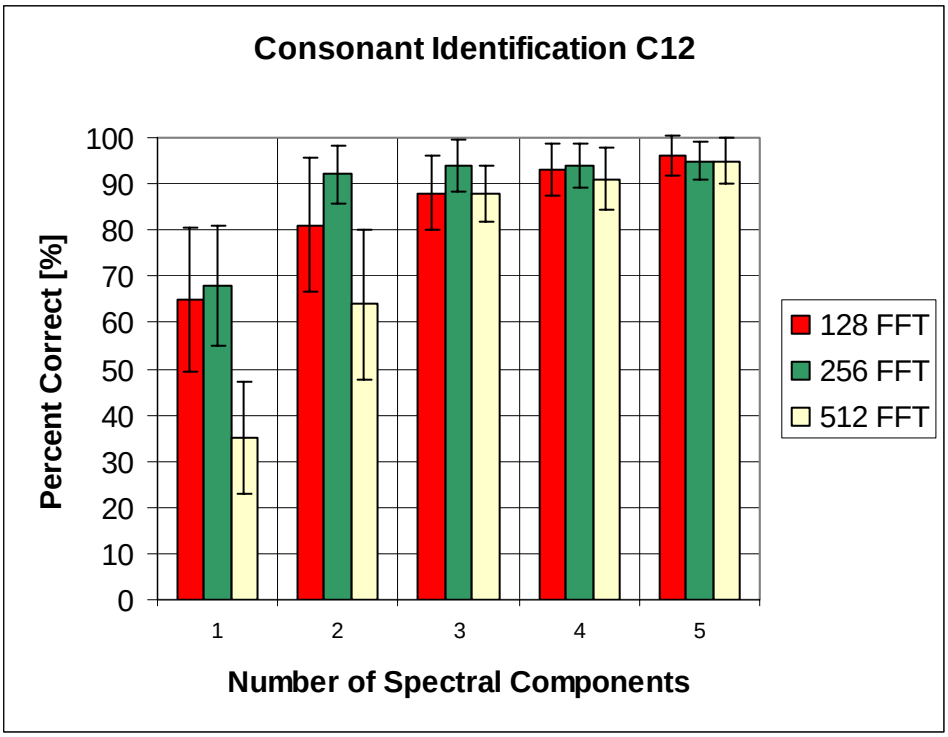


Fig. 5.4 Mean percentage correct scores of C12 "aCa" consonant test for three different analysis-synthesis window-lengths with one to five spectral components used in signal reconstruction.

Also visible from figure 5.4 is that consonant recognition scores were significantly better with the 128 and the 256 analysis-synthesis window lengths than with the 512 analysis-synthesis window length, if one or two spectral components were used in the reconstruction. Using only one spectral component, the correct answer scores with the 512 point analysis-synthesis window length was only 35%, whereas for the 128 and 256 point analysis-synthesis window lengths the scores were close to 70%. With two spectral components the 256 point analysis-synthesis window length was significantly better than for both the 128 and the 512 point analysis-synthesis window length. The 512 point analysis-synthesis window had the steepest increase in percentage of correctly recognized consonants by increasing the number of spectral components from one to three.

When three or more spectral components were used for reconstruction, consonant recognition scores were close to 90% for all of the three different analysis-synthesis window lengths.

The results over different consonant features such as voicing, nasality, sonorance, sibilance, frication, place, and manner of articulation according to Miller and Nicely [B92] which are adapted for German consonants [B20] are shown in Fig. 5.5-5.7.

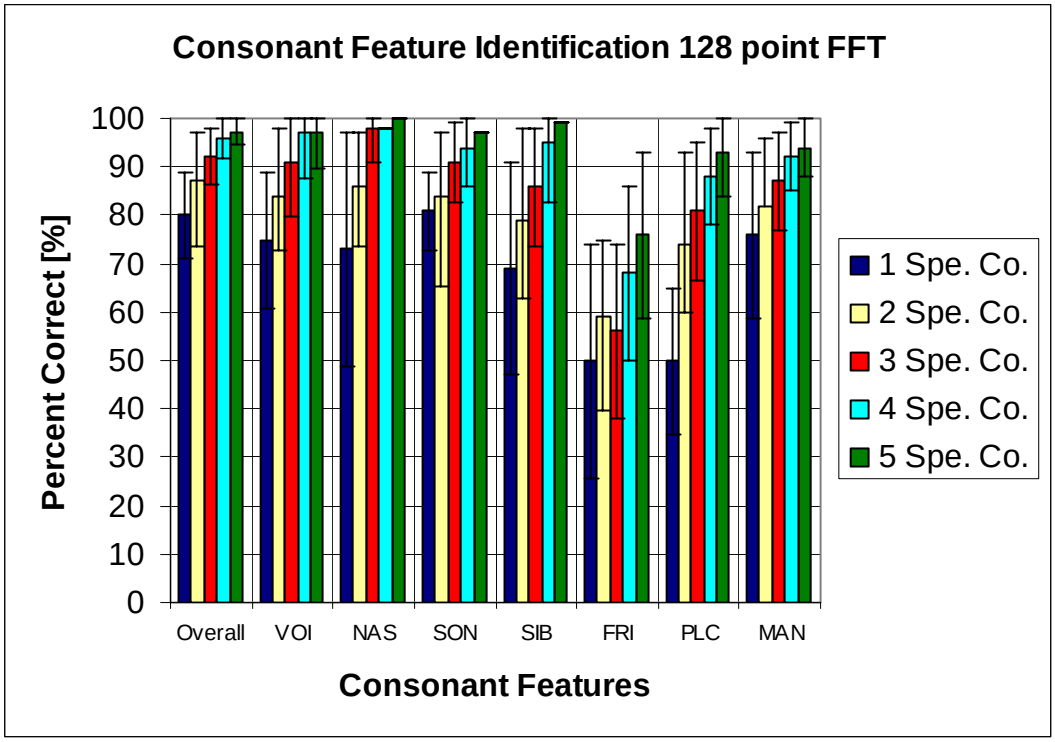


Fig. 5.5 Percent correct scores of the different consonant features voicing (VOI), nasality (NAS), sonorance (SON), sibilance (SIB), frication (FRI), place (PLC) and manner (MAN) for the 128 point analysis-synthesis window length.

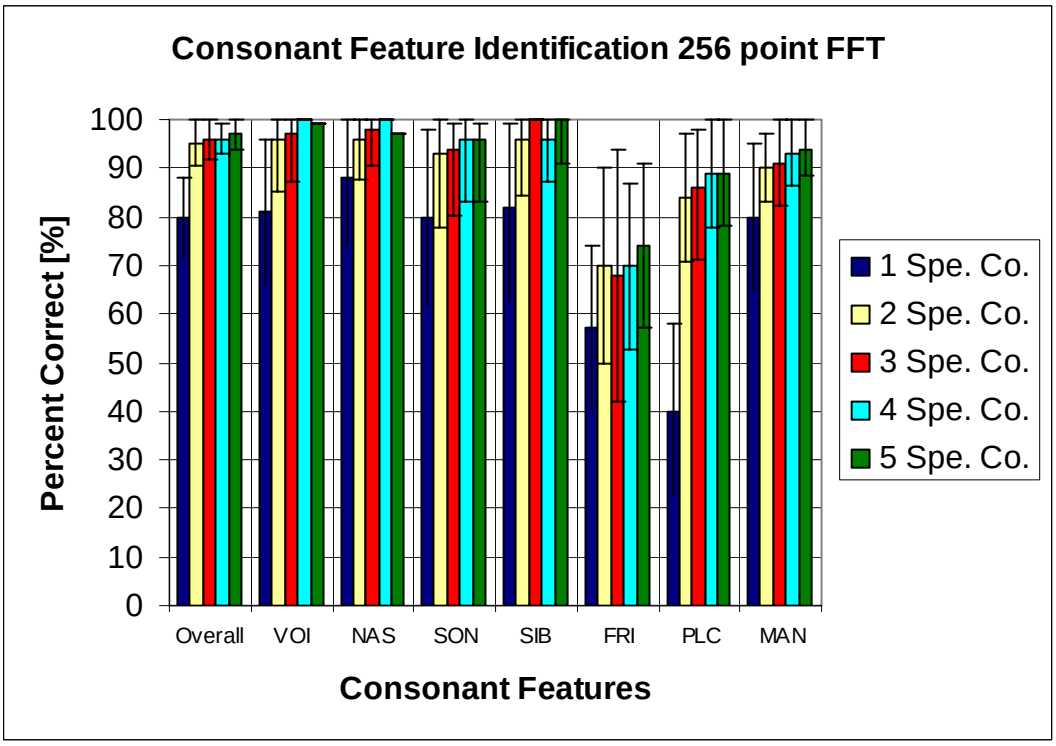


Fig. 5.6 Percent correct scores of the different consonant features voicing (VOI), nasality (NAS), sonorance (SON), sibilance (SIB), frication (FRI), place (PLC,) and manner (MAN) for the 256 point analysis-synthesis window length.

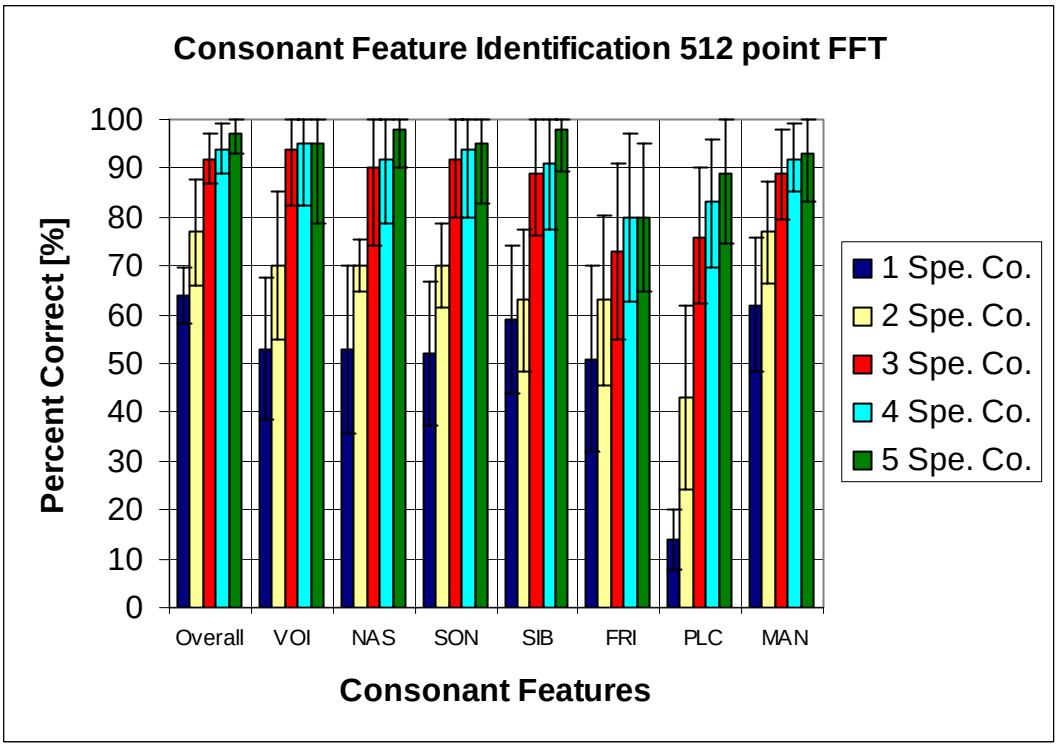


Fig. 5.7 Percent correct scores of the different consonant features voicing (VOI), nasality (NAS), sonorance (SON), sibilance (SIB), frication (FRI), place (PLC), and manner (MAN) for the 512 point analysis-synthesis window length.

It could be observed that the recognition of different consonant features can improve within the overall consonant recognition. This means that the consonant confusions, if such occurred, were scattered throughout the entire possible answer range, indicating that the tested subjects were rather confused by the sound of the spectrally reduced consonants.

The recognition of the frication feature (FRI) is significantly lower than for the recognition of all other consonant features, in particular for the 128 and 256 point analysis-synthesis window length. One interesting observation is that the frication feature identification seems to be easier for the 512 point analysis-synthesis window than for the two other analysis-synthesis windows, especially when using up to three spectral components in reconstruction. A comparison of the frication group identification for different analysis-synthesis window length is shown in Fig. 5.8.

The identification of the consonant place of articulation feature (PLC) is very problematic using the 512 point analysis-synthesis window with one and two spectral components. The same difficulty arises also with the 256 and 128 point analysis-synthesis window lengths using only one spectral component. The place feature identification improves significantly with growing number of spectral components used for reconstruction.

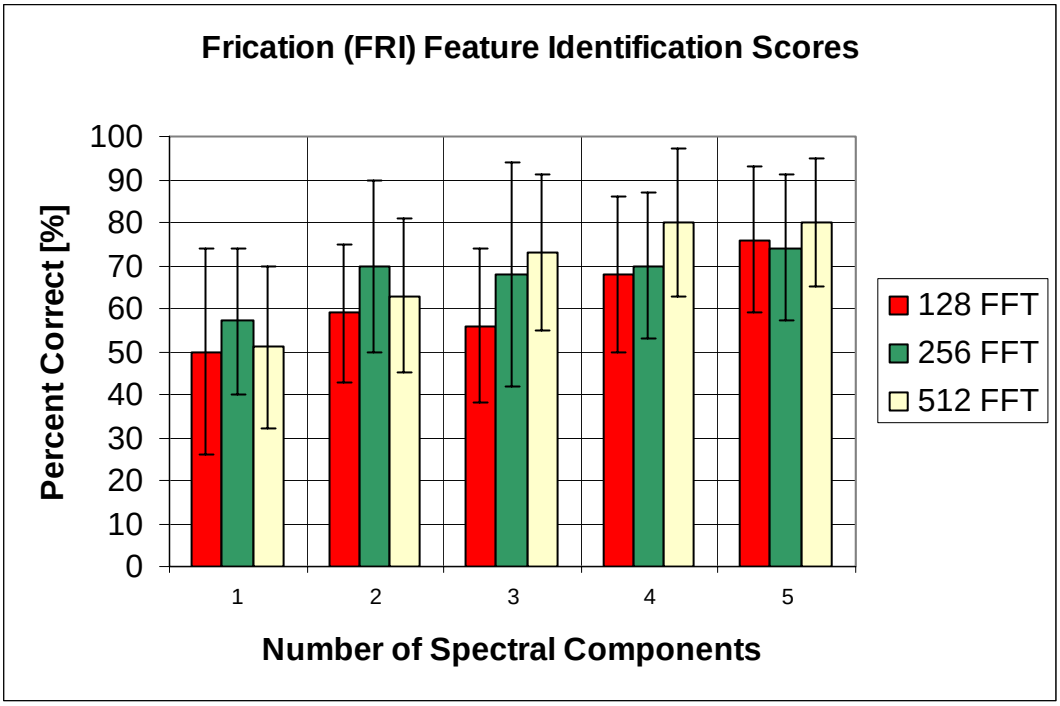


Fig. 5.8 A comparison of frication (FRI) consonant feature identification for different analysis-synthesis window lengths with one to five spectral components.

5.4.4 Vowel tests

The results for the German vowel identification over the three different analysis-synthesis window lengths with one to five spectral components are shown in figure 5.9. Vowels show the same tendency as consonants; with an increasing number of spectral components employed for signal reconstruction, the vowel identification scores increase. In contrast to the

consonant recognition (Fig. 5.4), only the 128 point analysis/synthesis window length provides a reasonable score of 65% when using only one spectral component.

By increasing the number of spectral components to two or more, the best score is always achieved with the 256 point analysis-synthesis window, which shows a very steep increase in vowel recognition scores when increasing the number of spectral components from one to two. Note that with the 128 point analysis-synthesis window length the best score does not exceed 85%, whereas with the 256 and 512 point analysis-synthesis window lengths vowel identification reaches up to 95% correct.

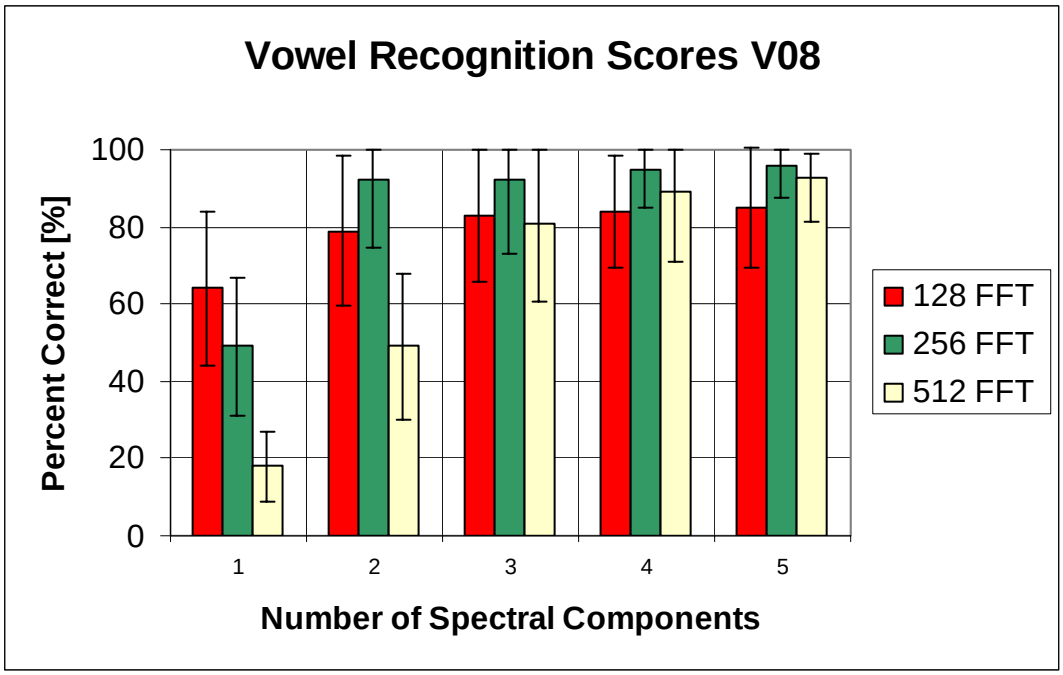


Fig. 5.9 Percent correct scores for German vowel identification over three different analysis-synthesis window lengths with one to five spectral components.

Results for the different vowel features for the first formant F1 and the second formant F2 over the different numbers of spectral components used in signal reconstruction are shown in Fig. 5.10-5.12.

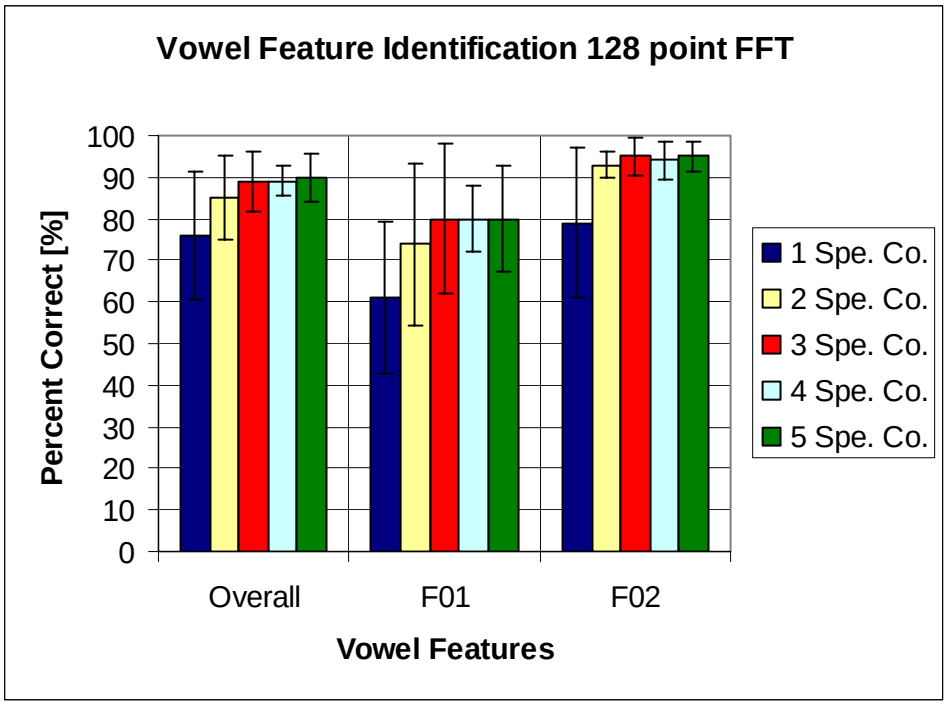


Fig. 5.10 Percent correct scores of the vowel features for the first formant F1 and the second formant F2 over a number of spectral components used in signal reconstruction for the 128 point analysis-synthesis window length.

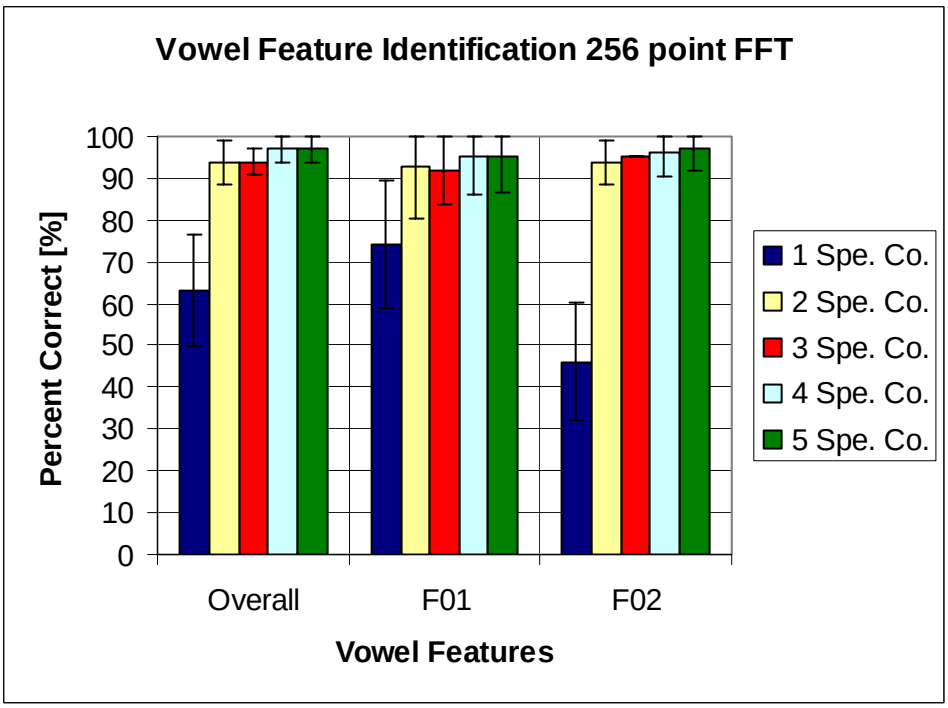


Fig. 5.11 Percent correct scores of the vowel features for the first formant F1 and the second formant F2 over a number of spectral components used in signal reconstruction for the 256 point analysis-synthesis window length.

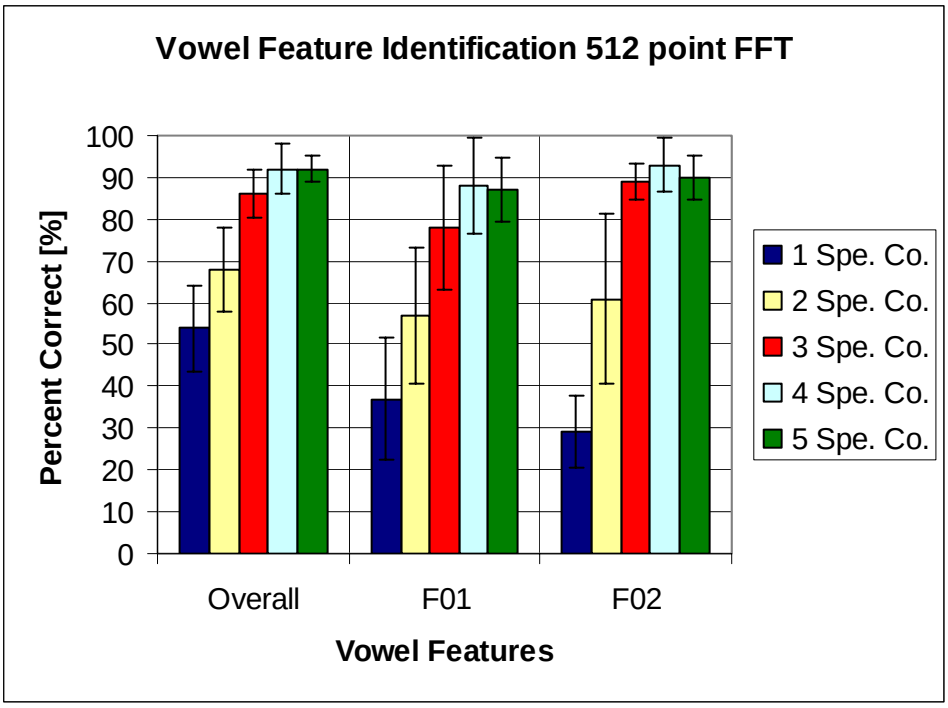


Fig. 5.12 Percent correct scores of the vowel features for the first formant F1 and the second formant F2 over a number of spectral components used in signal reconstruction for the 512 point analysis-synthesis window length.

For the 128 point analysis-synthesis window length, the first formant group identification does not increase further when increasing the number of spectral components to more than three. Using the 256 point analysis-synthesis window length with only one spectral component, the second formant vowel group is badly identified (45%). Saturation of a score of 95% is already achieved when using two spectral components.

Vowel group identification scores for the 512 point analysis-synthesis window strategy reach saturation when using three spectral components. With less than three spectral components, the scores are not satisfying.

5.5 Discussion and conclusions

The main result of the present study is that only a few spectral components are necessary for speech perception. This is in agreement with previous studies mentioned in chapter 3. Only one spectral component per 1.5 ms is required if the criterion for sufficient speech perception is defined as close to 100% sentence recognition. This follows from the result of the Oldenburg sentence test. Furthermore, some subjects showed relatively high individual correct answer scores (~70%; Fig 5.2b) with the 512 point analysis-synthesis window length (poor temporal resolution) even with only one spectral component in the reconstruction. However, it can not be excluded that these subjects make use of the sentence context (even if the sentences do not have a logical meaning). In contrary, for the subjects with very bad scores it could be that this kind of signal processing sounds too strange to be understood with the limited training time provided. If this is true, then the subjects with higher correct answer

rates were able to adapt more quickly to the specific signal processing scheme than subjects with lower (~25%) word identification rate in the sentences.

The second conclusion from these results is that there is possibly a minimal spectral component per time ratio which is required for good speech perception. This minimum spectral component per time ratio lies between two and four spectral components per 1.5 ms. It was not possible to determine the minimum spectral component per time ratio more precisely with the actual signal processing strategy.

Considering the learning saturation curves of one of the tested subjects in Fig. 5.3 it can be assumed that subjects can adapt to the present speech processing strategy. At first, correct-answer scores for consonant identification were about 25%. This corresponds exactly to the correct answer scores of the poorer German sentence identification group. The values of correct consonant-identification answer-scores rise to 70% (more than a 45% difference) after 15 repetitions of the test sequence. This value corresponds to the correct sentence identification answer scores of the better subject group (for comparison see Fig. 5.2 and Fig. 5.3). If these results are not only coincidence, then speculations about possible learning capacity of subjects for different signal processing strategies remain open. Unfortunately, it takes a long time to reach learning saturation, the saturation itself is difficult to determine, and there remains the possibility that the subject becomes used to only one particular speaker (pronunciation), as it is usually the case in clinical speech perception tests.

If we observe the absolute correct identification rates of consonants and compare them to correct consonant feature identifications, then the specific consonant feature identification rate can be better than the absolute consonant identification rate. This means that most confusions (incorrect identifications) are made within specific consonant groups. The only exception is frication (FRI). The two tested German consonants “F” and “S” belong to the frication feature. The interesting observation is that using the 512 point analysis-synthesis window length, the FRI feature identification (separation) is better than the identification of the same feature with a shorter window length. However, it is rather improbable that for the FRI feature identification, spectral information certainty would be more important than temporal information. The only reason the FRI feature identification with the 512 point analysis-synthesis window length is better could be that it sounds more natural. This, however, is only the case when three or more spectral-components are used. The results of the 256 point analysis-synthesis window length are in good agreement with the studies of Shannon *et al.* and Friesen *et al* [B118]. They reported that with normal hearing listeners nearly 100% of the voicing and manner information for consonants is received with only two spectral channels [B46].

The situation for vowel recognition is slightly different. The 128 point analysis-synthesis window length using only one spectral component shows significantly better vowel identification scores than both the 256 and the 512 point window lengths (see Fig. 5.9). Vowel identification with only one spectral component using 512 point analysis-synthesis is unacceptable (less than 20%). These poor vowel identification scores are probably the reason for lower sentence recognition with longer analysis-synthesis window length using one or two spectral components. However, vowel identification for the 256 point analysis-synthesis window length using two spectral components rises to ~95%, and for the 512 point analysis-synthesis window length up to ~90% using four spectral components. For the 128 point analysis-synthesis window length, vowel identification rates rise up to ~80% with the use of two spectral components and improves only to ~85% using five spectral components. This

shows that there could be an optimal spectral component rate-per-time for vowel identification, of approximately 2-4 spectral components per 3ms. This approximation is obtained from the following considerations. The spectrum of a generated signal is modified at every quarter of the synthesis window length (256 point synthesis window corresponds to 11.6 ms) due to the 75% overlap add of the re-synthesis frame. Only two central synthesis windows are relevant for spectral component addition due to synthesis frame multiplication with a Hamming window. The possibility that selected spectral components are close or the same in two subsequent analysis/synthesis windows is rather large, considering the assumption of a quasi stationary character of the signal (vowel case especially). This could be the reason for the above mentioned spectral component per time rate, since by using two spectral components, the possible number of reconstructed spectral peaks lies between two and four.

Vowel identification features predefined with the first formant F01 and the second formant F02 have the same character as the absolute correct answer responses. In the case of the 128 point analysis/synthesis window length, the F01 vowel feature is probably responsible for the relatively low correct identification score saturation grade (~80%) (see Fig. 5.9 and 5.10). It can be assumed that for primary sentence identification vowel recognition is of greater importance than consonant identification. It is also possible that the good sentence identification of some subjects with the 512 point analysis-synthesis window and only one used spectral component is the result of advanced consonant identification ability.

Within the present study, it was not possible to exclude all “warm up” effects nor was it possible to give a concrete answer to the question: What is the smallest number of spectral components required for close to 100% speech perception? Instead, there is no specific number of spectral components, but a specific spectral component per time rate required for nearly 100% speech perception. This spectral component per time rate has values between 2-4 spectral components per 1.5ms.

It was concluded that the optimal window lengths for further signal processing investigation are 128 and 256 points long, with preference on the 256 point analysis-synthesis window length. Given the vowel and consonant identification scores, the minimal number of spectral components used for signal reconstruction should be not smaller than two for the 256 point analysis-synthesis window length and not smaller than three for the 128 point analysis-synthesis window length.

Chapter 6

Saturday

"The new creature eats too much fruit. We are going to run short, most likely. "We" again – that is its word; mine, too, now, from hearing it so much."

M. Twain

Speech perception experiments with normal hearing subjects using spectral compression

6.1 Overview

Speech perception experiments using spectrally compressed, low pass filtered and spectrally reduced speech signals were performed on nine native German normal hearing adults. Signal processing schemes with linear spectral compression on both the FFT and the SPINC- auditory frequency scale were used to produce spectrally compressed speech signals with different compression ratios. After any processing a low pass filtering at 2000 Hz was applied to all employed speech signals in order to simulate an effect of a restricted residual hearing range. It was assumed that similar effects on speech perception could be observed by hearing impaired subjects with steeply sloping profound hearing loss in the higher frequency area. Spectral reduction was applied in order to reduce spectral overlapping and masking potentially caused by spectral compression.

A comparison was performed between speech perception of the normal hearing subjects using the different spectral compression schemes, with different spectral compression ratios, and with the scheme which implements only spectral reduction in combination with low-pass filtering. Sentence, vowel, and consonant recognition were investigated.

The results of the study demonstrate that the low-pass filtered and spectrally reduced signals dramatically decrease vowel and consonant identification scores of the normal hearing subjects (down to 55%). However, sentence recognition is less affected.

Vowel and consonant identification scores were improved using the linear spectral compression on both the FFT scale and the SPINC scale. However, sentence recognition decreased dramatically when using the linear spectral compression on the FFT scale with compression ratios equal to or larger than 1.6. Sentence recognition did not change significantly when the linear spectral compression on the SPINC scale or the spectral compression on the FFT scale with CR=1.3 were used.

The largest absolute improvement of consonant identification (+12%) was observed when the linear spectral compression on the SPINC scale was used with CR=1.3. The largest absolute vowel identification score improvement (+21%) was observed using the linear spectral compression on the FFT scale with CR=1.3.

Following the results of this study, it can be concluded that the linear spectral compression on the SPINC scale does not degrade speech intelligibility of the normal hearing subjects if spectral compression ratios smaller or equal to 1.3 are used. It can effectively increase consonant recognition through the additional high frequency information provided by its strong, non-linear compressive character in the high-frequency area. Speech comprehension of normal hearing subjects is preserved even if non-linear spectral operations on the FFT scale, such as the linear spectral compression on the SPINC scale, are performed.

6.2 Motivation

Since little is known about the capability of normal hearing and hearing impaired subjects of making use of spectrally compressed and shifted information, the motivation of this study was to investigate the ability of normal hearing subjects to make use of spectral information from higher spectral areas transposed into lower spectral regions. The two spectral compression schemes were selected for the present study based on the discussion of chapter 2 and 3. It is important to note that spectral reduction was tested in combination with spectral compression. To approximately simulate the restricted hearing area of a profoundly hearing-impaired subject, a low-pass filtering with cut-off frequency at 2 kHz was applied. From the results of this study, the most promising spectral compression schemes with their appropriate parameter settings will be derived.

6.3 Method

6.3.1 Signal processing and test parameter settings

The signal processing system based on the sinusoidal speech algorithm was used to process audio files from different speech tests. The main components of this system are the spectral analysis block, the spectral reduction block, the spectral compression block and the spectral reconstruction block (Fig. 6.1). For a more detailed description, see also chapter 4.

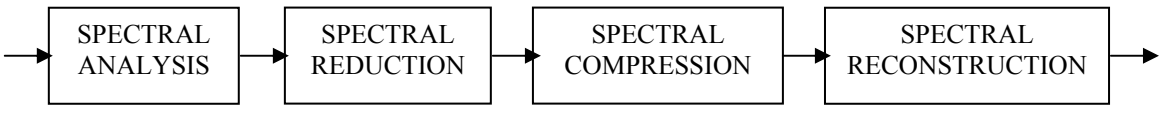


Fig. 6.1 Signal processing block diagram.

The spectral analysis block provides FFT analysis of the incoming audio file with a 256- point window length using 75% overlap and a Hanning window multiplication. The length of the analysis window was chosen after a preliminary speech test with normal hearing

German speaking adults (see chapter 5). The sampling frequency of the employed audio files was 22025 Hz resulting in a duration of the analysis window of 11.6 ms. Considering the employed overlap, a new FFT with spectral resolution of 86 Hz was calculated every 2.9 ms. To increase the spectral resolution of the system, additional frequency interpolation was carried out (see chapter 4). It used phase and approximate frequency values of the frequency bin corresponding to the spectral peak in two temporally-subsequent analysis windows. The frequency values were then chosen to fit the phase values.

The spectral reduction block provides selection of up to only eight spectral components in every FFT amplitude spectrum. The number of spectral components was chosen based on the result of the previous studies with spectrally reduced speech (see chapter 5). These studies showed that normal hearing German speaking adults require three or four spectral components for close to 100% speech comprehension. Eight spectral components were chosen to ensure good speech understanding in spectrally reduced and spectrally compressed speech. The reason for including the spectral reduction block is the prevention of possible masking effects which are expected from the spectral compression. As this system was designed for profoundly hearing-impaired subjects, effects of the possible spectral masking are very important, especially due to potentially broadened auditory filters within hearing-impaired subjects.

The spectral compression block can provide spectral compression of the incoming audio signal on the linear FFT scale or on the SPINC scale. The linear spectral compression is implemented by dividing the spectral component frequency values by the spectral compression ratio CR:

$$Fr_{OUT} = \frac{Fr_{IN}}{CR}, \tag{6.1}$$

where, Fr_{OUT} is the output frequency and Fr_{IN} the input frequency.

Spectral compression on the SPINC scale transforms the spectral component frequency values into the SPINC scale applying the spectral compression in *Spinc* units as defined in equation 6.1, applies spectral compression using the SPINC scale, and then transforms the compressed SPINC spectrum back to the FFT frequency scale (see chapter 3 and 4). The resulting input-output frequency relation is given by:

$$F_{OUT} = const * \tan\left(\frac{\arctan\left(F_{IN}/const\right)}{CR}\right) \tag{6.2}$$

where $const$ =1414.

For spectral compression on the FFT scale, three different linear compression ratios were used: CR = 1.3, 1.6 and 1.7. Two spectral compression ratio values were used to

perform the linear spectral compression on the SPINC scale: CR= 1.2 and 1.3. The investigated spectral compression ratio settings are given in Tab. 6.1. The nomenclature of the signal processing schemes employed in the present study is given in Tab. 6.2. The approximate values of the maximum linear spectral compression ratios were estimated in the preliminary speech tests. The condition for choosing the maximally tolerated compression ratios was at least 75% speech comprehension in sentences with non-low pass filtered signal. Fig. 6.2 shows the spectral compression input–output plots with linear and logarithmic axis scaling. For producing the reference signals, the original audio files were spectrally reduced and low-pass filtered, but no spectral compression was applied (CR=1.0).

The low pass filtering with 2000 Hz cut-off frequency was applied to every signal processing scheme to simulate a hearing loss. This was achieved by using only those spectral components for signal reconstruction whose values, after performing the spectral compression operation, were lower than 2000 Hz.

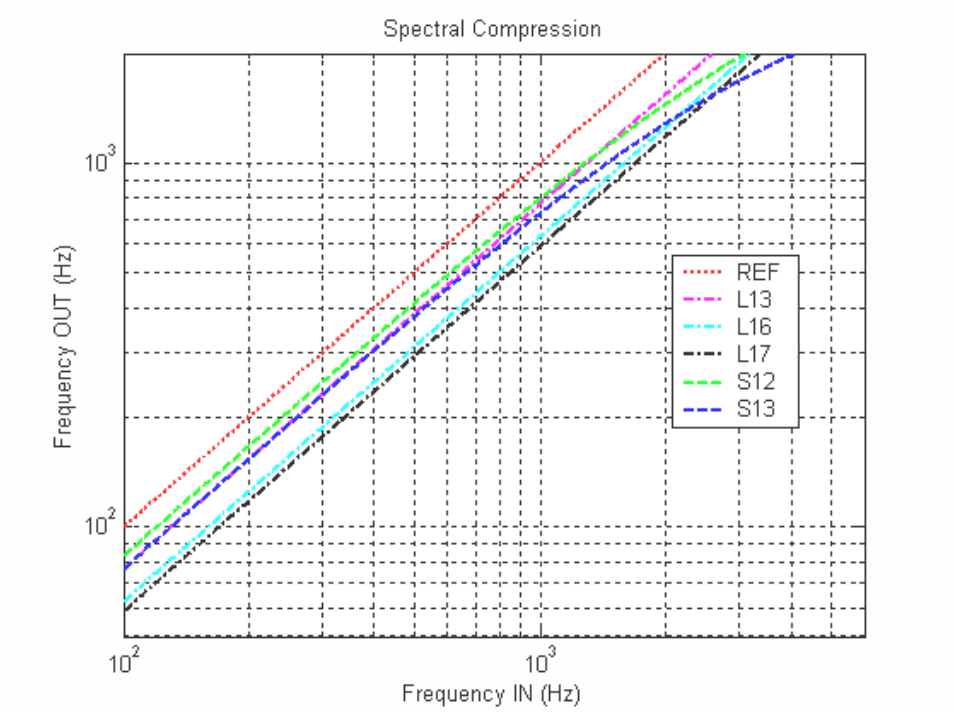
Signal reconstruction was performed using pure tone generation. The length of the reconstruction window was twice as long as the FFT calculation sequence (2.9ms*2= 5.8ms). In addition, a multiplication with a triangular synthesis window was applied to each synthesis frame (see chapter 4). This corresponding to a 50% synthesis overlap.

Parameter Settings	**LIN Compression**	**SPINC Compression**
Spectral Reduction	8 SC	8 SC
Spectral Compression (CR)	CR = 1.3; 1.6; 1.7	CR = 1.2; 1.3
Low-pass Filtering	2000 Hz	2000 Hz

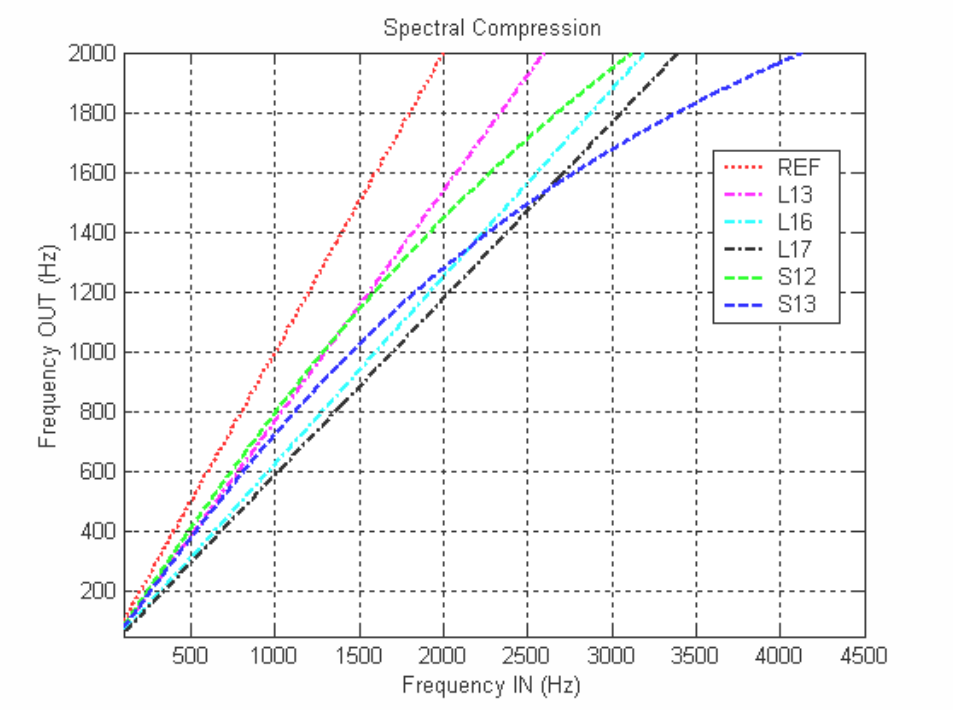
Table 6.1 Parameter settings for spectral compression and reduction.

Parameter settings	**Nomenclature**
Reference 8 Spectral Components; LP=2000 Hz	**REF**
LIN SC CR=1.3; 8 Spectral Components; LP=2000 Hz	**L13**
LIN SC CR=1.6; 8 Spectral Components; LP = 2000 Hz	**L16**
LIN SC CR=1.7; 8 Spectral Components; LP = 2000 Hz	**L17**
SPINC SC CR=1.2; 8 Spectral Components; LP = 2000 Hz	**S12**
SPINC SC CR=1.3; 8 Spectral Components; LP = 2000 Hz	**S13**

Table 6.2 Nomenclature of the employed signal processing schemes.



a)



b)

Fig. 6.2 Frequency input output curves for all tested spectral compression methods and reference. REF: reference, L13: linear spectral compression on the FFT scale with CR=1.3, L16: linear spectral compression with CR=1.6, L17: linear spectral compression with CR=1.7, S12: linear spectral compression on the SPINC scale with CR=1.2, S13: SPINC compression with CR=1.3, a) logarithmic frequency scales, b) linear frequency scales.

6.3.2 Performed speech tests

The German C12 consonant "aCa" logatome identification test, the German Vo8 vowel "dV" test and the Göttinger sentence test were performed with 9 normal-hearing native German speaking adults (three females and six males). All speech tests were performed in a random sequence.

The C12 test includes three different pronunciations of each consonant of the 12 German consonants and was randomly repeated four times for each parameter settings. The Vo8 test includes three different pronunciations of the eight German vowels, and was randomly repeated three times for each of the parameter settings. The Göttinger sentence test consists of 24 lists of 10 sentences each. For each of the parameter settings two different lists were tested. All test materials used recordings of a male speaker.

All processed *.wav files are mono and have a sampling frequency of 22.05 kHz and a 16 bit amplitude resolution.

All performed speech tests were carried out in a silent environment. The speech was presented at 65 dB RMS in a sound proof room at a distance of 1.5 m from a Philips Type 22AH586/16R active loudspeaker (playback was performed by a 16 bit PC sound card).

Different speech tests were performed in at least two sessions. In each session only one spectral compression method (linear or SPINC) was tested. Different values of the compression ratios were tested in a random sequence. Before each of the test sessions a ten minutes long training and session adaptation was performed. It consisted of a spectrally compressed speech signal spoken by a male speaker. During the adaptation phase, the subject could follow the text by simultaneous reading. Depending upon the compression method under investigation, linear spectral compression with CR=1.7 or SPINC compression with CR=1.3 was used. These two compression ratios are the largest settings tested for the two particular compression methods.

Audiograms of all subjects participating were recorded before conducting this test in order to ensure their normal hearing.

6.4 Results

6.4.1 Göttinger sentence test

The results of the sentence identification scores are given in Fig. 6.3 for the different spectral compression methods and compression ratio values. It was observed that using the linear spectral compression on the FFT with relatively large compression ratios (CR = 1.6 or 1.7), speech intelligibility in sentences was dramatically reduced from ~85% to ~55% or even ~40% (Fig. 6.3). It is also remarkable that even small differences in spectral compression ratios, as for example between CR= 1.6 (L16) and CR=1.7 (L17), can be critical. For these two cases, for example, the mean sentence identification score difference was approximately 15%.

No significant differences in sentence identification scores were observed between the reference (spectrally reduced and low-pass filtered signals) and the other three spectral

compression methods (L13, S12, S13). For each of these settings the mean sentence identification scores were close to 85%. The signal processing scheme involving spectral compression on the SPINC scale with CR=1.3 (S13) showed slightly lower values (~80%) than the other two schemes (L13 and S12). The difference was not statistically significant.

6.4.2 Consonant test

Mean consonant identification and absolute improvement scores are given in Fig. 6.4. The observed reference consonant identification score is ~55%. All implemented signal processing schemes indicated small but significant improvement of consonant recognition scores relative to the reference. The mean consonant improvement range was between 4 and 12%. The largest absolute improvement (+12%) was observed when spectral compression on the SPINC scale was used with CR=1.3. Both the spectral compressions on the SPINC scale with CR=1.2 and the linear spectral compression on the FFT scale indicated an absolute improvement of 8%.

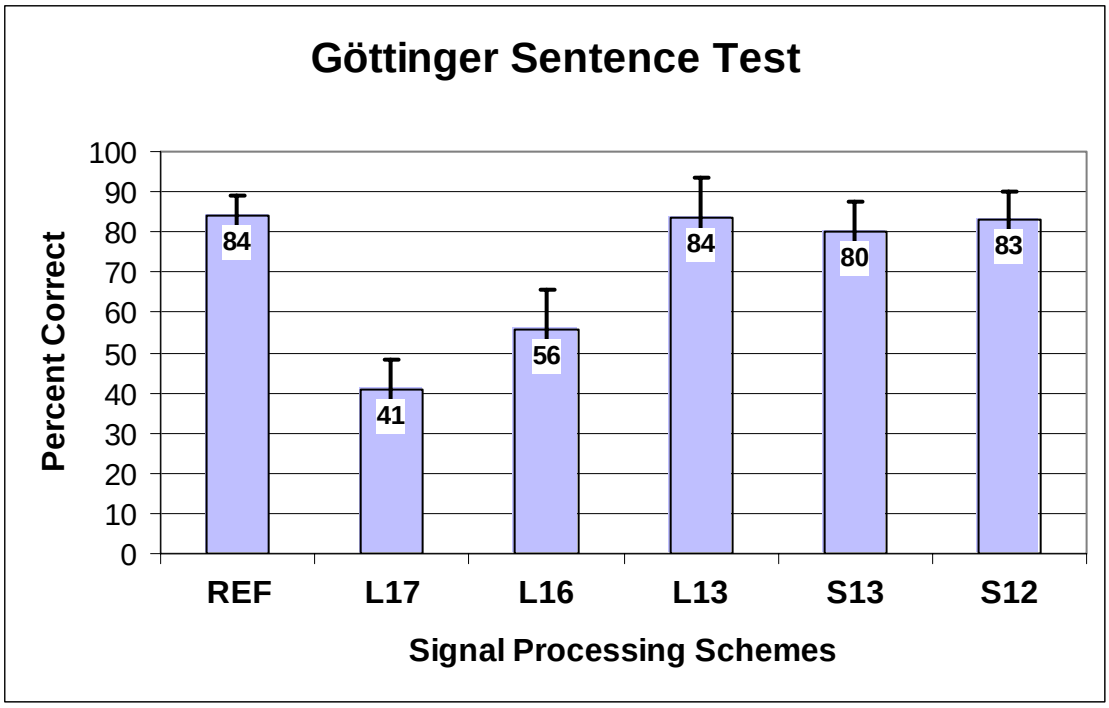
With the best consonant identification achieved with the SPINC spectral compression with CR=1.3 (~70%), some of the tested subjects reached even 80% correct consonant identifications. The tendency that a larger spectral compression provides better consonant identifications could be observed as well: both SPINC spectral compressions and the linear spectral compression with CR=1.7 showed better consonant identification than the linear compressions with CR = 1.3 and 1.6.

Consonant identification and their improvement scores for voicing, nasality, sonorance, sibilance, frication, place and manner, [B92] are shown in Fig. 6.5. Some of the consonant features used by Miller and Nicely, such as duration, are irrelevant for the German language [B20] and were therefore not considered.

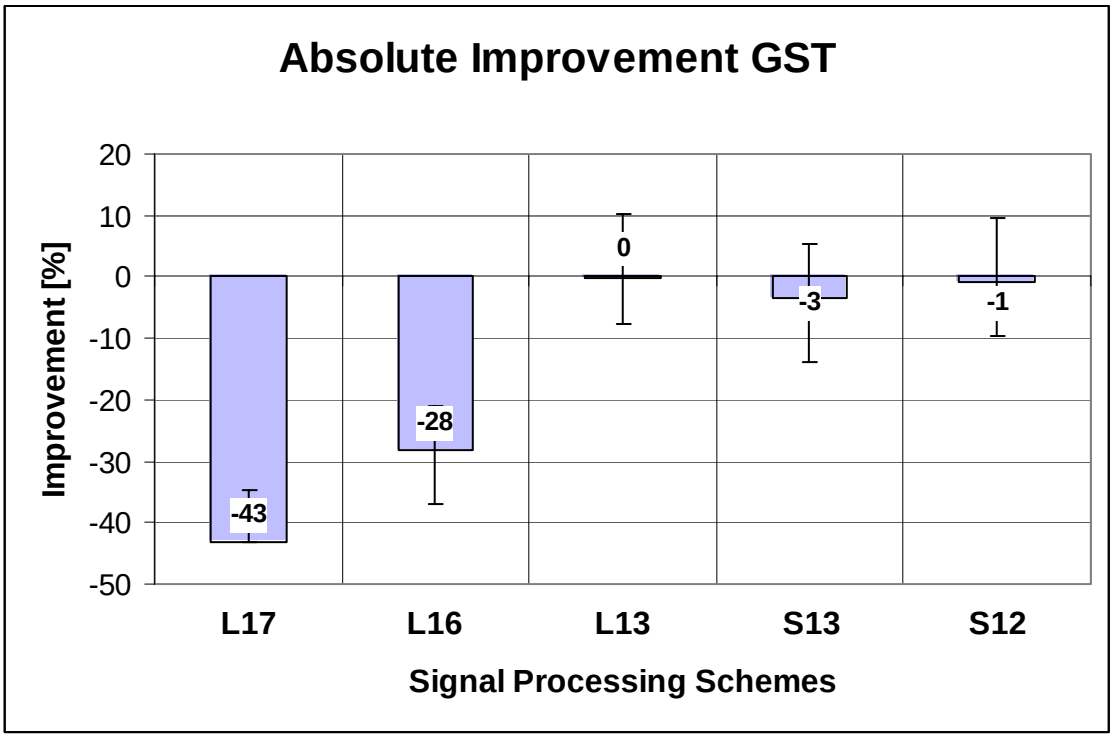
The identification scores of the VOI, NAS, SON and MAN consonant features were relatively good for all of the employed signal processing schemes including the reference. For these features there was always at least one spectral compression scheme which showed improved identification scores relative to the reference signal. For voicing, linear spectral compression on the FFT scale with (CR=1.6 and 1.7) showed slightly worse (-4%) identification scores (largest compression ratio) than the reference speech signal.

Identification scores of the SIB, FRI and PLC consonant features were rather poor and significantly worse than the average consonant identification scores. The identification scores of the SIB and FRI features were even very low. It could be shown that the FRI and the PLC identification was improved by spectral compression; however, the SIB identification worsened with spectral compression, only the compression scheme using the SPINC scale with CR=1.3 showed a significant (+24%) improvement of this feature. For all these three features the spectral compression on the SPINC scale with CR=1.3 showed the largest absolute improvements.

The NAS feature was the only one where the linear spectral compression on the FFT scale with CR=1.7 showed the largest absolute improvement of identification. The Overall identification parameter shows that the scattering of incorrect identifications was relatively low. That means that most of the incorrectly identified consonants were systematically confused with others. The Overall feature was also improved through use of spectral compression.



a)



b)

Fig. 6.3 a) Average results of the Göttinger sentence test (GST). Identification scores for five spectral compression methods and reference; b) Average absolute improvement of GST recognition scores calculated against reference signal. Data were collected over measurements of nine normal-hearing native German adults.

6.4.3 Vowel tests

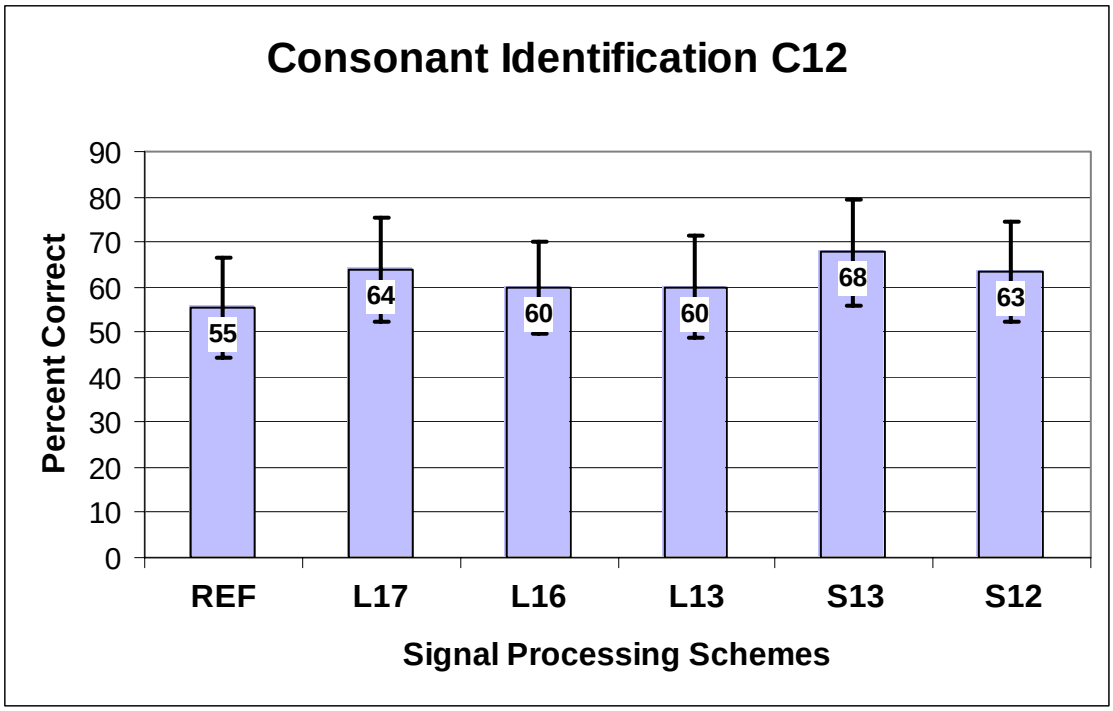
The German vowel identification scores are given in Fig. 6.6. The reference vowel identification indicated only 55%. All spectral compression schemes except the linear spectral compression on the FFT scale with CR=1.7 showed an improvement in vowel recognition.

The linear spectral compression on the FFT scale with CR=1.3 and both spectral compressions on the SPINC scale showed even a significant absolute improvement of +14 to +21% relative to the reference signal. The best vowel identification scores were achieved with the linear spectral compression on the FFT scale with CR=1.3 (76%) and the linear spectral compression on the SPINC scale with CR=1.2 (74%).

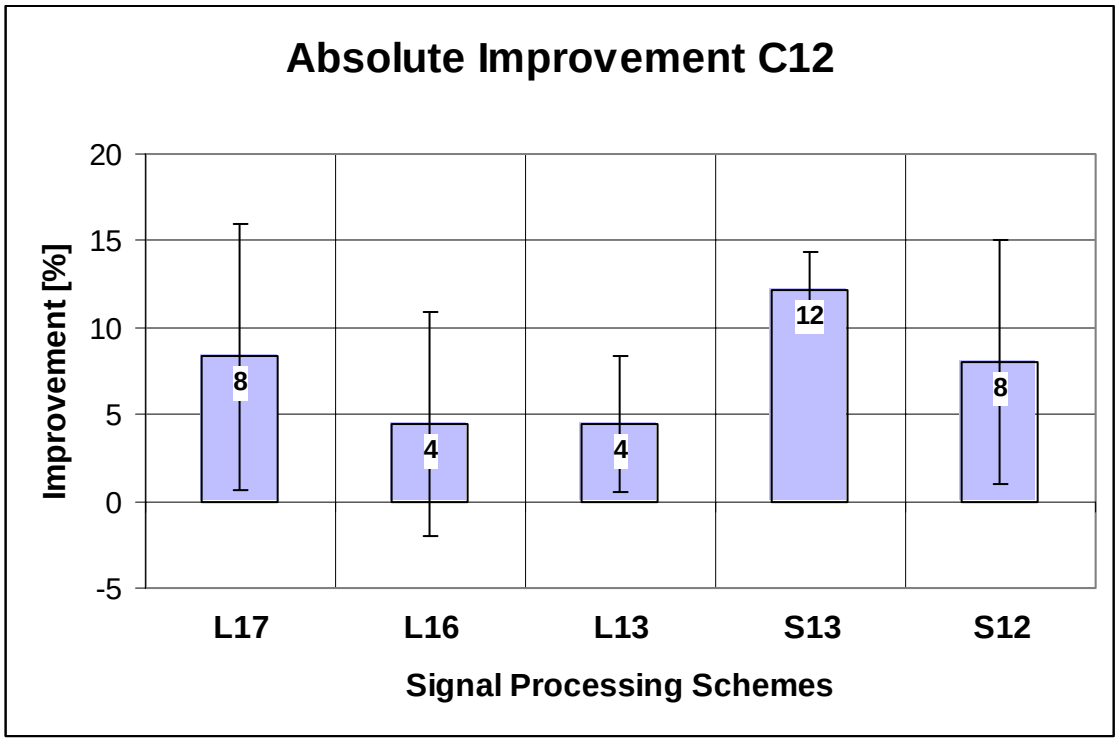
German vowel feature identification scores are given in Fig. 6.7. The identification of the first formant feature for the spectral compression on the SPINC scale with CR=1.2 and 1.3 and for the linear spectral compression on the FFT scale with CR=1.3 was approximately the same as for the reference. The identification scores for the linear spectral compression on the FFT scale with CR=1.6 and 1.7 were ~15% lower than for the reference. However, for the second formant feature all spectral compression schemes gave significantly better results than the reference. For all spectrally compressed signals, the identification of the second formant was approximately 25% better than the identification of the first formant feature. Only for the reference the first formant feature identification was better than the second formant feature identification.

The spectral compression methods with smaller low-frequency compressions (on the SPINC scale with CR=1.2 and 1.3 and on the FFT scale with CR=1.3) showed ~10% better absolute improvement within the first formant group identification than the linear compression on the FFT scale with CR=1.6 and 1.7.

Incorrect response-scattering represented by the Overall parameter is relatively small indicating mostly systematic incorrect consonant identifications. Scattering of data over vowel groups was much smaller than for consonants.

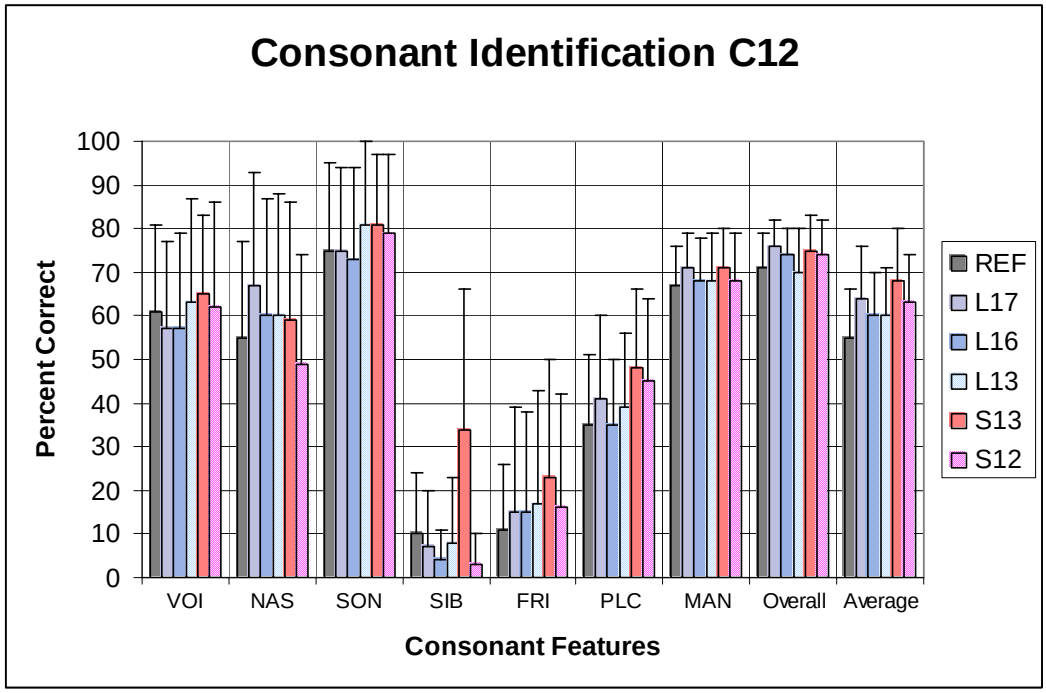


a)

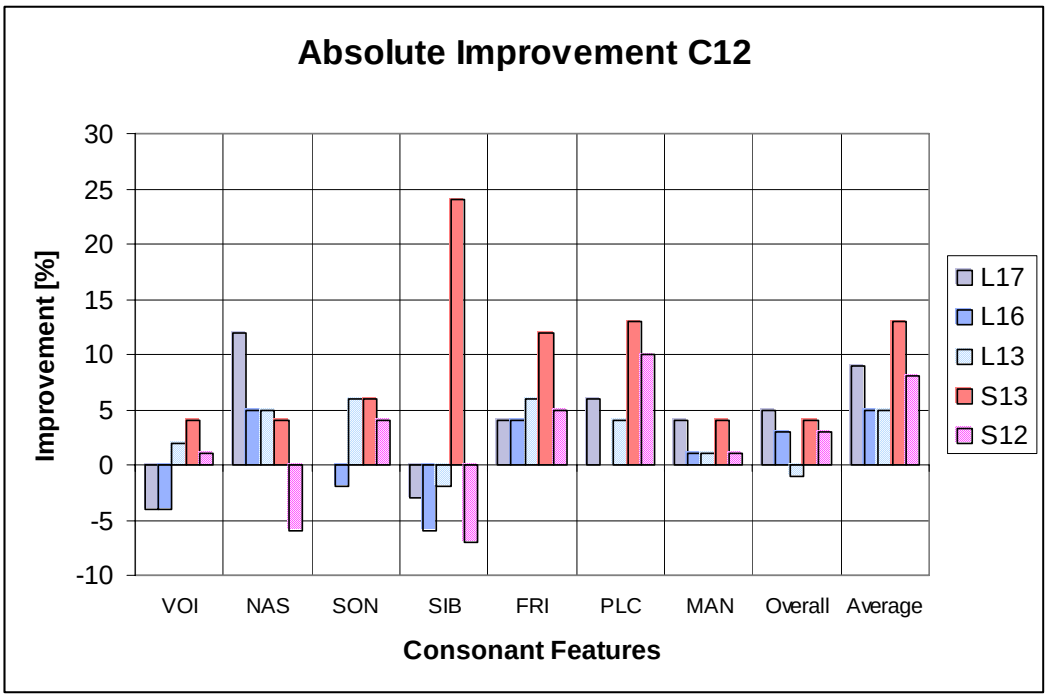


b)

Fig. 6.4 a) Average results of the C12 consonant test. Identification scores for five spectral compression methods and reference; b) Absolute improvement of consonant identification scores calculated relatively to the reference signal. Data were collected over measurements of nine normal-hearing native German adults.

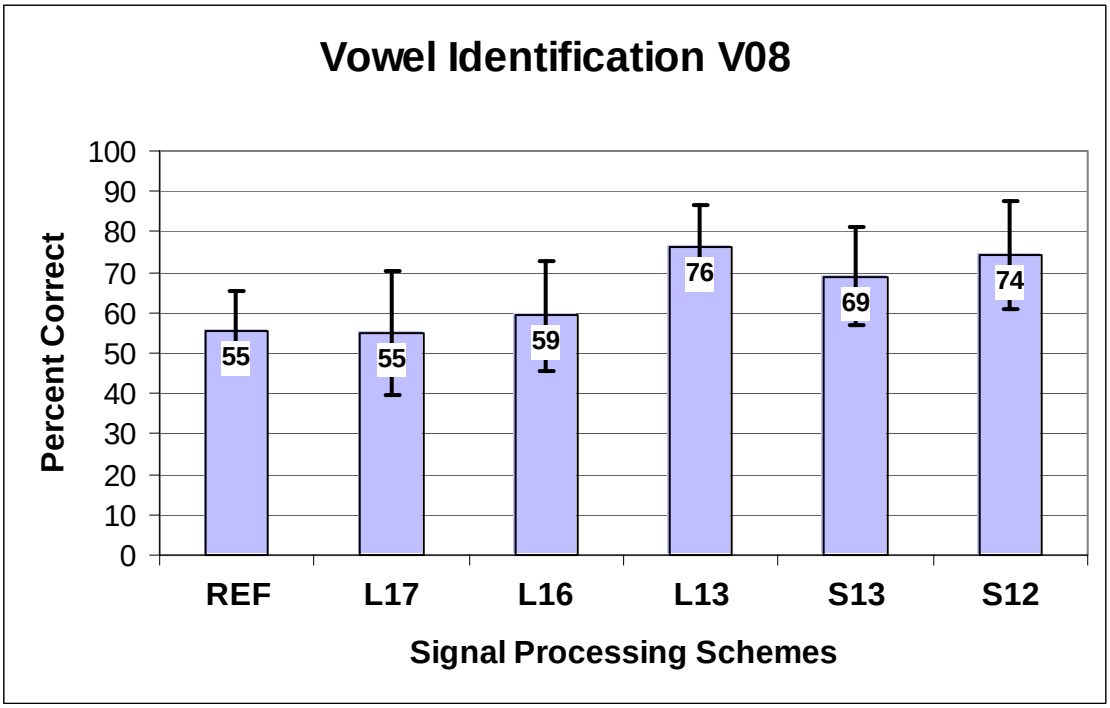


a)

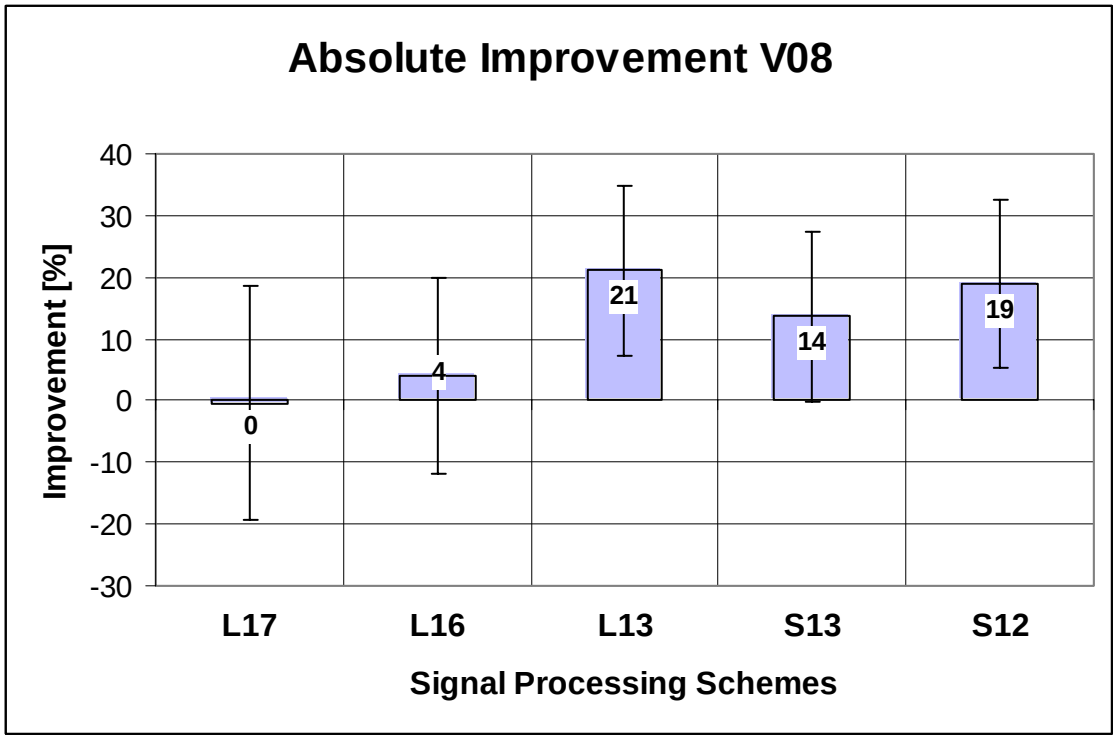


b)

Fig. 6.5 Information transmission analysis a) results of the C12 consonant test consonant feature (voicing VOI, nasality NAS, sonorance SON, sibilance SIB, frication FRI, place PLC and manner MAN [B92]. Identification scores for five spectral compression methods and reference. b) Corresponding absolute improvement. Data were collected over measurements of nine normal-hearing native German adults.

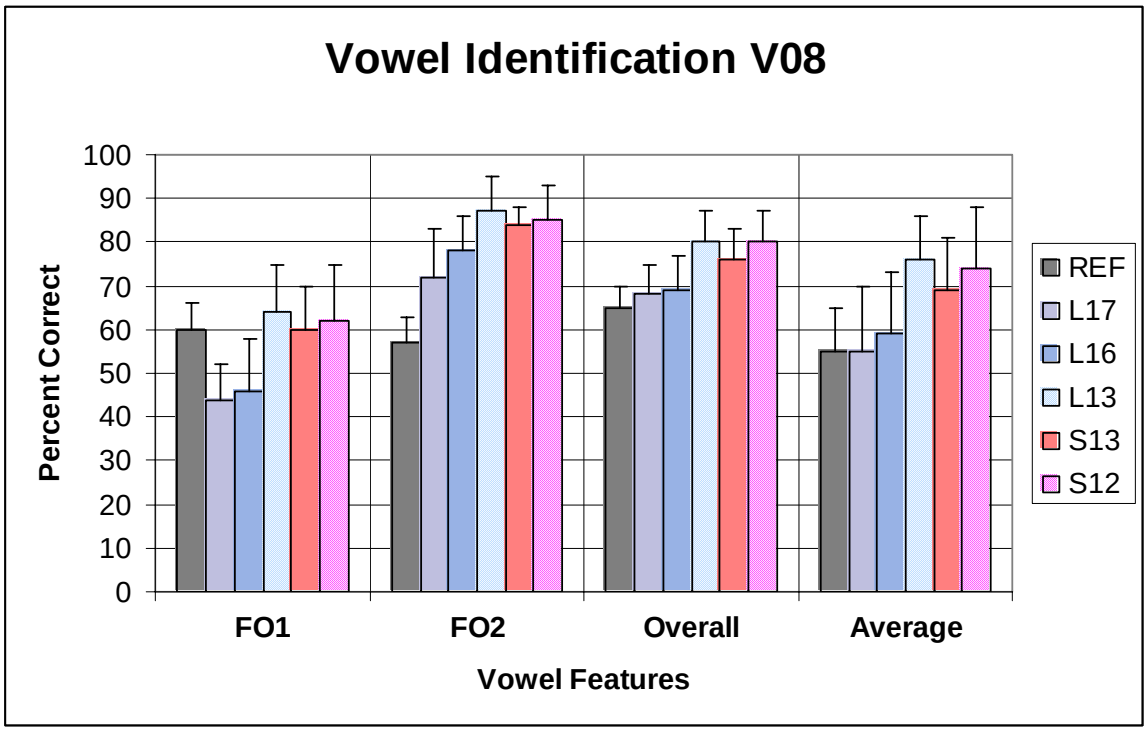


a)

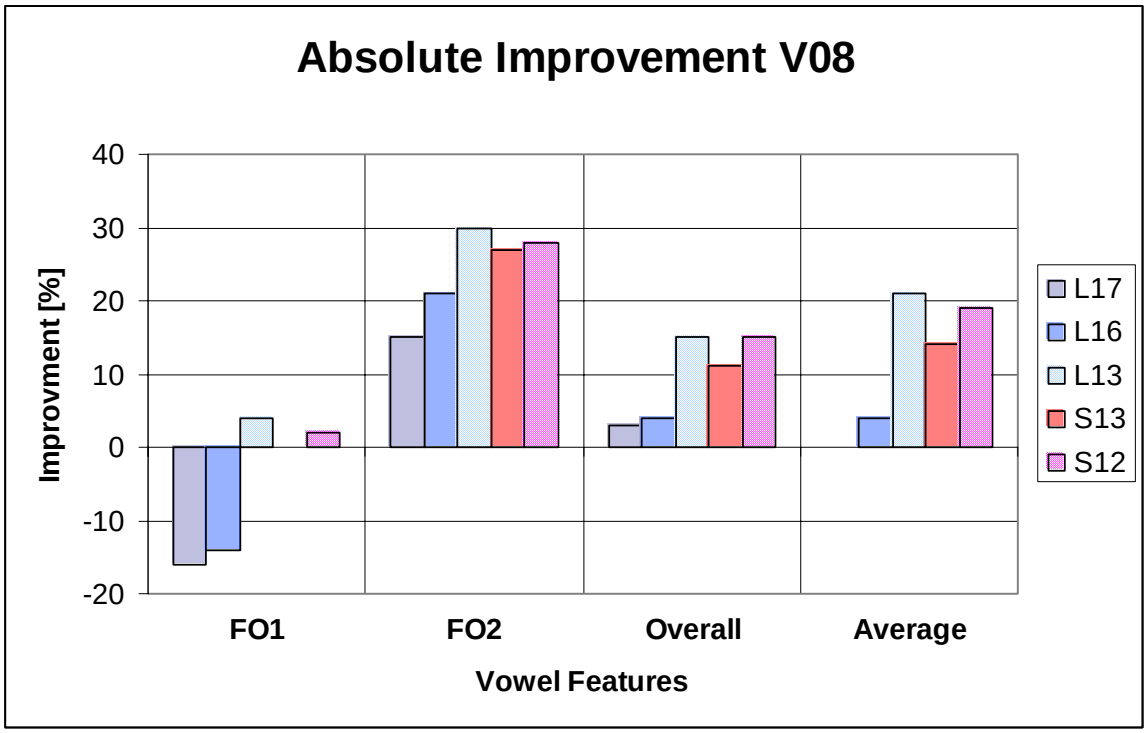


b)

Fig. 6.6 a) Results of the Vo8 vowel test. Identification scores for five spectral compression methods and reference and b) Corresponding absolute improvement calculated relative to the reference signal. Data were collected over measurements of nine normal-hearing native German adults.



a)



b)

Fig. 6.7 Information transmission analysis results a) average German vowel feature identification scores for five spectral compression methods and reference. b) Corresponding absolute improvement calculated relative to the reference signal. Data were collected over measurements of nine-normal hearing native German adults.

6.5 Discussion and conclusions

In comparison with the reference the results of the consonant and vowel identification tests were generally better for all spectral compression signal processing schemes (Fig. 6.4 and 6.6). The sentence identification scores, however, were unacceptably low for the linear spectral compression with CR=1.6 and 1.7 (see Fig 6.3.). The reason for this result could be a too strong frequency lowering of the first formant, which was observed during the vowel feature information transmission analysis (Fig. 6.7).

In spite of the observed improvement of the vowel and consonant identification, the recognition of the German sentences with the other processing schemes (L13, S12, S13) was approximately the same as for the reference signal (~83%). Note that the sentence recognition scores for these low spectral compression ratios (CR≤1.3) were still better than both vowel and consonant identification scores. Obviously, the subjects used context information for additional sentence recognition score improvement [B31].

Considering the improved consonant and vowel recognition which could be observed for every tested spectral compression signal processing scheme, it can be assumed that normal hearing subjects did use the spectrally transposed information. The spectral compression on the SPINC scale with CR=1.3 showed the largest consonant identification scores observed in this study. This finding is consistent with the fact that this spectral compression delivered the most information from the high frequency areas (see Fig. 6.2), which could be used for the additional consonant recognition improvement.

The measured reference scores of 55% for the low pass filtered (cutoff by 2000 Hz) and spectrally reduced consonant identification (Fig. 6.4) is in agreement with the study of Vickers *et al.* [B143]. They reported approximately 60% correct consonant identification scores for subjects with no dead regions and the same cutoff frequency.

Speculations of possible improvements of sentence recognition using spectral compression, when additional training would be offered to the subjects, are motivated by the observed improvements of vowel and consonant recognition scores. On the other hand, in view of the relatively high speaking rate of the employed sentence material, it is possible that the identification scores of the Göttingen sentences have already reached saturation.

The results of the information transmission analysis indicate that after a certain acclimatization phase or training, the consonant identification scores might demonstrate even greater improvement.

One surprising result was the overall observed improvement of the vowel identification scores. The same tendency was reported also by Sekimoto and Saito [B117]. They observed vowel identification improvements only for female speech. Within the present study, all used speech samples were spoken by a male speaker. A plausible explanation for vowel identification improvement could be the increased identification of the second formant vowel feature group (see Fig. 6.7). The importance of the second formant information for vowel recognition is described by Thomas [B129].

Generally observed improvements relative to the reference signal of the overall parameter for consonants and vowels (Fig. 6.5 and 6.7) show that the tested subjects were not confused by the spectral compression processing.

Both spectral compression schemes – linear on the FFT scale and linear on the SPINC scale - preserve speech intelligibility, if the spectral compression ratio does not exceed certain limits. This circumstance is encouraging for the use of both approaches to spectral compression as signal processing schemes for hearing impaired subjects with profound hearing loss or with steeply sloping high frequency hearing loss. However, it is to be expected that spectrally compressed speech sounds rather unnatural and therefore can be of benefit only for persons who can not benefit sufficiently from conventional high power hearing devices.

For additional comparison between the linear spectral compression on the FFT- and the SPINC-scale, speech recognition tests with normal hearing subjects using spectrally compressed and various low pass filtered signals would be of interest. The low pass filtering with cutoff frequencies smaller than 2 kHz which was employed in the present study would possibly better show the influences on speech intelligibility of the additional spectral information transposed from the "inaudible" higher spectral areas. However, only small spectral compression ratios ($CR \leq 1.3$) should be used.

Chapter 7

Sunday

**"Pulled through. This day is getting to be more and more trying.
It was selected and set apart last November as a day of rest."**

M. Twain

Speech perception experiments with hearing impaired listeners using spectral compression

7.1 Overview

Speech perception experiments using two different spectral compression schemes combined with and without spectral reduction were performed on six subjects with moderately severe to profound sloping sensorineural hearing loss.

The subjects with profound sensorineural hearing loss in the high frequency areas and moderately severe to severe hearing loss in the frequency area below 500 Hz indicated significant increase in German sentence and consonant recognition scores when using spectral compression on the SPINC scale. Their vowel identification scores, however, did decrease.

Two subjects with severe to profound steeply sloping hearing loss in the high frequency area and normal hearing to mild hearing loss below 500 Hz did not benefit from spectral compression and indicated a decrease in German sentence, consonant and vowel identification scores. Only one subject with the same hearing loss characteristics indicated a slight benefit in German sentence and vowel identification when linear spectral compression on the FFT scale was used. However, his consonant identification scores decreased.

Spectral reduction in combination with spectral compression resulted in increased consonant identification scores for two subjects. Clear improvement tendency in spectrally reduced and spectrally compressed consonant identification was observed using spectral compression on the SPINC scale. However, for vowel identification, spectral reduction combined with spectral compression turned out to be destructive. Spectral reduction only caused a decrease of the German sentence and vowel recognition for all tested subjects. However, it improved consonant recognition scores in three cases out of six.

For subjects with very poor speech perception capabilities a tendency of increasing relative speech perception in sentences was observed, caused by increasing the consonant identification provided by each of the tested spectral compression schemes.

The results of this study did not show a universal increase of speech perception neither for spectral compression on the FFT scale nor on the SPINC scale. However, based on the results of the present study, some suggestions for subject selection for both investigated spectral compression schemes can be made.

7.2 Motivation

In the previous study, the effects of linear spectral compression on the FFT scale and on the SPINC scale on the speech perception of normal hearing subjects have been described (see chapter 6). It was observed that by using spectral compression on the SPINC scale there was no significant decrease in speech perception in comparison with linearly compressed speech on a FFT scale using the same spectral compression ratios.

The better vowel identification scores, observed in the previous study, were contradictory to the expectations, that spectral compression increases only the perception of consonants. The improvement of vowel recognition for normal hearing subjects using non-linear spectral compression has also been observed by Sakamoto & Goto [B116].

In the present study, German sentences and consonant and vowel logatome identification scores for non-processed signals (reference) were compared with those produced using spectral reduction only, linear spectral compression on the SPINC scale, and linear spectral compression on the FFT scale (both combined with and without spectral reduction). The spectral reduction was introduced to avoid potential masking and overlapping of spectral segments narrowed through the spectral compression operations.

Six sensorineural hearing impaired subjects with high frequency hearing loss participated in the present study. At least two of the tested subjects had very little benefit from their hearing devices and used them only for lip-reading support. For most of the tested subjects the consonant identification was a real problem and they generally indicated better vowel recognition.

One of the hearing-impaired subjects indicated very good sentence recognition and was not expected to benefit from any spectral compression method. He was also the only one who indicated better consonant than vowel recognition scores with a non-processed (reference) signal.

Generally, the choice of hearing impaired subject groups, who would possibly benefit from spectral compression, remains unclear. Moreover, according to the experience of McDermott [B87], hearing-impaired subjects with severe to profound steeply sloping high frequency hearing losses are not ideal candidates for spectral compression strategies.

7.3 Method

7.3.1 Signal processing and test parameter settings

Audio files in *.wav format from the Innsbruck sentence test and German C12 a-C-a consonant and Vo8 d-V Vowel identification tests were processed using the signal processing algorithm based on the sinusoidal speech model (described in chapter 4). The present study used a simplified signal flow scheme consisting of spectral analysis, spectral reduction, spectral compression, and spectral reconstruction (Fig 7.1). Spectral reduction and spectral compression operations could be switched "on" or "off" by means of a particular setting of the signal processing parameters.

In the spectral analysis block, FFT calculations were performed using a 256 point window length, resulting in an effective duration of 11.6 ms (sampling frequency of 22050 Hz). The choice of the analysis window length was based on previous studies with normal hearing subjects (chapters 5 and 6). To increase spectral and temporal resolution of the spectral analysis block, 75% overlap and Hanning windowing were used. In addition, a frequency interpolation based on phase and approximate value data was carried out.

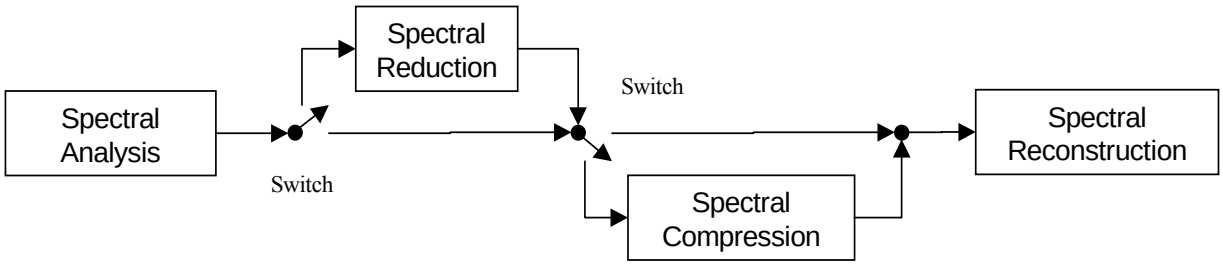


Fig. 7.1 Signal processing block diagram.

The spectral reduction block performs spectral peak identification and spectral peak selection. The number of spectral peaks used in this study was eight. The spectral reduction block could be switched off to produce spectrally compressed signals using the full input spectrum. In this case, each spectral component detected in the spectral analysis block was passed directly to the spectral compression block.

In the spectral compression block, linear spectral compression on the FFT-scale or on the SPINC-scale could be performed. The linear spectral compression on the FFT-scale was implemented as follows:

$$F_{OUT} = \frac{F_{IN}}{CR}, \tag{7.1}$$

where F_{OUT} is the output frequency, F_{IN} is the input frequency and CR is the compression ratio (see also chapter 2 and 4).

The linear spectral compression on the SPINC-scale is given by the equation:

$$F_{OUT} = const * \tan\left(\frac{\arctan(F_{IN}/const)}{CR}\right), \tag{7.2}$$

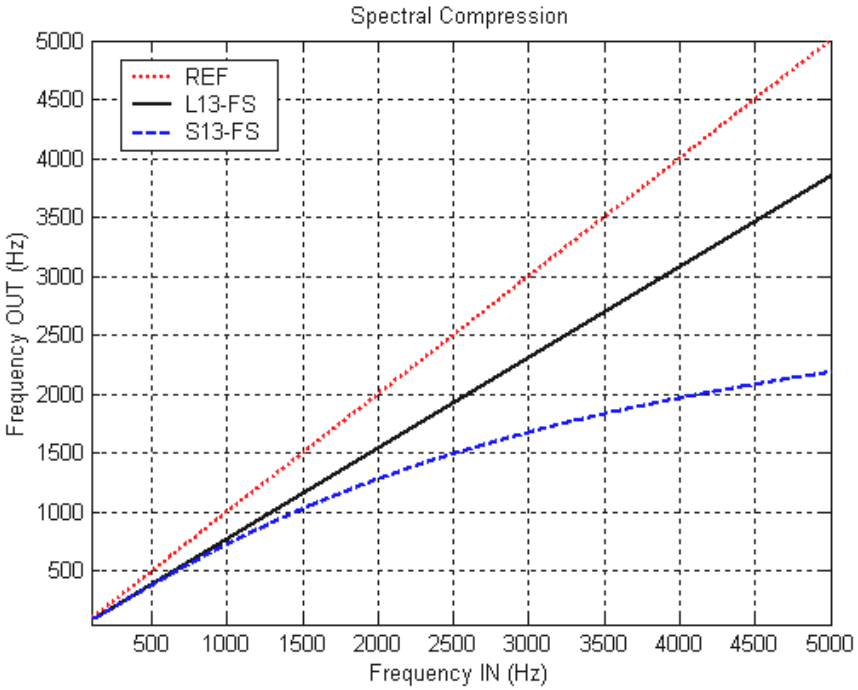
where F_{OUT} and F_{IN} are again the input and output frequencies, $const = \sqrt{2} * 1000$ is a empirical constant , and CR is the compression ratio (chapter 2 and 4).

A spectral compression ratio of CR=1.3 was chosen for both the FFT and the SPINC scale based on the previous study with normal hearing subjects (see chapter 6). Table 7.1 lists all signal processing schemes used in the present study. Higher compression ratios may cause dramatic reduction in vowel and sentence identification. The frequency input-output curves for the employed reference processing and the linear spectral compression on FFT- and SPINC-scales with CR=1.3 are given in Fig. 7.2 in normal and logarithmic scale. To generate a signal which is only spectrally reduced, the spectral compression block in Fig. 7.1 can be switched off.

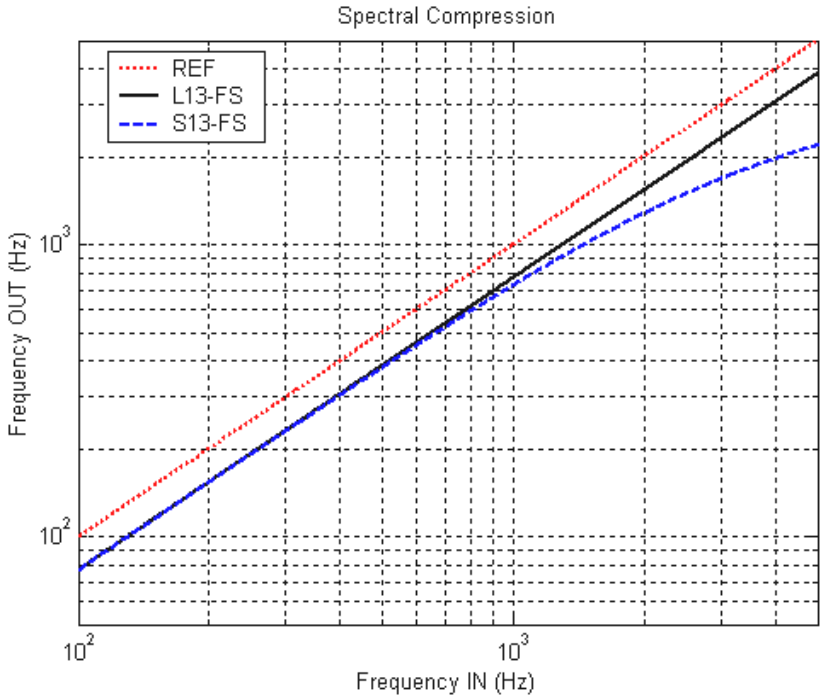
In the spectral reconstruction block, pure tones (sinusoids) corresponding to all frequency values produced by the spectral compression block are generated. The length of the reconstruction frame is twice as long as the length of the non-overlapped part of the analysis window. In the present study, a reconstruction frame length of 5.8 ms was used. The synthesis frame is multiplied with a triangular window of the same length and then added to the previously reconstructed frame with 50% overlap. This reconstruction method is a slightly modified version of the "Teilton" reconstruction technique described by Mummert [B95].

Signal processing scheme	**Nomenclature**
Reference non-compressed full spectrum signal	**REF**
Reference non-compressed with 8 spectral component used for reconstruction	**REF-8SC**
Linear spectral compression on the FFT-scale with CR=1.3 and full spectrum reconstruction	**L13-FS**
Linear spectral compression on the FFT-scale with CR=1.3 and 8 spectral components used for reconstruction	**L13-8SC**
Linear spectral compression on the SPINC-scale with CR=1.3 and full spectrum reconstruction	**S13-FS**
Linear spectral compression on the SPINC-scale with CR=1.3 and 8 spectral components used reconstruction	**S13-8SC**

Table 7.1 signal processing schemes used in the present study.



a)



b)

Fig. 7.2 Frequency input –output curves showing the investigated spectral compression methods and the reference. REF – reference, L13-FS –full spectrum linear compression on the FFT scale with CR=1.3, S13-FS – full spectrum linear compression on the SPINC-scale with CR=1.3: a) linear frequency display, b) logarithmic frequency display.

7.3.2 Performed speech tests and tested subjects

The German C12 consonant "aCa" logatome identification test, the German Vo8 vowel "dV" test and the Innsbruck sentence test were performed with 6 hearing impaired native German speaking adults. All speech tests were performed in a random sequence.

The C12 test includes three different pronunciations of each consonant of the 12 German consonants and was randomly repeated four times for each parameter settings. The Vo8 test includes three different pronunciations of the eight German vowels, and was randomly repeated three times for each of the parameter settings. The Innsbruck sentence test consists of 12 lists of 10 sentences each. For each of the parameter settings two different lists were tested. Consonant and vowel identification test materials used recordings of a male speaker. Innsbruck sentence test used recordings of a female speaker.

All processed *.wav files are mono and have a sampling frequency of 22.05 kHz and a 16 bit amplitude resolution.

All performed speech tests were carried out in a silent environment. The speech was presented at 65 dB RMS in a sound proof room at a distance of 1.5 m from a Philips Type 22AH586/16R active loudspeaker (playback was performed by a 16 bit PC sound card).

Different speech tests were performed in at least two sessions. In each session only one spectral compression method (linear or SPINC) was tested. Before each of the test sessions a ten minutes long training and session adaptation was performed. It consisted of a spectrally compressed speech signal spoken by a male speaker. During the adaptation phase, the subject could follow the text by simultaneous reading. Depending upon the compression method under investigation, linear spectral compression with CR=1.3 or SPINC compression with CR=1.3 was used.

As mentioned above six subjects with severe to profound sensorineural hearing loss were included in the present study. All subjects were temporally binaurally provided with Supero 412 Phonak hearing devices (max output 145 dB SPL, max gain 86 dB, prescription method - Phonak Digital Power, hook HE3 680). The audiograms, which were constructed from the better hearing loss levels of both ear audiograms, are shown in Fig. 7.3 for all tested subjects. Only two of the tested subjects had thresholds of hearing better than 100 dB SPL above 1 kHz (subjects EA and AA). All the subjects exhibited increased hearing loss levels for frequencies over 250 Hz. Two of the tested subjects indicated hearing loss levels larger then 60 dB for frequencies up to 250 Hz. Subjects "HB" and "EA" indicated profoundly steep sloping hearing-loss characteristics.

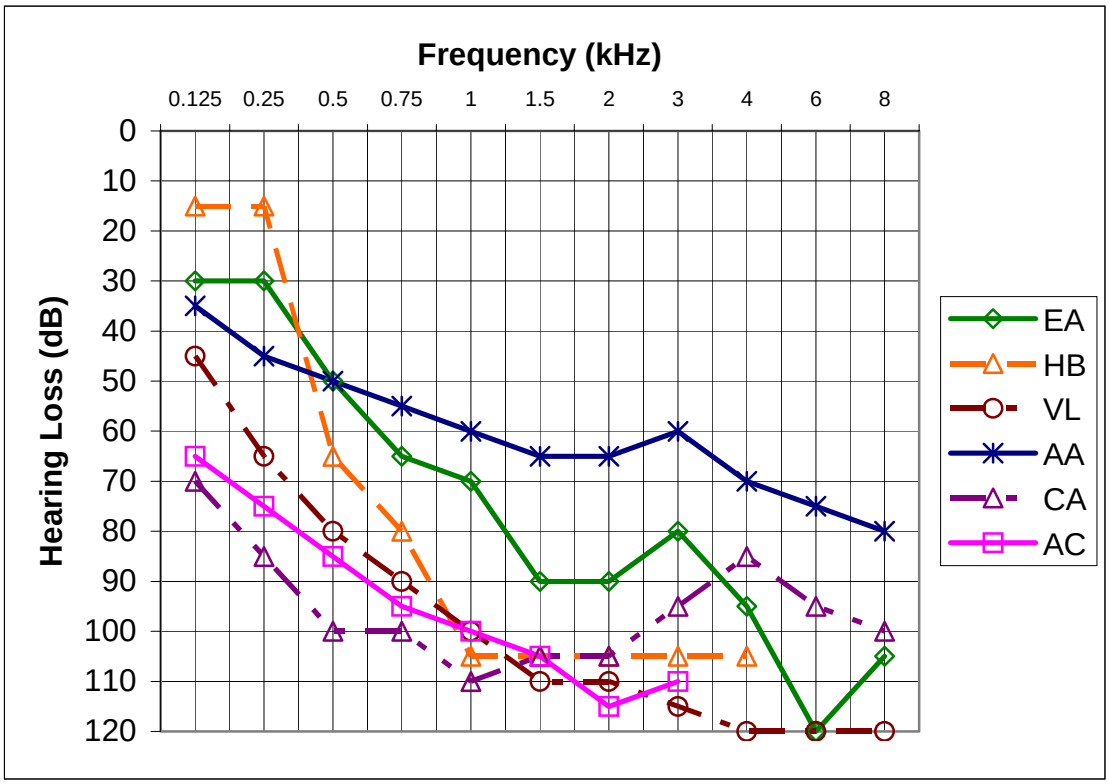


Fig. 7.3 Audiograms, constructed from the better hearing loss levels of both ear audiograms, for all tested subjects.

From other audiometric tests (not documented here), the subjects "EA" and "HB" indicated 80-95% German sentence recognition using the Innsbruck sentence test. The subjects "VL" and "AA" indicated nearly 50% and the subjects "CA" and "AC" indicated only 5-15% sentence recognition scores. All results were obtained in quiet. Subject "CA" was recently provided with a cochlear implant. The cochlear implant was switched off during the speech tests with spectrally compressed or reduced signals. Additional reference measurements for subject "CA" were performed using the cochlear implant only. Subject "AC" was a potential candidate for a cochlear implantation.

Two of the tested subjects were females and four were males. The age of the subjects was between 21 (AC) and 76 (AA) years (mean age= 44.5±20.8 STD).

The majority of the tested subjects indicated good lip-reading skills. All subjects were paid for their participation in this experiment.

7.4 Results

7.4.1 General observations

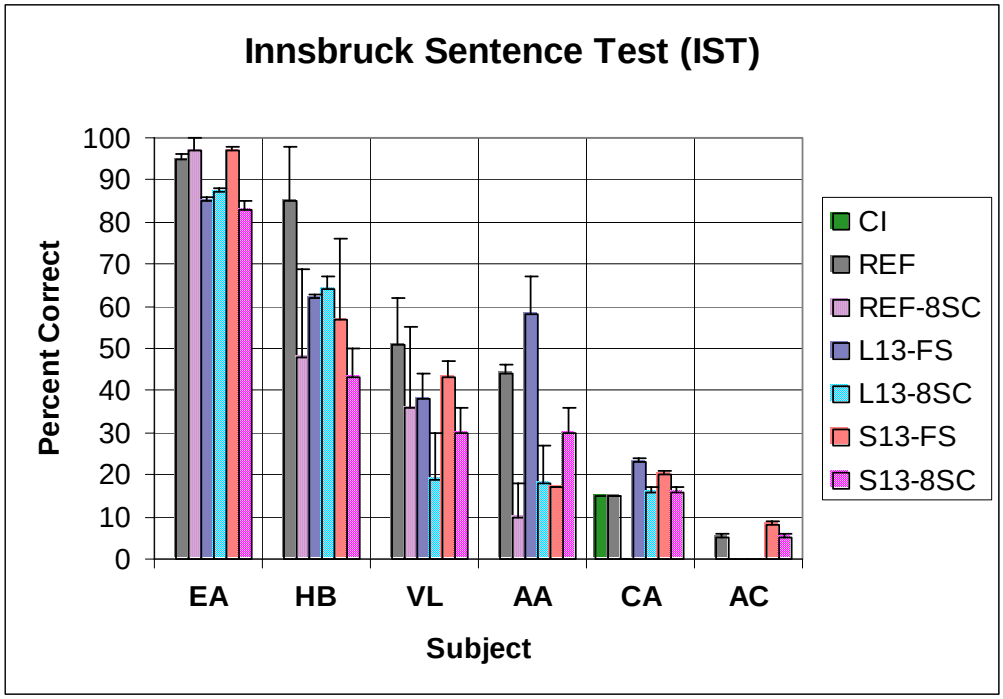
The results of the average correct identification scores and absolute improvement in percent against reference signal scores (REF) for the Innsbruck sentence test (IST), the C12 German

consonant identification test and the V08 German vowel identification test are all shown in Fig. 7.4-7.6.

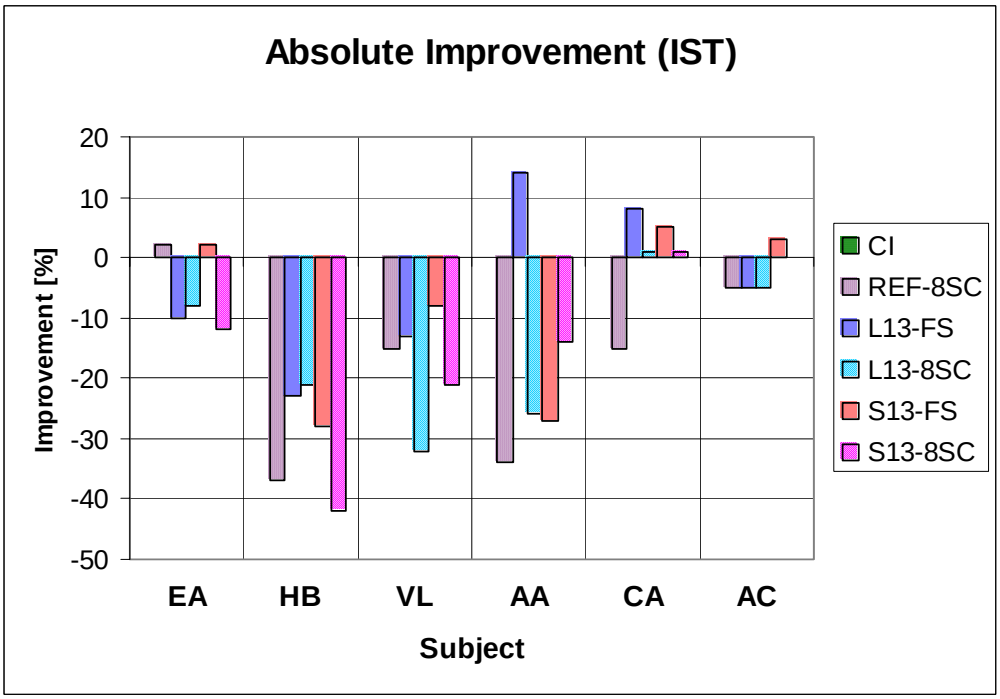
Subjects “AA”, “CA”, and “AC” indicated improvement relative to the reference signal for the IST sentence recognition. However, subject “AA” profited only from linear spectral compression on the FFT scale with full spectral reconstruction (L13-FS), subject “AC” profited only from linear spectral compression on the SPINC scale using full spectral reconstruction (S13-FS). For this latter subject, all other signal processing schemes failed totally (*i.e.* no sentence recognition at all). Subject “CA” indicated improved IST sentence recognition scores from both linear spectral compression on the FFT and on the SPINC scales. Subject “EA” indicated insignificant improvement of the IST sentence recognition scores for the spectrally reduced-only signal (REF-8SC) and for the full spectrum SPINC spectrally compressed signal (S13-FS). Subjects “HB” and “VL” indicated significantly worse IST recognition scores relative to the reference signal for any signal-processing scheme.

For the IST sentence recognition, spectrally reduced signals indicated worse recognition scores in 13 (72%) cases out of 18 (6 subjects x 3 signal processing schemes). In 5 (42%) out of 12 cases (6 subjects x 2 spectral processing schemes), linear spectral compression on the FFT scale was better than linear spectral compression on the SPINC scale; in 2 cases their score were equal, and in 5 cases linear spectral compression on the SPINC scale was better than spectral compression on the FFT scale (Fig. 7.4).

For consonant recognition, similar observations were made (Fig. 7.5). Subjects “AA”, “CA”, and “AC” clearly profited from almost every signal-processing scheme. For these subjects, the linear spectral compression on the SPINC scale combined with spectral reduction indicated better relative improvement scores than the corresponding linear spectral compression on the FFT scale. The three other subjects, “EA”, “HB”, and “VL” indicated decreased consonant recognition scores for every spectral processing scheme. Spectrally reduced signals gave worse recognition scores in 11 (61%) cases out of 18. In 6 (50%) out of 12 cases linear spectral compression on the FFT scale was better than linear spectral compression on the SPINC scale, and *vice versa* (Fig. 7.5).

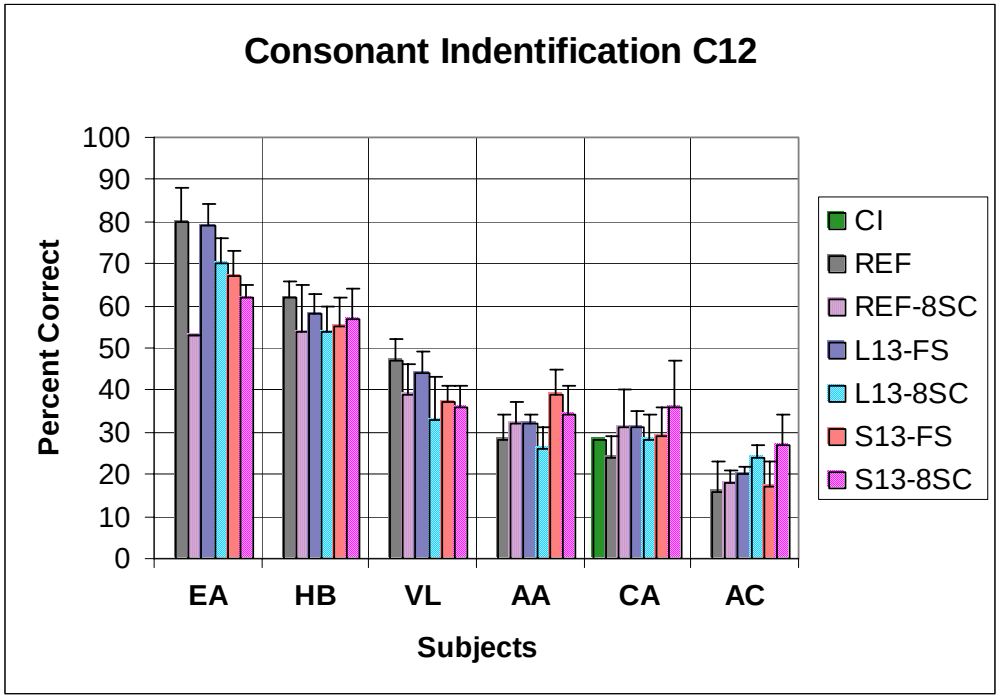


a)

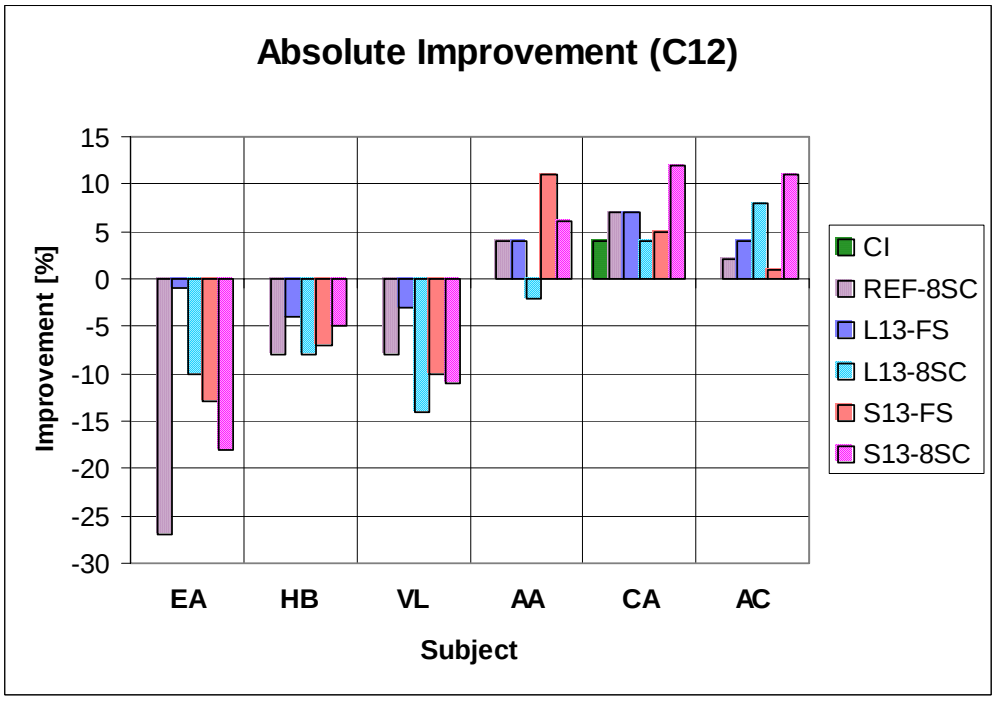


b)

Fig. 7.4 a) Average percent correct identification scores of the Innsbruck sentence test over all subjects for reference and each of the tested spectral processing methods; b) absolute improvement of identification scores in percent calculated against reference signal scores over all subjects for each of the tested spectral processing method.

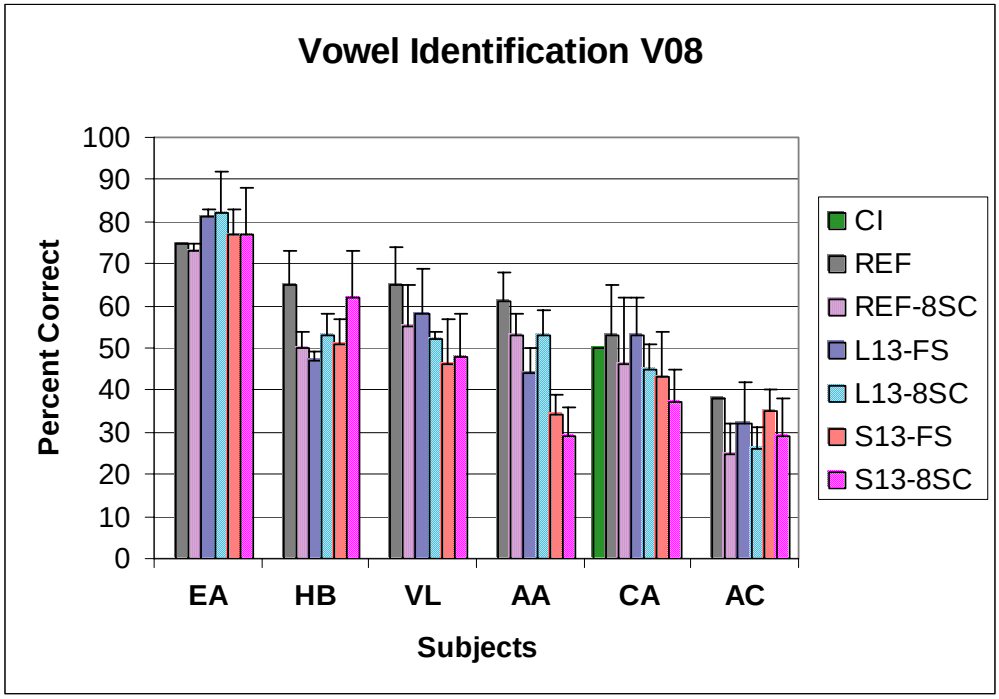


a)

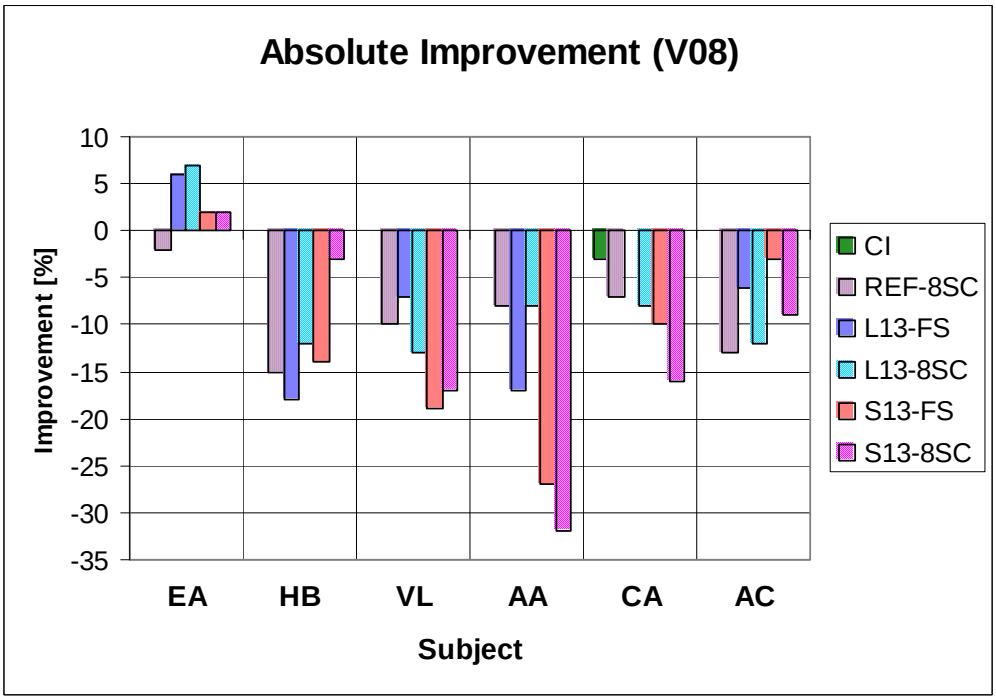


b)

Fig. 7.5 a) Average percent correct identification scores of the C12 German consonant identification test over all subjects for reference and each of the tested spectral processing methods; b) absolute improvement of identification scores in percent calculated against reference signal scores over all subjects for each tested spectral processing method.



a)



b)

Fig. 7.6 a) Average percent correct identification scores of the V08 German vowel identification test over all subjects for reference and each of the tested spectral processing methods; b) absolute improvement of identification scores in percent calculated against reference signal scores over all subjects for each tested spectral processing method.

For vowel recognition, only subject "EA" indicated increased improvement scores (Fig. 7.6). He generally profited from each of the spectral compression schemes. Linear spectral compression on the FFT scale indicated slightly better recognition scores than spectral compression on the SPINC scale. All other subjects indicated worse recognition scores for all spectral processing schemes inclusively cochlear implant vowel recognition scores for subject "CA". Spectrally reduced signals gave worse recognition scores in 12 (66%) cases out of 18. In 8 (67%) cases out of 12, the linear spectral compression on the FFT scale indicated better results for vowel recognition than the linear spectral compression on the SPINC scale.

7.4.2 Investigation of different subject categorization

Based on the full-spectrum reference sentence identification scores (Fig. 7.4a), subjects were categorized into three different groups. The first group, including subjects "EA" and "HB", shows good sentence recognition (>85%). The second group of subjects, including "AA" and "VL", shows moderate sentence recognition on the order of 50%. The third group with subjects "AC" and "CA" indicates very poor sentence recognition scores (less or close to 20%). Group related subject audiograms are shown in Fig. 7.7-7.9.

Absolute improvement scores of the three subject groups are given in Fig. 7.10-7.12 for the IST sentence recognition, German C12 consonant recognition, and German Vo8 vowel recognition. The first group did not indicate benefit in sentence recognition from any of the implemented signal processing schemes (Fig. 7.10). The smallest absolute worsening for the IST recognition occurred with the full spectrum spectral compression on the SPINC scale. For the average consonant recognition (Fig. 7.11), the full spectrum spectral compression on the FFT scale gave the least score decrease. For the group average vowel recognition (Fig. 7.12), the same happened for the spectrally reduced spectral compression on the SPINC scale. In comparison with the other subject groups, group I indicated on average the smallest decrease in vowel recognition for any of the implemented signal processing schemes. In addition for subject group I, spectral reduction in combination with spectral compression indicated better results than full spectrum spectral compression for vowel recognition only. For sentence and vowel recognition, the opposite was found, *i. e.* full spectrum spectral compression worse than spectrally reduced compression. The average results of group I on sentence, consonant and vowel identification spectrally reduced signals indicated decreased recognition scores compared with spectrally non-reduced signal in 6 (67%) cases out of 9 when compared with spectrum processing schemes. With full spectrum processing the linear spectral compression on the FFT scale was better than linear spectral compression on the SPINC scale in 4 (67%) cases out of 6.

The second subject group showed inconsistent results for IST and consonant recognition. For the first subject in group II (VL), all signal processing schemes decreased the recognition scores. For the second subject (AA), some of the processing schemes gave a benefit. In particular, the full spectrum linear compression on the FFT scale was advantageous for sentence recognition, and all processing schemes except spectrally reduced compression on the FFT scale gave better consonant recognition. Out of all subject groups, the second group indicated the largest decrease for vowel identification. In addition, the subjects from the second group did not profit from the use of spectral reduction in combination with spectral compression. In 8 (89%) out of 9 cases, the signal-processing

schemes without spectral reduction gave better scores than the signal processing schemes, in which spectral reduction was used.

The third subject group indicated a benefit from using spectral compression signal processing schemes for sentence and consonant identification. For vowels, however, all signal processing schemes resulted in decreasing recognition scores. In the case of consonant identification, spectral reduction in combination with the spectral compression on the SPINC scale produced better results than the full spectrum compression. In this case, the absolute improvement in the consonant recognition relative to the reference signal was larger than 10%. Considering all results of the third subject group, the spectral reduction indicated a decrease in seven cases out of nine in comparison to the full spectrum processing schemes. The linear spectral compression on the SPINC scale was better than the linear spectral compression on the FFT scale in 3 (50%) out of 6 cases.

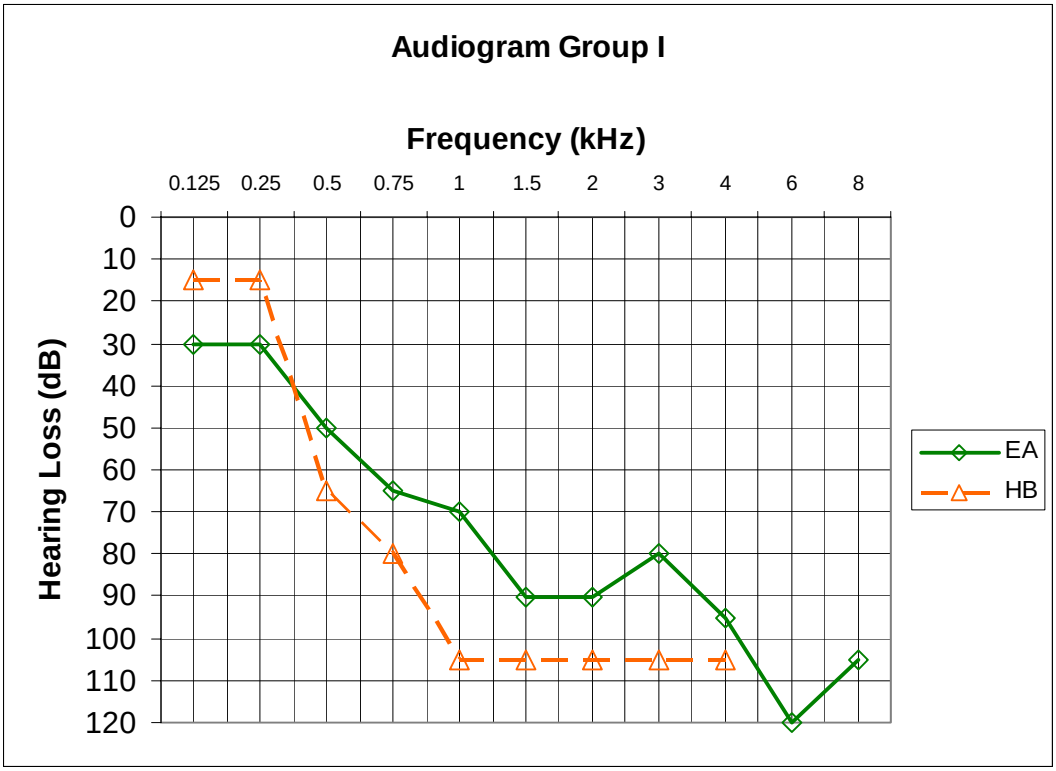


Fig. 7.7 Audiograms, constructed from the better hearing loss levels of both ear audiograms, for the first subject group.

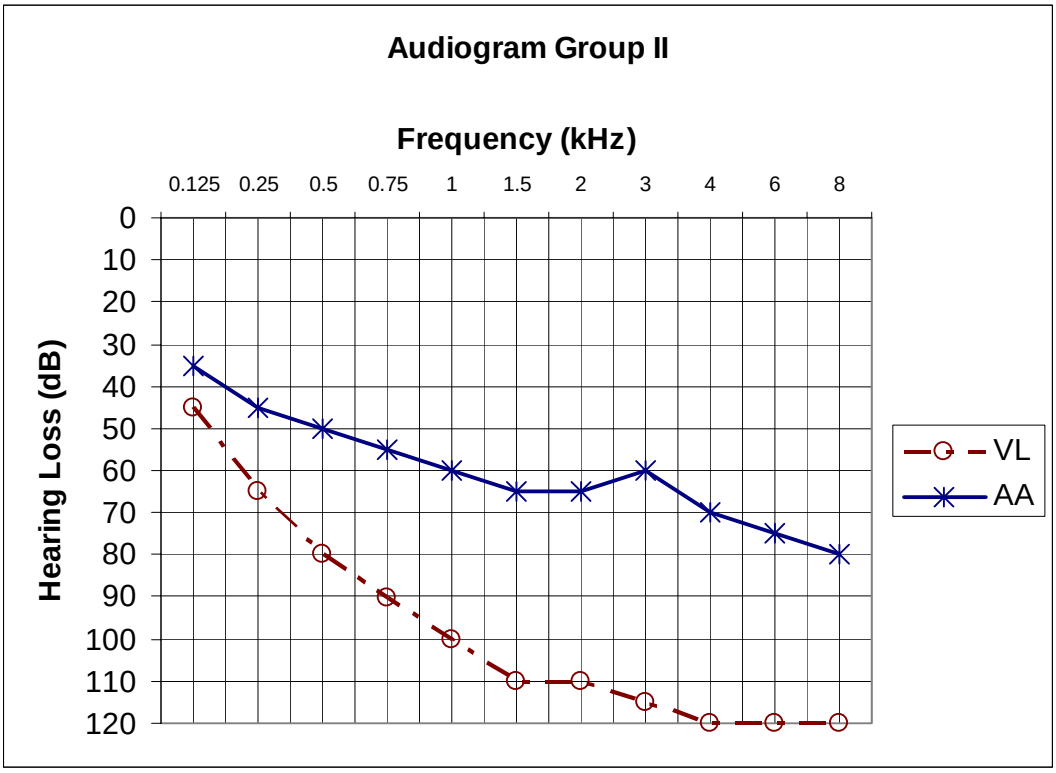


Fig. 7.8 Audiograms, constructed from the better hearing loss levels of both ear audiograms, for the second subject group.

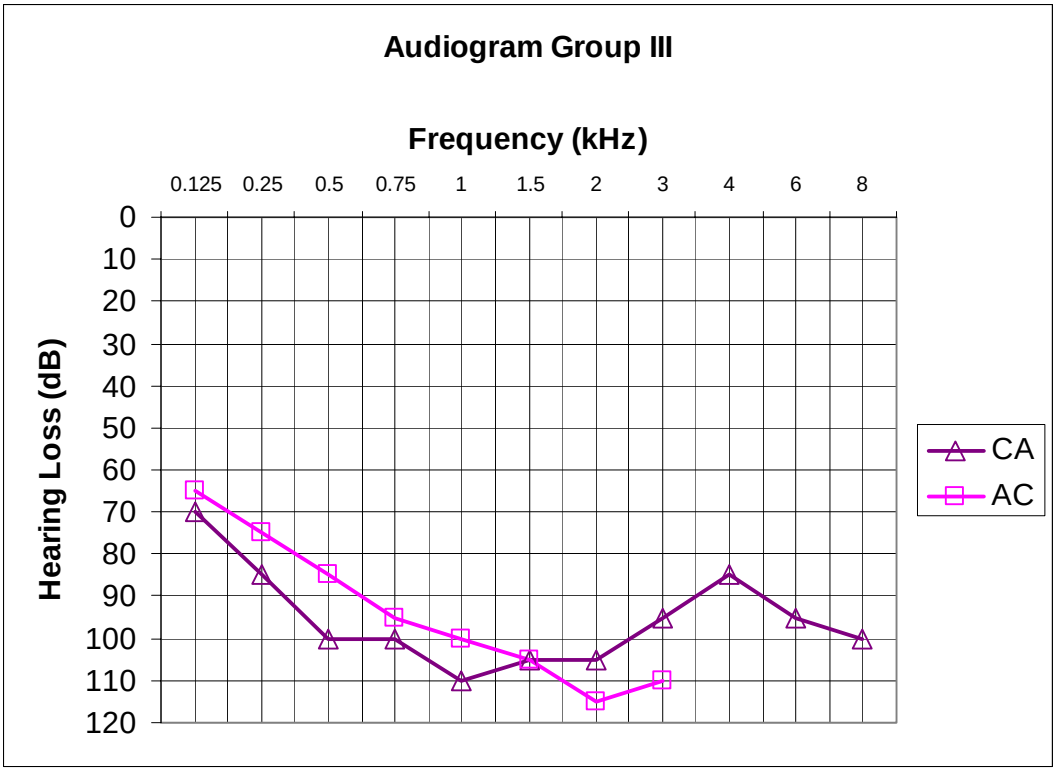
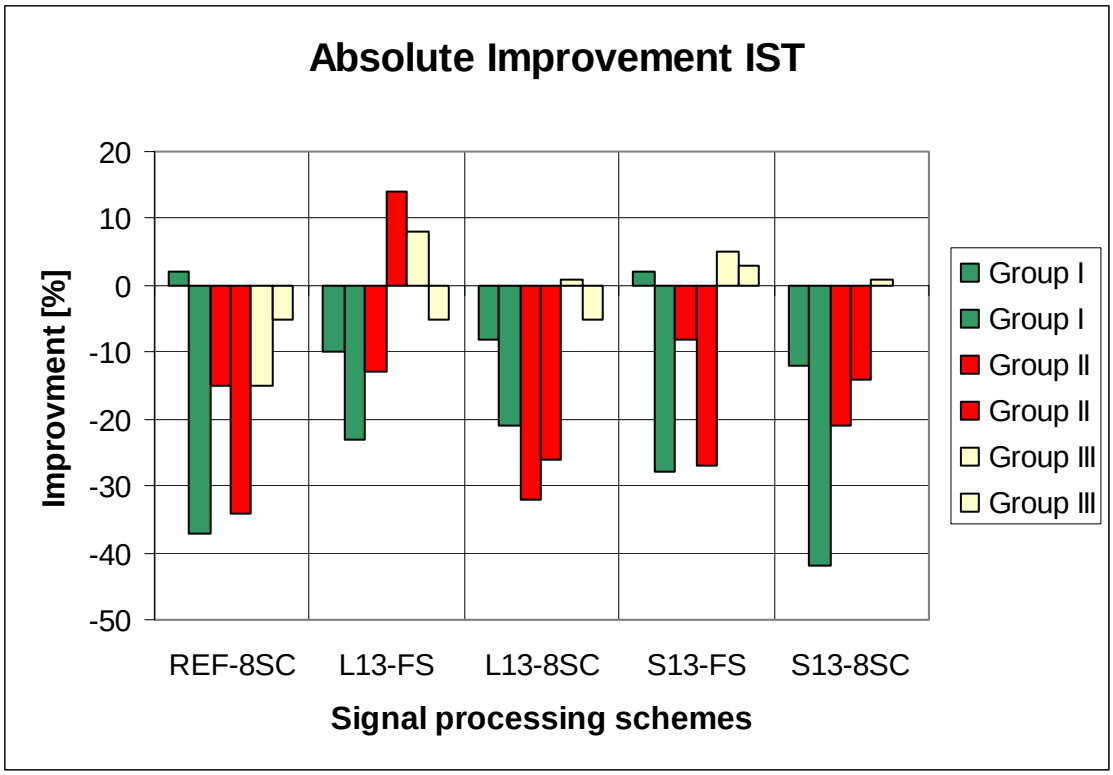
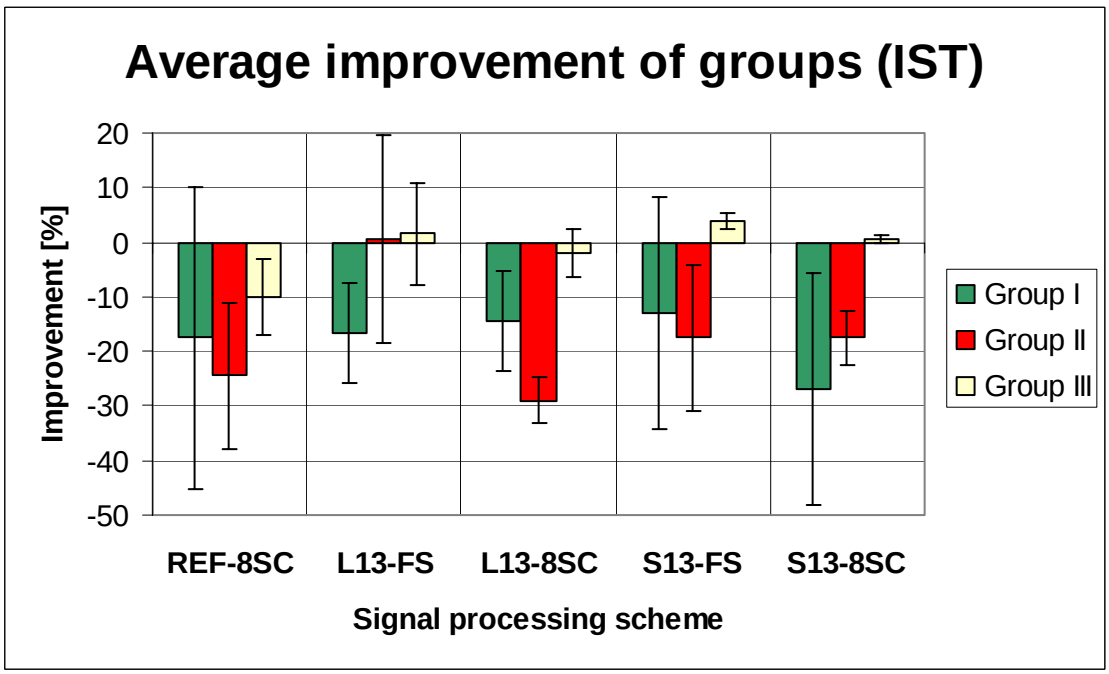


Fig. 7.9 Audiograms, constructed from the better hearing loss levels of both ear audiograms, for the third subject group.

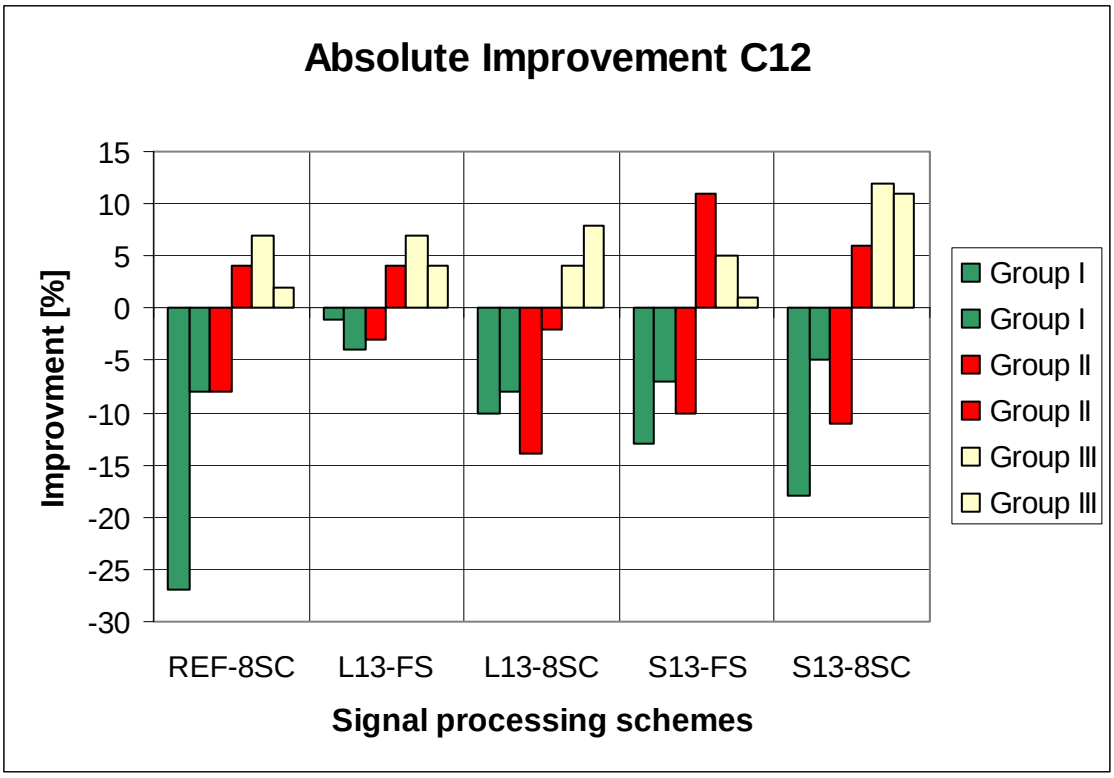


a)

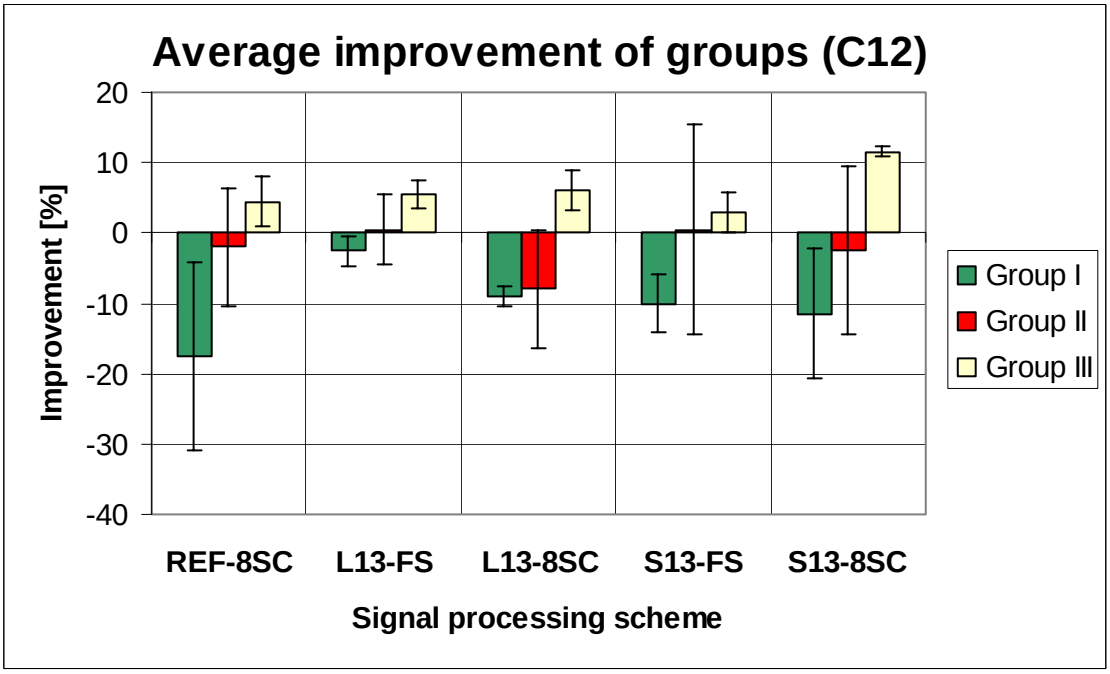


b)

Fig. 7.10 Absolute improvements against unprocessed signal for IST sentence recognition scores for the three subject groups.

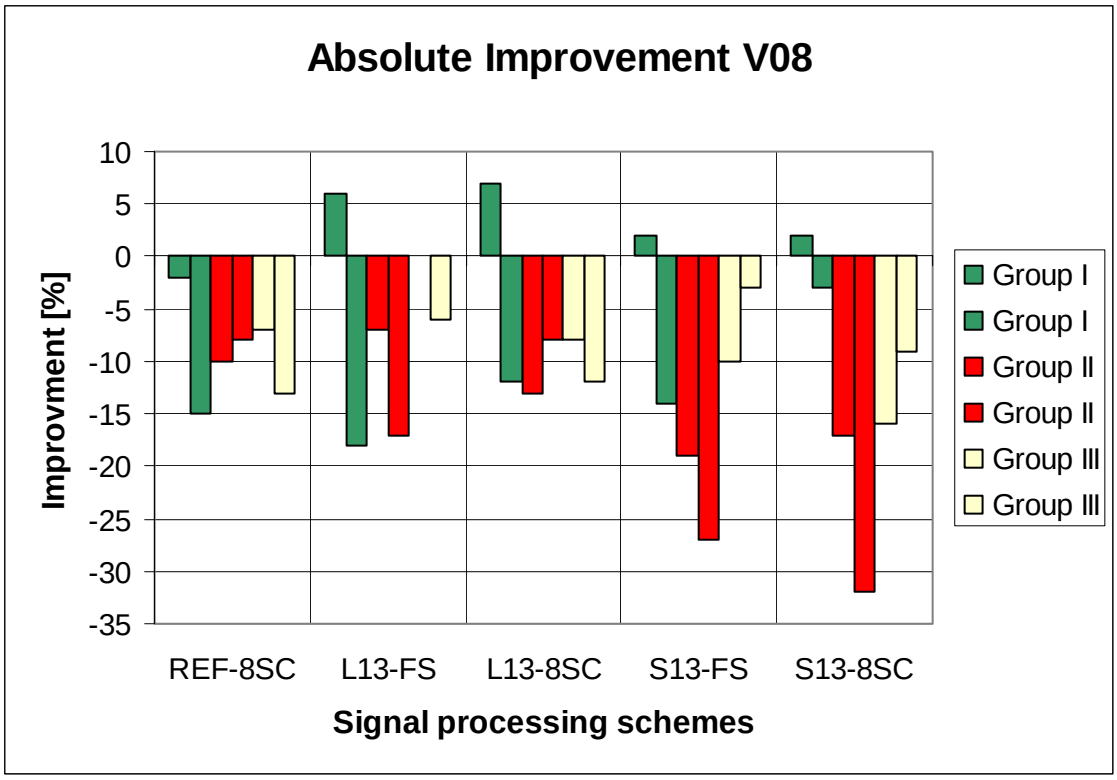


a)

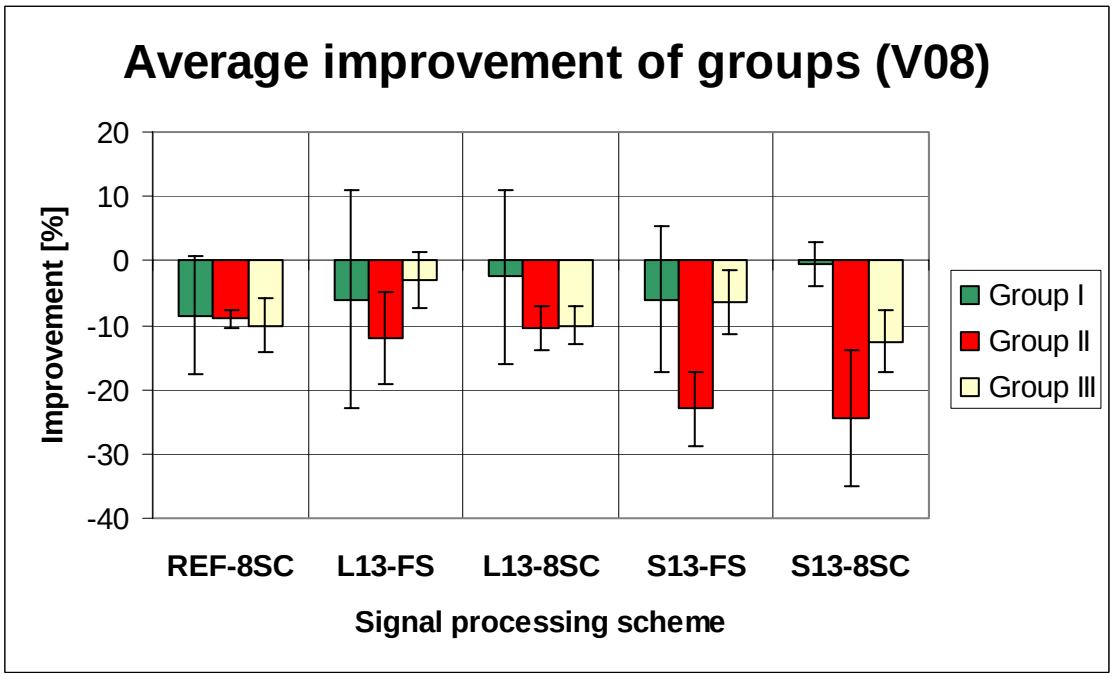


b)

Fig. 7.11 Absolute improvements for German C12 consonant identification scores for the three subject groups.



a)



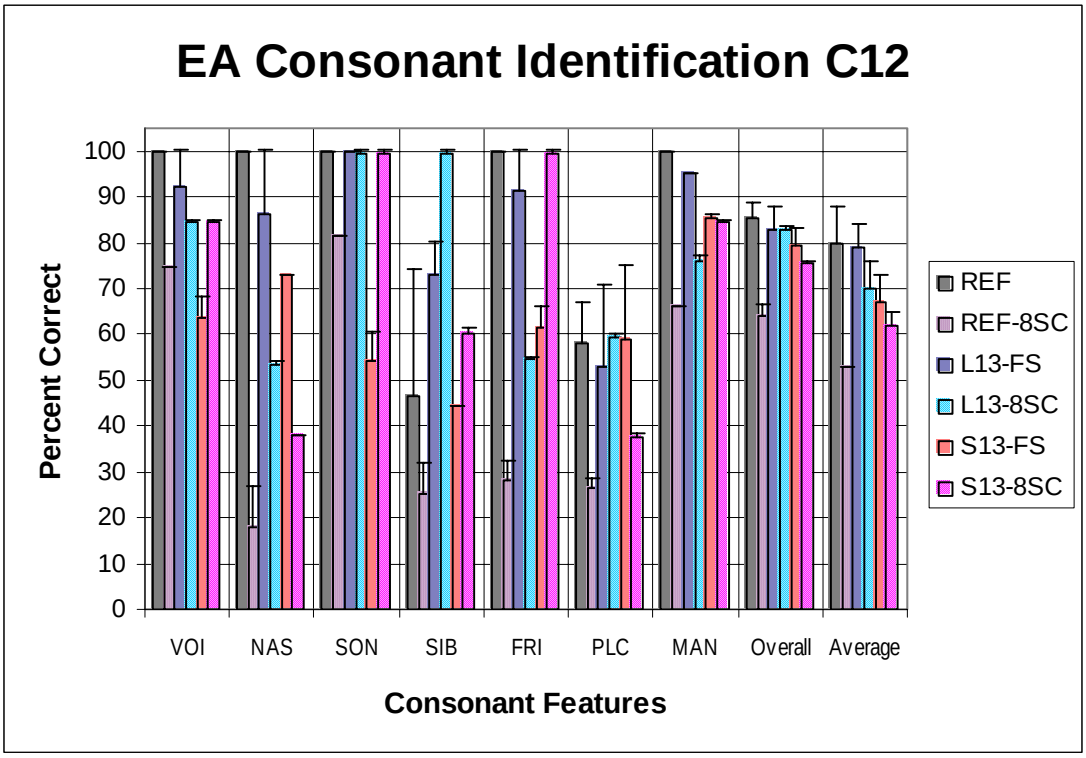
b)

Fig. 7.12 Absolute improvements for German Vo8 vowel identification scores for the three subject groups.

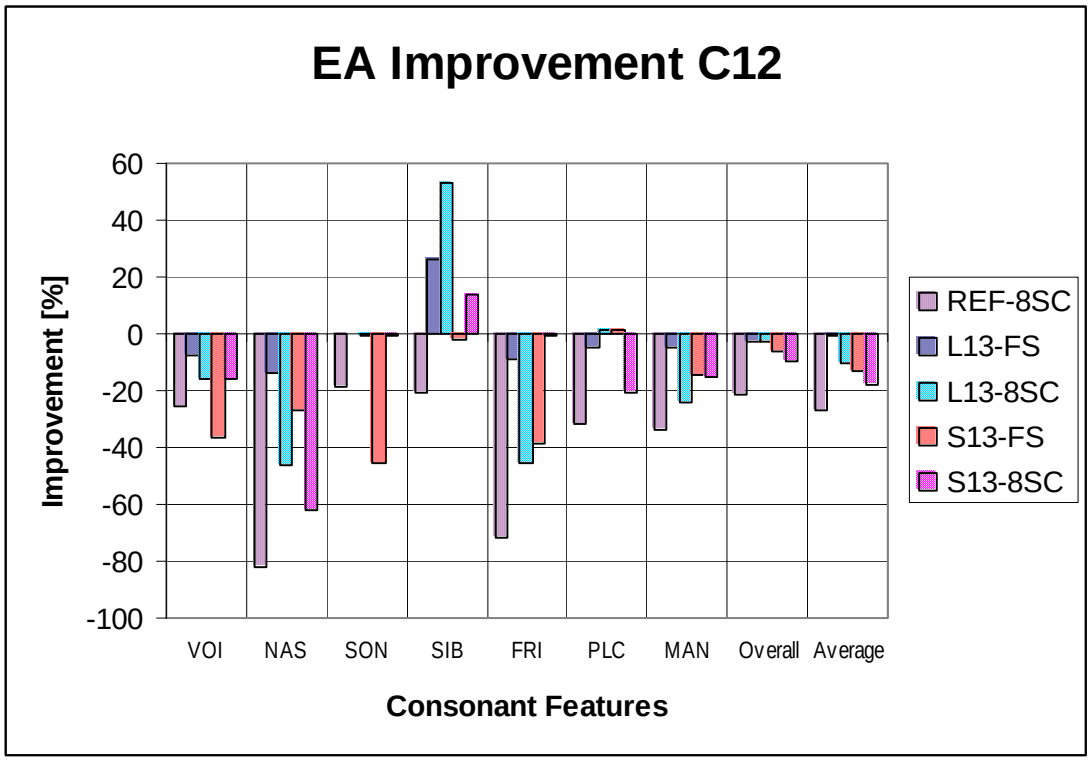
7.4.3 Analysis of consonant and vowel features

To investigate the impact of the employed signal processing schemes on consonant and vowel identification, a vowel and consonant feature analysis according to Miller and Nicely [B92] was performed. The consonant and vowel feature identification scores and the absolute improvement scores for each subject are given in Fig. 7.12 – 7.24 respectively.

Subject "EA" (Fig. 7.13) indicated a significant improvement in the sibilance consonant feature identification when a spectrally compressed signal was used, in particular for spectral compression on the FFT scale. The identification scores of all other consonant features decreased with any processing. Spectrally reduced signals compared with full spectrum signals processed with the same spectral compression scheme indicated decreased consonant feature identification scores in 19 out of 21 cases (90%) (7 consonant features x 3 signal processing schemes). The linear spectral compression on the FFT scale gave better identification scores in 9 out of 14 cases (64%) (7 consonant features x 2 signal processing schemes). The largest absolute decrease (80%) in the consonant feature identification was observed for nasality (NAS), when spectrally reduced only processing was used. The largest improvement (~55%) was observed for the sibilance consonant feature, when the spectrally reduced and on the FFT scale linearly spectral compressed signal was used.

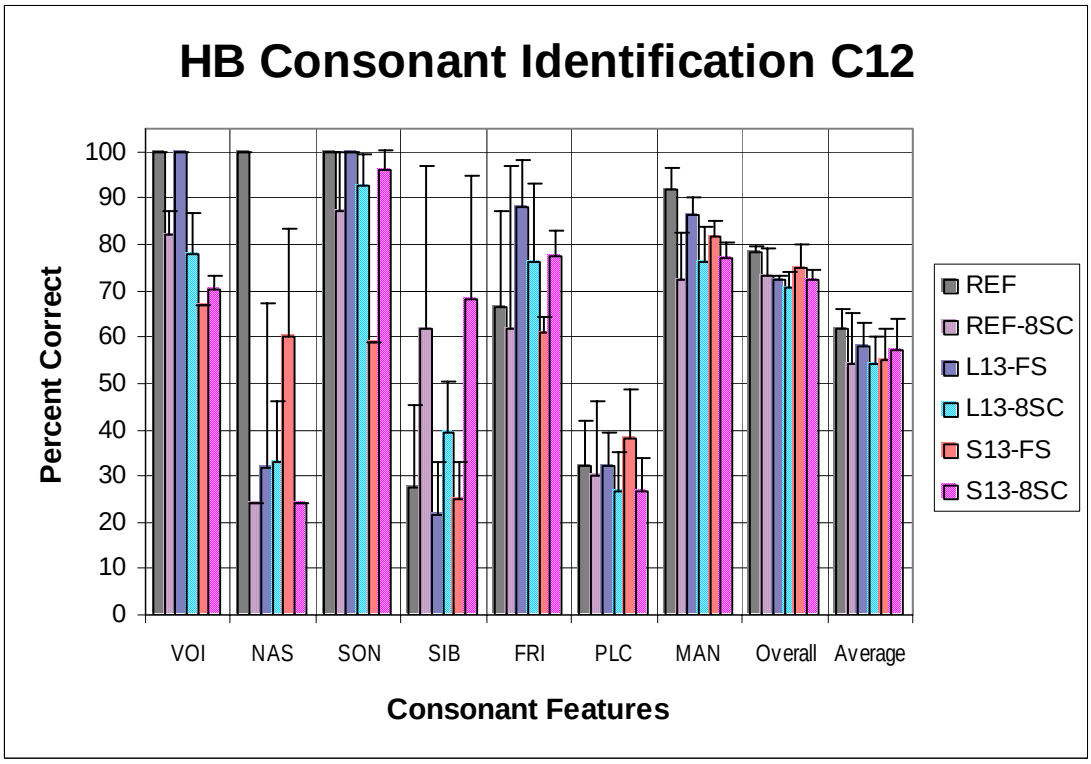


a)

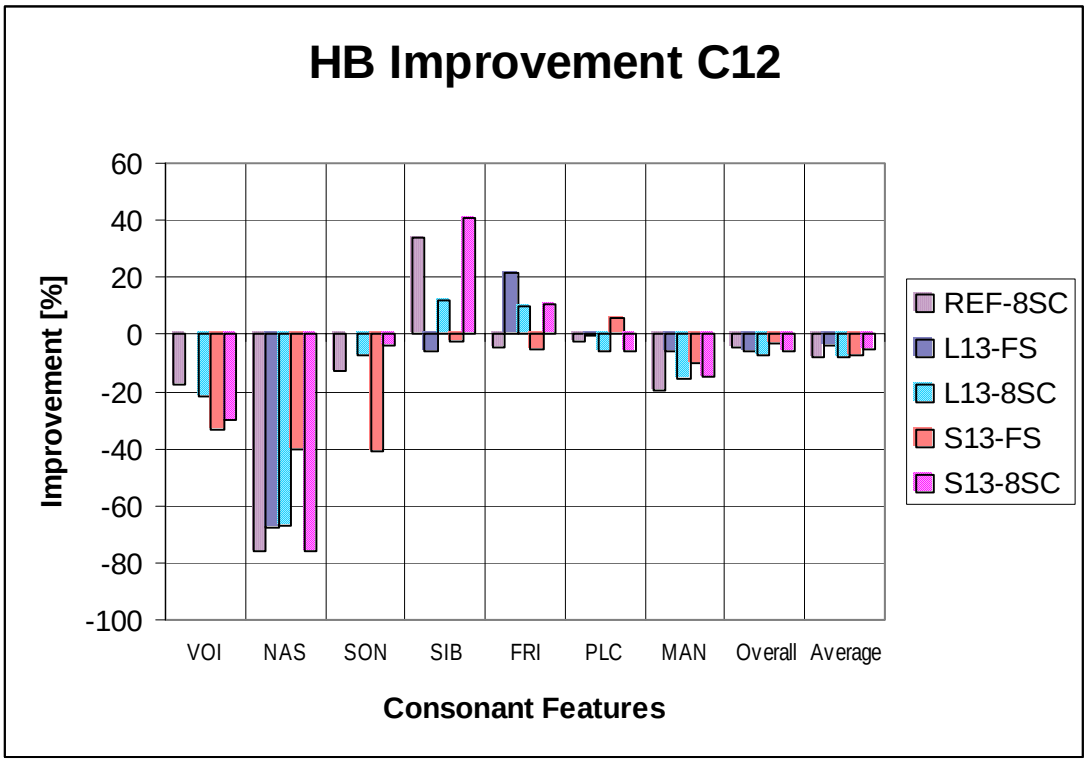


b)

Fig. 7.13 a) German C12 consonant feature recognition scores for subject "EA;" b) Absolute average improvement scores against reference.

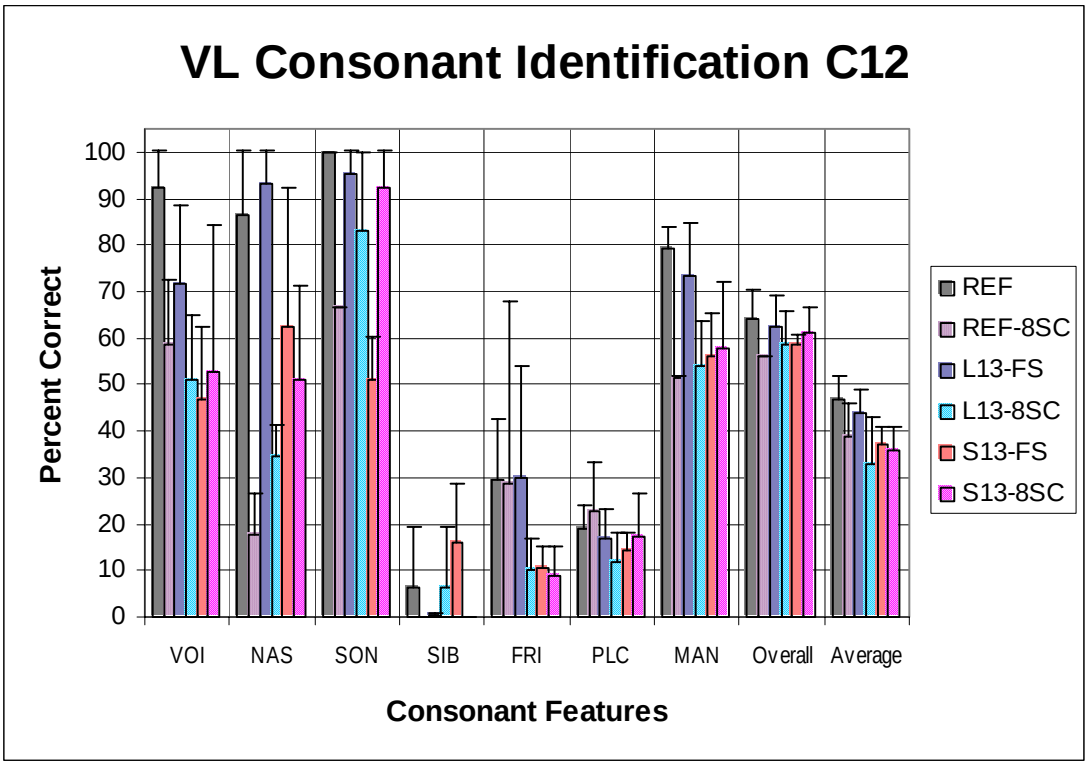


a)

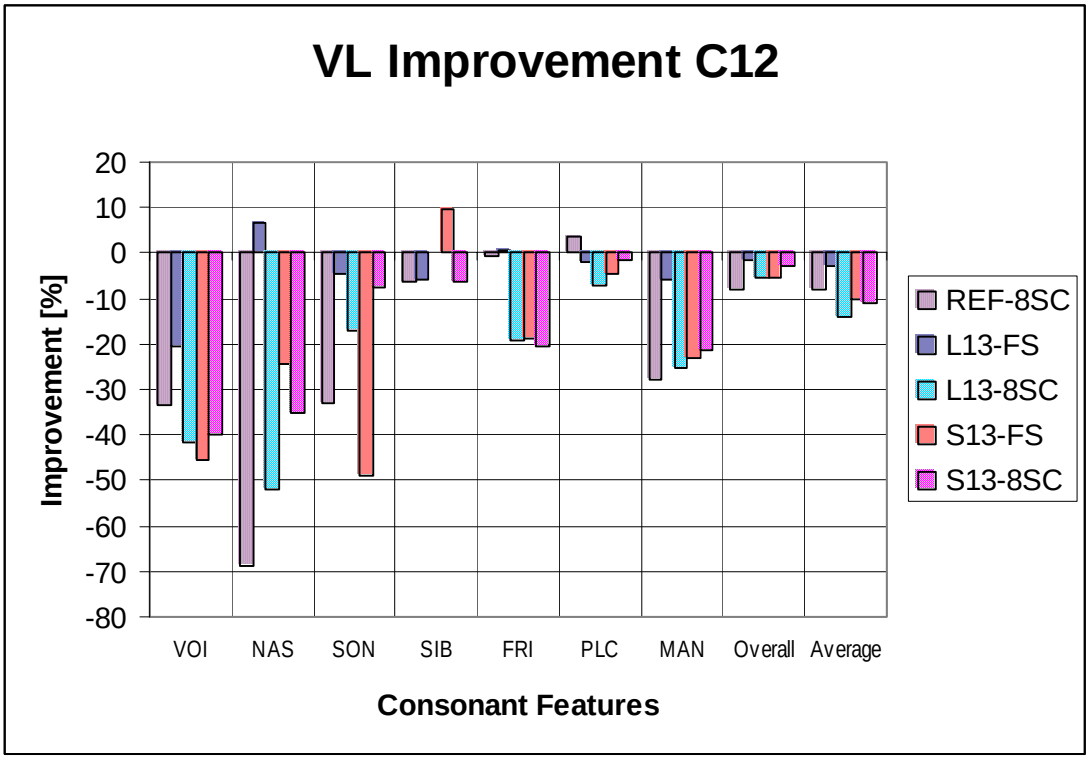


b)

Fig. 7.14 a) German C12 consonant feature recognition scores for subject "HB;" b) Absolute average improvement scores against reference.

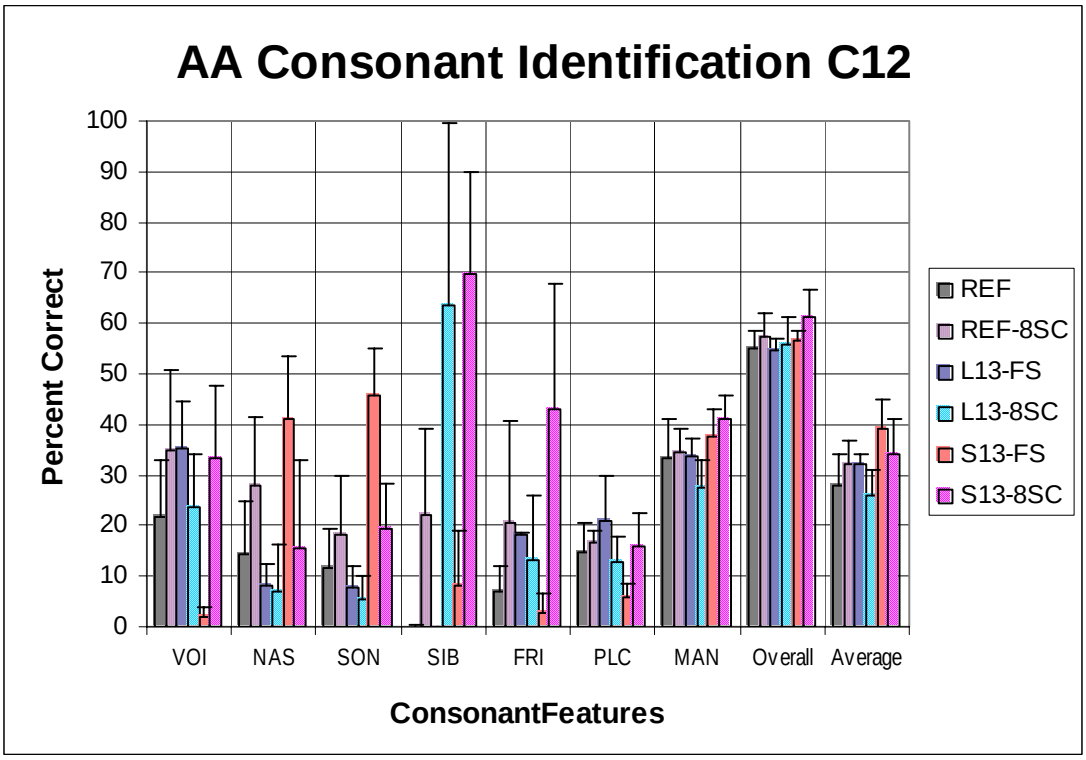


a)

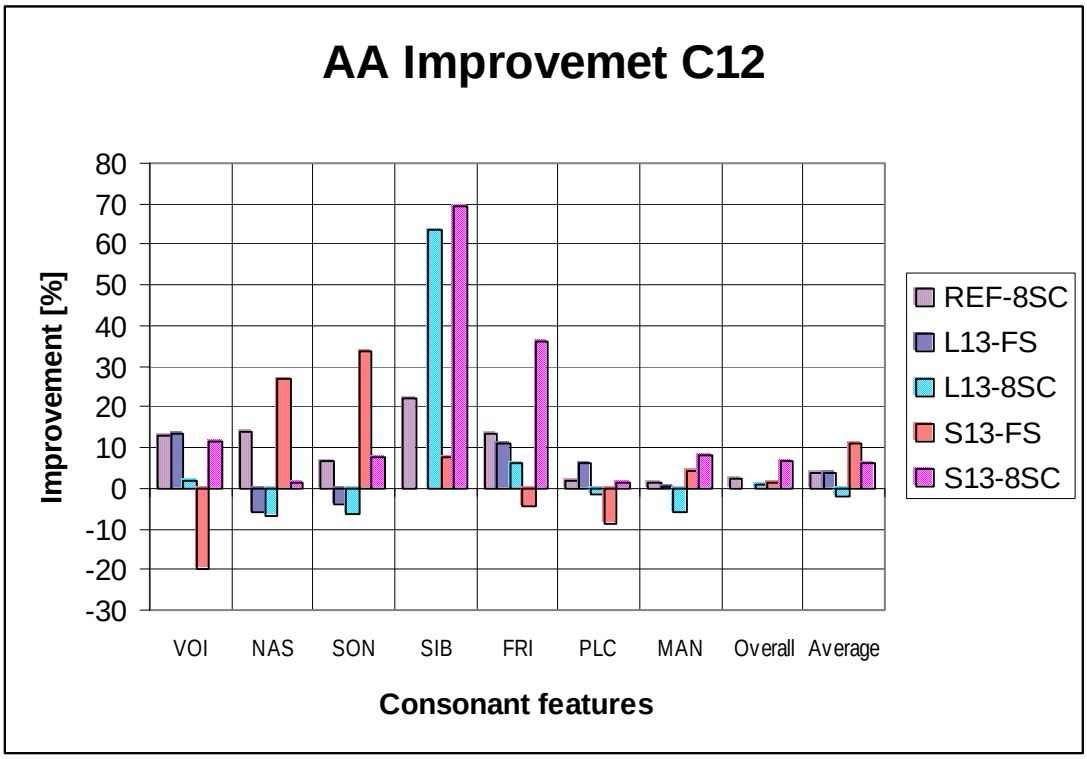


b)

Fig. 7.15 a) German C12 consonant feature recognition scores for subject "VL" b) Absolute average improvement scores against reference.

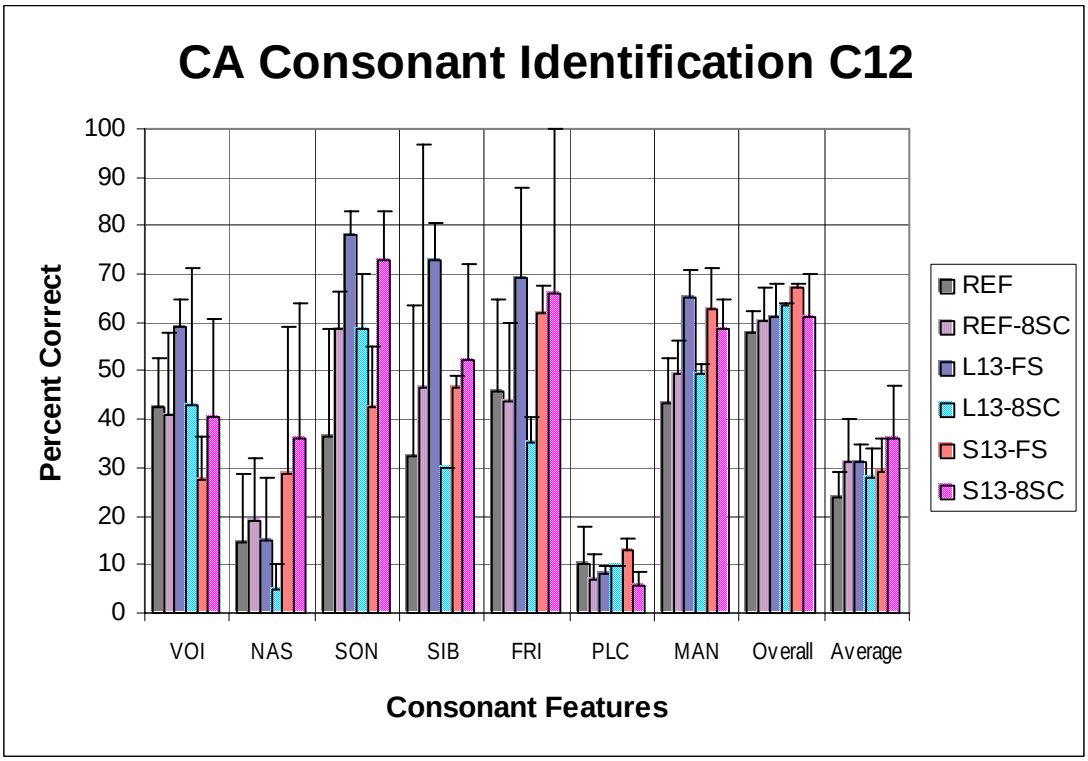


a)

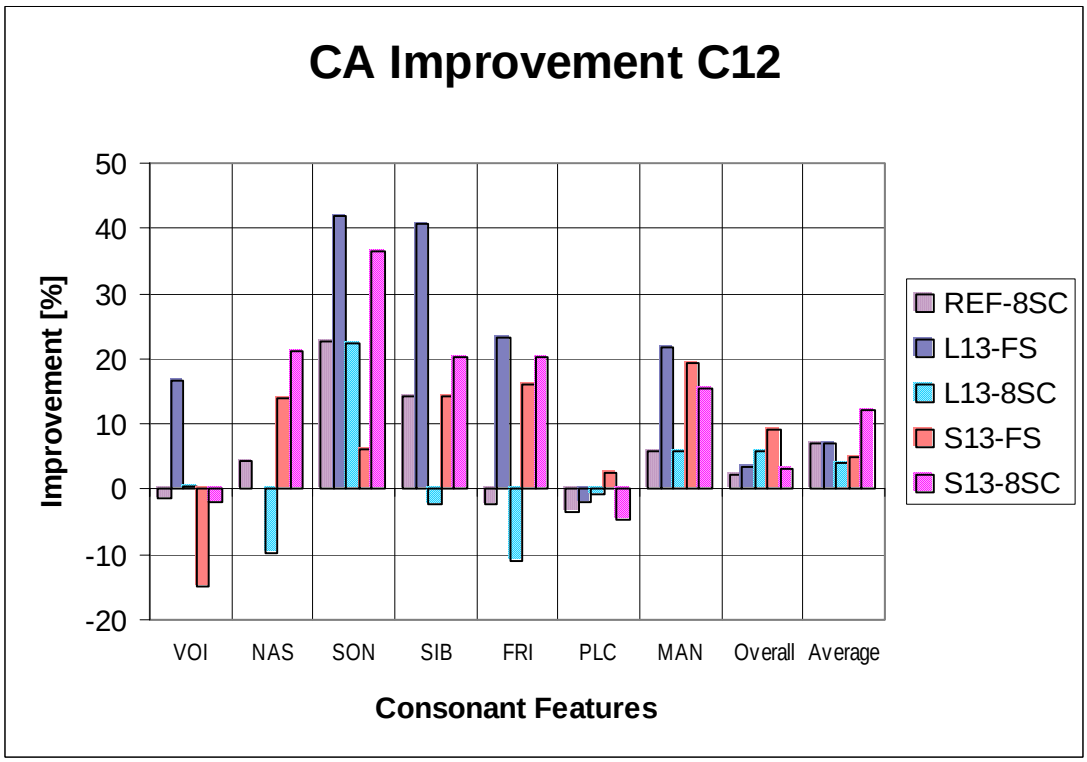


b)

Fig. 7.16 a) German C12 consonant feature recognition scores for subject "AA" b) Absolute average improvement scores against reference.

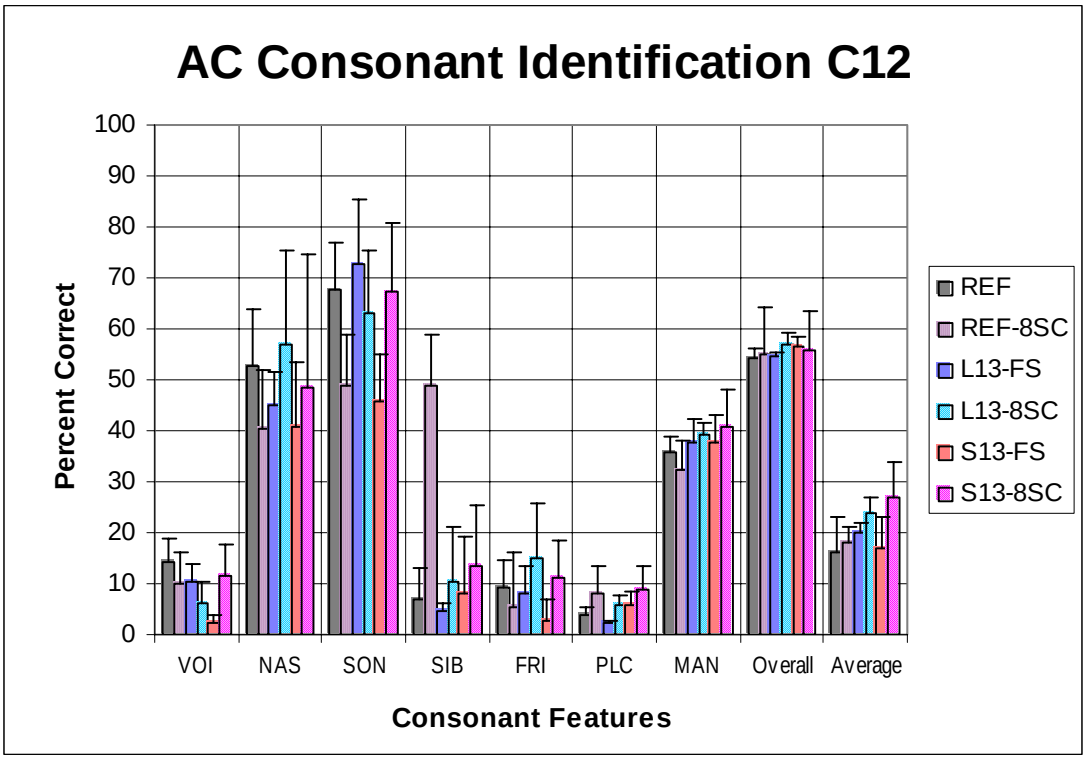


a)

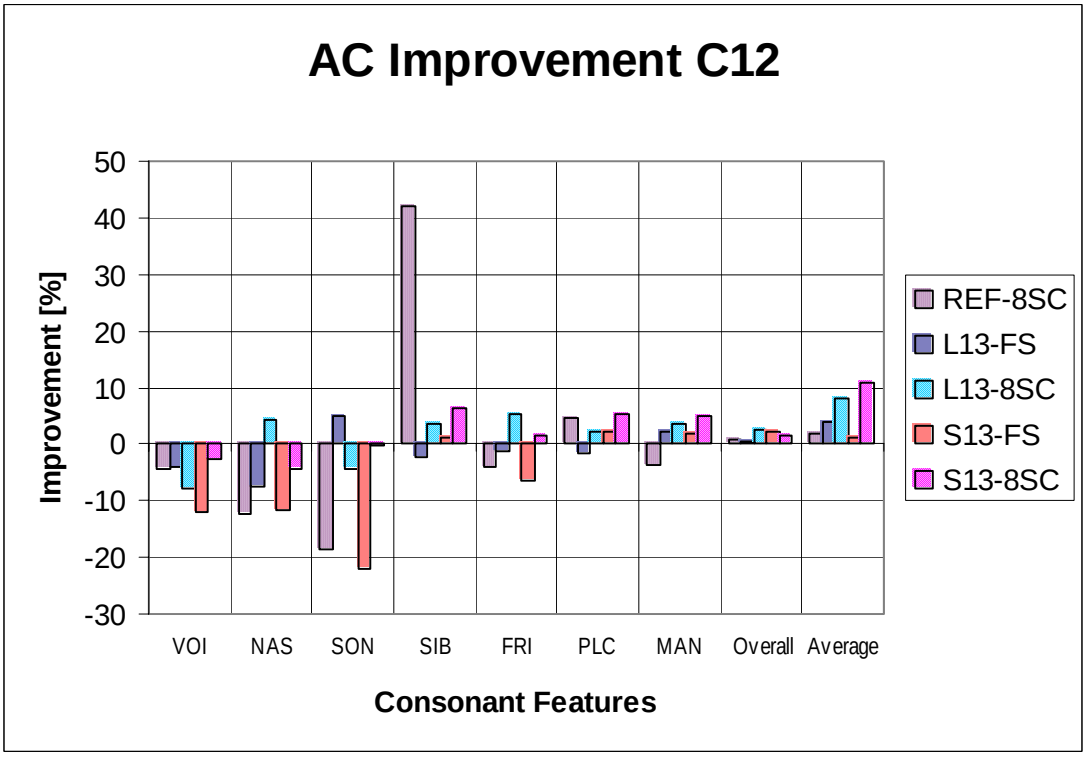


b)

Fig. 7.17 a) German C12 consonant feature recognition scores for subject "CA" b) Absolute average improvement scores against reference.



a)



b)

Fig. 7.18 a) German C12 consonant feature recognition scores for subject "AC" b) Absolute average improvement scores against reference.

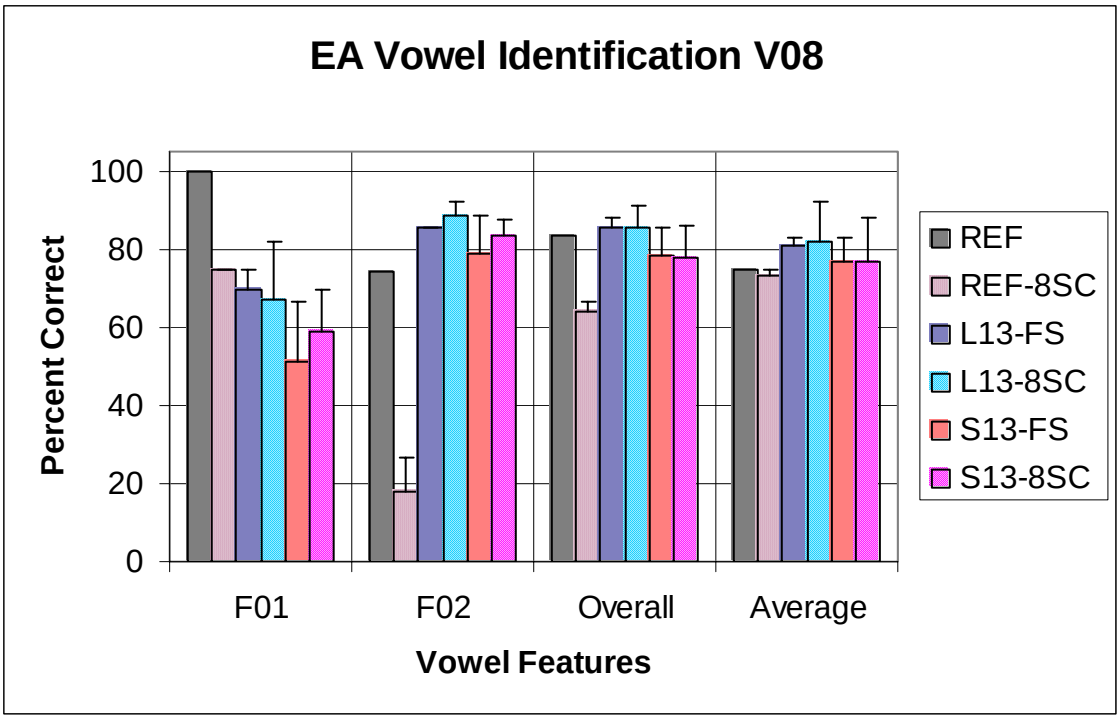
Subject "HB" indicated improvement for the sibilance and frication feature identification scores (Fig. 7.14). However, for the nasality feature identification, scores indicated a considerable decrease (~75%). Interestingly, the best scores for the sibilance feature as well as the worst scores for the nasality feature were observed under the conditions of spectrally reduced signal only, or spectrally reduced and SPINC spectrally compressed signal. In general, the spectrally reduced signal processing schemes indicated worse identification scores in 14 out of 21 cases (67%). Linear spectral compression on the FFT scale was better than the linear spectral compression on the SPINC scale in 6 out of 14 cases (43%).

Results from subject "VL" indicated a little improvement on the sibilance feature identification (~10%) for the full spectrum spectral compression on the SPINC scale (Fig. 7.15). The nasality feature was the one with the largest observed identification score decrease. The spectrally reduced signal processing schemes gave worse feature identification scores in 15 (71%) out of 21 cases. The linear spectral compression on the FFT scale indicated better consonant identification scores in 8 (57%) out of 14 cases.

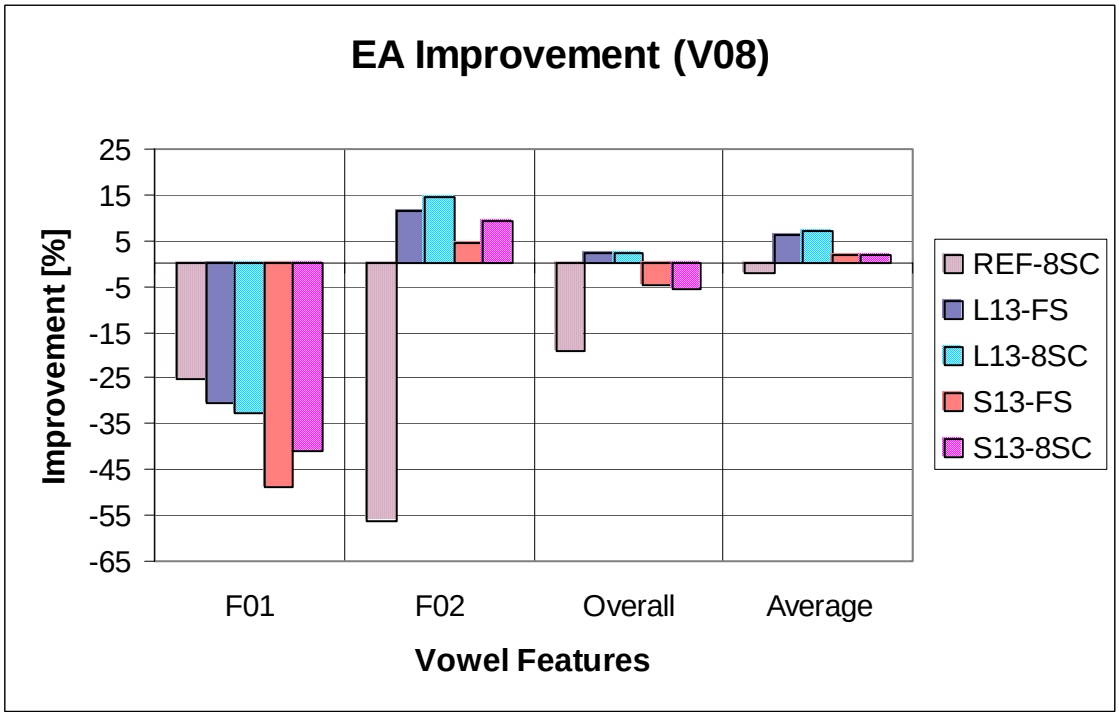
Subject "AA" showed an improvement in almost all the consonant feature identification scores. The only significant absolute decrease (20%) was observed for the voicing consonant feature, when the full spectrum linear spectral compression on the SPINC scale was used. The largest improvements in the consonant feature identification scores were observed for sibilance (~70%), frication (~35%), sonorance (~32%) and nasality (~28%). The spectrally reduced signal processing schemes indicated better results than the non-reduced schemes in 13 (62%) out of 21 cases. The linear spectral compression on the SPINC scale gave better results than the linear spectral compression on the FFT scale in 11 (79%) out of 14 cases.

Subject "CA" showed increasing identification scores of almost all consonant features. The only significant exception was voicing, where on the SPINC scale spectrally compressed processing gave a decrease of 15%. The place feature showed neither a significant improvement nor a decrease of the identification scores. Subject "CA" indicated the largest improvements in the consonant feature identification when the linear spectral compression on the FFT scales was used. The spectrally reduced signal processing schemes indicated better scores in 10 (48%) out of 21 cases. The linear spectral compression on the FFT scale was better than the spectral compression on the SPINC scale in 7 (50%) out of 14 cases.

Subject "AC" showed a significant decrease of voicing, nasality, and the sonorance consonant feature identification scores and a significant improvement in the sibilance identification. Changes in the frication, the place, and the manner feature identification were rather insignificant. The spectral reduction schemes gave better results in 14 (67%) out of 21 cases. Results from the linear spectral compression on the SPINC scale proved better than those of the linear spectral compression on the FFT scale in 7 (50%) out of 14 cases.

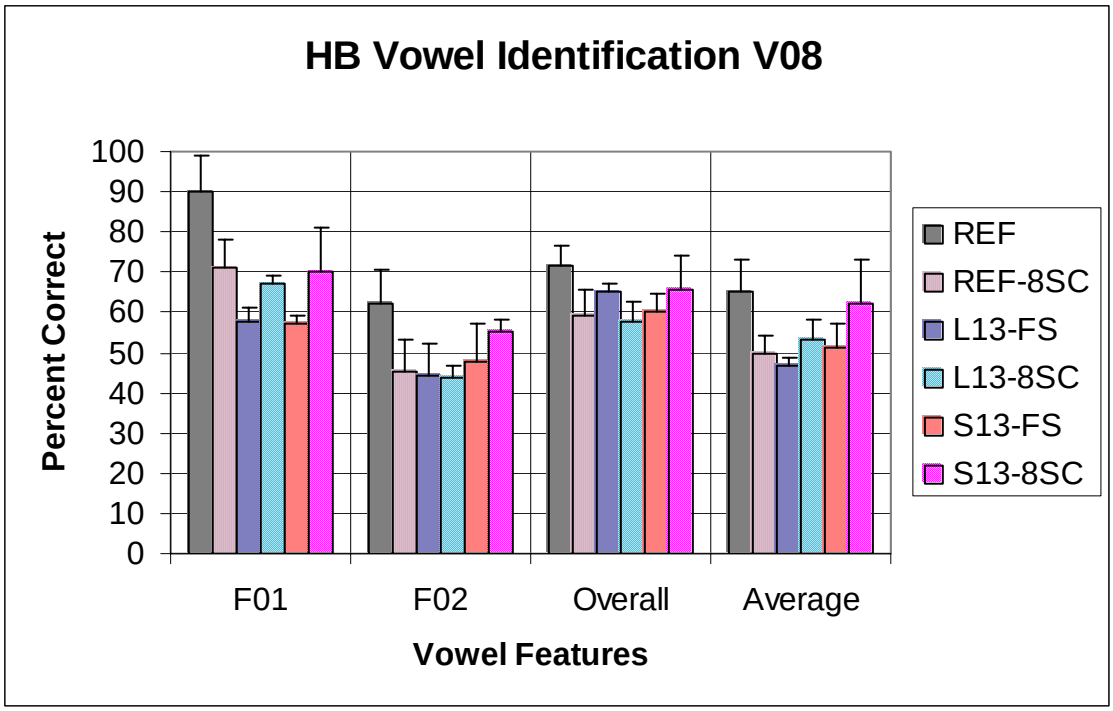


a)

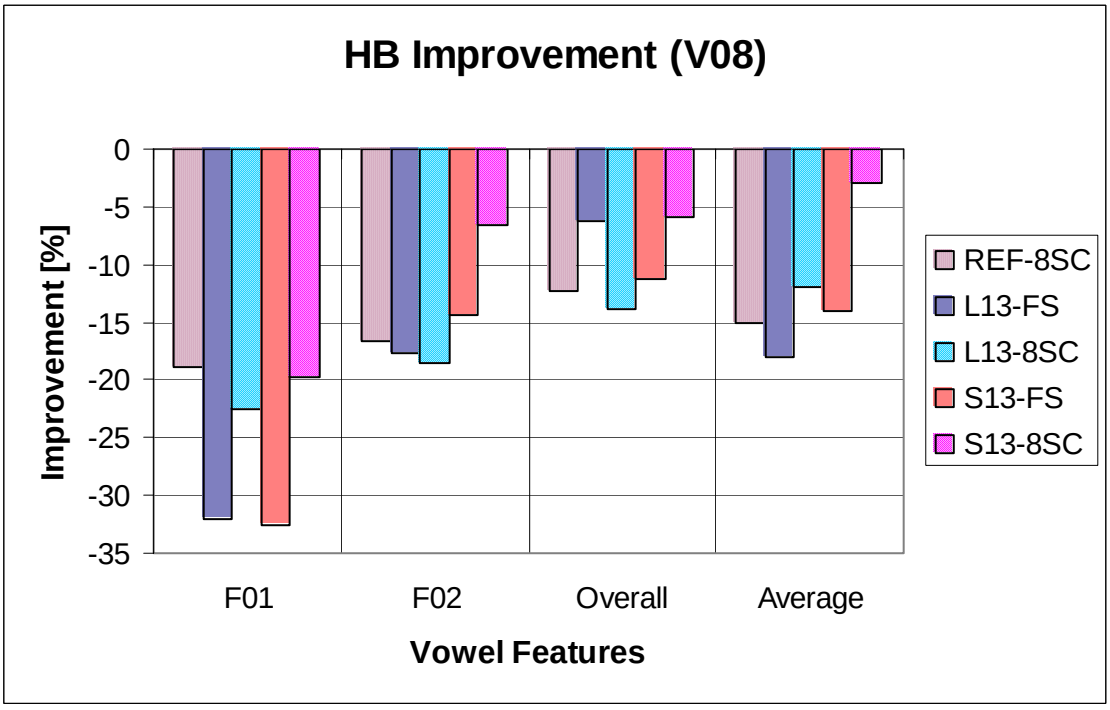


b)

Fig. 7.19 a) German Vo8 vowel feature recognition score values for subject "EA" b) Absolute average improvement scores against reference.

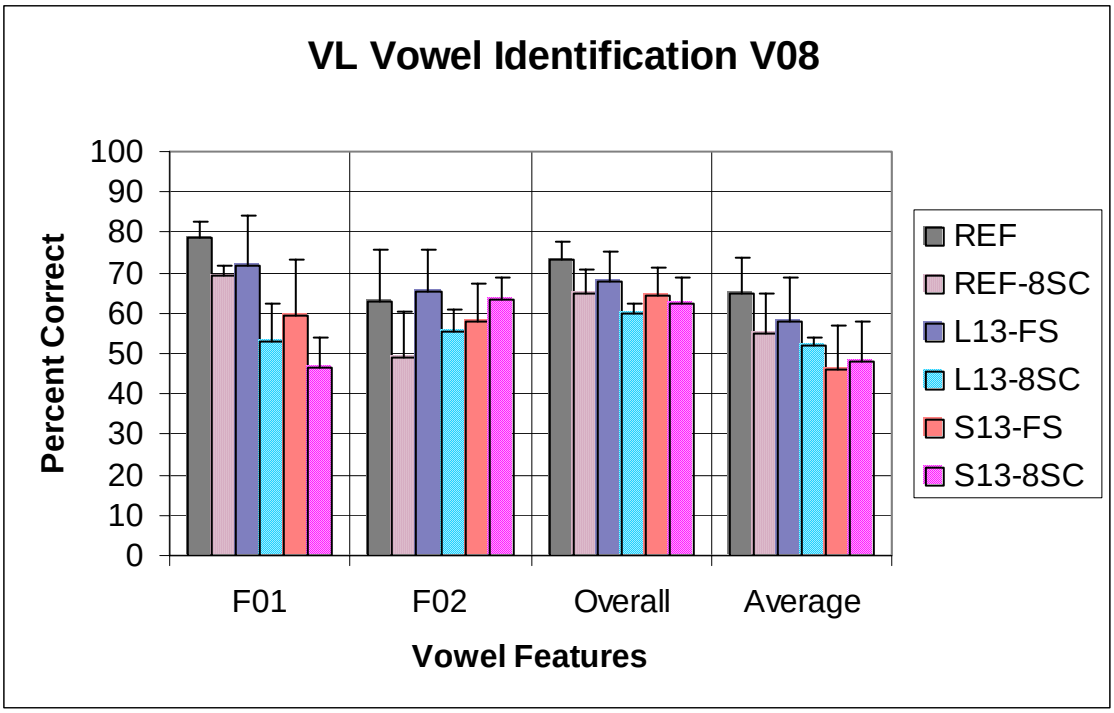


a)

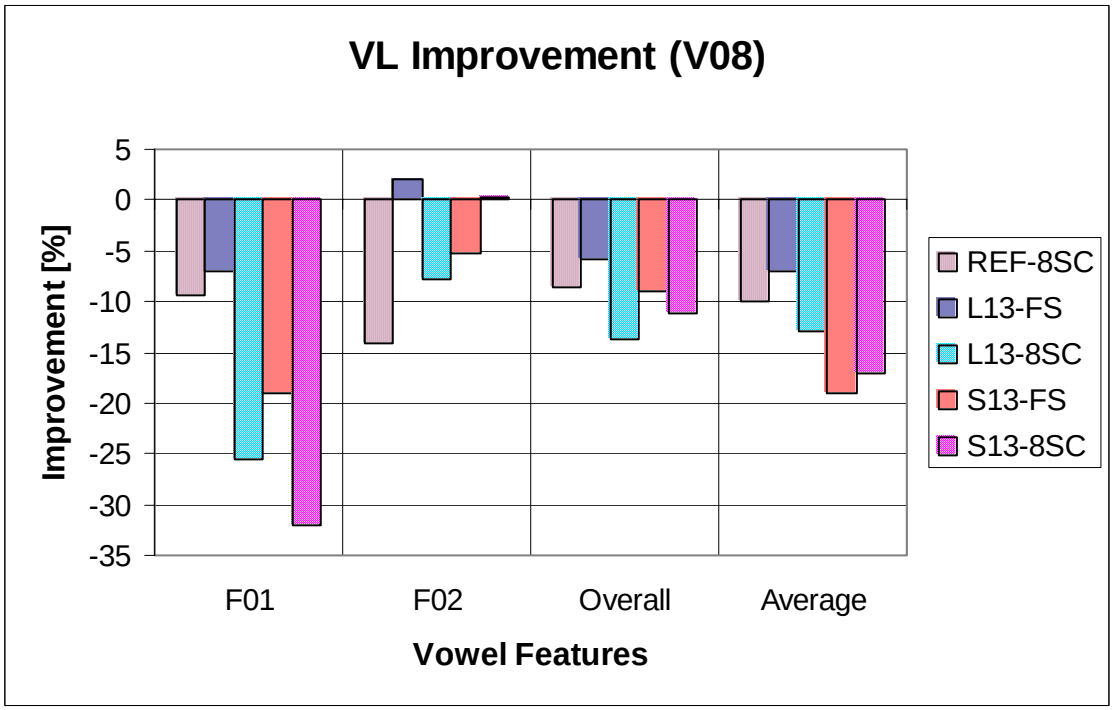


b)

Fig. 7.20 a) German Vo8 vowel feature recognition score values for subject "HB" b) Absolute average improvement scores against reference.

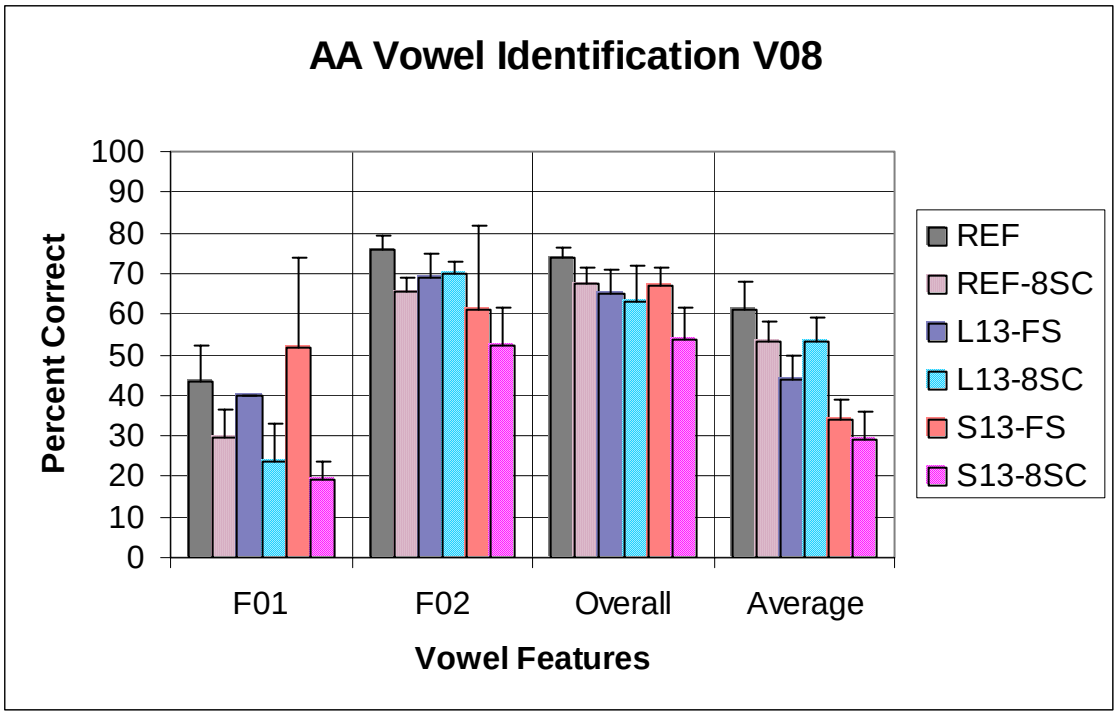


a)

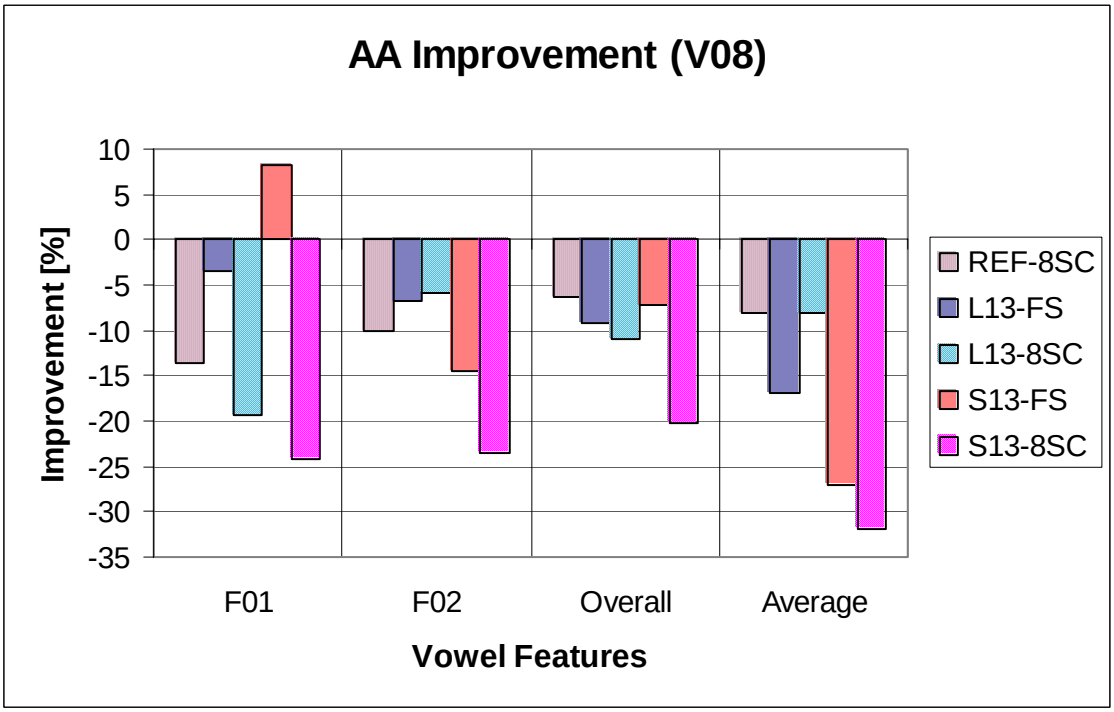


b)

Fig. 7.21 a) German Vo8 vowel feature recognition score values for subject "VL" b) Absolute average improvement scores against reference.

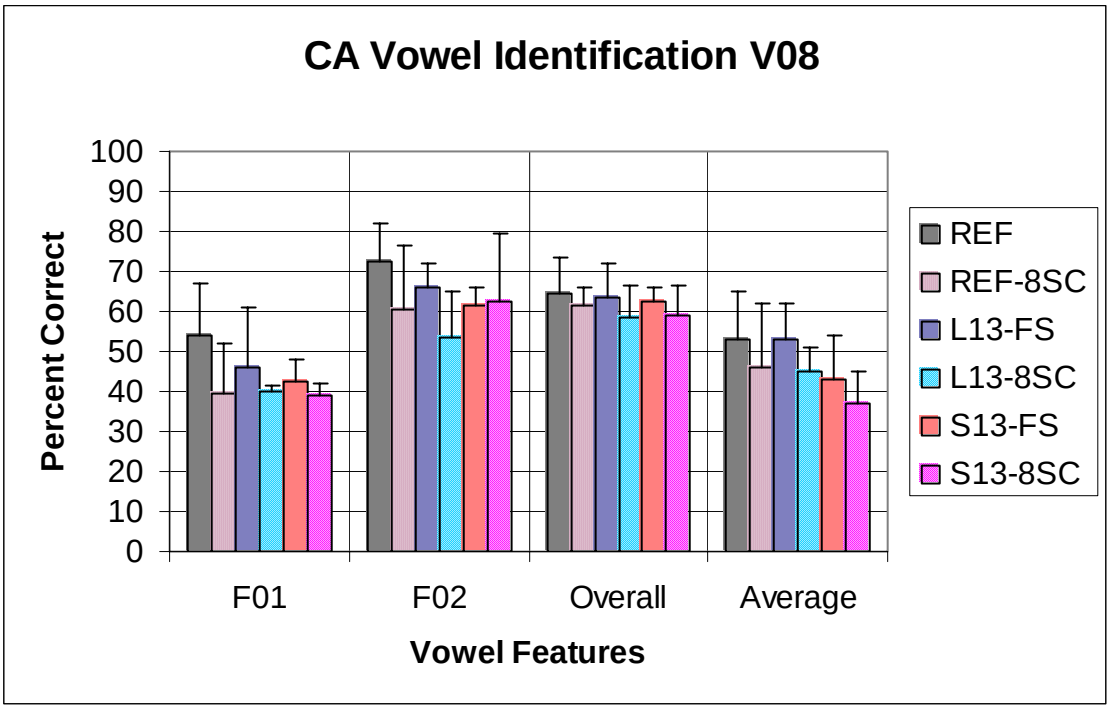


a)

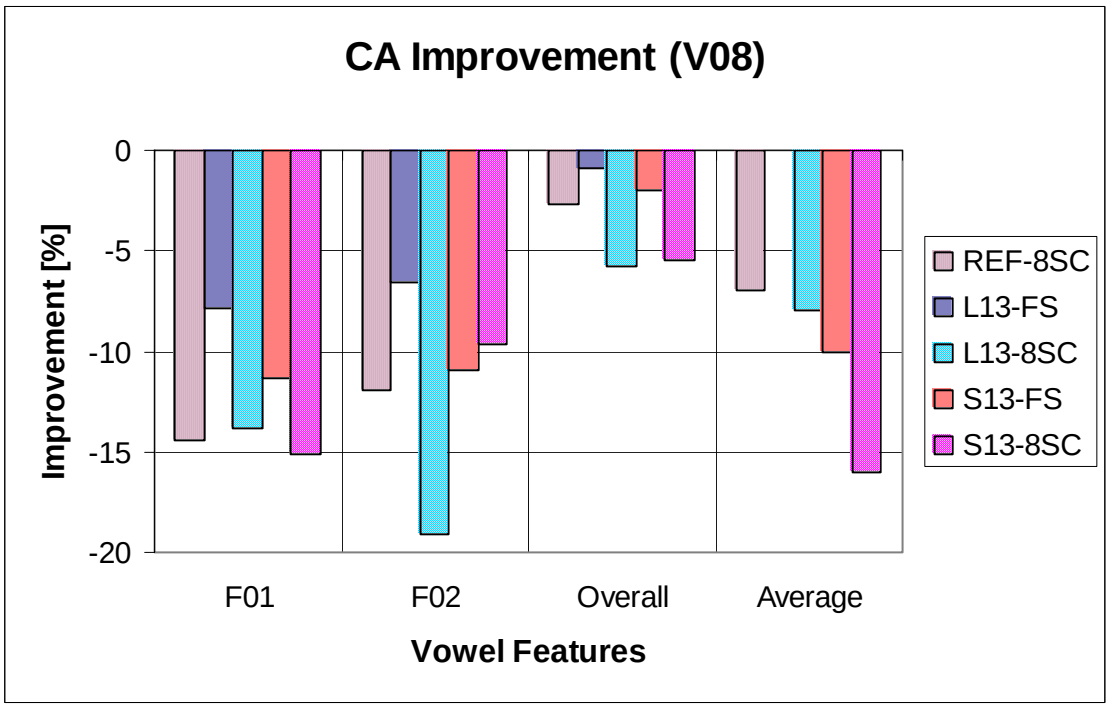


b)

Fig. 7.22 a) German Vo8 vowel feature recognition score values for subject "AA" b) Absolute average improvement scores against reference.

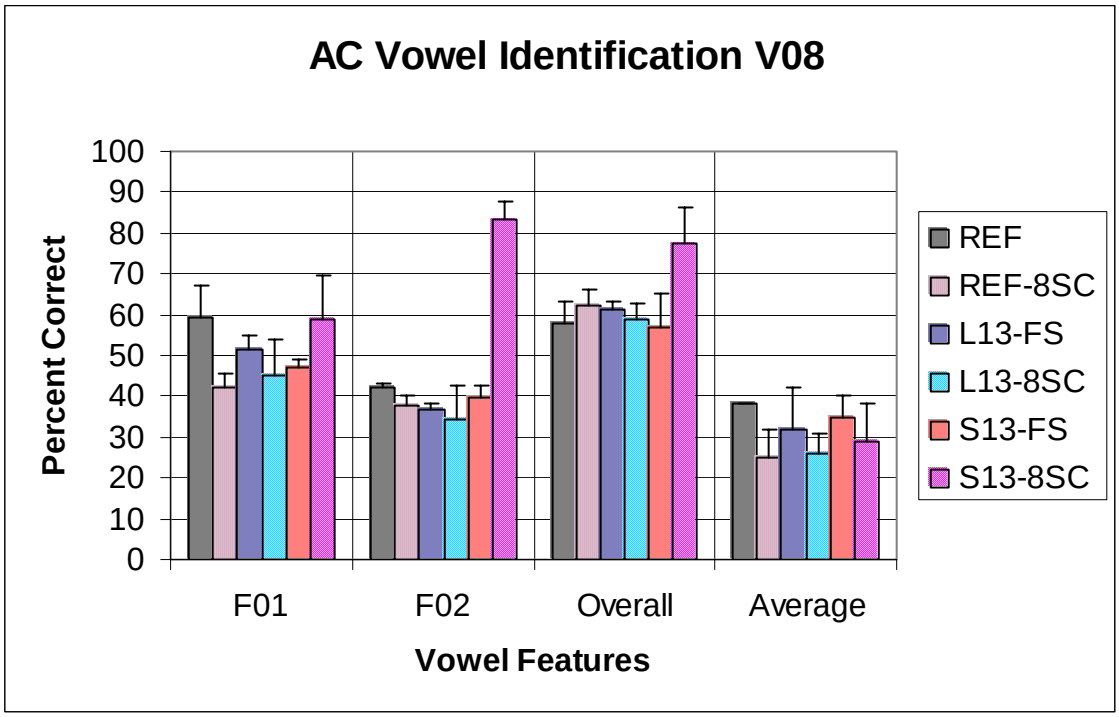


a)

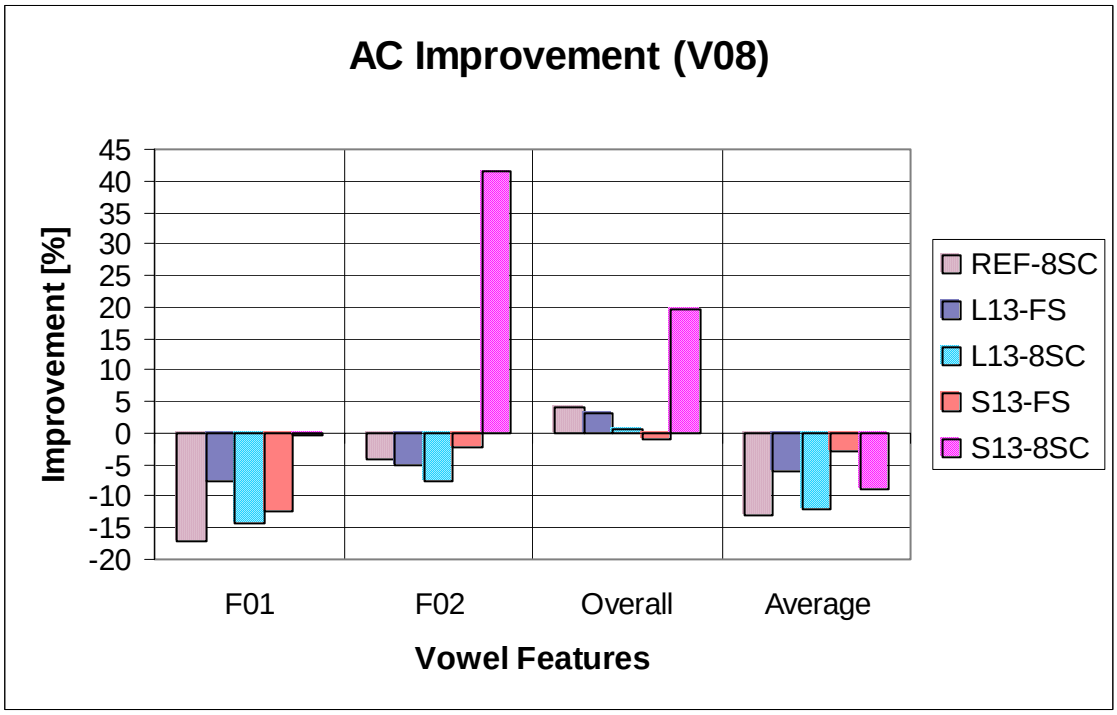


b)

Fig. 7.23 a) German Vo8 vowel feature recognition score values for subject "CA" b) Absolute average improvement scores against reference.



a)



b)

Fig. 7.24 a) German Vo8 vowel feature recognition score values for subject "AC" b) Absolute average improvement scores against reference.

Overall three subjects indicated decreases in voicing, nasality, sonorance, and manner consonant feature identification, when using the spectrally compressed signal. Voicing identification decreased for almost all of the six tested subjects. The sibilance and the frication features indicated an improvement in identification scores for all the tested subjects.

In the analysis of the vowel features, subject "EA" showed 100% identification of the first formant feature with reference processing. All signal processing schemes of the present study resulted in a decrease of the first formant feature identification scores. For the second formant both the linear spectral compression on the FFT scale and the spectral compression on the SPINC scale indicated a significant improvement (5-15%). The signal processing schemes with spectral reduction gave better results than the schemes without spectral reduction in 3 (50%) out of 6 cases (2 vowel feature groups x 3 signal processing schemes). Results for the linear spectral compression on the FFT scale were better than the linear spectral compression on the SPINC scale in 4 out of 4 cases (100%) (2 vowel feature groups x 2 signal processing schemes). For the reference signal, subject "EA" showed better identification of the first formant than of the second formant. The average vowel identification score value was close to the identification score of the second formant (~75%).

Subject "HB" showed a decrease in both formant identifications for all of the tested signal processing schemes. The subject also indicated better first formant than second formant identification (for all processing schemes). Again, the average vowel identification was close to the second formant identification. The signal processing schemes without spectral reduction showed better results than the schemes with spectral reduction in 3 (50%) out of 6 cases. The linear spectral compression scores on the SPINC scale were better than the linear spectral compression scores on the FFT scale in 3 out of 4 cases (75%).

Subject "VL" did achieve very slight improvement of the second formant vowel feature identification using full spectrum spectral compression on the FFT scale. For reference processing the subject also indicated better first formant than second formant identification. The average vowel identification was also close to the second formant identification (~65%). The signal processing schemes without spectral reduction showed better results than those without spectral reduction in 5 (83%) out of 6 cases. The linear spectral compression on the FFT scale was better than the linear spectral compression on the SPINC scale in 3 (75%) out of 4 cases.

In contrast to the previously described subjects, subject "AA" was significantly better in the identification of the second than the first formant feature. The subject also indicated an improvement in the first formant identification when the full spectrum spectral compression on the SPINC scale was used. The spectrally reduced signal processing schemes gave better scores only in 1 (17%) out of 6 cases. The linear spectral compression on the FFT scale was better than the linear spectral compression on the SPINC scale in 3 (75%) out of 4 cases. The average vowel identification score for subject "AA" (~60%) was close to the mean value of the first formant feature group identification (~45%) and the second formant identification score (~75%).

Subject "CA" was also better in the second formant feature identification, and indicated the same identification score relation as subject "AA" for the first and the second formant. However, the subject did not profit from any spectral compression scheme to improve the vowel feature identification. The signal processing schemes without spectral reduction were better than the ones with spectral reduction in 5 (83%) out of 6 cases. The linear spectral

compression on the FFT scale was better than the linear spectral compression on the SPINC scale in 3 (75%) out of 4 cases. In contrast to subject "AA", subject "CA" indicated an average vowel identification score (~54%) which was very close to the first formant identification score (~53%).

Subject "AC" showed better first formant than second formant identification. In contrast to subjects "EA", "HB", and "VL", his first formant identification score was significantly smaller (60%). Using the reduced spectrum spectral compression on the SPINC scale, subject "AC" indicated an insignificant decrease in first formant identification score (1%). With the same processing, the subject however profited significantly in the second formant vowel feature identification (+42%). This strategy unfortunately did not improve his average vowel identification. The spectrally reduced signal processing schemes indicated better scores only in 2 (33%) out of 6 cases. The linear spectral compression on the SPINC scale was better than the linear spectral compression on the FFT scale in 3 (75%) out of 4 cases. The average vowel identification score for the reference signal (~38%) was close to the second formant vowel feature group identification score (~42%).

Overall the signal processing schemes indicated an improvement of the second formant vowel feature identification score for three subjects out of six. An improvement of the first formant was observed only in 1 (17%) case out of 6.

7.5 Summary and conclusions

Spectral compression on the SPINC scale and on the FFT scale showed improvement in consonant identification in comparison with reference processing in 3 (50%) out of 6 cases. The absolute improvement using the spectral compression on the SPINC scale was larger than 10%. Using linear compression on the FFT scale, consonant recognition score improvement was not higher than 8%. This result shows that as expected, spectral compression can improve consonant recognition. However, this was not observed for all of the tested hearing-impaired subjects. A consonant recognition improvement could be observed only for the subjects from the second and the third subject groups. These groups consists of subjects with more than 80 dB hearing loss for the frequencies higher than 500 Hz. A surprising exception was only subject "AA" of the second subject group, who has a moderately severe pure tone average audiogram characteristics. The subjects with hearing loss levels smaller than 90 dB for the frequencies lower than 750 Hz did not profit from the spectral compression (first subject group). However, their consonant identification score values with the non-processed (reference) signal were already between 50-80%. The subjects from group III profited more from the spectrally reduced and spectrally compressed processing on the SPINC scale. This is an indication that the additional spectral information from the higher frequency areas, submitted by the spectral compressions on the auditory frequency scales, can lead to an improvement in the consonant identification for profoundly hearing-impaired subjects. The spectral reduction for the SPINC compressed signal was constructive and it improved the consonant recognition scores by 5-8% for the subjects from the third subject group. For all other subjects the spectral reduction in combination with the spectral compression for the consonant identification was rather destructive. For subject "CA" the spectral compression

on the SPINC scale resulted in even better consonant identification than in the case when he was using the cochlear implant[8]. The full spectrum linear compression on the FFT scale did not significantly decrease consonant recognition for the subjects from the first subject group and for subject "VL" from the second group.

The signal-processing schemes which employ spectral compression improved the sibilance consonant feature recognition for all subjects. However, the improvement for this feature only did not result in an increased average consonant identification score. The subjects who indicated an average consonant score improvement did also show an improvement in the nasality, the sonorance, and the manner feature identification. The subjects who did not profit from the spectral compression for consonant recognition indicated a decrease of the identification scores of the voicing, the nasality, the frication and the place feature. The reason for the decrease of identification for these features is possibly the narrowing and displacement of the formant structures in the consonant spectra. The linear spectral compression on the FFT scale caused a relatively small decrease or sometimes even a considerable improvement of these consonant feature identification scores for the subjects from all the three groups.

As expected, the spectral compression did reduce the vowel identification scores of the hearing impaired subjects (in 5 (83%) out of 6 cases). For almost all of the tested subjects it caused a decrease of the first formant identification scores. The reason for that could be that almost all tested subjects were capable to perceive the first formant in the reference signal. Four of the six subjects indicated better first than second formant identification. The reason for the better second formant identification for subject "CA" might be its smaller hearing loss for frequencies above 1.5 kHz. The reasons for such behaviour of subject "AA" are unclear. In 3 (50%) out of 6 cases spectral compression did show improvement in the second formant identification. This improvement was possibly caused by the better audibility of the second formant caused by shifting its frequency range into the residual hearing frequency areas. The improvement of individual feature identification does, however, not necessarily lead to an improvement of the average vowel identification scores, even if it is a considerable improvement, like in the case of subject "AC" (see Fig. 7.24). As the second formant is dominant for speech intelligibility [B47], [B129] it can be expected that after additional training, vowel identification score improvement can be achieved for subjects who indicated improved scores of the second formant identification.

A benefit from using spectral compression without using of spectral reduction (Fig. 7.10) for sentence recognition in quiet could be shown for the third subject group. Both subjects of group III showed a good improvement for linear compression on the SPINC scale. Subject "AA" from the second subject group indicated a good improvement in sentence recognition when the linear spectral compression on the FFT scale was used. This result correlates with the improved consonant recognition scores for the spectral compression on the SPINC and on the FFT scales. It may indicate that for subjects with very poor sentence recognition the speech perception is mostly correlated with the consonant recognition score values. On the other hand, subject "EA" from the first subject group indicated a small improvement in sentence recognition which correlated with the improved vowel identification

[8] Note that "CA" has only started to use the CI (the amazing fact is that this subject is extraordinarily good in lip-reading and plays in a symphonic orchestra).

score values. The two other subjects from the first and the second group, besides decrease of the IST sentence recognition, indicated also a decrease in both the average vowel and the consonant identification.

In general it can be stated that an improvement of speech perception when applying spectral compression can be expected for profoundly sensorineural hearing impaired subjects with moderately severe to severe hearing loss for frequencies below 500 Hz and very poor sentence recognition in quiet environment. This conclusion is in agreement with the experiance of McDermott mentioned in the introduction.

Subjects with profound steeply sloping high frequency hearing loss and with normal hearing or only slight hearing loss in the frequency area below 500 Hz cannot significantly benefit from the spectral compression. To verify this hypothesis, additional studies with profound hearing-impaired subjects have to be performed. Nevertheless, the linear spectral compression on the FFT scale with small spectral compression ratios can possibly improve vowel identification and after longer acclimatization even speech perception in sentences.

At this point, only speculations can be made about potential learning improvements of the speech perception using spectral compression. To answer this question, long-time studies with wearable hearing systems are required. The finding that the spectral compression on the SPINC scale with CR=1.3 did not indicate better results than linear spectral compression on the FFT scale for subjects of the first subject group does not mean that improvements with smaller compression ratios are not possible. In addition, it should be mentioned that no special fitting for spectral compression was performed. As spectral compression fitting methods are still not developed. One possible solution could be a stepwise increase of the spectral compression ratio during a longer period of acclimatization.

Spectral reduction of the spectrally compressed signal indicated better results only for consonant identification when spectral compression on the SPINC scale was used. The reason for this is possibly that the spectral compression on the SPINC scale affects the higher spectral regions more than lower ones. So for high spectral regions, spectral reduction was meaningful because it additionally improved consonant recognition scores. For the lower frequency areas, however, spectral reduction was rather destructive. This was observed by the decrease of the vowel and the IST sentence recognition for the spectrally reduced only signal. This result proposes a suggestion for the development of a signal processing scheme, which would perform spectral reduction only for higher frequency areas or only for consonants respectively noise like speech segments.

In this study a subject classification according to their IST sentence recognition was performed. This classification correlated with the subject hearing loss parameters especially in the low frequency areas. It is possible that this condition on the low frequency pure tone average (average hearing loss in dB for 125, 250 and 500 Hz), besides the profound sensorineural high frequency hearing loss, can be used as a criterion for subject selection for spectral compression. However, the number of tested subjects in the present study is too small to build up sufficient group statistics.

Chapter 8

Three months later

"The perplexity augments instead of diminishing. I sleep but little. It has ceased from lying around, and goes about on its four legs now."

M. Twain

Experiments with temporal modification of different speech segments

8.1 Overview

In this study, three preliminary tests of temporal modifications on the speech signal were performed with a limited number of normal-hearing adults in order to investigate the possibilities of employing temporal modifications in signal processing schemes for profound haring-impaired subjects. The first test was performed with one normal hearing adult in order to find approximate limitations for the temporal expansion and compression factors. The second test was performed with four normal hearing adults using the temporal modification factors determined during the first test. The different temporal modification factors were applied on different speech segments in order to investigate the most promising time modification schemes. The overall duration of the processed speech signal in the first and the second test was significantly longer than the overall duration of the original speech signal. In the third test which was performed on five normal hearing German speaking adults bidirectional temporal modifications of different speech segments were investigated in order to preserve the approximate overall duration of the original input in the processed output speech signals

8.2 Signal processing and parameter settings

Signal processing for the tests described in this chapter was applied via the use of the sinusoidal speech processing toolbox which allows temporal modifications of the speech signal without changing speech timbre (see chapter 4). The applied spectral processing parameters of the sinusoidal speech system were: analysis frame length of 1024 sample (46 ms), Hanning analysis window, and use of 64 spectral components (in the first experiment) respectively 512 spectral components (in the second and third experiment) in the signal reconstruction selected using the maximum peak picking criterion. The sampling frequency of the processed speech signals was 22050 Hz.

Different temporal prolongation and compression factors could be applied either to the whole signal or to particular speech segments, notably voiced and unvoiced segments.

The voiced/unvoiced speech segments were detected according to their spectral center of gravity (CG) (see equation 4.38). All speech segments with a center of gravity value smaller than 2.2 kHz were assumed to be voiced. Otherwise (CG≥2.2 kHz), they were assumed to be unvoiced.

8.3 Preliminary test of temporal modification factor intensity (Experiment I)

8.3.1 Motivation

In order to investigate the critical temporal modification factor intensity which does not decrease speech perception, the following constant temporal modification factors were applied to the acoustic signal: $\rho = 0.3$, 0.4, 0.5, 0.6, 1.0, 3.0, 4.0, 5.0, and 7.0. The overall duration of the processed speech signal after application of the time modification operation was equal to the original signal duration multiplied by the temporal modification factor.

8.3.2 Performed speech tests

The German C12 "a-C-a" consonant and V08 "d-V" vowel test materials were processed using the afore-mentioned temporal modification factors including three articulations of twelve consonants and eight vowels respectively. The logatome syllables were presented at 65 dB RMS in a sound proof room at a distance of 1.5 m from a Philips type 22AH586/16R active loudspeaker (playback was performed by a 16 bit PC sound card).

This test was carried out with one normal hearing German speaking adult.

8.3.3 Results

The vowel and consonant identification scores for the different time modification factors are given in Fig. 8.1. The identification scores of the prolonged logatome syllables do not show any decrease compared to the unprocessed signal ($\rho = 1.0$). Decrease of vowel and consonant identification occurs when temporal compression ($\rho<1.0$) is applied. The consonants seem to be more affected by large temporal compressions than the vowels. However, a small decrease in the shortened vowel identification occurs at a temporal modification factor of 0.6, where the consonant identification remains still unaffected. Using temporal modification factor values of ≤0.3, consonant identification decreases significantly (~60%).

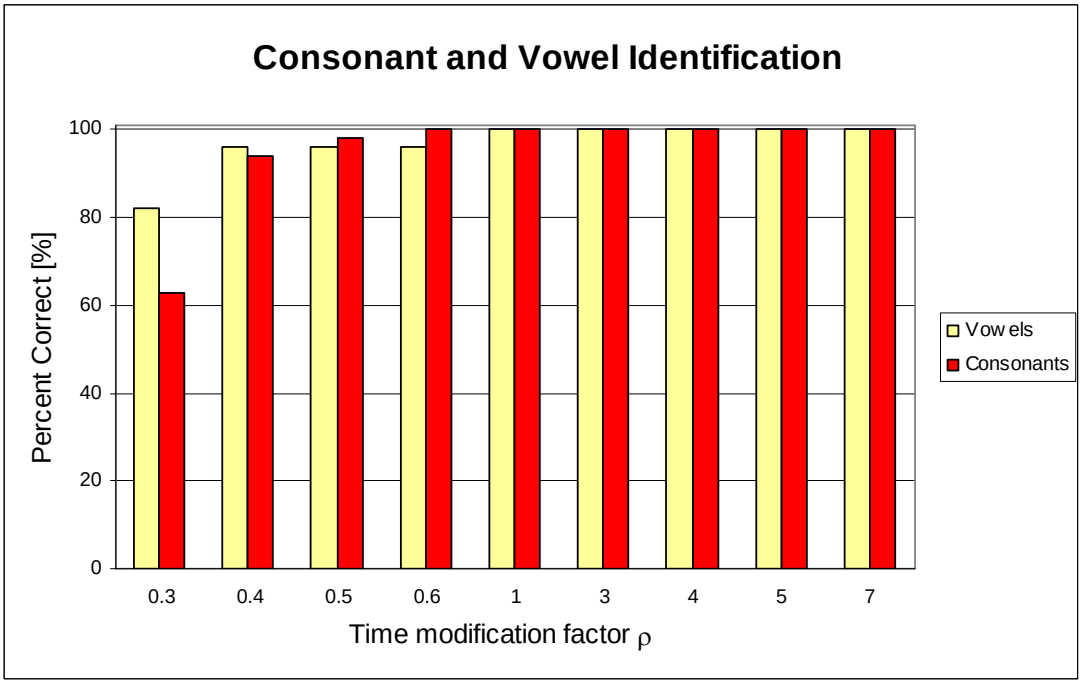


Fig. 8.1 Consonant and vowel identification scores of the temporally modified speech signal using materials from the German C12 consonant and V08 vowel logatome tests.

8.3.4 Discussion and conclusions

The fact that this study was performed with only one well-trained listener must be taken into account. This means that the estimated values of the temporal modification factors which can be applied on the consonant and vowels have to be considered only as approximation, thus providing a guideline for the following two studies.

Another factor which has to be considered is that the speaker rate of the different speech tests used in this study and in the studies mentioned in the literature (see chapter 3) can differ from each other. The temporal modification factor values have to be adjusted according to the speaker rate and are dependent on a particular speech material. As there is no developed technique for the measurement of the speaker rate speed (which can even differ for different languages and speakers) then the only possibility to avoid misleading results is the usage of the same speech material for all studies to be compared.

Contrary to the results of the present preliminary study, Nagafuchi [B96] observed slightly decreasing speech identification capabilities of young adults observed under increasing temporal expansion of the speech signal.

The largest acceptable temporal expansion observed in the present study can be explained by the tested person's listening experience on the temporally modified signals and the fact that the test consists of relatively short consonant or vowel logatomes. Also, there are no suprasegmental cues present in the test material. The situation for longer speech segments such as words and sentences is different. From the authors' own experience, recognition of sentences with a temporal modification factor $\rho = 5.0$ (the overall signal duration is five times longer than the original signal duration) is almost impossible.

As one of the requirements for the potential signal processing scheme is the approximate equality of the input-output signal duration, an important result of the time modification investigation is the estimation of the largest acceptable temporal compressions of any of the speech segments. This would allow a build-up of "temporal reservoirs", which subsequently can be used for the compensation of temporal expansion of other speech segments. If no temporal compression can be applied, then the build-up of such "temporal reservoirs" would be impossible, and therefore the desired temporal modification signal processing scheme under the condition of duration preservation would not be feasible.

Based on the above considerations, a minimal temporal modification factor value of $\rho = 0.5$ and a maximal value of $\rho = 3.0$ were chosen for signal processing in the following studies with temporal modifications of different speech segments.

8.4 Perception of temporally modified speech in sentences (Experiment II)

8.4.1 Motivation

In order to investigate the perception of temporally modified speech in sentences, four normal hearing adults were tested using speech material which was processed with non-uniform temporal modification. The following temporal speech modifications were tested: whole sentence prolongation, voiced segments prolongation, unvoiced segments prolongation, whole sentence shortening, voiced segments shortening, unvoiced segments shortening (see table 8.1). The temporal modification factor for the temporal prolongation was $\rho = 3.0$ (three times longer), and for temporal shortening $\rho = 0.5$ (two times shorter). These values of the time modification factor were chosen according to the results of the study described in the previous section.

Signal Processing Scheme	**Notation**	**Temporal modification factor ρ**
The whole sentence is temporally prolonged	**LL**	ρ=3.0
Only the voiced sentence segments are prolonged	**VL**	ρ=3.0
Only the unvoiced sentence segments are prolonged	**UL**	ρ=3.0
The whole sentence is temporally shortened	**SS**	ρ=0.5
Only the voiced sentence segments are temporally shortened	**VS**	ρ=0.5
Only the unvoiced sentence segments are temporally shortened	**US**	ρ=0.5

Table 8.1 Parameter settings and nomenclature for temporal modifications in experiment II.

8.4.2 Performed speech tests

The Oldenburg sentence test was carried out with the six temporally modified schemes of table 8.1 using random sentence lists. As a reference, the non-processed sentences were tested as well. This test took place in a sound proof room with a 1.5 m distance from a Westra type LAB-1001 active loudspeaker. During testing all listeners were exposed to Oldenburg noise of a constant level of 65 dB. The test attempted to determine, the signal to noise ratio (SNR) at which sentence recognition scores of approximately 50% are achieved. The test procedure involved adaptive variations of the speech signal loudness. The test was applied on four normal hearing adults in multiple test sessions. Three of the tested subjects

(1-3) were experienced in speech test participation and were familiar with the Oldenburg sentence test material. The estimation of the SNR value was performed five times for each subject and each signal processing scheme included the reference signal.

8.4.3 Results

The differences between the unprocessed speech (reference) SNR and the estimated SNR values of different temporal modification schemes ($SNR_{reference}$-$SNR_{processed}$) are shown in Fig. 8.2-8.6. Note that positive SNR differences indicate better understanding with the actual processing scheme, whereas negative SNR differences indicate more difficulties in speech perception.

The whole sentence prolongation (LL) and the unvoiced segment prolongation (UL) achieved lower SNR values than the reference signal (Fig. 8.2). However, in all cases the SNR differences were relatively small (<2.0 dB). Some subjects also showed a lower level of the SNR for the signal processing scheme with expanded voiced segments (VL) (Fig. 8.3 and 8.4).

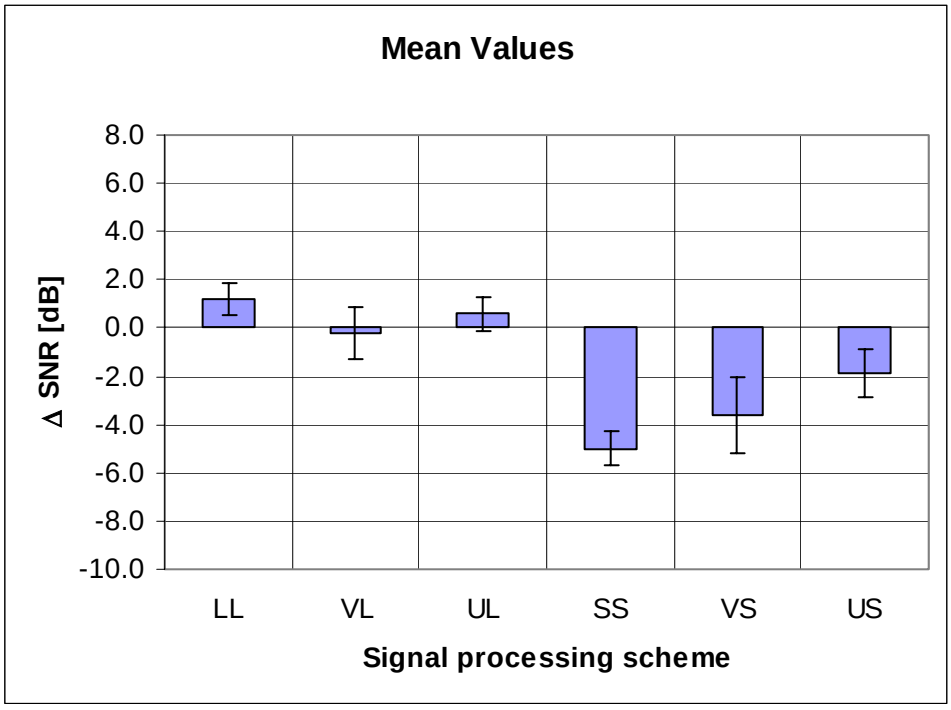


Fig. 8.2 Mean values of SNR differences between the reference signal and the temporally modified signal for 50% sentence recognition.

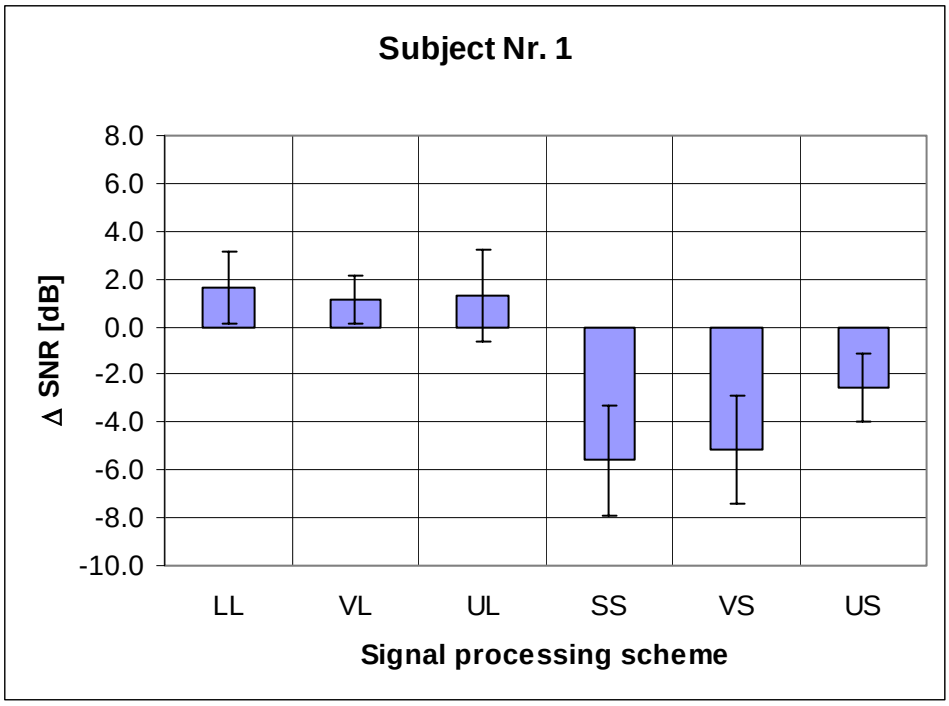


Fig. 8.3 Mean performance of subject Nr 1 achieved for SNR differences between the reference signal and the temporally modified signal for 50% sentence recognition.

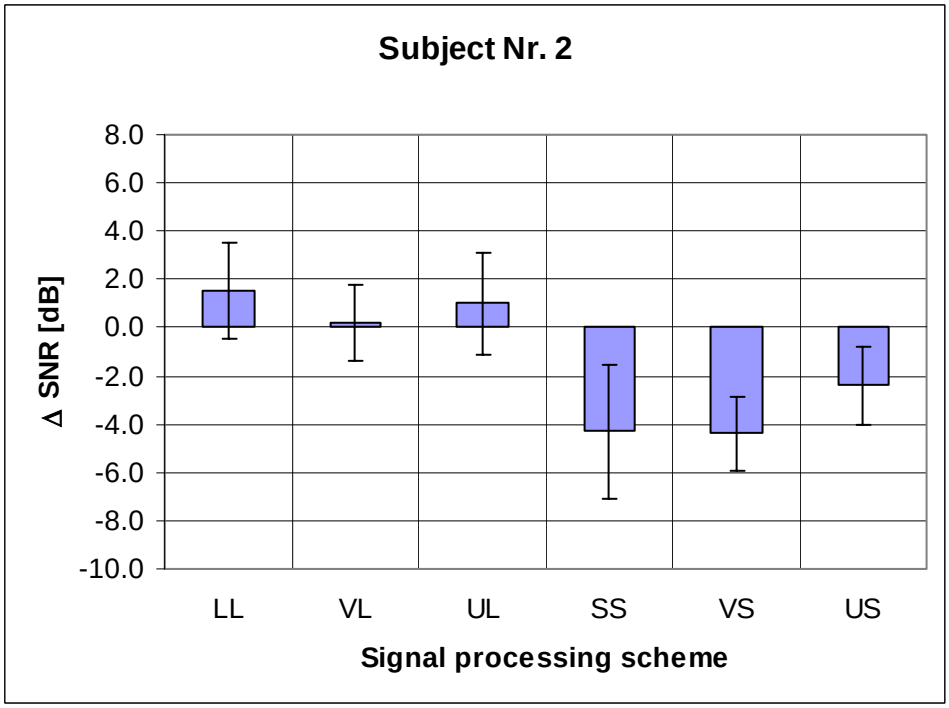


Fig. 8.4 Mean performance of subject Nr 2 achieved for SNR differences between the reference signal and the temporally modified signal for 50% sentence recognition.

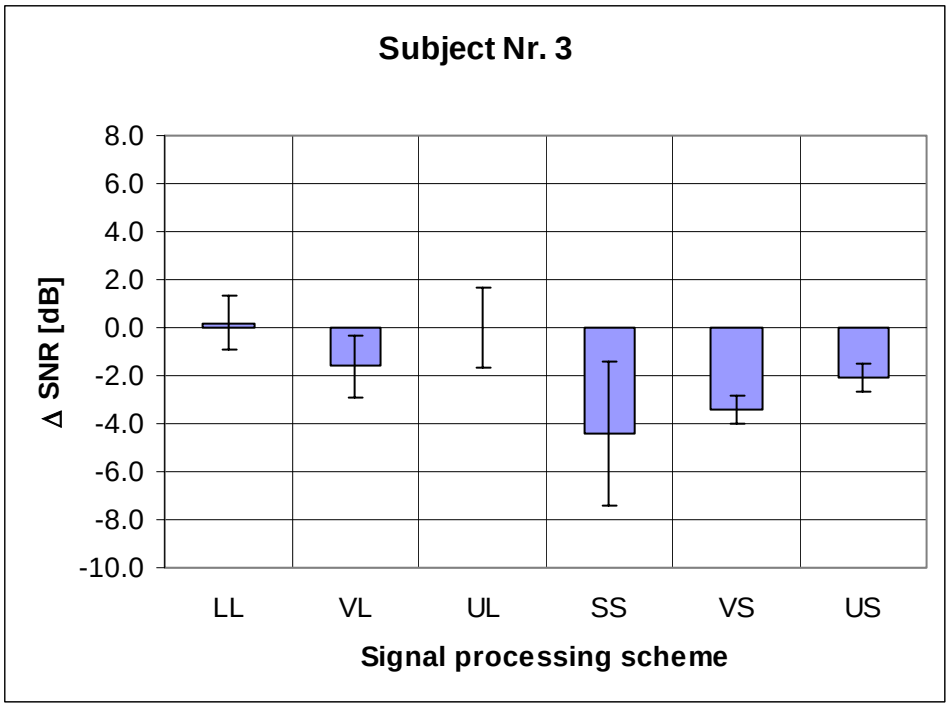


Fig. 8.5 Mean performance of subject Nr 3 achieved for SNR differences between the reference signal and the temporally modified signal for 50% sentence recognition.

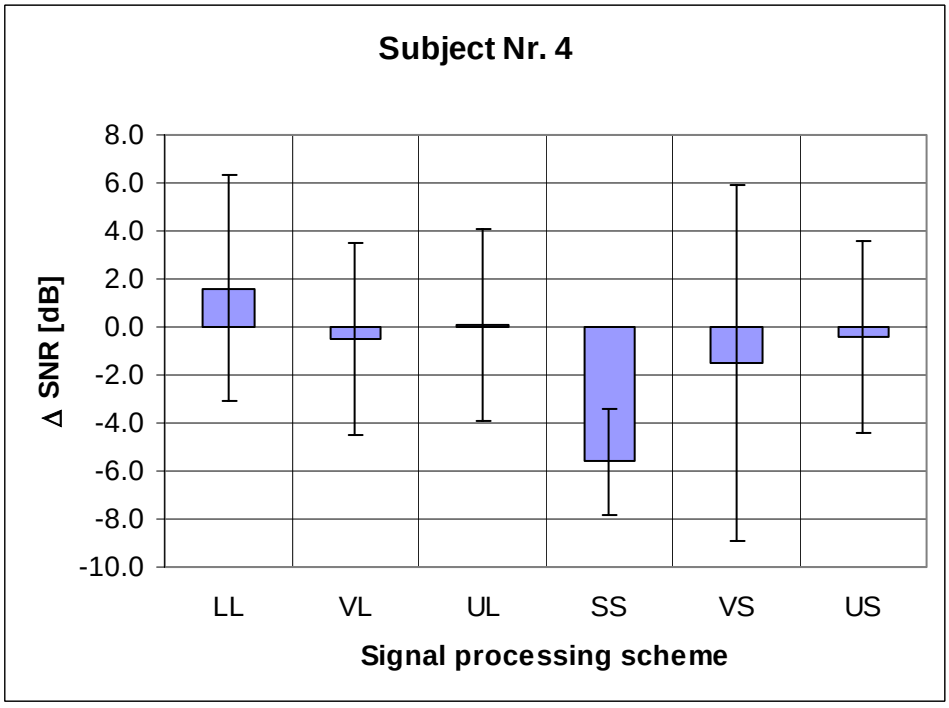


Fig. 8.6 Mean performance of subject Nr 4 achieved for SNR differences between the reference signal and the temporally modified signal for 50% sentence recognition.

All tested subjects achieved decreased sentence identification results for the shortened speech (SS) (Fig. 8.2-8.6). The most difficult signal processing scheme was obviously the one involving whole sentence temporal shortening. The signal processing schemes which perform voiced segment shortening (VS), also achieved relatively poor results. However, the unvoiced segment shortening is not as critical, yet showed poorer results than the reference signal.

The negative effects on the speech recognition caused by the signal shortening are larger than the positive effects of speech recognition caused by the prolongation of the speech signal.

8.4.4 Discussion and conclusions

The results of the present study show that the speech signal recognition is decreased for temporal shortening of the speech signal. Also prolongation of the speech signal or any of its segments can improve speech recognition in a noisy environment. However, the increase of speech intelligibility observed by temporal prolongation is relatively small with respect to the decrease of speech intelligibility caused by temporal shortening.

According to the results of this study, it can not be predicted whether speech prolongation of the voiced segments and shortening of the unvoiced segments or prolongation of the unvoiced segments and shortening of the voiced segments can improve speech comprehension relative to the reference signal.

It was therefore decided to continue the preliminary study with normal-hearing subjects using bidirectionally temporally modified speech signals.

8.5 Perception of bidirectionally temporally modified speech in sentences (Experiment III)

8.5.1 Motivation

In order to investigate bidirectionally temporally modified speech perception in sentences, signal processing schemes providing approximately equal input –output signal duration were tested on five normal hearing adults using speech material with temporally non-uniformly modified segments. The following temporal speech segment modifications were investigated: temporal prolongation of voiced segments ($\rho = 1.5$) combined with temporal shortening of unvoiced segments ($\rho = 0.5$), and temporal prolongation of unvoiced segments ($\rho = 3.0$) combined with temporal shortening of voiced segments ($\rho = 0.5$). Table 8.2 lists the two processing strategies.

Signal Processing Scheme	Notation	Temporal modification factor ρ	
		Voiced	Unvoiced
The voiced sentence segments are prolonged and unvoiced sentence segments are temporally shortened	**VL/US**	ρ=1.5	ρ=0.5
The voiced sentence segments are shortened and unvoiced sentence segments are temporally prolonged	**VS/UL**	ρ=0.5	ρ=3.0

Table 8.2 Parameter settings and nomenclature for bidirectional temporal modification resulting in equal duration of the input and output signal.

Because most speech segments of the employed test material were classified as "voiced" (according to the center of gravity classification), a smaller value of the temporal modification factor ($\rho = 1.5$ instead of $\rho = 3.0$) was used for the voiced speech segment prolongation.

8.5.2 Performed tests

The speech materials and the test settings were the same as in the study described in the previous section. However, two different testing orders were applied:

- During the first testing-part (IIIa), SNR measurements for the reference and the two different temporal parameter settings were carried out with two normal-hearing subjects in a random sequence.
- In the second testing-part (IIIb), the average SNR values of nine subsequent sentence lists consisting of 10 sentences each were determined for the reference and the two

processing strategies in different test sessions with five normal-hearing subjects. The order of the signal processing strategies was randomized.

8.5.3 Results IIIa

The differences between the SNR values of the reference and the two temporal modification schemes ($SNR_{reference}$-$SNR_{processed}$) for 50% speech perception are shown in Fig. 8.7.

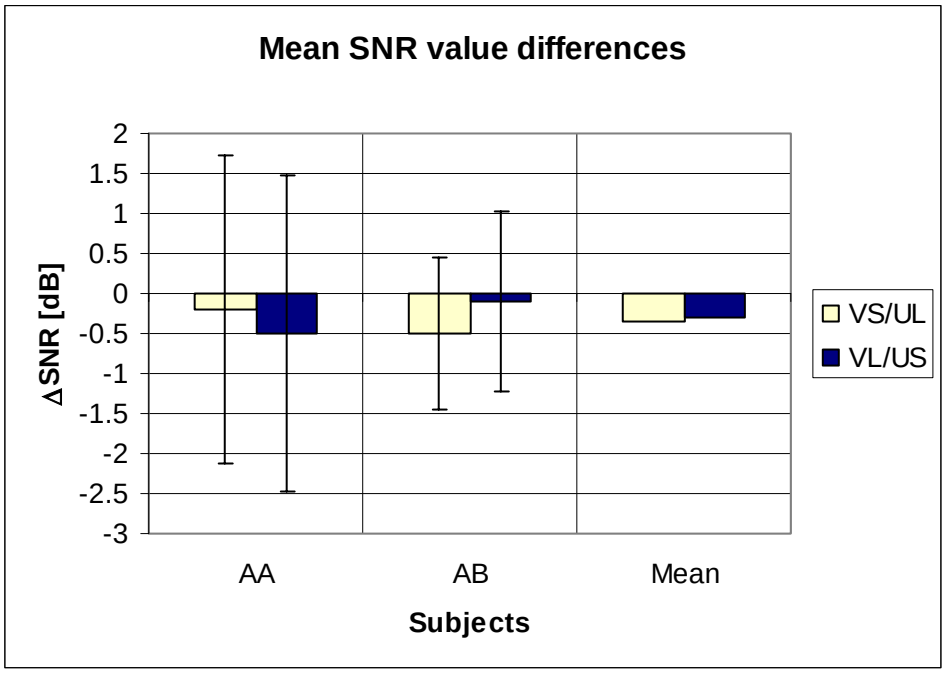


Fig. 8.7 Differences between reference signal SNR and achieved SNR values of the VS/UL and VL/US temporal modification schemes for two normal hearing subjects.

The unmodified speech (reference) showed slightly better results than the temporally modified speech. However, the differences between achieved SNR values for the reference and the modified speech are insignificant. Also the difference between the voiced-long/unvoiced-short (VL/US) and voiced-short/unvoiced-long (VS/UL) schemes was insignificant. However, the deviation areas of the SNR difference values were rather large (±2 dB for subject AA and ±1 dB for subject AB (see Fig. 8.7)).

8.5.4 Results IIIb

The differences between the SNR values of the reference and the two temporal modification schemes ($SNR_{reference}$-$SNR_{processed}$) for 50% speech perception measured in the subsequent nine sentence lists are shown in Fig. 8.8.

All tested subjects showed negative SNR difference values, indicating that the unmodified signal was easier to understand. The achieved mean SNR values for the reference signal were significantly better (~-2 dB) than the measured SNR values for both temporal modification schemes.

In three out of five cases, prolongation of the voiced speech segments in combination with shortening of the unvoiced speech segments (VL/US) indicated better results than the prolongation of the unvoiced speech segments in combination with shortening of the voiced speech segments (VS/UL) (see Fig. 8.8). However, the mean SNR values of the VL/US and the VS/UL schemes differed only insignificantly.

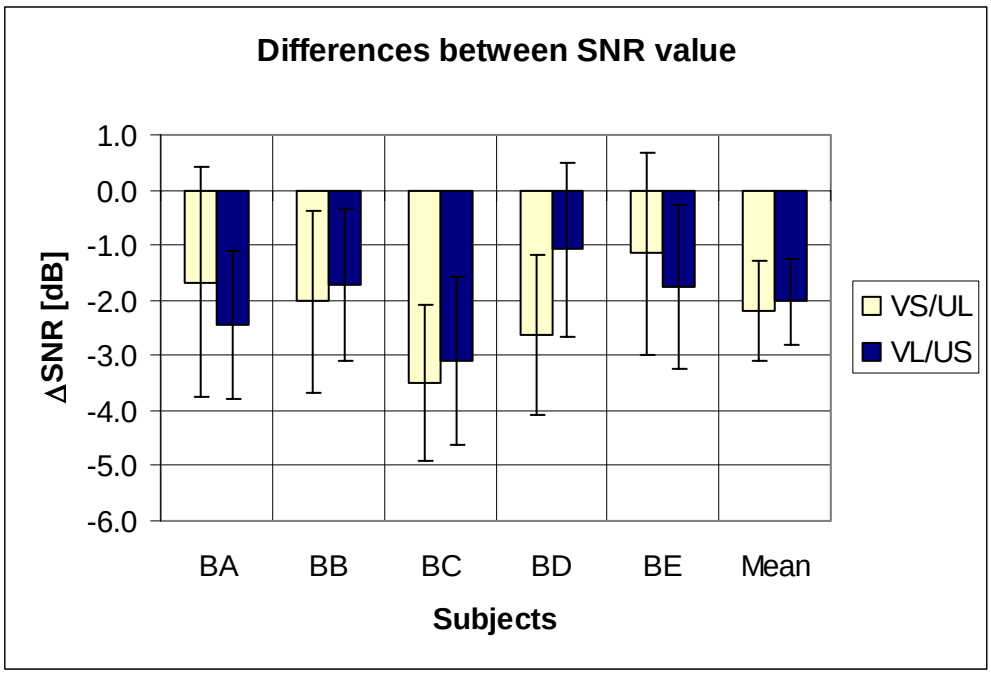


Fig. 8.8 Differences between reference signal SNR and achieved SNR values of the VS/UL and VL/US temporal modification schemes for five normal-hearing subjects measured in the nine subsequent test lists.

8.5.5 Discussion of test III and general conclusions

From the results of the studies IIIa and IIIb, it seems that the bidirectionally temporally modified speech signal is similarly or rather less intelligible than the unmodified speech. The reason for this is most probably that the shortening of any particular speech segment causes a worsening of the speech signal comprehension which is greater than the speech perception improvement provided by the prolongation of the opposite speech segments. This result shows that the temporal modification schemes which provide close to one input/output signal length duration ratio cannot improve speech signal comprehension. Thus further investigations into the combination of temporally modified and spectrally compressed speech comprehension mechanism seen unnecessary.

Temporal stretching of different speech segments without simultaneous temporal shortening of any other speech segments can possibly improve speech comprehension (see study II). However, the results of this study provide only general tendencies of the effects of temporal modifications of speech signals on speech comprehension.

As mentioned earlier, results of different temporal modification studies performed with different speech materials are difficult to compare. Suprasegmental factors such as rhythm,

prosodic elements, or even dialectical variations of the used speech materials are possible reasons for differing results. More detailed investigations into different speech segment classifications and the effects caused by their respective temporal modifications could be a topic for future studies.

Chapter 9

Ten years later

"They are boys; we found it out long ago. It was their coming in that small, immature shape that puzzled us; we were not used to it."

M. Twain

Summary and conclusions

The purpose of the present thesis was the development, implementation, investigation and evaluation of different signal processing strategies for profoundly hearing impaired patients. Based on a simple model of profound hearing impairment, earlier signal processing approaches known from the literature and signal processing strategies employed for electric hearing (cochlear implants), the following three signal processing strategies were proposed:

- spectral reduction
- spectral compression, and
- temporal modifications.

A sinusoidal speech signal processing system was implemented as a baseline tool for all processing strategies and investigations. It was shown that this sinusoidal speech system can be successfully employed for the implementation of all proposed spectral and temporal modifications. It even proved to be very well suited for all investigations carried out within the present work. A potential problem for the implementation of the system in a hearing device is that it requires a high spectral analysis resolution (at least 1024 point FFT analysis length), otherwise it produces musical noise in the reconstruction of high-frequent signal components. A possible solution for this phenomenon is the sinusoidal noise speech reconstruction approach. Additional studies on the prevention of musical noise in the reconstructed signal could however be performed.

Two different spectral compression schemes were investigated and evaluated on normal hearing and hearing-impaired subjects: linear spectral compression on the physical (FFT) frequency scale, and linear spectral compression on an auditory (SPINC) scale. The linear spectral compression on the SPINC scale is a completely new approach proposed in this thesis. Both spectral compression schemes were successfully tested with different compression factors on normal hearing and profoundly sensorineural hearing impaired subjects, and the approximate values of the spectral compression ratios were established. For both the linear spectral compression on the FFT scale and for the linear spectral compression on the SPINC scale, the applied spectral compression ratios should not be larger than 1.3. For subjects with normal hearing in the lower frequency area and profound hearing impairment in

the high frequency regions, the linear spectral compression on the FFT scale with CR=1.3 achieved improved vowel recognition. For subjects with profound hearing loss in the whole audible frequency range, the linear spectral compression on the SPINC scale with CR=1.3 achieved improvement in consonant recognition and as a result also in sentence recognition. To generalize these findings, additional studies involving more hearing-impaired subjects and wearable spectral compression devices are necessary. For the further studies it is proposed to implement the spectral compression in combination with a small upward spectral shift for compensation of the low frequency lowering.

In the course of the experiments with spectral compression, a selection criterion for the identification of hearing impaired subjects which might benefit from spectral compression was proposed as well. This selection criterion involves the calculation of the pure tone average over the measured low audiometric frequency values (125, 250 and 500 Hz). This special candidate selection criterion should also be investigated with a larger number of hearing impaired subjects.

As one of the aspects relates to signal processing strategies used in cochlear implants, spectral reduction of the speech signal was investigated. The studies described in this thesis showed that the speech signal can be reduced to a very limited number of spectral components while still being understandable. From the study described in chapter 5 it was concluded that there is a minimum spectral component per time unit ratio which is required for sufficient speech perception. The value of this minimum spectral component per time unit ratio was found to be approximately one spectral component per 1.5 ms. It was also observed that the value of the spectral component per time ratio for proper vowel identification is larger than for sufficient consonant identification. This is in agreement with studies mentioned in chapter 3, which demonstrated that vowel comprehension reacts more sensitive to the reduction of spectral content.

Applying spectral reduction in combination with spectral compression to the hearing impaired subjects showed that in most cases subjects preferred the spectrally non-reduced speech signal over the spectrally reduced signal. However, in combination with the linear spectral compression on the SPINC scale the spectral reduction proved to be advantageous for reducing the high frequency spectral content. This finding can be explained with the increased high-frequency component density caused by the large local spectral compression ratio values provided by the spectral compression on the SPINC scale.

The combined use of spectral-compression hearing devices with cochlear implants might be advantageous. This would require additional studies with wearable spectral compression systems which are presently unavailable.

The studies on the temporal modification of the speech signal showed that the slowing down of the whole signal or different speech segments could improve speech perception of normal-hearing subjects. Unfortunately, the maximal temporal expansion restricted by the required signal input-output time matching, is not sufficient to achieve any promising results. In addition, for the implementation of temporal compressing and stretching in a real-time system, a special, presently non-existing algorithm for the detection of the speech rhythm and its changes would be required.

Appendix

Pure-Tone Average in dB HL	Degree of Hearing Loss
≤15	Normal hearing
16-25	Slight hearing loss
26-40	Mild hearing loss
41-55	Moderate hearing loss
56-70	Moderately severe hearing loss
71-90	Severe hearing loss
≥90	Profound hearing loss

Tab. A1 Typically used categories to describe the degree of hearing loss based on the pure-tone average

Abbreviations

AGC	Automatic gain control
BARK	Critical-band rate
BWF	Narrow frequency bandwidth factor
CG	Centre of gravity
CI	Cochlear implant
CNC	Consonant - vowel nucleus - consonant
C12	Consonant test
CR	Compression ratio
ERB	Equivalent rectangular bandwidth
FFT	Fast Fourier transformation
FRI	Frication
F01	First formant
F02	Second formant
IFFT	Inverse fast Fourier transformation
IST	Innsbrucker sentence test
LPC	Linear prediction coefficients
LPF	Low-pass filtered
LP	Low-pass
LV	Long vowel
MAC	Minimal auditory capability
MAN	Manner
NAS	Nasality
NFBNG	Narrow frequency band noise generator
PLC	Place
PPIG	Polynomial phase interpolation generator
PTA	Pure tone average
PTG	Pure tone generator
REF	Reference
SC	Spectral compression
SIB	Sibilance
SiNoSp	Sinusoidal noise speech
SiSp	Sinusoidal speech
SiVo	Sinusoidal voice
SNR	Signal to noise ratio
SON	Sonorance
SPINC	Spectral pitch increment
SPL	Sound pressure level
SS	Shortened speech
SV	Short vowel
UL	Unvoiced long

US	Unvoiced short
VL	Voiced long
VOI	Voicing
VS	Voiced short
V08	Vowel identification test

Notation

$A(t_m)$	Reconstructed signal at a discrete time sample $t_m = 1,\ldots,L_{SF}$
A_k	Amplitude of the k-th reconstructed spectral component
$\widetilde{A}_k(t_m)$	Interpolated amplitude
$\alpha(M^*)$	Phase interpolation coefficient
$\beta(M^*)$	Phase interpolation coefficient
BN_{FFT}^{OUT}	FFT bin number of the spectrally compressed spectral component
CR	Spectral compression ratio
CR_{Lin}	Constant spectral compression ratio
$CR(Fr)$	Frequency dependent spectral compression ratio
CR_{SPINC}	Spectral compression ratio in the SPINC scale
ΔFr	Resolution of the FFT analysis
$Fr_{Sampl.}$	Sampling frequency
Fr_{OUT}	Output frequency
Fr_{IN}	Input frequency
Fr_{CutOff}	Cutoff frequency
$Fr_{CentreF}$	Centre of flipping frequency
Fr_{Shift}	Spectral shift value.
Fr_{appr}	Approximate frequency
Fr^{Tr}	Frequency value from the spectral component tracking frame
$\mathrm{Fr}_{\mathrm{Frame}}$	Frame formation frequency
Φ_{OUT}	Output spinc value
Φ_{IN}	Input spinc value
L_{FFT}	FFT frame-length
L_{SF}	Length of the synthesis frame
N_{mod}	Modified frame-length
$OF_{Analysis}$	Analysis frame overlap factor
OF_{Synth}	Overlap factor of the synthesis frame formation
$OF^{max}{}_{Synth}$	Maximal overlap factor of the synthesis frame
$OF_{Analysis}$	Overlap factor of the signal analysis frame
s	FFT frequency bin factor
$\widetilde{s}(t_m)$	Value of the reconstructed frame at the sample t_m
$SR_{original}$	Original range of frequencies
$SR_{compressed}$	Spectral range after performing the spectral compression
θ	Phase angle
Θ_k	Phase of the actually reconstructed spectral component
$\Theta^{Tr}{}_k$	Phase taken from the spectral component tracking frame
Θ_k	Phase value of the k-th reconstructed spectral component
$\widetilde{\Theta}_k(t_m)$	Interpolated phase value
T	Time period between two subsequent overlapping signal frames

T_{pitch}	Pitch period
$t_0(m)$	Onset time
V_{FFT}^{n}	Value of the n-th bin of the FFT spectrum
$\|X(f_i)\|$	Magnitude FFT spectrum at bin *i*
$W(t_m)$	Value of the windowing function at the corresponding discrete time

Bibliography

[B1] *Cochlear Prostheses* Churchill Livingstone Inc., 1990.

[B2] Adelman. Signal processing apparatus. [US4419544]. 1983. USA.
Ref Type: Patent

[B3] Aebi C., "Zeitliche Modification von Sprachsignalen." Dipl. El. Ing. Diploma Thesis, ETH Zurich, 2002.

[B4] AVR Communications Ltd. Frequency Transposing Hearing Aid. AVR Communications Ltd. 283,971[5,014,319]. 1991. USA. 1988.
Ref Type: Patent

[B5] Bashford J.A. and Warren R.M., "Effects on spectral alternation on the intelligibility of words and sentences," *Perception & Psyhoohysics*, vol. 42, no. 5, pp. 431-438, 1987.

[B6] Beasley D.S., Schwimmer S., and Rintelman W.F., "Intelligibility of time-compressed CNC monosyllables," *Journal of Speech and Hearing Research*, vol. 15 pp. 340-350, 1972.

[B7] Bench J., "The Auditory Response," *Perinatal physiology* 2 ed. New York: Plenum Publishing Corporation, 1978, pp. 751-760.

[B8] Berlin C.I., "Unusual forms of residual high-frequency hearing," *Seminars in Hearing*, vol. 6, no. 4, pp. 389-395, 1985.

[B9] Biondi. Method and device for making natural sounds audible to hearing impaired. [DE1762185]. 1970.
Ref Type: Patent

[B10] Blesser B., "Speech perception under conditions of spectral transformation: I. phonetic characteristics," *Journal of Speech and Hearing Research*, vol. 15 pp. 5-41, 1972.

[B11] Boothroyd A., *Speech Acoustics and Perception* PRO-ED, Inc., 1986.

[B12] Boothroyd A., Mulhearn B., Gong J., and Ostroff J., "Effects of spectral smearing on phoneme and word recognition," *J.Acoust.Soc.Am.*, vol. 100, no. 3, pp. 1807-1818, 1996.

[B13] Breeuwer M. and Plomp R., "Speechreading supplemented with formant-frequency information from voiced speech," *J.Acoust.Soc.Am.*, vol. 77, no. 1, pp. 314-317, 1985.

[B14] Clark G.M., Black R., Dewhurst D.J., Forster I.C., Patrick J.F., and Tong Y.C., "A multiple-electrode hearing prosthesis for cochlear implantation in deaf patients," *Medical Progress Through Technology*, vol. 5 pp. 127-140, 1977.

[B15] Daniloff R.G., Shriner T.H., and Zemlin W.R., "Intelligibility of Vowels Altred in Duration and Frequency," *J.Acoust.Soc.Am.*, vol. 44, no. 3, pp. 700-707, 1968.

[B16] Davis W.E., "Proportional Frequency Compression in Hearing Instruments," *The Hearing Review*, vol. Feb pp. 34-42, 2001.

[B17] De Filippo C.L. and Scott B.L., "A Method for Training and Evaluating the Reception of Ongoing Speech," *J.Acoust.Soc.Am.*, vol. 63 pp. 1186-1192, 1978.

[B18] Dean M.R. and McDermott H.J., "Preferred Listening Levels: The Effect of Background Noise for Moderate-to-Profoundly Hearing-Impaired Aid Users," *Scandinavian Audiology*, 2000.

[B19] Dillier N., "Mikroelektronische Hörprothesen." PD Universitätsspital Zürich, 1995.

[B20] Dillier N. and Spillmann T., "Deutsche Version der Minimal Auditory Capability (MAC)-Test-Batterie: Anwendungen bei Hörgeräte- und CI-Trägern mit und ohne Störlärm," 1992, pp. 238-263.

[B21] Djourno A. and Eyries C., "Prosthese auditive per excitation electrique a distance du nerf sensoriel a l'aide d'un," *Presse Medicale*, vol. 35 pp. 14-17, 1957.

[B22] Dorman M.F., Lindholm J.M., and Hannley M.T., "Influence of the First Formant on the Recogniton of Voiced Stop Consonants by Hearing-Impaired Listeners," *Journal of Speech and Hearing Research*, vol. 28 pp. 377-380, 1985.

[B23] Dorman M.F. and Loizou P.C., "The Identification of Consonants and Vowels by Cochlear Implant Patients Using 6-Channel Continuous Interleaved Sampling Processor and by Normal-Hearing Subjects Using Simulations of Processors with Two to Nine Channels," *Ear & Hearing* , vol. 19, no. 2, pp. 162-166, 1998.

[B24] Dorman M.F., Loizou P.C., and Fitzke J., "The Identification of Speech in Noise by Cochlear Implant Patients and Normal-Hearing Listeners Using 6-Channel Signal Processors," *Ear & Hearing*, vol. 19, no. 6, pp. 481-484, 1998.

[B25] Dorman M.F., Loizou P.C., and Rainey D., "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," *J.Acoust.Soc.Am.*, vol. 102, no. 5, pp. 2993-2996, 1997.

[B26] Dorman M.F., Loizou P.C., and Rainey D., "Speech intelligibility as a function of the number of channels of simulation for signal processors using sine-wave and noise-band outputs," *J.Acoust.Soc.Am.*, vol. 102, no. 4, pp. 2403-2411, 1997.

[B27] Dorman M.F., Soli S., Dankowski K., Smith L.M., McCandless G., and Parkin J., "Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant," *J.Acoust.Soc.Am.*, vol. 88, no. 5, pp. 2074-2079, 1990.

[B28] Doyle J.H., Doyle J.B., and Turnbull F.M., "Electrical stimulation of eighth cranial nerve," *Arch Otolaryngol*, vol. 80 pp. 388-391, 1964.

[B29] Dreschler W.A. and Plomp R., "Relations between psychophysical data and specch perception for hearing impiared subjects. I," *J.Acoust.Soc.Am.*, vol. 68 pp. 1608-1616, 1980.

[B30] Dreschler W.A. and Plomp R., "Relations between psychophysical data and specch perception for hearing impiared subjects. II," *J.Acoust.Soc.Am.*, vol. 78 pp. 1261-1270, 1985.

[B31] Dubno J.R., Ahlstrom J.B., and Horwitz A.R., "Use of context by young and aged adults with normal hearing," *J.Acoust.Soc.Am.*, vol. 107, no. 1, pp. 538-546, 2000.

[B32] Dubno J.R. and Dorman M.F., "Effects of spectral flattening on vowel identification," *J.Acoust.Soc.Am.*, vol. 82, no. 5, pp. 1503-1511, 1987.

[B33] Dudley H., "Remaking Speech," *J.Acoust.Soc.Am.*, vol. 11, no. 2, pp. 169-177, 1939.

[B34] Dupret J.P., "EMILY: Une nouvelle approche de la correction auditive," *Les Cahiers de l'Audition*, vol. 5, no. 2, 1991.

[B35] Dupret J.P. and Levefre F. Electronic Device for Processing a Sound Signal. [US5077800]. 1991. USA.
Ref Type: Patent

[B36] Eisenberg L.S., Shannon R.V., Schaefer-Martinez A., and Wygonski J., "Speech recognition with reduced spectral cues as a function of age," *J.Acoust.Soc.Am.*, vol. 107, no. 5, pp. 2704-2710, 2000.

[B37] Erber N.P., "Evaluation of spectral hearing aids for deaf children," *Journal of Speech and Hearing Disorders*, vol. XXXVI, no. 4, pp. 527-537, 1971.

[B38] Erber N.P., "Speech-Envelope Cues as an Acoustic Aid to Lipreading for Profoundly Deaf Children," *J.Acoust.Soc.Am.*, vol. 51, no. 4, pp. 1224-1227, 1972.

[B39] Ericsson. A method of improving the intelligibility of a sound signal, and a device for reproducing a sound signal. Ericsson. [WO 00/75920]. 2000.
Ref Type: Patent

[B40] Fairbanks G., Everitt W., and Jaeger R., "Method for Time or Frequency Compression-expantion of Speech," *Trans.IRE-PGA*, vol. AU2 pp. 7-11, 1954.

[B41] Fairbanks G. and Kodman F., "Word intelligibility as a function of time compression," *J.Acoust.Soc.Am.*, vol. 29 pp. 636-644, 1957.

[B42] Festen J.M. and Plomp R., "Relations between auditory functions in impaired hearing," *J.Acoust.Soc.Am.*, vol. 73 pp. 652-662, 1983.

[B43] Fishman K.E., Shannon R.V., and Slattery W.H., "Speech Recognition as a Function of the Number of Electrodes Used in the SPEAK Cohlear Implant Speech Processor," *Journal of Speech, Language, and Hearing Research*, vol. 40 pp. 1201-1215, 1997.

[B44] Flynn M.C., Dowell R.C., and Clark G.M., "Aided Speech Recognition Abilities of Adults With a Severe or Severe-to-Profound Hearing Loss," *JSLHR*, vol. 41 pp. 285-299, 1998.

[B45] Franck B.A.M., van Kreveld-Bos C.S.G.M., Dreschler W.A., and Verschuure H., "Evaluation of spectral enhancement in hearing aids, combined with

phonemic compression," *J.Acoust.Soc.Am.*, vol. 106, no. 3, pp. 1452-1464, 1999.

[B46] Friesen L.M., Shannon R.V., Baskent D., and Wang X., "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J.Acoust.Soc.Am.*, vol. 110, no. 2, pp. 1150-1163, 2001.

[B47] Fu Q.J. and Shannon R.V., "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," *J.Acoust.Soc.Am.*, vol. 105, no. 3, pp. 1889-1900, 1999.

[B48] Fu Q.J., Shannon R.V., and Wang X., "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *J.Acoust.Soc.Am.*, vol. 104, no. 6, pp. 3586-3596, 1998.

[B49] Garvey W.D., "The intelligibility of speeded speech," *J.exp.Psychol.*, vol. 45 pp. 102-106, 1953.

[B50] Gelfand S.A., *Essentials of Audiology* Thieme Medical Publishers, Inc., 1997.

[B51] Gisselsson L., "Experimental investigation into the problem of humoral transmission in the cochlea.," *Acta Oto-laryngol.*, vol. 82 pp. 16, 1950.

[B52] Glasberg B.R. and Moore B.C.J., "Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech," *Scandinavian Audiology Suppl.*, vol. 32 pp. 1-25, 1989.

[B53] Goldstein E.B., *Sensation and Perception*, Fifth Edition ed. Brooks/Cole Publishing Company, 1999.

[B54] Gopnik A., Meltzoff A.N., and Kuhl P.K., *The scientist in the crib: minds, brains, and how children learn*, 1 ed. New York: William Morrow and Company, Inc., 1999.

[B55] Grant K.W., Ardell L.H., Kuhl P.K., and Sparks D.W., "The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects," *J.Acoust.Soc.Am.*, vol. 77, no. 2, pp. 671-677, 1985.

[B56] Gravel J.S. and Chute P.M., "Transposing Hearing Aids for Children," *New Technologies* 1996, pp. 253-271.

[B57] Hertrich I. and Ackermann H., "A vowel synthesizer based on formant sinusoids modulated by fundamental frequency," *J.Acoust.Soc.Am.*, vol. 106, no. 5, pp. 2988-2990, 1999.

[B58] Hill F.J., McRae L.P., and McClellan R.P., "Speech Recognition as a Function of Channel Capacity in a Discrete Set of Channels," *J.Acoust.Soc.Am.*, vol. 44, no. 1, pp. 13-18, 1968.

[B59] Hochmair E.S., "An implantable current source for electrical nerve stimulation," *IEEE Transcriptions on Biomedical Engineering*, vol. 27 pp. 278-280, 1980.

[B60] Hogan C.A. and Turner C.W., "High-frequency audibility: Benefits for hearing-impaired listeners," *J.Acoust.Soc.Am.*, vol. 104, no. 1, pp. 432-441, 1998.

[B61] Hurtig R.R. and Turner C.W. Hearing aid with proportional spectral compression and shifting of audio signals. [WO 99/14986]. 1999.
Ref Type: Patent

[B62] Ito M., Tsuchida J., and Yano M., "On the effectiveness of whole spectral shape for vowel perception," *J.Acoust.Soc.Am.*, vol. 110, no. 2, pp. 1141-1149, 2001.

[B63] Johansson B., "The use of the transposer for the management of the deaf child," *J.internat.Audiol.*, vol. 5 pp. 362-372, 1966.

[B64] Kates J.M., "Speech Enhancement Based on a Sinusoidal Model," *Journal of Speech and Hearing Research*, vol. 37 pp. 449-464, 1994.

[B65] Klatt D.H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *J.Acoust.Soc.Am.*, vol. 59, no. 5, pp. 1208-1221, 1976.

[B66] Klatt D.H., "Software for a cascade/parallel formant synthesizer," *J.Acoust.Soc.Am.*, vol. 67, no. 3, pp. 971-995, 1980.

[B67] Kluender K.R. and Lotto A.J., "Effects on first formant onset frequency on [-voice] judgments result from auditory processes not specific to humans," *J.Acoust.Soc.Am.*, vol. 95, no. 2, pp. 1044-1052, 1994.

[B68] Klumpp R.B. and Webster J.C., "Intelligibility of Time-Compressed Speech," *J.Acoust.Soc.Am.*, vol. 33 pp. 265-267, 1961.

[B69] Korl S., "Automatische Geräuschklassifizierung zur Anwendung in Hörgeräten." Dipl. El. Ing. Diploma Thesis, ETH Zurich, 1999.

[B70] König G. and Eichler H., "Ein neues Verfahren zur Hörverbesserung bei hochgradiger Perzeptionsschwerhörigkeit durch Hörapparate," *Arch.Ohr.Nas.Kehlkopfhk.*, vol. 165 pp. 326-331, 1954.

[B71] Kurtzrock G., "The Effects of Time and Frequency Distortion Upon Word Intelligibility." University of Illinois, 1956.

[B72] Lafon J.C. Improvement of hearing instruments. [EP0054450]. 1984. Ref Type: Patent

[B73] Lafon J.C., "Transposition et Modulation," *Bulletin d'Audiophonologie, Annales Scientifiques de l'Universitè de Franche-Comptè*, vol. XII (3&4) pp. 201-252, 1996.

[B74] LarsenW.J., *Human embryology* New York: Churchill Livingston Inc., 1993.

[B75] Launer S., "Loudness Perception in Listeners with Sensorineural Hearing Impairment." Universität Oldenburg, 1995.

[B76] Liberman A.B., Delattre P., and Cooper F.S., "Some cues for the distin and voiceless stops in initial position," *Lang.Speech*, vol. 1 pp. 153-167, 1958.

[B77] Liberman A.B., Delattre P., Gerstman L.J., and Cooper F.S., "Tempo of frequency change as a cue for distinguishing classes of speech sounds," *Journal of Experimental Psychology*, vol. 52, no. 2, pp. 127-137, 1956.

[B78] Ling D., "Three experiments on frequency transposition," *Amer.Ann.Deaf*, vol. 113 pp. 283-294, 1968.

[B79] Ling D. and Doehring D.G., "Learning limits of deaf children for coded speech," *Journal of Speech and Hearing Research*, vol. 12 pp. 83-94, 1969.

[B80] Ling D. and Druz W.S., "Transposition of high frequency sounds by partial vocoding of the speech spectrum: its use by deaf children," *The Journal of Auditory Research*, vol. 7 pp. 133-144, 1967.

[B81] Loizou P.C., "Mimicking the Human Ear," *IEEE*, vol. Signal Processing Magazine September 1998 pp. 101-129, 1998.

[B82] Loizou P.C., Dorman M.F., and Tu Z., "On the number of channels needed to understand speech," *J.Acoust.Soc.Am.*, vol. 106, no. 4, Pt. 1, pp. 2097-2103, 1999.

[B83] Mazor M., Simon H., Scheinberg C., and Levitt H., "Moderate frequency compression for the moderately hearing impaired," *J.Acoust.Soc.Am.*, vol. 62, no. 5, pp. 1273-1278, 1977.

[B84] McAulay R.J. and Quatieri T.F., "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE*, vol. ASSP-34, no. 4, pp. 744-754, 1986.

[B85] McAulay R.J. and Quatieri T.F., "Pitch estimation and voicing detection based on a sinusoidal speech model," *IEEE Internation Conference on Acoustic, Speech, and Signal Processing*, vol. 1, no. S1, pp. 249-252, 1990.

[B86] McDermott H.J. Auditory perception with sloping high-hearing frequency loss, and the limited benefit of frequency transposition. IHCON Lake Tahoe, CA. 2000. 23-8-2000.
Ref Type: Conference Proceeding

[B87] McDermott H.J. 2001.
Ref Type: Personal Communication

[B88] McDermott H.J. and Dean M.R., "Speech perception with steeply sloping hearing loss: Effects of frequency transposition," *British Journal of Audiology*, 2000.

[B89] McDermott H.J., Dorkos V.P., Dean M.R., and Ching T.Y.C., "Improvements in Speech Perception With Use of the AVR TranSonic Frequency-Transposing Hearing Aid," *Journal of Speech, Language, and Hearing Research*, vol. 42 pp. 1323-1335, 1999.

[B90] McDermott H.J. and Knight M.R., "Preliminary Results with the AVR ImaCt Frequency-Transposing Hearing Aid," *J Am Acad Audiol*, 2000.

[B91] Michelson R.P., "The results of electrical stimulation of the cochlea in human sensory deafness," *Ann-Otol-Rhinol-Laryngol.*, vol. 80 pp. 914-918, 1971.

[B92] Miller G.A. and Nicely P.E., "An Analysis of Perceptual Confusions Among Some English Consonants," *J.Acoust.Soc.Am.*, vol. 27, no. 2, pp. 338-352, 1954.

[B93] Moore B.C.J., *An Introduction to the Psychology of Hearing*, 4th ed. Academic Press, 1997.

[B94] Moore B.C.J., *Cochlear Hearing Loss* London: Whurr Publishers Ltd, 1998.

[B95] Mummert M, "Sprachcodierung durch Konturierung eines gehörangepassten Spectrogramms und ihre Anwendung zur Datenreduktion." Technische Universität München, 1997.

[B96] Nagafuchi M., "Intelligibility of Distorted Speech Sounds Shifted in Frequency and Time in Normal Children," *Audiology*, vol. 15 pp. 326-337, 1976.

[B97] Nawab S.H., Quatieri T.F., and Lim J.S., "Signal Reconstruction from Short-Time Fourier Transform Magnitude," *IEEE*, vol. ASSP-31, no. 4, pp. 986-998, 1983.

[B98] Nelson D.J., "Cross-spectral methods for processing speech," *J.Acoust.Soc.Am.*, vol. 110, no. 5, pp. 2575-2592, 2001.

[B99] Ochai Y., Saito S., and Sakai Y., "Ariticulation Study of Speech Qualities in Rotational Synchronous Distortion," *Mem.Fac.Eng.Nagoya University*, vol. 7 pp. 40-48, 1955.

[B100] Oeken F.W., "Can the hearing of patients suffering from high-tone perceptive deafness be improved by frequency transposition?," *J.internat.Audiol.*, vol. 2 pp. 263-266, 1963.

[B101] Pavlovic C.V., "Band Importance Functions for Audiological Applications," *Ear & Hearing*, vol. 15, no. 1, pp. 100-104, 1994.

[B102] Perwitzschky E., "Ein neues Prinzip der Hörverbesserung," *Zschr.Hals.Nas.Ohrenhk.*, vol. 12 pp. 593-602, 1925.

[B103] Peterson G.E. and Barney H.L., "Control Methods Used in a Study of the Vowels," *J.Acoust.Soc.Am.*, vol. 24, no. 2, pp. 175-184, 1952.

[B104] Peterson, G. and Lehiste, I., "Revised CNC lists for auditory tests," *Journal of Speech and Hearing Disorders*, vol. 27 pp. 62-70, 1962.

[B105] Pimonow L. Speech Transformer. [FR1309425]. 1961. France.
Ref Type: Patent

[B106] Pimonow L., "The aplication of synthetic speech to aural rehabilitation," *J.Aud.Res.*, vol. 3 pp. 73-82, 1963.

[B107] Press W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P., *Numerical recipes in C: the art of scientific computing*, 2nd ed. ed. Cambridge University Press, 1992.

[B108] Quatieri T.F. and McAulay R.J., "Speech Transformations Based on a Sinusoidal Representation," *IEEE*, vol. ASSP-34, no. 6, pp. 1449-1464, 1986.

[B109] Quatieri T.F. and McAulay R.J., "Peak-to-RMS Reduction of Speech Based on a Sinusoidal Model," *IEEE*, vol. 39, no. 2, pp. 273-288, 1991.

[B110] Quatieri T.F. and McAulay R.J., "Shape Invariant Time-Scale and Pitch Modification of Speech," *IEEE Transcriptions on Signal Processing*, vol. 40, no. 3, pp. 497-510, 1992.

[B111] Remez R.E., Rubin Ph.E., and Carrell Th.D., "Speech Perception Without Traditional Speech Cues," *Science* , vol. 212 pp. 947-950, 1981.

[B112] Rosen S., "Temoral information in speech: acoustic, auditory and linguistic aspects," *Phil.Trans.R.Soc.Lond.B*, vol. 336 pp. 367-373, 2002.

[B113] Rosen S., Faulkner A., and Wilkinson L., "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," *J.Acoust.Soc.Am.*, vol. 106, no. 6, pp. 3629-3636, 1999.

[B114] Rosen S., Walliker J.R., Fourcin A., and Ball V., "A microprocessor-based acoustic hearing aid for the profoundly impaired listener," *Journal of Rehabilitation Research and Development* , vol. 24, no. 4, pp. 239-260, 1987.

[B115] Rosenhouse J., "The Frequency Transposing Hearing Aid: A New Prototype," *The Hearing Journal*, no. February, pp. 14-15, 1989.

[B116] Sakamoto S., Goto K., Tateno M., and Kaga K., "Frequency compression hearing aid for severe-to-profound hearing impairments," *Auris Nasus Larynx*, vol. 27 pp. 327-334, 2000.

[B117] Sekimoto S. and Saito S., "Nonlinear frequency compression speech processing based on the parcor analysis-synthesis technique," *Ann.Bull.RILP*, vol. 14 pp. 65-72, 1980.

[B118] Shannon R.V., Zeng Fan-Gang, Kamath V., Wygonski J., and Ekelid M., "Speech Recognition with Primarily Temporal Cues," *Science*, vol. 270, no. 13, pp. 303-304, 1995.

[B119] Shannon R.V., Zeng Fan-Gang, and Wygonski J., "Speech recognition with altered spectral distribution of envelope cues," *J.Acoust.Soc.Am.*, vol. 104, no. 4, pp. 2467-2476, 1998.

[B120] Shigeru. Frequency-Compression Hearing Aid. [JP57055700]. 1982. Japan. Ref Type: Patent

[B121] Smoorenburg G.F., "Speech processing hearing aids for the profoundly hearing impaired," *Psychoacoustics, Speech and Hearing Aids*, pp. 83-93, 1996.

[B122] Springer A.M., "Lässt sich das restgehör für tiefe Frequenzen dazu verwenden, dass durch Frequenztransposition der Hörbereich erweitert wird?," *Zschr.Laryng.*, vol. 41 pp. 420-422, 1962.

[B123] Strong and Palmer. Speech Coding Hearing Aid System Utilizing Formant Frequency Transformation. [US4051331]. 1977. USA.
Ref Type: Patent

[B124] Studebaker G.A., Pavlovic C.V., and Sherbecoe R.L., "A frequency importance function for continuous discourse," *J.Acoust.Soc.Am.*, vol. 81, no. 4, pp. 1130-1138, 1987.

[B125] Tato J., "Die sensibilisierte Sprachaudiometrie," *Acta Oto-laryngol.*, vol. 51 pp. 600-614, 1960.

[B126] Terhardt E., "The SPINC Function for Scaling of Frequency in Auditory Models," *Acustica*, vol. 77 pp. 40-42, 1992.

[B127] Terhardt E., *Akustische Kommunikation* Springer, 1998.

[B128] The Math Works, *Using MATLAB*, 6 ed. The MATH WORKS, 2000.

[B129] Thomas I.B., "The Influence of First and Second Formants on the Inelligibility of Clipped Speech," *J.Audio Eng.Soc.*, vol. 16, no. 2, pp. 182-185, 1968.

[B130] Thomson-CFS. Signal processing method and device for hearing correction of hearing impaired. [EP1006511]. 2000.
Ref Type: Patent

[B131] Tiffany W.R. and Bennett D.A., "Intelligibility of Slow-Played," *Journal of Speech and Hearing Research*, vol. 4 pp. 248-258, 1961.

[B132] Turner C.W., "The limits of high-frquency amplification," *The Hearing Journal*, vol. 52, no. 2, pp. 10-14, 1999.

[B133] Turner C.W. and Hurtig R.R., "Proportional Frequency compression of speech for listeners with sensorineural hearing loss," *J.Acoust.Soc.Am.*, vol. 106, no. 2, pp. 877-886, 1999.

[B134] Turner C.W., Siu-Ling Chi, and Flock S., "Limiting Spectral Resolution in Speech for Listeners With Sensorineural Hearing Loss," *Journal of Speech, Language, and Hearing Research*, vol. 42 pp. 773-784, 1999.

[B135] Twain M., *A tramp abroad* Penguin Putnam Inc., 1997.

[B136] Tyler R.S., Summerfiled Q., Wood E.J., and Fernandes M., "Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners," *J.Acoust.Soc.Am.*, vol. 72 pp. 740-752, 1982.

[B137] van Rooij J.C.G.M. and Plomp R., "Auditive and Cognitive factors in speech perception by elderly listeners. II: Multivariate analyses," *J.Acoust.Soc.Am.*, vol. 88 pp. 2611-2621, 1990.

[B138] van Schijndel N.H., Houtgast T., and Festen J.M., "Effects of degradation of intensity, time, or frequency content on speech inelligibility for normal-hearing and hearing-imaired listeners," *J.Acoust.Soc.Am.*, vol. 110, no. 1, pp. 529-542, 2001.

[B139] Van Tassel D.J., Kirby V.M., and Widin G.P., "Speech waveform envelope cues for consonant recognition," *J.Acoust.Soc.Am.*, vol. 82, no. 4, pp. 1152-1161, 1987.

[B140] Velmans M. Aids for deaf persons. [US3819875]. 1974. USA.
Ref Type: Patent

[B141] Velmans M., "Evaluation of a new frequency transposing hearing aid and/or speech training aid for sensory neural deaf," *Int.J.Rehab.Research*, vol. 7, no. (2), pp. 196-198, 1984.

[B142] Velmans M. and Marcuson M., "The Acceptability of Spectrum-Preserving and Spectrum-Destroing Transposition to Severely Hearing-Impaired Listeners," *British Journal of Audiology*, vol. 17 pp. 17-26, 1983.

[B143] Vickers D.A., Moore B.C.J., and Baer Th., "Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies," *J.Acoust.Soc.Am.*, vol. 110, no. 2, pp. 1164-1175, 2001.

[B144] Vigneron and Lamotte. Method and device for division of audible frequencies and suppressionof distortion in the output signal. [FR2364520]. 1971. France.
Ref Type: Patent

[B145] Walliker J.R. and Smith D. Signal Processing Hearing Aids for the Profoundly Deaf. U.K. EPI Group. 1986.
Ref Type: Serial (Book,Monograph)

[B146] Warren R.M., Riener K.R., Bashford J.A., and BrubakerB.S., "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Perception & Psyhoohysics*, vol. 57, no. 2, pp. 175-182, 1995.

[B147] Wyrsch S., "Adaptive Subband Signal Processing for Hearing Instruments." PhD Diss. ETH No. 13577, ETH Zurich, 2000.

[B148] Zwicker E. and Fastl H., *Psychoacustics*, 2nd ed. Springer-Verlag, 1999.

Curriculum Vitae

Education

06.1999 – 06.2003	Doctor of Technical Sciences (PhD), Swiss Federal Institute of Technology (ETH), Zurich Switzerland
09.1995 – 05.1999	Diploma in Physics (MD Physics), Swiss Federal Institute of Technology (ETH), Zurich Switzerland
	Diploma Thesis: "Dose Dependence of Resolution in Microtomography"
	Scholarships from the ETH and the Latvian Maritime Foundation
09.1994 – 08.1995	Exchange studies at the ETH Zurich
	Scholarships from the European Physical Society and the SOROS Foundation
09.1991 – 09.1995	BSc Physics University of Latvia, Riga
09.1991 – 09.1995	Latvian Maritime Academy, Riga Studies at the Faculty of Maritime Transportation
09.1980 – 06.1991	1st Gymnasium of Riga, Latvia Graduated with honors in Latvian literature
09.1980 – 06.1991	P. Juriana Music School, Riga, Latvia Graduated Cello and Bassoon Classes

Professional experience

06.1999 – 06.2003	Research assistant Laboratory for Experimental Audiology, Dept. Otothinolaryngology Head and Neck Surgery, University Hospital Zurich, Switzerland
08.2002 – 02.2003	Programmer (Temporary employment within the Flood team), Swiss Re Zurich, Switzerland

02.2001 – 08.2001	Physics teacher Juventus School Zurich, Switzerland
10.1999 – 02.2000	Teaching assistant Signal and Information Processing Laboratory (ISI ETH)
06.1997 – 07.1999	Measurement assistant Institute of Biomedical Engineering (IBT ETH)
12.1996 – 07.1997	Teaching assistant Institute of Communication Engineering (ETH):
12.1996 – 02.1997	Teaching assistant Power Electronic Systems Laboratory (PES ETH):
06.1994 – 09.1994	Cadet (operation area: Persian Gulf, Indian Ocean) M/T “Ryvingen”, Limmassol Acomarit AG
06.1993 – 08.1993	Cadet (operation area: Baltic See, North See) M/T “Janis Sudrabkalns”, Riga, Latvian Shipping Co
05.1992 – 08.1992	Cadet (operation area: North Atlantic Ocean) M/V “Diplot”, Riga, Latvian Fishing Co