Politecnico di Milano

Scuola di Ingegneria dell'Informazione

Polo territoriale di Como

Master of Science in

Computer Engineering

Characterization of the features of clear speech: an acoustic analysis of the influence of speech-processing settings in cochlear implants

Supervisor: **Prof. Christian Forlani**

Assistant Supervisors: **Prof. Gabriella Tognola**
Dott. Alessia Paglialonga

Master Graduation Thesis by: **Luca Feliciani**
Student Id. number: **731864**

Academic Year 2009 – 2010

Politecnico di Milano

Scuola di Ingegneria dell'Informazione

Polo territoriale di Como

Corso di Laurea Specialistica in

Ingegneria Informatica

Caratterizzazione delle proprietà del clear speech: una analisi acustica dell'influenza delle impostazioni di elaborazione del parlato negli impianti cocleari.

Relatore: **Prof. Christian Forlani**

Correlatori: **Prof. Gabriella Tognola**
Dott. Alessia Paglialonga

Tesi di Laurea di: **Luca Feliciani**
Matricola: **731864**

Anno Accademico 2009 – 2010

A me, alla mia Famiglia

Ringraziamenti

a tutti quelli che mi hanno sostenuto e sopportato

alla Frasba dal Lac splendida scoperta di questi anni a Como

Sommario

La capacità di percezione del suono è un fattore importante per la definizione del benessere personale. La corretta comprensione del parlato è poi fondamentale nelle attività quotidiane e nelle relazioni di ogni giorno. Per questo il perfezionamento delle apparecchiature acustiche per ovviare alle problematiche dei soggetti audiolesi, risulta tra i campi di sviluppo maggiormente rilevanti nelle applicazioni biomediche. I traguardi raggiunti dalle tecnologie attuali riguardanti gli apparecchi acustici e gli impianti cocleari, forniscono gli strumenti per una sempre migliore definizione del suono.

Ma un problema riscontrato nei pazienti che fanno uso di queste apparecchiature, rimane ancora oggi la scarsa naturalezza della percezione del parlato. I vari campi di ricerca e di sviluppo si sono concentrati maggiormente sulla definizione delle tecniche di acquisizione ed elaborazione del segnale, per compensare le carenze percettive, mancando di approfondire quali possano essere le caratteristiche oggettive che rendano il parlato intellegibile. Una nuova branca di ricerca si sta orientando su questo studio andando ad analizzare le caratteristiche temporali e spettrali di quello che viene definito il parlato 'chiaro' (o iperarticolato) che risulta maggiormente intellegibile rispetto al parlato colloquiale comunemente usato.

Questa Tesi affronta la ricerca in letteratura di studi che abbiano preso in esame queste caratteristiche, e definisce un insieme di quelle maggiormente rilevanti per l'incremento dell'intelligibilità. Sulla base delle caratteristiche selezionate, viene poi sviluppato un confronto tra dei segmenti di clear speech (parole e frasi) e dei segnali processati a partire dagli originali, attraverso un simulatore di impianto cocleare.

Questo simulatore riproduce il segnale inviato al nervo acustico, così come viene percepito da un soggetto impiantato. Il confronto delle caratteristiche rilevate, tra il segnale originale e i segnali processati, fornisce una misura della bontà di elaborazione di un impianto cocleare.

I risultati ottenuti dimostrano come questa elaborazione di segnale vada a conservare in maniera ottimale le caratteristiche selezionate, e

definiscono un parametro di ottimizzazione per il simulatore stesso, fornendo una misura del numero di canali di processamento capace di mantenere le proprietà del clear speech. Questo valore risulta essere in linea con le scelte implementative delle aziende produttrici degli impianti cocleari ed è interpretato quindi come una conferma della bontà dell'elaborazione e delle peculiarità delle caratteristiche estratte.

Abstract

The ability to hear sound is an important factor in the definition of personal well-being. The correct understanding of speech is also essential in the daily activities and relationships. For this reason the improvement of hearing systems to address the problems of deaf people, appears among the most significant development in the fields of biomedical applications. The achievements of the current technologies relating to hearing aids and cochlear implants, provide the tools for better sound definition.

The problem in patients who use these devices, still remains the lack of naturalness of speech perception. The various fields of research and development have focused more on defining the technical acquisition and on signal processing to compensate the lack of perception, without insight into what may be the objective characteristics that make speech intelligible. A new branch of research is targeting on this study by analyzing the temporal and spectral characteristics of what is called the 'clear' speech (or hyper-articulated) that is more intelligible than conversational speech commonly used.

This Thesis addresses the research of these characteristics, through the studies in literature, and defines a set of these of significant importance for the speech intelligibility. Based on the selected features, is then developed a comparison between segments spoken clearly (words and phrases) and the signals processed starting from original, through a cochlear implant simulator.

This simulator reproduces the signal sent to the auditory nerve, as perceived by an individual implanted with a cochlear implant. A comparison of the characteristics found between the original signal and processed signal, gives a measure of the quality of the processing of a cochlear implant

The results show how that signal processing optimally preserves the selected features and provides a parameter optimization for the simulator, founding a number of processing channels capable of maintaining the properties of clear speech. This value is in line with the implementation choices of the manufacturers of cochlear implants

and is therefore a confirmation of goodness of the development of the peculiarities of the features extracted.

Table of contents

Figures captions

Tables captions

Chapter 1

Introduction

The correct perception of the sounds that surround us is of fundamental importance, and also is a relevant factor to define our well-being. Hearing impaired people need devices able to improve their listening skills. The tools used to fix or restore the ability of perception are hearing aids and cochlear implants, in which the sound quality is the most important factor of quality assessment. These implants are complex devices that combine many forms of signal processing. The expansion of this market in recent years has led the producers to focus their efforts to the evolution of these digital products, exploiting their operating possibility and limiting their lack of quality. Indeed, despite the efforts and improvements in the prosthesis, the performance is often poor and hearing impairment patients complain of 'lack of naturalness' of the processed speech.

A branch of research to solve this problem, still only partly explored, is based on the study of the characteristics that makes the speech more intelligible, and on their enhancement in the hearing prostheses. The implementation of new algorithms and tools based on the study of the input signal, and more specifically speech, can contribute in improving the intelligibility, making the digital hearing aids more reliable and accurate.

This Thesis wants to intervene in this area of research, going to explore the temporal and spectral characteristics that determine the intelligibility of a speech that is commonly named as 'clear speech’.

From the academic literature available has been make a research on what features can quantify the differences in intelligibility between this style and the most common style of speech, known as 'conversational'. From this assessment, it is extracted a set of characteristics most relevant to prove an effective influence on the best speech perception. In recognition of these features follows the implementation of their extraction from signal segments (sentences, words, syllables of phonemes) by reference to the methods suggested in the studies analyzed to verify the variation of these parameters on the dataset used. Has been then implemented a simple signal processing to simulate the perception of speech in a patient with cochlear implant, and has expanded the original dataset, with the signals processed by the simulator. By extracting the relevant features in speech clear, from these processed signals, and comparing the results with those obtained from the original signal, it was possible to make an assessment of the influence of signal processing in a cochlear implant, for the features defined as important in preserving the clear speech intelligibility.

This experiment covers an area of study yet unexplored concerns the influence of signal processing in hearing implants on speech perception, according to an analysis of temporal and spectral characteristics, and not only in the perceptual The results obtained can used as a significant detection of the relevance of the extracted features, as well as the goodness of the cochlear mortgage, and pose a basis for the design of acoustic devices that can significantly improve speech perception.

1.1 Motivation and goals

There are many studies and many approaches proposed, aimed at improving the perception of sounds, particularly speech, for patients with hearing aids. Some approaches focus on the acquisition of the signal by studying the best configuration of the microphones used on the instruments and their calibration and interaction, or the management of gains. Other measures relate to the processing of input signal to reduce background noise and potential feedback, or in the compression and expansion for frequency bands considered most influent on the perception of the signal.

Another idea recently taken into account for the perception improving is based on the study of the definition and processing of speech perceptual features, for the aspects concerning the lack of naturalness of the signal received by the hearing aids users.

This new open research scenarios aim to study the speech signal, exploring in detail both the temporal and spectral characteristics, verifying the intelligibility difference on normal-hearing and hearing-impaired listeners. In particular, speech intelligibility is highly influenced by the so-called ‘speech style’. In everyday communication different speech styles are used more or less consciously, depending to the talkers and the environment condition. For instance: in presence of ambient noise there is an increase in vocal level; in a small room, speech can be produced faster than in an auditorium; a relaxed speech is significantly different from emotional speech. Trough the study of speech characteristics we can distinguish several styles that differ by some features. Usually we talk about ‘conversational’ speech as the style used naturally, in quiet environments with people with normal hearing ability and knowledge of the language, but there are some other different speaking styles used in different situations: the style used with young people is known as ‘infant-directed’ speech or ‘motherease’, the one used to give a command to a machine is called ‘computer-directed’ speech, the tendency to increase different features of the voice when speaking in noise is called ‘Lombard’ speech, and ‘clear’ speech (e.g. hyper-articulated) is the one used in some particular situation as with hearing non mother tongue people, or in order to be better understood by listeners who are moderately impaired in their ability to understand speech, due to a hearing impairment.

As defined by Smiljanic and Bradlow (Smiljanic & Bradlow, 2009), “*Clear speech is an intelligibility-enhancing speaking style that talkers naturally and spontaneously adopt when listeners have perceptual difficulty due to, for instance, a hearing loss or a different native language.”*
Previous analysis have demonstrated a significant difference in the level of intelligibility between different styles, showing increases in the range of 10–20 percentage points for clear (CLR) speech over conversational (CNV) speech for a range of speech materials and listening conditions and for a variety of listener populations (Liu, Rio, Bradlow, & Zeng, 2004), for both normal-hearing and hearing-impaired listeners in every environment (average clear speech gain was 20 and 26 percentage points for normal-hearing and hearing-impaired listeners, respectively), demonstrating that sentences spoken “clearly” are significantly more intelligible than those spoken “conversationally” (Picheny, Durlach, & Braida, 1985; Payton, Uchanski, & Braida, 1994; Uchanski, Choi, Braida, Reed, & Durlach, 1996)

The reviewed studies focus exclusively on the overall influence of these characteristics on intelligibility, making comparisons between portions of clear speech against portions of conversational, or specifically analyzing the variation of a given feature on the perception of speech. From the resulting data can be derived a general definition of the differences between the two types of speech, very useful in identifying possible interventions on the signal, but not for a specific scope. No study focuses on the evaluation of the theory of clear speech applied to an area of considerable interest and utility, which is that concerning the hearing prosthesis, analyzing the modification introduced by these devices on the characteristics observed.

The aim of this Thesis, however, is precisely to fill this gap, by performing an initial analysis of the impact of the signal processing carried in a cochlear implant, to those characteristics that are fundamental in maintaining the perception of 'clarity' of clear speech, and then in its intelligibility.

It was then carried out a search of the characteristics of the clear speech, compared to conversational speech that results more influent on intelligibility. The studies regarding the temporal and spectral characteristics of speech are assessed, both for sentences and words, both in its smaller components, such as segments, syllables and vowels and consonants. A selection of most relevant and effectively verifiable characteristics was examined and was used for the implementation of an extraction toolbox, from a dataset of words and phrases spoken clearly by an Italian speaker.

These features represent a parameter for the detection of the intelligibility of speech signal. To this end, was implemented a cochlear implant signal simulator which is used to process the original dataset, changing their parameters, in order to obtain the original and degraded signals and verify the reliability of features previously noted.

The signals created by the simulator are, as the original ones, subjected to the extraction of features detected. The results of these calculations are compared to determine the goodness the analysis made, and also to check prior knowledge about the processing of the signal through a cochlear implant.

The accurate definition of the parameters can then be used in a future implementation of a deteriorated signal simulator which approximates the signal received by hearing impaired listeners and carriers of hearing aids, to be used to reduce the number of audiometric tests on patients in studies on the perception of speech intelligibility, thus limiting also the difficulties of interpretation of the results arising from the inevitable difference of the ability to listen to each patient, and subjective evaluations.

The numerous audiological and non-audiological variables, which may interact in complex ways for different patients, in fact, make it difficult for the clinician to predict benefit of a tested approach. This research can be useful even to understand how to quantify the influence of these factors and how they interact with one another for individual patients.

As a further development this application could provide a confirmation of the studies considered in the evaluation of perceptual characteristics of clear speech intelligibility, setting the stage for a definition of a signal processor that increases the perception of the goodness of a conversational speech, bringing it closer in that of a clear speech.

1.2 Original contributions

This thesis studies the contribution of different features extracted from a signal represented a clear speech, selected from the studies on the sector, as those most influential on intelligibility. The parameters and the values that define the levels of influence are studied separately and for each feature was implemented a specific algorithm. This approach allows a refined assessment of each feature taken separately, and an overall assessment.

The examination which the features are subject to is a comparison between original signal and the one processed through a cochlear implant simulator. These signals are by definition less intelligible as the original but the study on features allows checking which of these are more affected by the process and then on which one can define a course of action to improve simulation algorithms, to give users an acoustic phrotheses that affect the intelligibility to a lesser extent, thus improving the performance.

This type of experiment has not yet been made by any study and the results can provide a basis for further knowledge of these characteristics, as well as the development of various fields of application concerning the signal processing in hearing aids.

1.3 Thesis outline

The text of the thesis is structured as follows.

In Chapter 2 is presented the hearing aids and cochlear implants state of the art, specifying the method of operation and areas of intervention for improving their performance.

In Chapter 3 provides a detailed definition of the clear speech providing references and clarifying the considerations that make it more intelligible than to a conversational speech. In Section 3.2 define the temporal and spectral properties, indicating the differences and improvements introduced on the characteristics examined.

Chapter 4 shows the implementation of the selected features as suggested in the studies reviewed, and as revised for the purposes of argument, defining the toolbox of extraction, the Clear Features Extractor (CFE).

Chapter 5 defines the material used and the test protocol needed to display the results. Section 5.1 describes the speech material used. Section 5.2 describes the implementation of the cochlear implant signal simulator, called Cochlear Implant Speech Generator (CISG) that is used to process the original signals. Section 5.3 describes the segmentation of which must be submitted for input to be passed to the CFE and the Section 5.4 defines the analysis of the obtained segments. Section 5.5 describes the protocol of analysis for the comparison of results from the extraction of the characteristics from original signal and processed signals.

Chapter 6 shows the most relevant results of the work.

Chapter 7 discusses the results obtained, highlighting the peculiarities of the features and providing suggestions about the characteristics analyzed, and present the conclusions.

Chapter 8 finally closes the thesis, giving a view on possible future developments.

Chapter 2

Hearing aids and cochlear implants

2.1 Hearing aids

A hearing aid is an electronic device that amplifies sound, used by people with impaired hearing. The device consists of a microphone, a battery power supply, an amplifier, and a receiver. The microphone receives sound waves directed toward the person with hearing loss, then converts the sound waves to electrical impulses that are amplified with the aid of the power supply, and the receiver converts the electrical impulses back into sound vibrations.

In the early sixties analog hearing aids with electronic components were put on the market, in which michrophone, amplifier, battery and receiver was contained in a box of 10 cm side. The main problems of these analog devices were due to their dimension and the few possibility of signal processing.

The first digital products came to market in the late '80s introducing the Digital Signal Processing (DSP), obtaining a poor commercial success, mainly due to the overall size of the devices and the short battery life. Over time, hearing aid technology has evolved and has been increasingly appreciated in the market, in its major forms, ranging from behind-the-ear to completely-in-the-canal, as shown in Figure 2.1

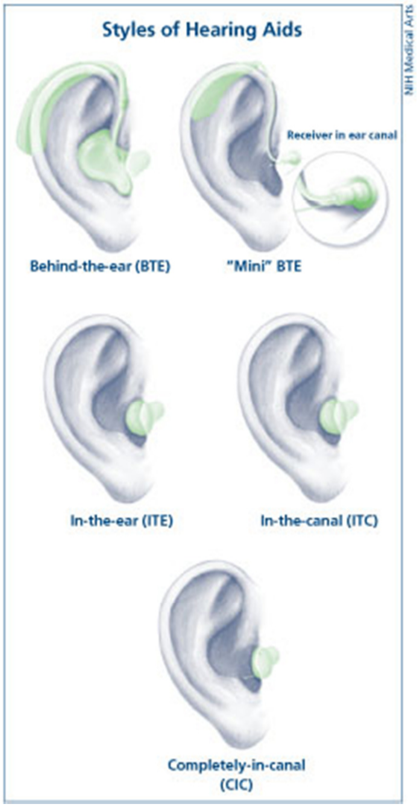


Figure 2.1: Different types of hearing aids on the market. From the Behind-the-ear to the Completely-in-canal the benefit clearly representating in this figure is the size reduction. Taken from NIH Medical Arts

But the success of these devices is not only due to the reduction of their size, the introduction of digital technology has allowed to increase the processing capability end enhancement of the speech signal, supplying to the inevitable deterioration that the analog-to-digital conversion introduce.

Important researches have concentrated on the use of DSP introduced in digital apparatus as signal generators or as managers of the adaptation of the microphone input signal. For example, a method that can be applied is to detect the various components of the signal and to synthetize their reproduction, so as to eliminate the noise. This methodology involves a capacity of automatic recognition and synthesis of the segments themselves, not yet defined at a level which can be used on acoustic equipment, working in real time.

Another utilization of the DSP in the hearing aids is concentrated on the noise reduction. For example, one approach to reduce the influence of noise in the acoustic equipment includes a signal processing in its temporal and spectral components. The simpler mechanisms in noise suppression provide a subtraction in the frequency bands where the noise occurs, or a phase cancellation. Both of these types of intervention affect, however, markedly on the signal: the band removal affects even the useful signal without improving the signal to noise ratio, while the phase cancellation requires the knowledge of the exact waveform noise, not only its spectral properties.

Another approach is to increase for properties of speech, for example by detecting the presence for vowels and acting on their spectral properties by changing the power balance, thereby making more noticeable these segments. Such a development does not take in account for changes to the signal by acting only on parts of it that are not necessarily helpful to improve the intelligibility.

A non-exausting list of benefits associated to the introduction of digital can be summarized as follows as exposed by Ricketts (Ricketts, 2009):

- *Gain Processing*, relating to a greatly increased flexibility and control of compression processing of the input signal gain and the introduction of expansion, the opposite of compression, to reduce theintensity of low-level environmental sound;
- *Digital Feedback Reduction*, i.e. the use of advanced feedback cancellation and notch filters to reduce and eliminate feedback caused by the proximity to objects and movements of the jaw;
- *Digital Noise Reduction*, i.e. the reduction of stationary noise when detected, by filtering specific bands, with the possibility of improving the intelligibility and also to reduce even indirected noise to the microphones;
- *Directional microphones* and DSP, the use of directional microphones and their calibration by DSP with the possibility to modify sensitivity and directionality depending on the situation;
- *Digital Hearing Aids as Signal Generator*, relating to the ability of generate sound signal by the DSP that can be used for testing hearing ability and adapting it to the characteristics of the patiencte;
- *Digital Speech Enhancement*, representing the possibility to increase the caracteristics of relative segments of sound, analizing it in both temporal and spectral domain.

An example of gain processing is shown in Figure 2.2 in which two simple curves for the recovery of flat and high frequency hearing loss are represented. For each patient similar curves must be designed dependently by their response to stimuli tests required to verify the specific hearing loss.

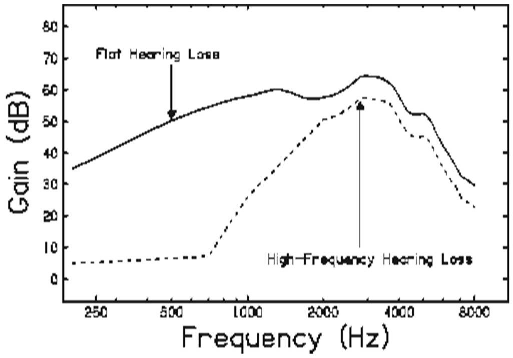


Figure 2.2: Gain Processing curves. In this example, gain ranges from 40-65 dB for a relatively flat hearing loss (solid line). Very little gain is provided for the low pitches in the case of a high pitch hearing loss (dashed line).

A schematic implementation of the state of art of these signal processing blocks on a high-end hearing aid is presented in Figure 2.3 in which is shown a simple classification system, used to extract and select particular features that guide the processing of the signal.

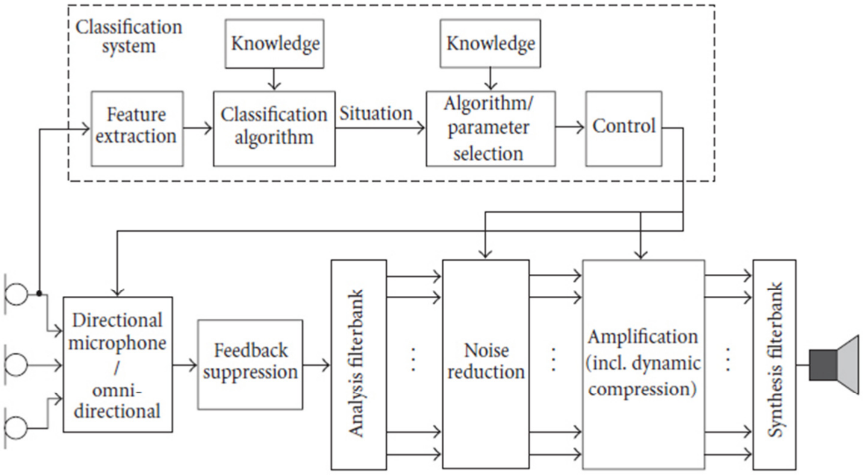


Figure 2.3: Processing stage of a high-end hearing aid. The signal acquired by the directional microphones is processed to suppress eventual feedback and a noise reduction with a specific amplification of useful bands is computed using parameters that derive from a classification system, that recognize the signal and extract the right features information. Finally the signal is resynthesized and emitted into the ear. Taken from (Hamacher, Chalupper, & Eggers, 2005)

2.2 Cochlear implant

A cochlear implant is an electronic device that is surgically implanted into the cochlea of a deaf individual. A transmitter placed outside the scalp sends signals to a receiver under the scalp, which in turn transmits an electrical code to the auditory nerve.

A microphone is located behind the ear to collect the sound waves that are transmitted through a speech-processor. The speech-processor analyzes the sound waves and relays data back to electrodes in the implanted device. The patient receives electrical pulses at the level of the auditory nerve fibers that can be distinguished as hearing sensations. Although the implant does not transmit speech in the same manner as it would be perceived by a person with normal hearing, it allows the individual to perceive and distinguish sounds that would not otherwise be audible to him or her and to use those sounds along with other environmental cues to improve communication.

So a cochlear implant is an artificial electronic ear that can restore hearing in profoundly deaf people, and is used when the hearing aids do not get the desired result (i.e. when the cochlea is damaged). It is also defined as ‘artificial cochlea’, or ‘bionic ear’ and it is a tool that replaces the pathological cochlea by sending directly to the auditory nerves electric stimuli that represent language and ambient noise.

Figure 2.4 highlights the common components for all models of cochlear implants: a directional microphone captures the sound signals, a processor compute from them a spatial-temporal pattern of stimulation of nerve fibers, a transmission system communicate to the implant different parameters and stimulation modes, the implant then adjusts the electrical parameters of electrodes placed along the cochlea in order to produce the desired pattern. Electrodes are placed along the cochlea according to the tonotopic distribution of frequencies: electrodes at the base convey information of high-frequency sounds and, vice versa, electrodes at the apex convey information of low-frequency sounds.

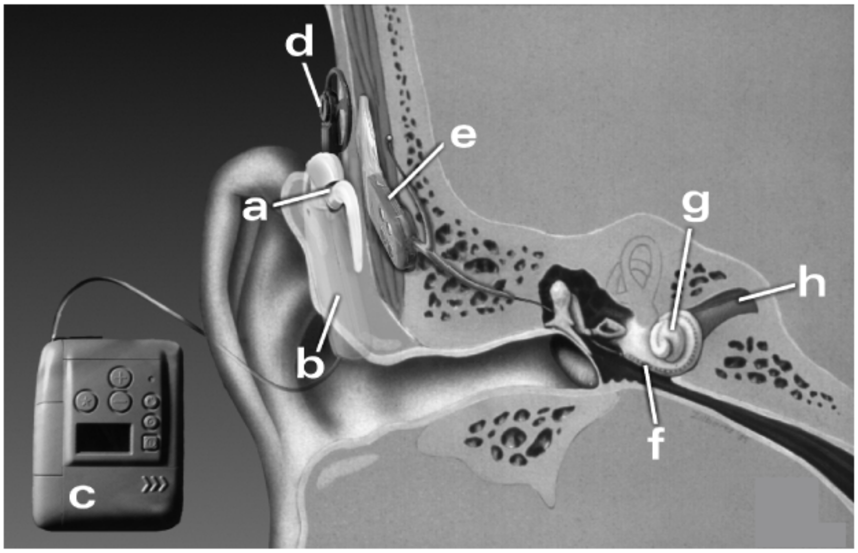


Figure 2.4: Typical main components of a cochlear implant (Nucleus®, Cochlear Corporation): (a) microphone; (b) speech processor 'behind the ear' (ESPrit); alternatively: (c) speech processor 'body worn' (SPrint); (d) Transmitting coil; (e) receiver / stimulator, (f) array of electrodes, (g) cochlea, (h) the auditory nerve.

As with hearing aids, the use of digital technology allows for very large data processing capability. In addition to the aforementioned benefits and areas of research open to hearing aids, in cochlear implants is the signal processing which contributes essentially for proper operation. In modern systems different strategies for signal processing are applied, and the approaches used most recently for hearing on the market are all based on a filter-bank and can be related to the four described below, although on different devices can have different trade names:

- CA (Compressed Analog) or SAS (Simultaneous Analog Stimulation) as it evolves: the acoustic signal is first compressed by the AGC (Automatic Gain Control), approximating the dynamic range of response to electrical stimulation, after which the signal is filtered in bands applying a contiguous frequency independent gain for each band. The waveforms resulting from the filtrating are then sent simultaneously to the electrodes (working in monopolar configuration), in a similar manner. The pattern of stimulation obtained with CA and SAS strategies still holds, after processing, most of the information of the un-processed signal, including fine structure, allowing the brain to receive and interpret signals minimally processed. However, this simultaneous stimulation (necessarily monopolar) results in a strong interaction between the stimulation channels which can create distortions in the information spectrum and degrade or compromise the understanding;

- CIS (Continuous Interleaved Sampling): in contrast to CA, the CIS generates at the interface with the nerve, non-simultaneous biphasic impulses. The amplitudes of the pulses are modulated by the envelopes of the signals processed by a bank of band pass filters. The envelopes are extracted by rectifying and low pass filtering (typically with cut to 200 or 400 Hz) by applying the same filtering AGC. The pulse trains are sent to electrodes at a repetition rate ranging from ~200 to ~1000 pps (pulses-per-second), constant in time and for each channel. In the CIS, N filters are used (typically 8-12), that correspond to N stimulating electrodes
- SPEAK (Spectral Peak): from the envelope of the signal, determined as described above for the CIS but on a higher number of filters (typically 22), the strategy makes a further compression stage, selecting (for each cycle of stimulation) only the envelopes with the highest energy (above a threshold) and then stimulating a subset of available channels (typically 6-10). The number of stimulation channels actually used is not constant and depends on the spectral content of the signal under test and the threshold that governs the selection. As a result of this variable number of sites of stimulation, also changes the repetition rate of electrical impulses, creating adaptive manner in the best compromise between spectral and temporal resolution.
- N-of-M. Very similar to the SPEAK, it differs essentially from the fact that the selection of maximum spectral produces a fixed number (N) of channels of stimulation among the M available. N is a clinical parameter settings based on patient characteristics, the pulse repetition rate is then fixed once N is set.

Figure 2.5 is shown the processing of a signal in a cochlear implant using the CIS strategy, the signal is divided in sub-bands that matches the cochlear frequency, and for each sub-band a train of impulses is generated. The relative amplitudes of the current pulses delivered to the electrodes reflect the spectral content of the input signal. For instance, if the speech signal contains mostly high frequency information (e.g., /s/), then the pulse amplitudes of the fourth channel will be large relative to the pulse amplitudes of channel 1-3.

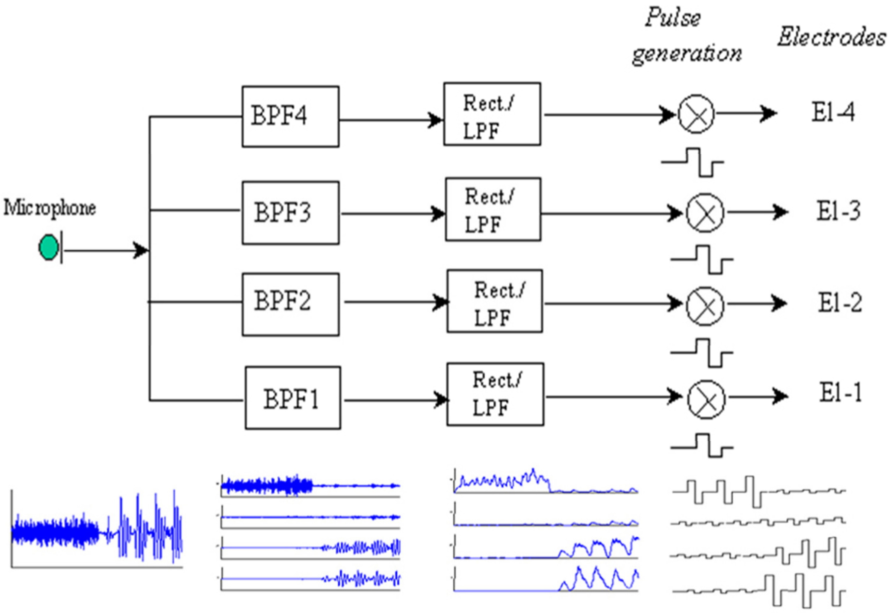


Figure 2.5: A simplified implementation of the CIS signal processing strategy using the syllable ‘’sa’’ as an input signal. The signal first goes through a set of four bandpass filters that divide the acoustic waveform into four channels. The envelopes of the bandpassed waveforms are then detected by rectification and low-pass filtering. Current pulses are generated with amplitudes proportional to the envelopes of each channel and transmitted to the four electrodes through a radio-frequency link. Note that in the actual implementation the envelopes are compressed to fit the patient's electrical dynamic range. Taken from (Loizou, 2006)

The success of cochlear implants can be attributed to the combined effort of scientist from various disciplines including bioengineering, physiology, otolaryngology, speech science, and signal processing. Each of these disciplines contributed to various aspects of the design of cochlear prostheses. Signal processing, in particular, played an important role in the development of different techniques for deriving electrical stimuli from the speech signal.

The signal processor in a cochlear implant is responsible for breaking the input signal into different frequency bands or channels and delivering the filtered signals to the appropriate electrodes. The main function of the signal processor is to decompose the input signal into its frequency components. The designers of cochlear prosthesis are faced with the challenge of developing signal-processing techniques that mimic the function of a healthy cochlea. Designers of cochlear prosthesis were faced with the challenge of developing signal-processing techniques that would mimic the function of a normal cochlea.

Chapter 3

Acoustic analysis of Clear Speech

3.1 What is clear speech?

The production of clear speech is guided by some ‘rules’ or assumptions that characterize this style, as the request of “read the materials as if you were talking to someone who is hearing impaired, not a native speaker of your language” or “speak clearly and precisely”. Usually a speaker that wants to make his speech clear, articulates all the phonemes precisely and accurately, reduces his speech rate, increases slightly the pauses between phrases, and modestly increases vocal volume (Uchanski R. M., 2005). Following these rules the speaker can produce a speech that is more intelligible and, as demonstrated by the studies of Picheny et al (Picheny, Durlach, & Braida, 1985), Payton et al (Payton, Uchanski, & Braida, 1994) and Uchanski et al (Uchanski, Choi, Braida, Reed, & Durlach, 1996), possesses some peculiar acoustic features.

An example of conversational and clear speech is shown in Figure 3.1, in which is evident the differences in duration and an increase in the length of each phoneme and a higher definition of their spectro-temporal caracteristics.

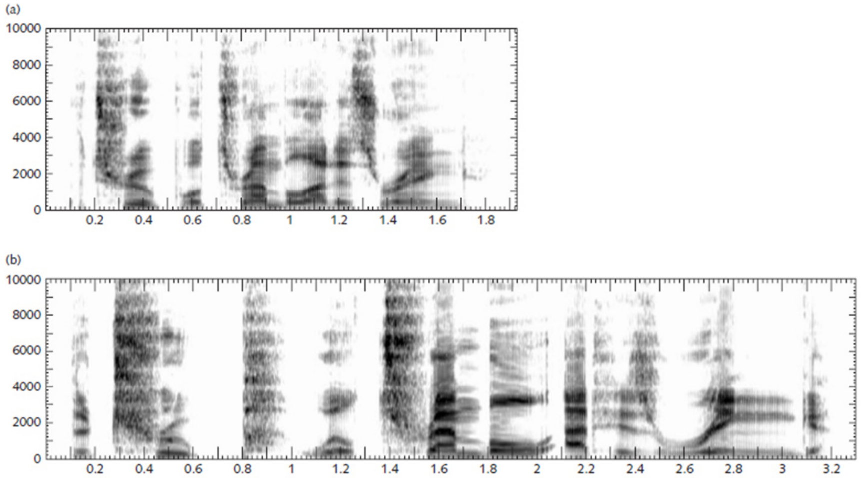


Figure 3.1: Spectrogram of the sentence “The troop will tremble at this ring”, spoken by a male talker in both a) conversational and b) clear speaking style. Taken from (Uchanski R. M., 2005)

The main features that are peculiar of clear speech and that contribute most to its improved intelligibility with respect to conversational speech will be reviewed and discussed later in full detail.

As an example, some differences that can more easily be observed between clear and conversational speech are briefly recalled here and have been detected in studies carried out by Picheny et al (Picheny, Durlach, & Braida, 1986) and Krause and Braida (Krause & Braida, 2009). For example, the number and types of phonological phenomena are different in clear and conversational speech. In colloquial speech, the vowels are changed or reduced, and the stop burst (release of air following plosive consonants) are often not issued. Conversely, in clear speech the vowels are changed to a lesser extent, and stop bursts are always released. Moreover, the intensity of obstructive sounds, especially stop consonants, is much greater (up to 10 dB greater) in clear speech than in conversational language.

Finally, the most obvious factor: a speaker speaks more slowly using a clear speech style, than when he speaks in a colloquial one. This difference is not only characterized by the increase and lengthening of pauses (as in a speech just spoken more slowly), but the duration of individual speech sounds is also increased. This increase is not entirely uniform, and is a function of the identity of the phoneme and its acoustic environment associated. Because of this difference, speech is more clearly understood by both normal-hearing listeners and listeners with hearing loss (in the study by Picheny et al (Picheny, Durlach, & Braida, 1985),the average increase in recognition performance was 20 and 26 percentage points for normal hearing and listeners with hearing loss, respectively).

A study of Krause and Braida (Krause & Braida, 2004) employed “natural” clear speech, training talkers to produce clear speech with the same speaking-rate as conversational speech. The talkers were able to produce “fast” clear speech that had the same speaking-rate as conversational speech. Perceptual results still showed significantly high intelligibility for “fast” clear speech than same-rate conversational speech. This result is shown in Figure 3.2 in which the temporal difference from conversational and clear speech is not visible, but some spectral difference is still noticeable.

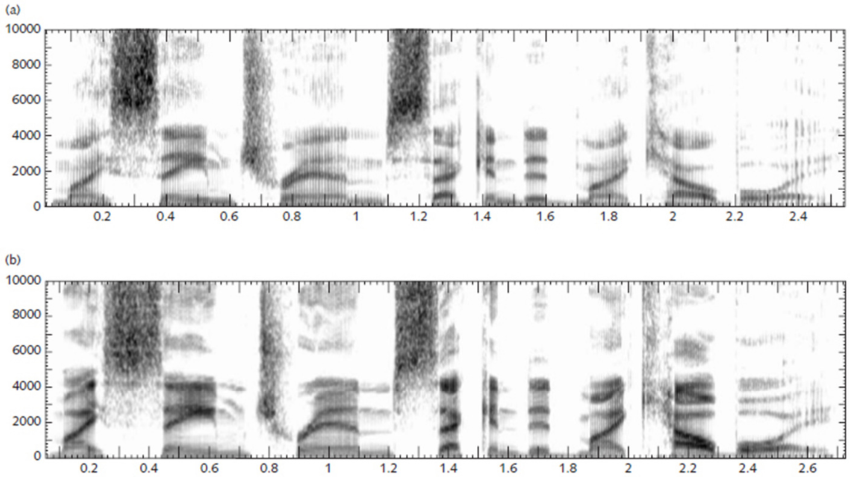


Figure 3.2: Spectrogram of the sentence "My sane trend seconded my cowboys" spoken by a male talker in a) conversational speaking style and b) clear speaking style at normal rate. Taken from (Uchanski R. M., 2005)

This suggests that clear speech has some inherent acoustic properties, independent from speaking of rate, that contribute to its improved intelligibility. The next section will identify, describe and analyze in detail the different acoustic properties, both temporal and spectral, of clear speech, approaching their relations with speech intelligibility. These characteristics are analyzed in detail.

3.2 Acoustic analysis

There are many acoustic characteristics of speech that could be considered as relevant to clear speech, because clear and conversational speech vary on a wide range of acoustic and phonetic dimensions.

The relative importance of each characteristic, in its contribution to the intelligibility of speech, is not equal and there is no strong evidence yet available that can determine which single factors are the most important for the differences in intelligibility. In the present work, the selection of the parameters included in the analysis was done taking into account how important each feature and parameter is in relation to intelligibility, also considering the tendency for many of the acoustical characteristics to co-vary, thus often making it difficult to rigorously determine the relationship between intelligibility and each single acoustics characteristic (Picheny, Durlach, & Braida, 1986).

One usual and useful approach to this problem, proposed first by Picheny et al (Picheny, Durlach, & Braida, 1986) and the adopted by Krause and Braida (Krause & Braida, 2004; 2009), Drullman et al (Drullman, Festen, & Plomp, 1994a; 1994b; Drullman, 1995), Liu et al (Liu & Zeng, 2006) and Greenberg and Arai (Greenberg & Arai, 2004), is to develop a model for the production of the acoustic speech waveform that permits the synthetic manipulation of parameters that are thought to contribute to intelligibility increase, and that can be varied independently. Following this approach, these authors defined a selection of parameters in both the time domain and the frequency domain. These parameters were chosen considering the difference between the clear and conversational speech, at different levels of detail:

i) *global*, i.e. the temporal and spectral macro changes;
ii) *phonological*, i.e. the feature changes of phonemes that occur when a word is actually spoken in a sentence; and
iii) *phonetic*, i.e. the changes in the acoustic properties of individual sounds, such as for example the ratio between the power of a consonant and his follower vowel, or the changes in vowel formants.

The study of the main effects of these characteristics on speech intelligibility can be the base for different studies regarding their role and the possible degree of interaction among them.

Taking into account the classification and results by Krause and Braida (Krause & Braida, 2009) and Picheny et al (Picheny, Durlach, & Braida, 1986), in this study the parameters and features of clear speech were separated in two different macro categories:

i) *temporal measures*, which relate to variations in the modulation waveform and phonological features of speech in the temporal domain;
ii) *spectral measures*, involving changes in frequency content and energy of the speech signal, either in its entirety or in its different phonetic components.

3.2.1 Temporal features

A complete features analysis for the clear speech must take in account their temporal properties that have attracted the interest of researchers, for three main reasons as explained by Rosen (Rosen, 1992). First, form psychoacoustic studies, there is a general consensus that place-frequency mechanisms on their own cannot account for many aspects of the perception of pitch, and by implication the perception of intonation in speech. Second, temporal information is important both for the perception of melodic pitch and for the auditory representation of spectral shape as suggested by theoretical models derived from physiological evidence. Third, Rosen (Rosen, 1992) report the information of an evident success of the large number of patients who have received single-channel cochlear implants. Such systems deliver an electrical signal based on the speech waveform to a single electrode placed in or near the cochlea, this allowing no pace-based frequency analysis. Therefore, this approach at the speech signal want to be a complementary one to the standard Fourier based spectral approach.

3.2.1.1 Fluctuation based grouping

The analysis of the structure of speech, as suggested by Rosen (Rosen, 1992), can approach the signal as a composition of three main temporal features based on dominant fluctuation rates: envelope, periodicity and fine structure. The subdivision of the temporal features takes in account this approach and permits to expose the relevant characteristics of each fluctuation group.

Envelope

The envelope of speech is defined as the fluctuation in amplitude at rates between 2 and 50 Hz. As an example, the envelope waveforms extracted from six sentences pronounced by a male speaker, are shown in Figure 3.3. It's possible from this representation, to recognize the different segment of each speech signal.

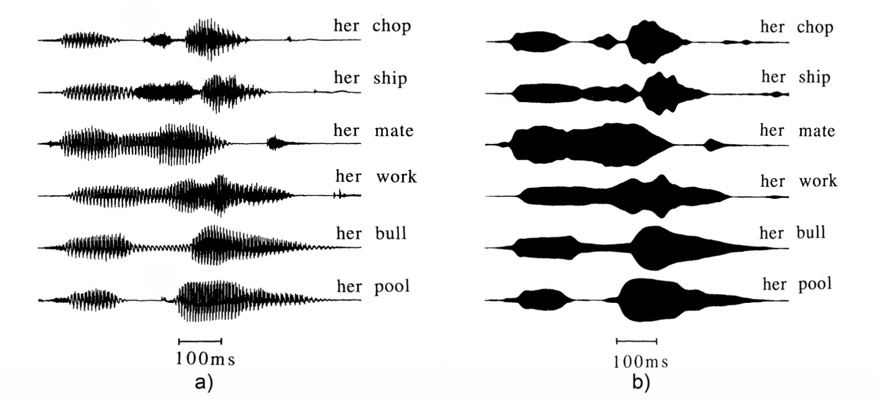


Figure 3.3: a) Speech pressure waveforms of six phrases uttered by a male speaker, b) envelope waveforms obtained from those of figure (a) by full-wave rectification and low-pass filtering at 20 Hz. This processing allows to preserve much of the envelope information and, also, to eliminate that occurring at high fluctuation rates, as can be observed in the loss of the release bursts in the envelope waveforms of 'chop' (top waveforms) and 'pool' (bottom waveforms). Taken from (Rosen, 1992)

The temporal envelope cue, as demonstrated in different studies (Houtgast & Steeneken, 1985; Drullman, 1995; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), has been shown to be sufficient for speech recognition in quiet. These cues convey linguistic information about manner of articulation, voicing, vowel identity and prosody. More specifically, the speech envelope can be described by a set of acoustic features that can be correlated to auditory properties:

- *intensity* (loudness), that can be defined as the mean amplitude of the envelope;
- *duration* (length), that represents the time of sustain of the envelope;
- *rise time* (attack), defined as the time interval from envelope onset, when the envelope reaches the first maximum in of its amplitude;
- *fall time* (decay), defined as the time interval from the last maximum of the envelope amplitude, to zero.

For clear speech it is possible to notice a greater intensity of the sound pressure in the range of 5-8 dB, measured at the time of recording (Picheny, Durlach, & Braida, 1986), naturally given by a greater attention paid by a speaker who wants to speak 'clearly'. Also, Picheny et al in the same study measured the average speaking rate and found a decrease from 160-200 words/min in conversational to 90-100 words/min in clear and then an increase in duration. Specific research on the influence of temporal envelope on speech intelligibility demonstrated that smearing the envelope by low pass filtering affects the intelligibility of consonants more than the intelligibility of vowels (Drullman, Festen, & Plomp, 1994a; 1994b). In the same research Drullman et al measured that at a critical SNR, intelligibility is virtually unaffected by envelope filtering when modulations either above 16 Hz or below 4 Hz, are reduced. Similar results for the 4 Hz limit were found by Xu and Pfingst(Xu & Pfingst, 2008). Thus, throughout the different studies, the optimal crossover modulation frequency for speech, that seems crucial to provide good intelligibility, lies in the range of 8 – 10 Hz (Drullman, Festen, & Plomp, 1994b).

Periodicity

Periodicity of speech gives information about the source of excitation in speech production. The speech sounds can be differentiated for their in:

- *periodic*, which is produced by the vocal cord and the vocal tract and typically, includes the vowels and the voiced consonants. The fluctuation of this sounds can be defined as ranging between 50 and 500 Hz;
- *aperiodic*, which is produced without a vocal component and it's dominated by noises, includes some consonants as the plosive and fricatives. These signals fluctuate at rates from a few kHz up to 5-10 kHZ.

An example is shown in Figure 3.4, in which are exposed the waveforms of different consonants at the same time scale. The periodicity of the consonant /m/, /w/ and /b/ is recognizable by the particular waveform fluctuation, completely absent in aperiodic sounds.

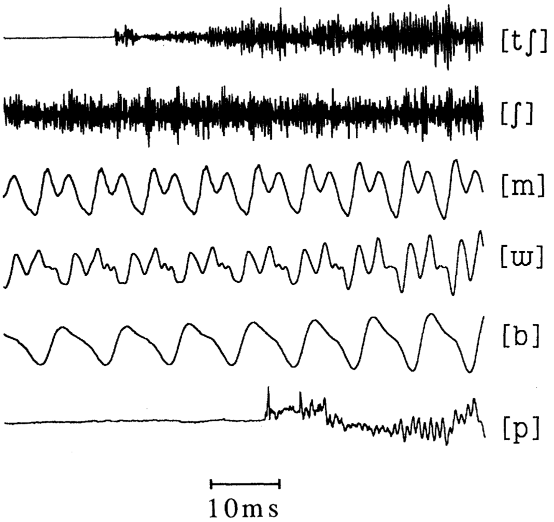


Figure 3.4: Examples of portions of the speech pressure waveform previously shown in Figure 3.1, chosen so as to illustrate periodicity information. In this time scale, the regularity in voiced sound /m/, /w/ and /b/ (periodic), and the irregular waveform in the other voiceless sound (aperiodic) can be easily observed. Taken from (Krause & Braida, 2004)

As exposed by Rosen (Rosen, 1992), periodicity conveys information about voicing and manner, and intonation and stress. Changes in the rate of periodic fluctuation are reflected in changes of fundamental frequency that characterize the speaker. Smilljianic and Bradlow (Smiljanic & Bradlow, 2009) have found that clear speech have more stable global temporal properties, that can be associated with the stability of the vocal sound fluctuation.

Fine structure

The fine structure of speech acoustically informs about the spectrum (amplitude and phase) and contains the formant pattern of the spoken sound. Its fluctuation varies between 600 and 10000 Hz. It gives information about articulation and vowel quality, and about voicing and manner. An example of the composition of the fine structure is presented in Figure 3.5 comparing the fine structure fluctuation with some lowest band of the same waveform.

Drullman et al (Drullman, Festen, & Plomp, 1994a; 1994b) measured the intelligibility scores of speech, in the case of complete suppression of the envelope modulation, and demonstrated that the fine structure alone supplies insufficient information for the recognition of the speech.

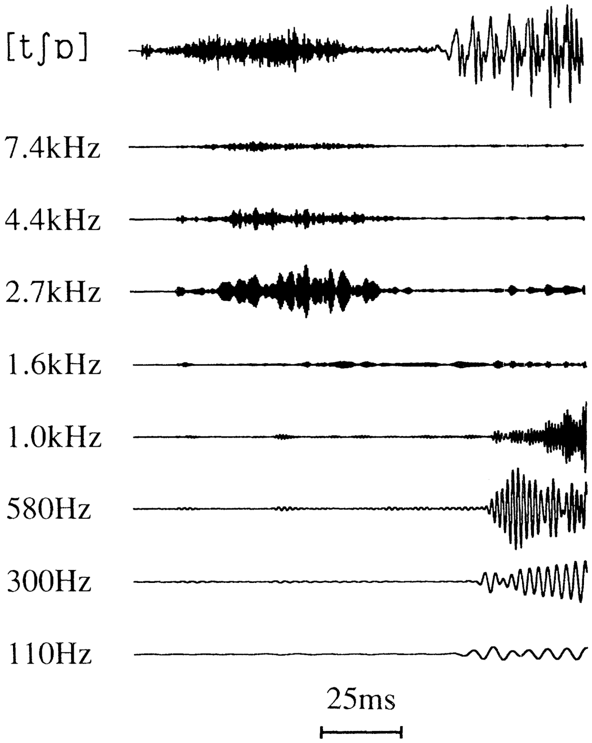


Figure 3.5: Examples of portions of the speech pressure waveform of the initial part of word 'chop', chosen so as to illustrate fine structure information. Note that the envelope features relating to the release burst, rise time and duration of the frication can only be found in fine structures frequencies (>600Hz), and not in the lower ones. Taken from (Rosen, 1992)

A significant result in the analysis of this structure of the speech was reached by Drullman (Drullman, 1995) and Liu et al (Liu, Rio, Bradlow, & Zeng, 2004; Liu & Zeng, 2006) who concluded that a better encoding of temporal envelope (2:50 Hz) and fine structure (500:10000 Hz), improve speech perception in quiet and noise respectively. Drullman (Drullman, 1995) made more accurate observations of these fluctuation and shown that peaks amplitude are most important than toughts, for intelligibility.

3.2.1.2 Phonetical based grouping

A useful categorization, that permit to evaluate the influence of the different features of speech on different psychoacoustic measures, is the one proposed by Picheny et al (Picheny, Durlach, & Braida, 1986) that divide the examined features in global, segmental and phonological properties;

Global

They are the general characteristics of speech that can be observed in an entire sentence. They include:

- *speaking rate*, represent the number of words per minute;

- *pause frequency* and *duration*, measured counting the occurrence and the average duration in a sentence;
- *temporal envelope modulations*, i.e. the fluctuation of the waveform in a segment of a sentence with respect to the intensity mean.

Picheny et al (Picheny, Durlach, & Braida, 1986) found lower speaking rate and an increase of pause frequency and duration. In the same study was measured an increased modulation depth in the temporal envelope for frequencies below 3-4 Hz, result that was confirmed by Krause and Braida (Krause & Braida, 2009). Greenberg and Arai (Greenberg & Arai, 2004) measured the intelligibility varying the low frequency modulation and founded that intelligibility depends on the integrity of low frequency modulation spectrum that lies below 8 HZ, and that frequencies under 1.5 Hz play an important role in decoding the signal under reverberant conditions. For the lower frequency region Smilijianic and Bradlow (Smiljanic & Bradlow, 2009) obtain in their study, higher peaks in 1 to 3 Hz region, and strong component in 2 to 4 Hz region.

Segmental

They are the characteristics of different segments of a sentence, like words or syllables or phonemes. They include:

- *segment duration*, it is the time of an analyzed segment. As discussed before, in clear speech an increment in duration of each phoneme is notable.
- *Voice Onset Time* (VOT), is the length of time that passes between when a stop consonant is released and when voicing begins
  sound insertion, occurs exclusively in clear speech content words.

As results from the study of Picheny et al (Picheny, Durlach, & Braida, 1986) , replaced and confirmed by Krause and Braida (Krause & Braida, 2009) and Smilianic and Bradlow (Smiljanic & Bradlow, 2005), the segment duration of vowels is strongly associated with increased intelligibility.

In Figure 3.6 are explained the VOT for different plosive consonants, the acoustic events of the consonant-vowels (CV) composed of a stop consonant followed by a vowel are the following form:

i) a period of silence preceding the release of the consonant in which pressure is built up (*closure*) in which the oral cavity is blocked completely at a certain place (e.g. for [b], the lips are closed to block the oral cavity; for [d], the tip of the tongue touches the part above the upper teeth (alveolar ridge) to create a blockage).;

ii) the release of the constriction at which fricative noise is generated (*blockage*) in which the oral cavity is held blocked, as air from the lungs continues to come into the cavity. Therefore the air pressure inside the cavity increases. It is typically followed by a period of aspiration for voiceless stops;

iii) onset of voicing for the following vowel (*release*) in which the blockage is released. Since the air pressure inside the cavity is now higher than the pressure outside, air rushes out and creates an “explosion” (hence the name plosive).

The duration of each of these events were measures for the CV. The first segment is called the silence duration or “stop gap”, the second element is called the voice onset time (VOT), and the third element is the vowel duration. The VOT is measured from the beginning of the burst until the time in which voicing of the vowel begins for both voiced and voiceless consonants.

The essential part of a plosive consonant is the blockage stage, as usually a plosive is placed between two vowels, so the closure stage coincides with the production of the preceding vowel, and the release stage coincides with the production of the following vowel. Now, if the vocal cords start to vibrate in the blockage stage, before the release of the plosive, the consonant is voiced. If the vocal cords start to vibrate at about the same time as the consonant is released, it becomes voiceless unaspirated. If the vibration only starts significantly after the release, the plosive is voiceless aspirated. In the study of clear speech a difference in the VOT are measured. Study of Picheny et al (Picheny, Durlach, & Braida, 1986) got a larger VOT for voiceless plosives and these results are confirmed by the following studies by Krause and Braida (Krause & Braida, 2009) and Smilianic and Bradlow (Smiljanic & Bradlow, 2005).

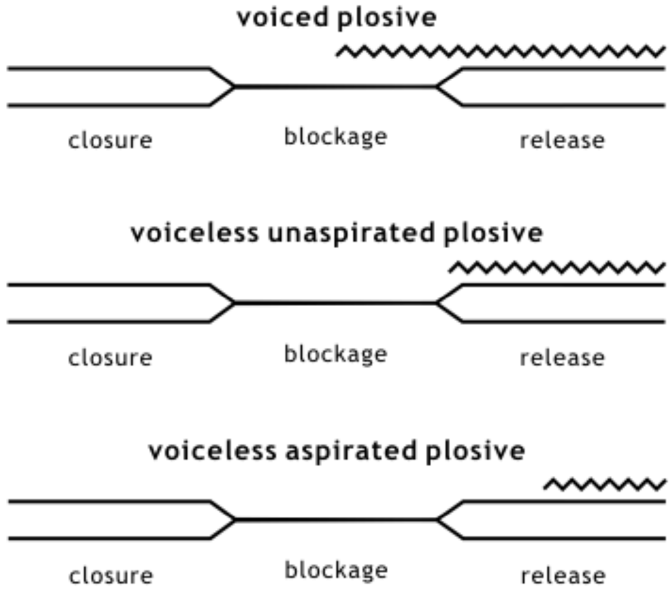


Figure 3.6 Voice Onset Time of voiced, voiceless unaspirated and voiceless aspirated plosives. Voiced plosives have a VOT noticeably less than zero, meaning the vocal cords start vibrating before the stop is released. Unaspirated voiceless plosives have a VOT at or near zero, meaning that the voicing of a following sonorant (such as a vowel) begins at or near to when the stop is released. Aspirated plosives followed by a sonorant have a VOT greater than this amount, the length of the VOT in such cases is a practical measure of aspiration.

Phonological

This group includes different particular phenomena, not present in all the languages, related to speech articulation:

- *vowel modification*, like the substitution in syllables and vowels that tend to becomes unstressed and toneless (i.e. “schwa”-like);
- *burst elimination,* at the release of a stop consonant, as for plosive consonant in sentence-final position (a bust is the noise produced at the open of the cavity);
- *alveolar flap* (the particular sound like the ‘r’ in English word ‘free’) for specific consonant preceded and followed by particular vowels;

In clear speech there is a reduction of all this phenomena, excluding for sound insertion (Picheny, Durlach, & Braida, 1986; Krause & Braida., 2004). All this features are not used in the intelligibility test, because they vary largely from different talkers, then a dependence on intelligibility are not yet proved.

3.2.2 Spectral features – phonetical based grouping

One of the most important properties of the normal auditory system is that it acts as a frequency analyzer. Therefore, when exploring the relationship between the perceptual attributes of speech sounds and their acoustic structure, most emphasis is placed on the frequency spectrum and in all the other caracteristics related to a frequency analysis. So, for example, we talk about the frequencies of the formants in a vowel, or the multi-harmonic nature of voiced speech.

There exist a lot of methodologies to analyze the spectral content of the clear speech and a list of all the studied features would not be exhaustive. The most important are the phonetical characteristics that can be used in a comparison between different speaking styles.

The analysis of these parameters gives some important results that can be resumed organizing the spectral features in global and segmental properties as proposed by Smiljianic and Bradlow (Smiljanic & Bradlow, 2009):

Global

For these features the spectral characteristics are analyzed on the speech signal in each whole:

- *fundamental frequency* F0, as an average for the occurrence of all the voiced segments;
- *distribution of spectral energy*, i.e. the energy distribution between frequencies;
- *frequency cutoffs*, i.e. the influence of different filtering frequencies on intelligibility;
- *spectro-temporal representation*, i.e. the response in time at the variation of different spectral parameters.

For the fundamental frequency the results from Picheny et al (Picheny, Durlach, & Braida, 1986) and Krause and Braida (Krause & Braida, 2009) give an higher average and larger range as shown in Figure 3.7, and Klasmeyer (Klasmeyer, 1997) founds an increase in the perception of difference between two equal signals with fundamental frequency difference of at least a factor of four.

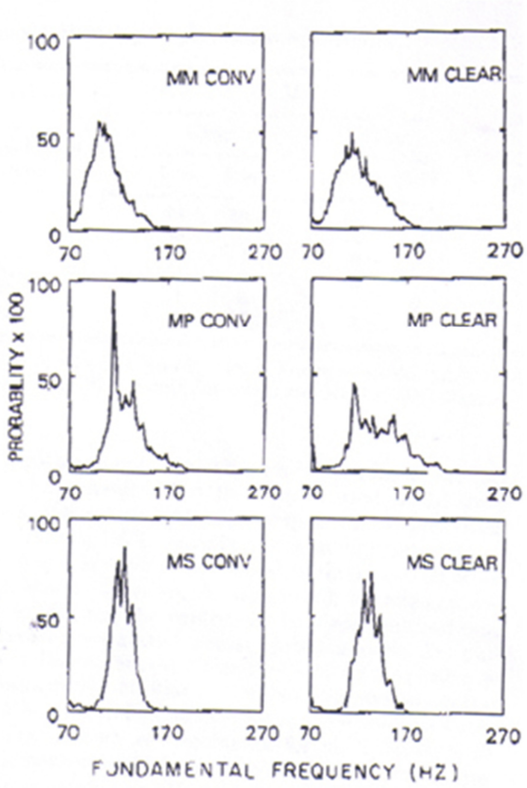


Figure 3.7: Fundamental frequency (F0) data. Left graphs show F0 histograms for the conversational speaking mode for three speakers, the right graphs shows F0 histograms for the clear speaking mode. Taken by (Picheny, Durlach, & Braida, 1986).

The distribution of spectral energy, for Picheny et al (Picheny, Durlach, & Braida, 1986), Krause and Braida (Krause & Braida, 2009) and Smiljianic and Bradlow (Smiljanic & Bradlow, 2005), results higher at higher frequencies and in particular it implies more intense frequency components around 1000Hz of the long term spectra (distribution of spectral energy over the course of the utterance).

Low and high frequency cutoffs affect sound quality as demonstrated in different studies (Punch & Beck, 1980; 1986; Tecca & Goldstein, 1984), and low cutoff frequency and spectral slope are important in hearing aid sound quality judgment by the results obtained by Gabrielsson (Gabrielsson, 1998) that expose how individual with hearing loss preferred listening to speech with a flat response in the low frequency and with a +6dB/oct increase between 1000 and 4000 Hz.

For the spectro-temporal representation, Greenberg an Arai (Greenberg & Arai, 2004) has demonstrated that a detailed one is not required for intelligibility, that seem to be dependent on both the magnitude and phase of the modulation spectrum.

Segmental

The segmental properties include:

- *fundamental frequency* F0, calculated for all the voiced segments (vowels and voiced consonants);
- *vowel formant movements*, the vowel steady state and/or the transitions of first (F1) and second (F2) formant frequencies as the frequency variation of the formant during the vowel utterance;
- *root mean square* of fricative and plosive consonant;
- *vowel space*, the area between vowel categories as defined by F1 x F2 dimensions
- *short-term spectra*, spectrum of the signal around a particular point in time
- *consonant-vowel ratio* (CV-ratio), calculated simply by dividing the consonant power by the power of adjacent vowel.

An example of vowel formant movements is shown in Figure 3.8 in which the modification of vowels formant for different phonemes is expressed by the movement of formants F1 and F2 values. Smiljianic and Bradlow (Smiljanic & Bradlow, 2005) expose how an acoustic strengthening at the segmental level (having more 'extreme' formant values) can make the phoneme more 'transparent' in the acoustic signal.

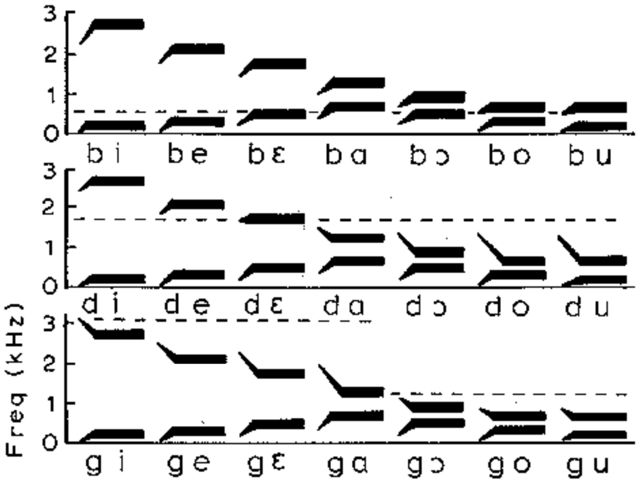


Figure 3.8: Vowel formant transition. Exposed for different vowels, used with different consonants, the transition is the increase of decrease of frequency value in the vowel. Note the distance between the two formants that differs between vowels.

The vowel space, that is naturally wider for vowels, compared to consonants (Korczak & Stapells, 2010), is shown in Figure 3.9. It represents, on the F1 value and F1-F2 value axis, the distance between different vowels and its representation as the articulatory place in which are produced. The link between an expanded vowel space and an intelligibility increase is not fully established, but Smiljianic and Bradlow (Smiljanic & Bradlow, 2005) assert that vowel hyperarticulation seems to be a valid enhancement strategy for increase speech intelligibility. Vowel lengthening and vowel expansion along both F1 and F2 dimensions contributes significantly to increase intelligibility, while enhanced vowel formant dynamic features did not (Smiljanic & Bradlow, 2005) .The studies of Picheny et al (Picheny, Durlach, & Braida, 1986) and Krause and Braida (Krause & Braida, 2009) reveal that talkers who are naturally more intelligible tend to produce more expanded vowel spaces with an increase between F1 and F2 formant frequencies.

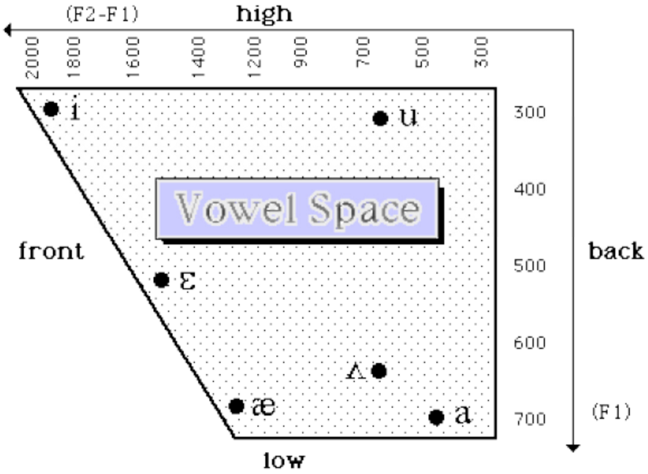


Figure 3.9: Vowel space. Represent the area between vowel categories as defined by F1 × F2 dimensions. In clear speech this area seems to be larger. This vowel hyperarticulation seems to be a valid enhancement strategy for increase speech intelligibility.

In the same studies, and with the same results for Smiljianic and Bradlow (Smiljanic & Bradlow, 2005) it is shown that in the short term vowel spectra there is an increase rate of F2 transition (longer duration of formant transition, narrower formant bandwidth) and an increase of energy in the second and third formant.

From the analysis of the vowel ratio made by Picheny et al (Picheny, Durlach, & Braida, 1986) results an increase in the consonant-vowel ratio potentially given by an increase in the consonant power.

In the next Chapter from all this features a set of most important for the intelligibility of clear speech are selected, and for all these parameters an explanation of their extraction algorithm is given.

Chapter 4

Clear Features Extractor

A set of caracteristics described and analyzed in Chapter 3, have been selected to implement and to develop a specific toolbox called Clear Features Extractors (CFE). This toolbox is intended as a tool that automatically extracts relevant features from speech and, as such, allows to compare different speech samples to evaluate to what extent they resemble 'clear speech'. In particular, this toolbox was used in this work to analyze speech processed by a cochlear implant and to evaluate how the choice of different cochlear implants settings can influence the clear features of the speech perceived by an implanted patient.

The following set of features was selected and implemented as the most important for the evaluation of intelligibility, and as those that can be extracted automatically, without a manual computation of some parameters.

This section shows the methods of extraction of individual characteristics and features, relevant to the recognition of clear speech. The analysis follows the subdivision exposed in Section 3.2, between the temporal and spectral features.

4.1 Temporal features

4.1.1 Temporal Envelope

As suggested by Krause and Braida (Krause & Braida, 2004) to extract the envelope function the following algorithm is used:

i) subdivide the signal in seven different channel using a bank of seven second-order octave bandwidth Butterworth filters, with center frequencies in the range of 188-7938 Hz, chosen using the Greenwood function (Greenwood, 1990) which is based on physiological human data;

ii) compute the Hilbert transform (Drullman, Festen, & Plomp, 1994a);

iii) low-pass filter by fourth-order Butterworth filter with 60 Hz cutoff frequency, to extract the envelope in each sub-band;

iv) to obtain the envelope of the entire signal, just sum every sub-band envelopes.

The Greenwood function describes the distribution of frequencies along the cochlea and is expressed by the formula:

$$F = A(10^{ax} - k)$$

were x is the distance from the stapes in the cochlea, A and a are constants (for man): A=165.4 (to yield frequency in Hertz) and a= 0.06 (if x is expressed in millimeters or a =2.1 if x is expressed as a portion of a basilar membrane length), and the integration constant k is fixed at the value k=0.88 to the lower frequency limit of 20 Hz for man, and a graphical representation is shown in Figure 4.1.

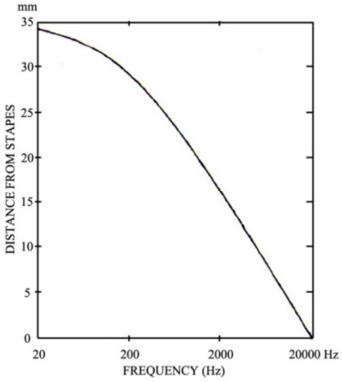


Figure 4.1: Greenwood function. Based on research on human cadaver and similar results from six other species, it provided a convenient mathematical expression for a cochlear frequency-position map. The different constants are scaled and normalized to support the applicability to the living human cochlea.

For a mathematical description of the Hilbert transform and of the Butterworth filter implementation, see Appendix A.

In Figure 4.2 the algorithm for the envelope extraction is presented. The Hilbert transform was used, instead of the rectifying method, because it requires less processing on the signal that can alter its characteristics and produces more accurate estimates of the envelope. Use of higher envelope cutoff frequencies, however, yields envelopes close to those extracted by the Hilbert transform.

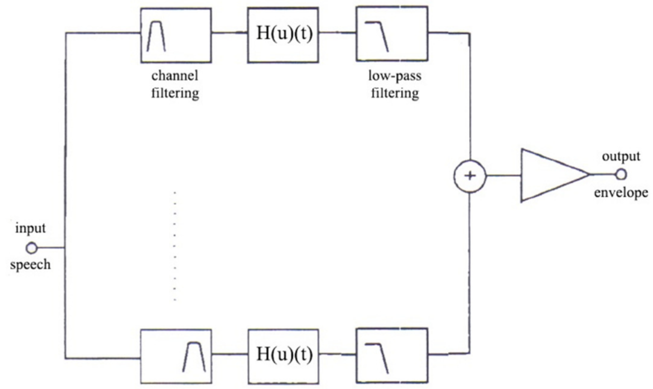


Figure 4.2: Envelope extraction algorithm. The input signal is divided in seven channels with Butterworth filters with central and cutoff frequencies given by the Greenwood function. The envelopes are extracted by a Hilbert transform and then smoothed by a low-pass filter. The resulting signals are added to give the total envelope.

In Figure 4.3 is shown an example of sub-band envelope extraction from the Italian word "frase". In the top-left panel is represented the waveform of the word, in the other panels are represented the envelope extracted from the same word, for the seven different channels.

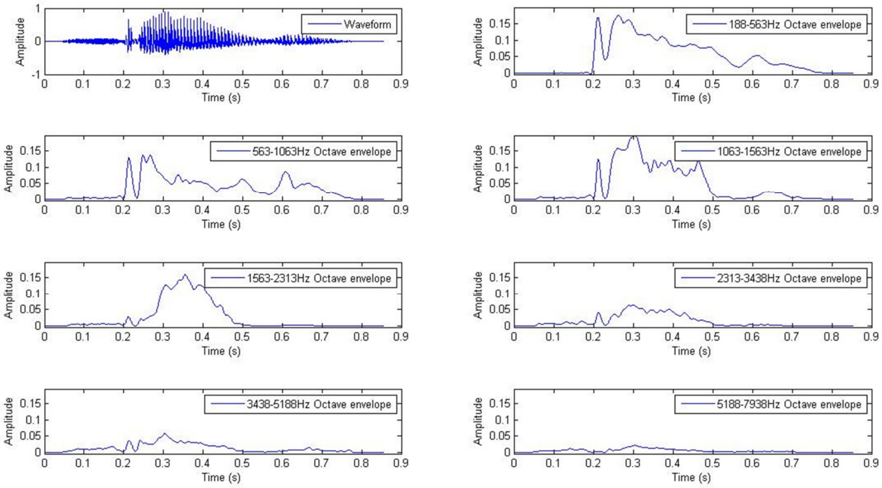


Figure 4.3: Envelope extraction for Italian word "frase" in seven sub-bands from 188 Hz to 7938 Hz. The speech sample was filtered in different bands with a Butterworth filter bank, and then the envelope were extracted in each band by using the Hilbert transform

4.1.2 Modulation Index

The modulation index was estimated from the temporal envelope as follows:

i) after the computation of the temporal envelope for each sub-band downsample the intensity envelope to obtain a signal in the range of 0-220 Hz;
ii) compute power spectra with a Fast Fourier Transform;
iii) then, as shown by Houtgast and Steeneken (Houtgast & Steeneken, 1985) normalize each value of the frequency in the envelope spectra is by the mean of the global envelope function;
iv) obtain a 1/3-octave representation of the spectra by summing components over 1/3-octave intervals with center frequencies ranging from 0,4 to 20Hz.

In Figure 4.9 the algorithm for the modulation index is presented, as explained before.

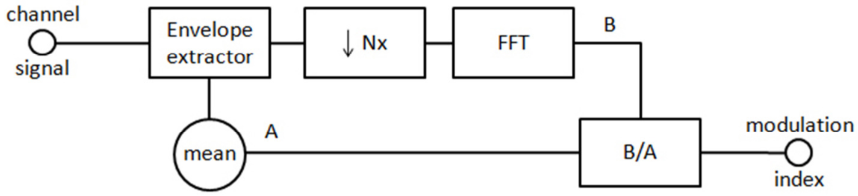


Figure 4.4: Modulation index algorithm. From the input signal from the channel subdivision is extracted the envelope. From the envelope is computed the mean and the signal is downsampled to 220 Hz. The FFT is computed and the resulting values are normalized with the envelope mean, obtaining the Modulation Index.

In Figure 4.5 the modulation index of an octave-band-filtered-speech is shown as estimated by Houtgast and Steeneken. On the left-hand side is shown the envelope of one-minute speech signal represented at one-second intervals; on the right-hand side is shown a representation on 1/3 octave band of the envelope spectrum, normalized for the envelope intensity mean.

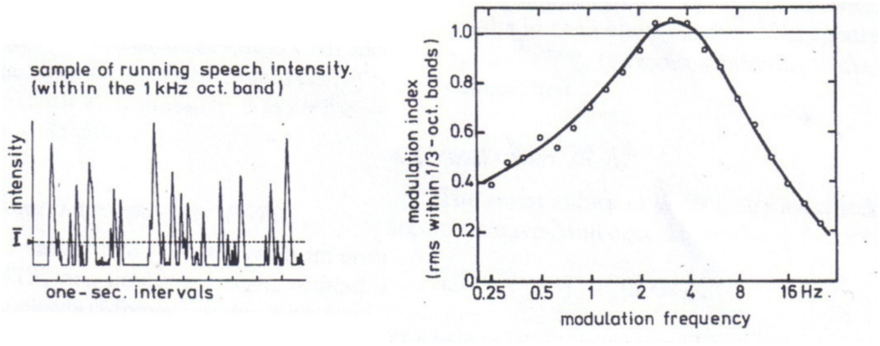


Figure 4.5: The fluctuation of the intensity envelope of octave band filtered speech (left panel) quantified in 1/3-octave band, normalized with respect to the mean value of the envelope (Ī the left panel). The ordinate of this envelope spectrum is interpreted as the modulation index. Taken from (Houtgast & Steeneken, 1985)

Following is given a pseudocode representation of the 1/3-octave band sum, used in different algorithm in this work:

```
SUM OVER 1/3-OCTAVE INTERVALS
x(f) = Values to sum
f_central = Center frequency of the first octave band
f_max = Maximum octave center frequency
while f_central<f_max
    for i=1:3
            comment: 1/3 octave center frequency, lower and upper limit
            f_octave_third ← f_central × $10^{0.1(i-1)}$
            f_lower ← f_ octave_third / $\sqrt{2}$
            f_upper ← f_octave_third × $\sqrt{2}$
        res ← sum(f_lower<x<f_upper)
        return (res)
    end
    comment: Next octave center frequency
    f_central ← f_central × 2
end
```

4.2 Spectral and energy features

4.2.1 Fundamental frequency

As exposed by Krause and Braida (Krause & Braida, 2004) the fundamental frequency of a speech segment is averaged on the occurrence of each voiced portions of speech signal:

i) each segment is filtered by an Hamming window with the same length of the signal;
ii) the FFT of the windowed signal is computed;

iii) the maximum value on the range between 50 – 300 Hz (given by the estimate of speaker fundamental frequency) is taken as the fundamental frequency of the voiced signal;

iv) the fundamental frequency range is computed as the difference between the maximum and minimum value founded for the vocal segment of the signal.

As an example, in Figure 4.6 is exposed the waveform, spectrum computation for a singular vowel /a/ in Italian word ‘caldo’ pronounced by a male speaker.

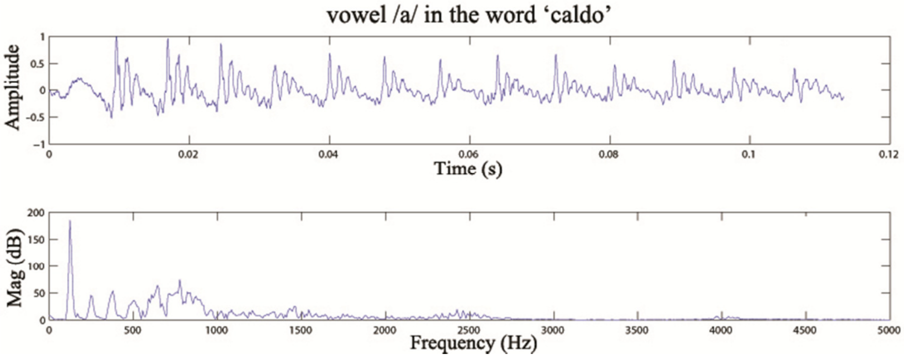


Figure 4.6 Waveform (top-panel), spectrum (bottom-panel) for the segment representing the vowel “a” extracted from the word “caldo” pronounced by a male Italian speaker. The spectrum band is limited to 5000 Hz.

4.2.2 Vowel formants

To find the formants of a vowel segment, Linear Prediction Coding (LPC) is used as suggested by O’Shaugnessy (O'Shaugnessy, 2008). LPC models the signal as if it were generated by a signal of minimum energy being passed through a purely-recursive IIR filter. To find the formant frequencies from the filter, we need to find the locations of the resonances that make up the filter. This involves treating the filter coefficients as a polynomial and solving for the roots of the polynomial from 0 Hz up to half the sample frequency. A number of two coefficient for kHz, as suggested by O’Shaugnessy (O'Shaugnessy, 2008), is sufficient for the estimation of the three formants needed. Then limiting the search of the local maxima in the range 0 – 5 kHz the number of sufficient coefficient is fixed to 10.

For a description of the LPC method see Appendix A.

In Figure 4.7 is shown the LPC computation (bottom-panel) for the same segment of previous figure.

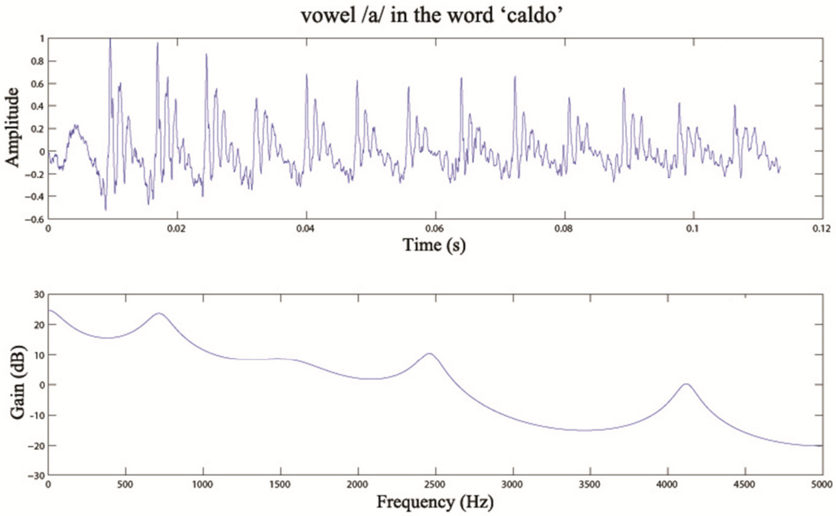


Figure 4.7: Waveform (top-panel)and LPC computation (bottom-panel) for the segment representing the vowel “a” extracted from the word “caldo” pronounced by a male Italian speaker. The LP filter representation is limited to 5000 Hz. The visible peaks identify the formant of the considered vowel segment.

4.2.3 Vowel space

The vowel space is estimated as the area between the first two formant dimensions for each vowel recognized in the speech. The F1 and F2 values are extracted with the vowel formants extraction algorithm exposed in the previous Section. Averaging all the occurrence of the same vowel in a sentence or in a word, the vowel space is built and expressed by a graphic representation.

An example is shown in Figure 4.8 that represents the vowel space computation of the sentence “è rimasto solo al mondo” pronounced by an Italian male speaker.

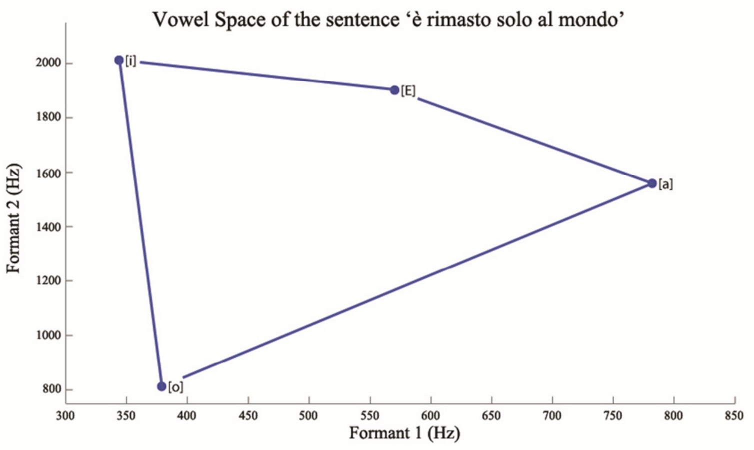


Figure 4.8: Vowel space for italian sentence “è rimasto solo al mondo”. Each vowel is represented as the F1-F2 dimension computed as the average of all the occurrence in the sentence.

4.2.4 Long-term spectra

To consider the spectral property of a signal Krause and Braida (Krause & Braida, 2004) propose to analyze the entire signal computing the long-term spectra of the speech as follows:

i) subdivide the speech taking different segments of the signal, using a 25,6 ms non-overlapping Hamming windows;
ii) compute the FFT of each segment;
iii) compute the mean of the resulting spectrums;
iv) finally, a 1/3-octave representation is obtained summing components over 1/3-octave intervals with center frequencies ranging from 62.5 to 8000 HZ.

In Figure 4.4 the algorithm for the long-term spectra is presented, as explained before

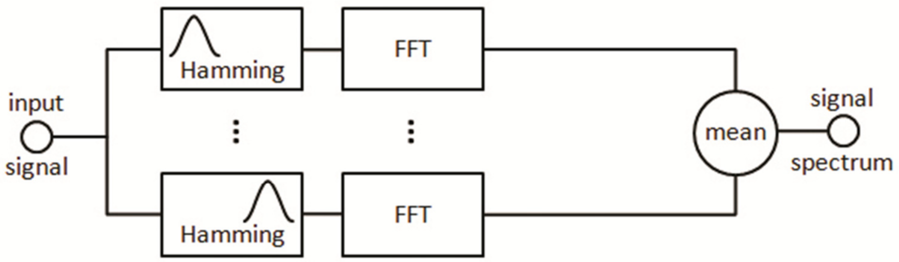


Figure 4.9: Long-term and short-term spectra algorithm. Signal is filtered with a sliding Hamming window with no-overlap, and the spectrum is computed as the mean of the resulting FFT computation.

In Figure 4.10 is shown the results obtained by Krause and Braida (Krause & Braida, 2004) in which are visible the difference between the conversational style, the clear style at slow rate and the clear style in normal (conversational) rate.

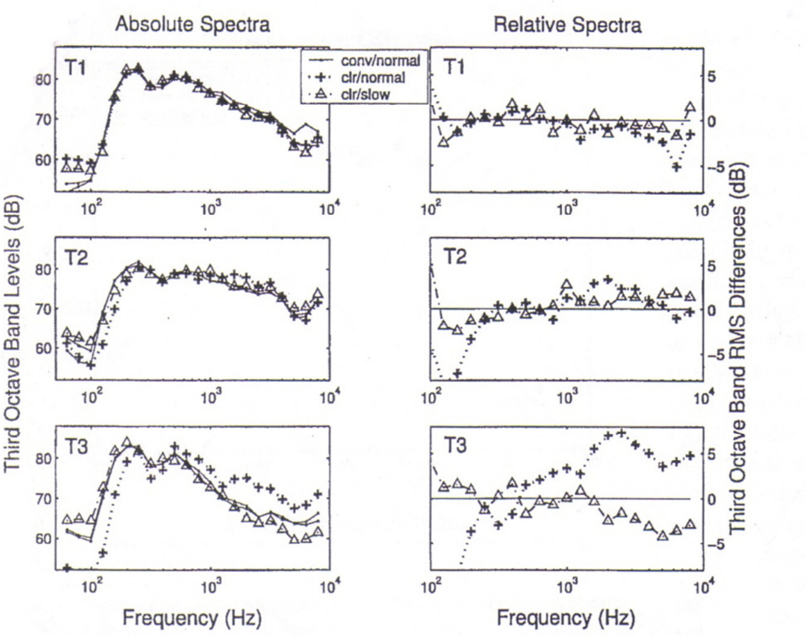


Figure 4.10: Third-octave band rms data. Left Graphs show absolute spectra for three talkers. Corresponding graphs on the right show spectral differences obtained by subtracting the conv/normal spectrum from the clear/normal and clear/slow spectra, which depicts the relative distribution of spectral energy between conversational and clear speech. Taken from (Krause & Braida, 2004).

4.2.5 Short-term spectra

As for the long- term spectra, Krause and Braida (Krause & Braida, 2004) propose a more detailed analysis for particular segments of speech as the one representing the vowel, computing a short-term spectra as follows:

i) passes the signal in a preemphasis filter with a slope of +6 dB/octave in order to boost the higher frequencies;
ii) subdivide the signal taking different segments, using a 25,6 ms sliding Hamming windows shifted in 1 ms intervals;
iii) compute the FFT of each segments;
iv) normalize each segment by its root mean square;
v) finally obtain a 1/3-octave representation summing components over 1/3-octave intervals with center frequencies ranging from 62.5 to 8000 HZ.

The representation of the algorithm is the same as for the long-term spectra in Figure 4.9.

4.2.6 Consonant-vowel power and ratio (CV-ratio)

As specified by Picheny et al (Picheny, Durlach, & Braida, 1986) the relative power of a particular point in time is calculated as energy of the signal weighted with a 20 ms Kaiser window, computed every millisec. For the various classes of speech sounds, different measures are taken. The maximum value is used for vowels, plosives and fricatives; the midpoint value is used for semivowel and nasal. Then the same method is used to compute the power of postvocalic plosive, prevocalic nasal and semivowel consonants.

Calculating the power as explained before, the CV-ratio is computed as the ratio between a plosive consonant power in the nearest adjacent vowel either preceding or following the consonant.

Chapter 5

Materials and methods

5.1 Speech material

In speech audiometry the test material should be chosen according to criteria that take into account the statistical and phonotattical phonetics of the language, the frequency of use of the words, the plausibility and reasonableness of the proposed sentences, their profiles and intonations, and so on. The material here used includes twenty two-syllable words and three sentences with a number of syllables ranging between 9 and 13, chosen from the set of words and sentences prepared by Turrini et al (Turrini & Cutugno, 1993).

Stimuly are spoken by an Italian male speaker and are part of the speech material typically used in clinical speech audiometry. The recordings were made in a recording studio located in Padua, in July 1996. The male speaker is an expert selected phonetics, a researcher in the field of pronunciation of the sounds of Italian. In each recording the audio speaker in production, was monitored by a second speaker for the correction of errors in pronunciation.

The recording was done directly in the following chain of digital equipment: a Neumann Condenser Michrophone and a Tascam DAT Recorder.

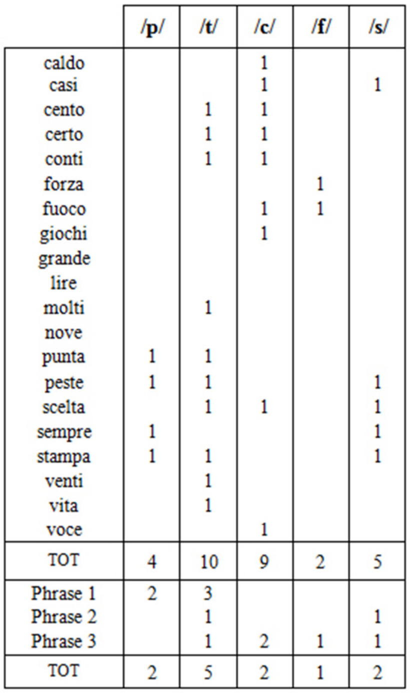
The system is calibrated with a signal frequency of 1000 Hz to 78.4 dB RMS. The speaker has uttered the word lists clearly avoiding suspensive intonation. For phrases, intonation given to each sequence was typical of the prosodic structure of declarative form of Italian. All recorded signals were analyzed with the Computerized Speech Laboratory system KAY 4300 with automatic procedures that measure the performance of energy, directly in dB SPL, depending on the time and calculates the integral values of RMS of pre-defined portions of signal. The equalization of the levels, when necessary, was done through a software program for digital signal processing audio (Sound Forge 4.0, Sonic Foundry).

The analysis of words is carried out seeking for the presence of occlusive and fricative non-vocalic consonants (respectively / p /, / t /, / c / and / f /, / s /), used at the beginning of the word or within syllables. In addition, the presence of vowels was analyzed that reflects the average occurrence in the

Italian language. For the sentences three periods are chosen that contained all detectable plosive and fricative consonants and avoid as much as possible the presence of diphthongs, which can create confusion in vowels recognition:

- Sentence 1: “abbiamo preparato una torta”
- Sentence 2: “è rimasto solo al mondo”
- Sentence 3: “Franco è andato via di corsa”

The complete list of words and phrases examined is shown in Table 5.1 where it is highlighted the presence of the occlusive and fricative non-vocalic consonants.



	/p/	/t/	/c/	/f/	/s/
caldo			1		
casi			1		1
cento		1	1		
certo		1	1		
conti		1	1		
forza				1	
fuoco			1	1	
giochi			1		
grande					
lire					
molti		1			
nove					
punta	1	1			
peste	1	1			1
scelta		1	1		1
sempre	1				1
stampa	1	1			1
venti		1			
vita		1			
voce			1		
TOT	4	10	9	2	5
Phrase 1	2	3			
Phrase 2		1			1
Phrase 3		1	2	1	1
TOT	2	5	2	1	2

Table 5.1: List of words and phrases used by specifying the presence of occlusive and fricative consonants to be analyzed. Sentence 1: “abbiamo preparato una torta”; Sentence 2: “è rimasto solo al mondo”; Sentence 3: “franco è andato via di corsa”.

5.2 Speech processing

To evaluate the features extracted from the clear speech, an algorithm that simulates the signal processing by a cochlear implant is applied to the material available.

5.2.1 Cochlear Implant Speech Generator (CISG)

For the evaluation of processing strategies for cochlear implants different techniques have been proposed over the years, to simulate the perception of acoustic signals that the hearing impaired that wears a cochlear implant perceives, in response to electrical stimulation. This approach has been used, for example, to study innovative processing algorithms without having to intervene on the real prosthesis but simulating the algorithms on personal computers and creating sounds with which can be heard and evaluated by subjects with normal hearing for the recognition of the processed and synthesized signal.

These 'acoustic simulations' are widely used in studies on speech for cochlear implants, and they have proved to be powerful tools to quantify the effect on the implant performance after changing the parameters and settings of strategy or reduction of information content (e.g., spectral channels), it is shown that the best performance of implanted subjects in psychophysical tests of recognition are comparable to those of normal hearing subjects listening to acoustic simulations, in terms of equality of signal processing.

Therefore the results obtained with the use of these synthesized acoustic simulations represent a reference level for the performance of the implanted subject, and the introduction of these simulations also eliminates a large number of factors to interindividual variability that occur necessarily between different patients with implant, factors for example related to several causes of non-ideality of the interface with the nerve and also to unpredictability of efficiency in the patient, of the route of transmission and interpretation nervous of the electrode stimulation.

The method of synthesis of acoustic simulations adopted in this study refers to the algorithm proposed by Shannon et al (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), with the choice of processing parameters based on the characteristics of a particular cochlear implant under consideration, the Nucleus ® 24 by Cochlear Corporation. The method implemented follows these steps:

i) decompose the input signal in different channels, through a band pass filtering with second-order Butterworth filters with center frequencies related to the tables prepared by Greenwood (Greenwood, 1990) and used in the selected cochlear implant;
ii) to the output waveform of each filters the Hilbert transform is applied;
iii) the results of the Hilbert transforms are then filtered through a low pass fourth order Butterworth filter with cutoff frequency of 60 Hz in order to extract the envelopes on each channel;
iv) the resulting envelopes are used to modulate noise bands calculated by filtering white noise through the same band pass filters used for the envelope;

v) after the modulation, noise bands are added together;

vi) The resulting modulated noise is filtered through a Butterworth filter of fourth order low-pass cutoff frequency of 4000 Hz;

vii) finally the RMS value is adjusted to be equal to the original one.

The Figure 5.1 is a representation of this algorithm.

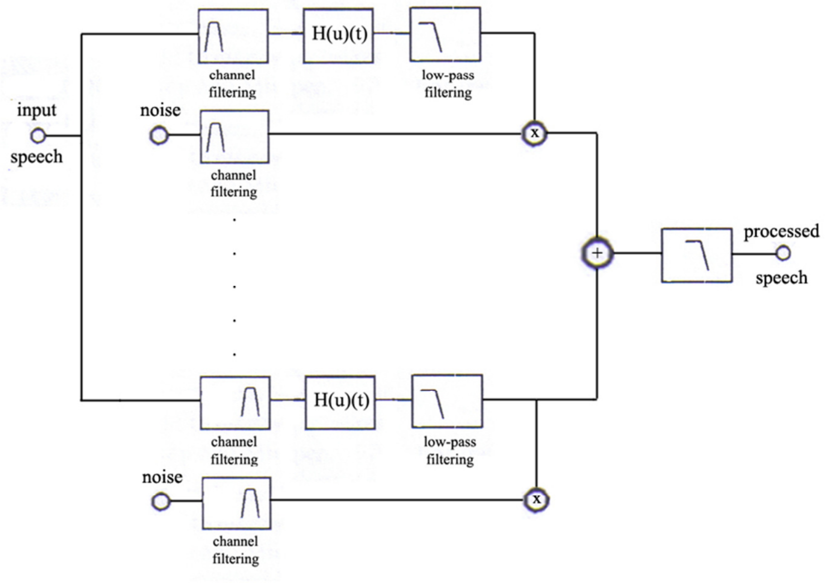


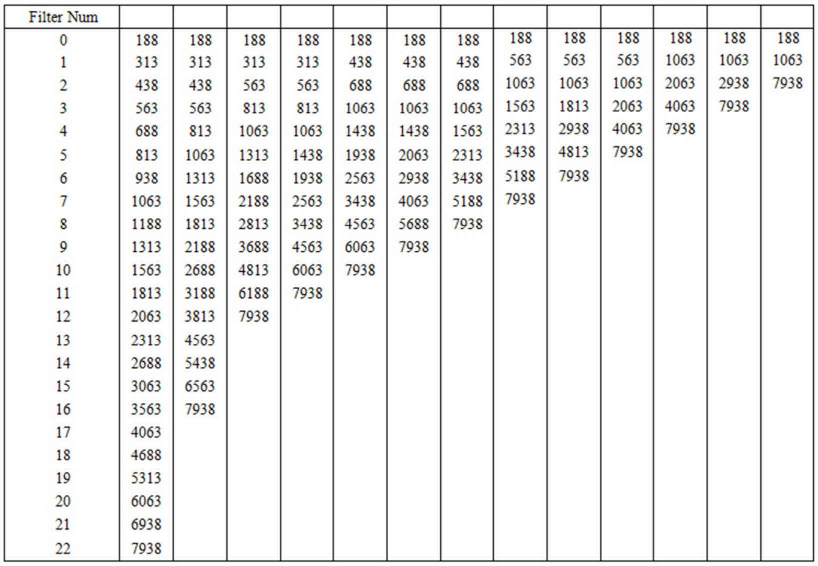
Figure 5.1: Cochlear implant speech simulator. From the input signal are extracted the envelopes for each channel selected from the Greenwood formula. A white noise is filtered in the same channel and each output is modulated by the correspondent envelope. The modulated noises are then added together and filtered with a low-pass Butterworth filter with cut-off frequency of 4000 Hz to eliminate the processing artifacts.

The described method replaces the fine structure of the original signal, with a fine structure generated by the modulation of white noise. This choice is also justified by the results obtained from Drullman (Drullman, 1995) that indicates that fine structure cue play a less important role than envelope cues in the perceived intelligibility of a speech signal.

5.2.2 Allocation Table

Regarding the allocation of frequency bands for each channel the tables for the allocation of frequencies of Nucleus ® 24 implant are taken as reference. They cover a wide variety of situations and can fit to strategies with a number of channels from 1 up to the limit 22. The overall bandwidth is clearly limited, except in special cases determined by the needs of individual patients, from frequencies 188 and 7938 Hz and the definition of the bands of the filters is based on psychophysical bases (as the distribution of critical bands) and involves a distribution of frequencies along the stimulation array, that is linear below 1000 Hz and logarithmic above that threshold. The bandwidths are constant at low frequency and increases at higher frequencies, implying that the last band is nearly 10 times larger than the first.

Below in Table 5.2 are represented the most significant bands for a number of channels from 22 to 2.



Filter Num													
0	188	188	188	188	188	188	188	188	188	188	188	188	188
1	313	313	313	313	438	438	438	563	563	563	1063	1063	1063
2	438	438	563	563	688	688	688	1063	1063	1063	2063	2938	7938
3	563	563	813	813	1063	1063	1063	1563	1813	2063	4063	7938	
4	688	813	1063	1063	1438	1438	1563	2313	2938	4063	7938		
5	813	1063	1313	1438	1938	2063	2313	3438	4813	7938			
6	938	1313	1688	1938	2563	2938	3438	5188	7938				
7	1063	1563	2188	2563	3438	4063	5188	7938					
8	1188	1813	2813	3438	4563	5688	7938						
9	1313	2188	3688	4563	6063	7938							
10	1563	2688	4813	6063	7938								
11	1813	3188	6188	7938									
12	2063	3813	7938										
13	2313	4563											
14	2688	5438											
15	3063	6563											
16	3563	7938											
17	4063												
18	4688												
19	5313												
20	6063												
21	6938												
22	7938												

Table 5.2: Allocation tables for filter bank in acoustic simulations. Are presented those used by the channel numbers from 22 to 2. The frequencies listed indicate the boundaries between adjacent bands, the lowest frequency is 188 Hz in each case and the highest is 7938 Hz until the subdivision in 22 filter bands

5.3 Speech segmentation

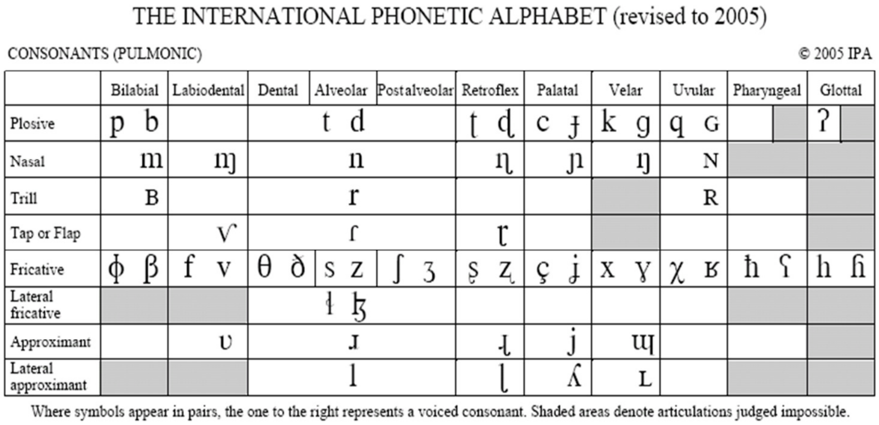
The speech signal can be considered to be the output of a linear system. Depending on the type of input excitation (source), different classes of speech sound, are produced. The most defined two are: voiced and unvoiced. If the input excitation is noise, then unvoiced sound such as /s/, /t/, etc., are produced, and if the input excitation is periodic then voiced sounds such as /a/, /i/ etc., are produced. In the unvoiced case, noise is generated either by forcing air through a narrow constriction (e.g., production of /f/) or by building air pressure behind an obstruction and then suddenly releasing that pressure (e.g., production of /t/). In contrast, the excitation used to produce voiced sounds is periodic and is generated by the vibrating vocal cords.

In this work a similar representation is used, focusing on the difference between vowels and consonants. Every word and sentence from the speech material exposed in Section 5.1 is analyzed and segmented into their phonetic components in order to be processed by the Clear Features Extractor. Through the free digital audio editor Audacity ® 1.3.12-beta (Unicode), the segmentation was performed in intervals of samples for the various components of speech.

The categories of speech segments under consideration, and their recognition name in the algorithm, were:

- *silence* (“silence”), i.e. the absence of signal at the beginning, at the end and during the signal;
- *vowels* (“vowel”), i.e. the Italian basic vowels: / a /, / e /, / E /, / i /, / o /, / O /;
- *voiced consonants* (“vocal”) i.e. that consonants that are produced even with the voice :/ m /, / n /, / l /, / r /;
- *plosive and fricative voiced consonants* (“plovoc”), i.e. the percussive consonants that concern a following emission of voiced signal: / b /, / d /, / g /, / v /, / z /;
- *plosive fricative non-voiced consonants* (“unvoiced”), i.e. the purely noisy consonants: / p /, / t /, / c / and / f /, / s /;
- *diphthongs* (“dipht”), i.e. two adjacent vowel sounds occurring within the same syllable (as example ‘ua’ in the Italian word ‘quando’ or the rapidly succession of a word with a final vowel and another with a beginning vowel in a sentence).

This categorization is in according to the International Phonetic Alphabet as in Table 5.3.



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Table 5.3: International Phonetic Alphabet and the subdivision in voiced and unvoiced consonants

The signals were then manually subdivided, relieving the initial sample and final sample of each segment which are always adjacent. In Figure 5.2 is shown the segmentation of the Italian word “certo”, it is clear the difference between the speech segment categories. For the un-voiced plosive consonants /c/ is evident the absence of periodic component, as for /t/ that differs for its duration. The vowel /e/ and /o/ are characterized by a visible periodic fluctuation and a greater intensity. The voiced trill consonant /r/ have similar characteristic of periodicity as the vowel but with introduction of noise due to the emission mode.

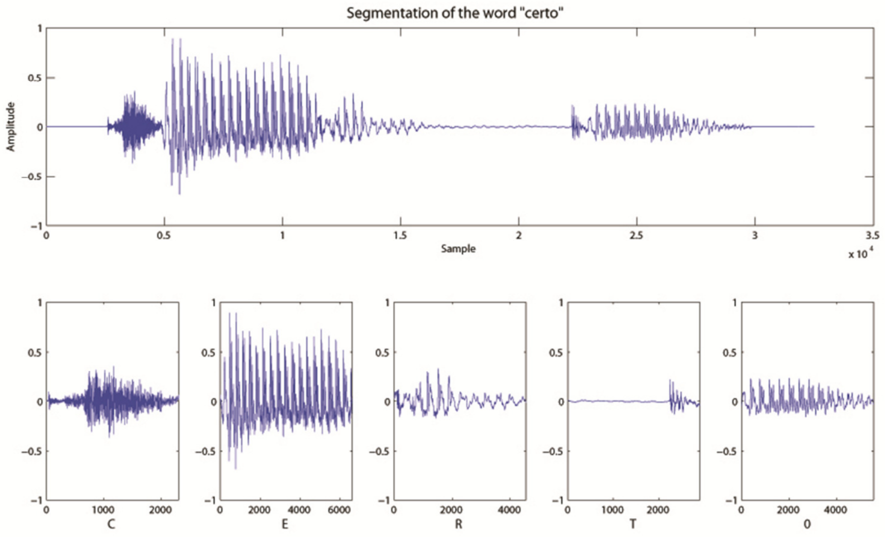


Figure 5.2: Example of manual segmentation of the Italian word “certo” pronounced by a male speaker. For each utterance, it’s recognizable its duration in sample and its intensity. For the consonant /t/ a short segment of silence was introduced in this realization to permit its recognition. All the other silence segments are not visible.

For processing in the Clear Feature Extractor, the values found were manually entered into a struct which also includes the global values of the signal and the processed results.

5.4 Segments analysis

The segmented signal were computed as shown in Figure 5.3 in which the various segments of a signal are processed to extract different features. In this representation the silence end diphthong segments are excluded because there are not interested characteristics to extract from.

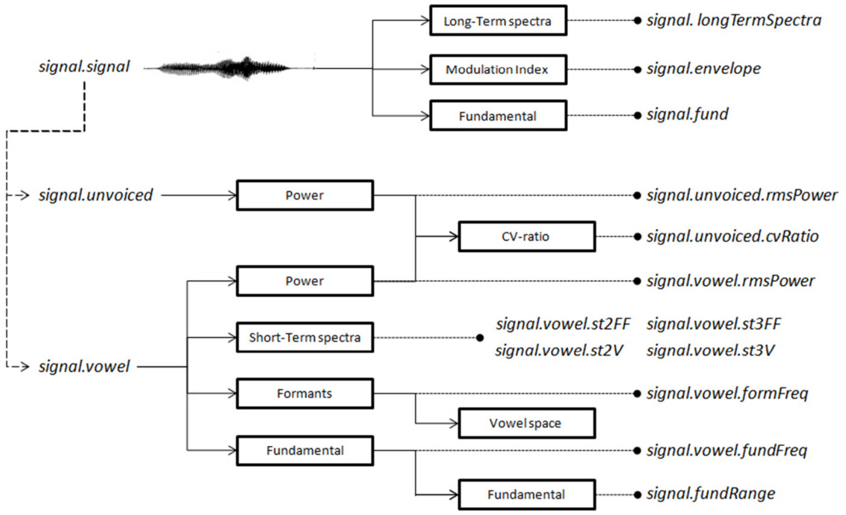


Figure 5.3: Schematic diagram of the signal processing features and resulting. Only the relevant segments of a signal are processed, the silence, plovoic voiced and dipth segments are excluded from the processing for their irrelevance in the determination of characterized features.

Therefore, a careful analysis of each considered segment and the extracted features is performed as explained below:

5.4.1 Global

The overall analysis is performed concerned the calculation of characteristics on the global signal:

Long-term spectra: the long-term analysis of the spectrum for each signal representing a word is made as shown in Section 4.2.4. The resulting pattern is analyzed in third-octave bands between 62.5 to 20000 Hz. A subsequent and more specific analysis is then performed in the band between 1000 and 3000 Hz, which, according to the study of Greenberg and Arai (Greenberg & Arai, 2004) and later confirmed by Kain et al (Kain, Amano-Kusumoto, & Hosom, 2008), is of major importance for the preservation of the intelligibility of clear speech. These results are used for the comparison between signals processed with the CISG, as shown in Section 6.2

Modulation index: the values of the modulation index ranged in frequency between 0.4 and 10 Hz are extracted by the method described in Section 4.1.2. This processing allows to control changes in the fluctuation of the index, that are characteristics of the level of intelligibility of clear speech that are relevant for this range of frequency and more specifically in the band below 4 Hz, as stated in the results obtained by Krause and Braida (Krause & Braida, 2004), Greenberg and Arai (Greenberg & Arai, 2004) and Drullman et al (Drullman, Festen, & Plomp, 1994a; 1994b). These results are used for the comparison between the signals processed in the cochlear simulator, as shown in Section 0.

Fundamental frequency: for a correct estimation of the fundamental frequency for each word, an analysis on the whole data set of words available is taken, verifying what the fundamental frequency of the speaker is. A signal containing all the words used is built, and from this one the fundamental is get, taken the highest peak of the frequency spectrum. For our speaker the fundamental frequency results of 120Hz. Then, for every word the fundamental is estimated in a range between 50 and 300 Hz around the frequency derived from the overall analysis. The frequency of the highest peak of the spectrum for each word is taken as the fundamental of the word itself and evaluated for results as outlined in Section 5.5.3.

Then, phonological analysis of the various parts of the signal is computed, as subdivision shown in Figure 5.3, for vowel segments, unvoiced segments and vocal segments.

5.4.2 Unvoiced

Power : the power value of signal segments representing fricatives and plosive consonants of the word in question, are extracted as explained in Section 4.2.6. The result is analyzed individually both as a comparison between the powers of the consonants in the same word and as a comparison between the results of the various signals processed by the CISG, wanting to check the increase recorded by Picheny et al (Picheny, Durlach, & Braida, 1986), as shown in Section 0. In addition, the power of the consonants in question is used in the calculation of the CV-ratio, explained below.

CV-ratio : the power ratio of the adjacent consonant and vowel in the same syllable is analyzed checking the increase recorded by Picheny et al (Picheny, Durlach, & Braida, 1986) and the changes introduced by the signals processed by the CISG, as shown in Section 0.

5.4.3 Vowel

Power: the power value of signal segments representing the vowels of the word are extracted as explained in Section 0. The power is used in the CV ratio as explained above.

Fundamental: the fundamental frequency is estimated as described in Section 4.2.1 whereas the value obtained from analysis of the fundamental of all the words spoken by the speaker. The search is then limited between 50 and 300 Hz. Obtained value is used to calculate the range of frequency variation that, as described by Picheny et al (Picheny, Durlach, & Braida, 1986), in clear speech must show an increase. The analysis is explained in Section 0.

Formants: the extraction of the formants, obtained as described in Section 4.2.2, is useful in determining the vowel space and to calculate the power value of the second and third formant. The parameters of F1 and F2 of all the original and processed signals are then used as explained in Section 5.5.5, while the frequency values of F2 and F3 of the original signal, are used in the detection of their power as explained below.

Short-term spectra : the short term spectra for each signal is calculated as described in Section 4.2.5, and used in the power estimation of second and third formant, as explained below.

Second and third formant power: the power values of F2 and F3 obtained as described above, are calculated by taking the frequency values closer to those of the corresponding original signal from the short term spectra, calculated per octave, . These results are then processed as described in Section 6.7.

5.5 Analysis protocol

Every word and every sentence of the dataset used, segmented as explained as above, is then analyzed using the CFE described in Chapter 4 and then used to create the signals resulting from processing of the CISG, as described in Section 5.2.1. This processing is performed for different numbers of channels (2, 4, 8, 10, 12 and 16) and any resulting processed signal is analyzed again using the CFE. The comparison between the results of the original signal and the signals processed in different channels, for each feature, is explained below. For some features, a first global analysis has identified the numbers of channels for a more interesting comparison, allowing to select only the original signal and the signals processed by the number of channels equal to 2 (which shown no degradation of the characteristics considered) and 12 proved the be the optimal number of channels. This limitation on the processed signal is specified in the characteristics, when used.

5.5.1 Long-Term Spectra

From the original signal and the processing of the CISG the spectrum in the band between 1000 and 3000 Hz is obtained. A comparison between the signals processed in this range, is performed by normalizing the values found with the sum of the values of the range, for the original signal and processes for 2 and 12 channels, computed as follows:

$$\mathrm{LTnorm}\ (\mathrm{n}) = \frac{\mathrm{LT}(\mathrm{n})}{\sum \mathrm{LT}(\mathrm{n})}$$

where LTnorm (n) is the long-term spectra normalized value at frequency n, LT(n) is the long-term spectra value at frequency n and the sum is calculated for all n with frequencies in the range of 792-3160 Hz as the results of the third-octave elaboration.

In this way is visible a trend that takes into account the specific amount of signals power only in the interested band and it possible to make a direct comparison between the various processed signals. The results of this analysis, and of a specific one on words 'giochi' and 'punta', are shown in Section 6.2.

In Figure 5.4 is presented an example of the two representations for the long-term spectra of the word 'caldo'. The left panel shown the representation between 0 and 5000Hz in which is visible the anormal configuration of power of the signal processed with two bands. The right panel, instead, shown a representation of the range 1000-3000 Hz with the computation explained before, and the difference in signal power is more defined.

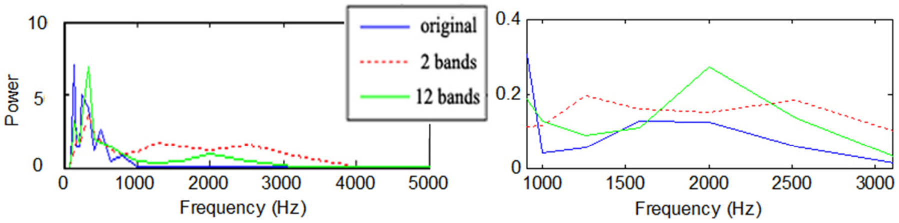


Figure 5.4: Representation of results for the Long-term spectra computation. In the left panel is shown the absolute value for the original signal, the 2 channels and 12 channels processed signals, in the range 0 – 5000 Hz for a word of the dataset.

5.5.2 Modulation Index

The comparison between the original signal and the signals processed at 2 and 12 channels, is done globally on all processed signals and exposed specifically for three words ('caldo', 'certo', 'peste') as an examples of typical patterns. The results of this analysis are shown in Section 0.

In Figure 5.5 is shown a representation of these results for a word in the dataset.

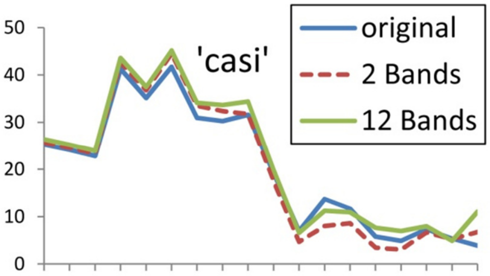


Figure 5.5: Modulation index representation for the word 'casi', for the original signal, the two channels and twelve channels processed signals.

5.5.3 Fundamental

The search for the fundamental frequency are computed to all the processed signals and exposed as the absolute value, compared to the value obtained by the original signal. The computation is performed on the entire signal for each processed version for each word. The results of this analysis are shown in Section 0.

In Figure 5.6 is shown a representation of these results for a word in the dataset.

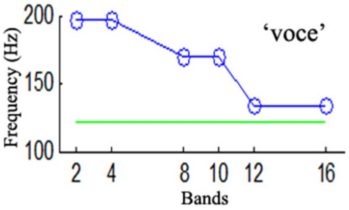


Figure 5.6: Fundamental frequency elaboration for the word 'voce' in the original signal (green line) and different channels processed signals (blue points).

5.5.4 Fundamental range

The frequency range of the fundamental is computed as the range between the fundamental frequency of the occurrence of vowels in each word, taking in account all the signal of the dataset for each processing number of channels. These results are compared to the fundamental frequency range of the words of the original dataset. Results are exposed in Section 0.

5.5.5 Vowel space

The vowel of the words or phrases analyzed, are represented graphically by a Cartesian graph which has on the x-axes the values of F1, and in the y-axes those of F2. Each vowel is then represented by a point in the vowel space. Any multiple occurrences of the same vowel are represented by the average of values obtained. This procedure is applied to the original signal and to all the signals processed with different channels. The results are shown in Section 0.

In Figure Figure 5.7 is shown a representation of the vowel space for a sentence in the dataset.

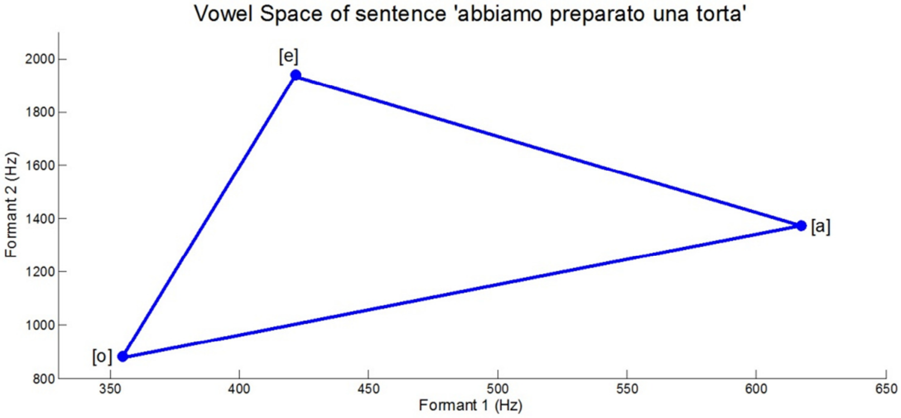


Figure 5.7: Vowel space of the sentence 'abbiamo preparato una torta'. This is the representation of the original signal.

5.5.6 Second and third formant power

The values for the second and third formants of voicing segments of analyzed words are compared for all the processed signals. The analysis takes into account only the values obtained without further processing and the representation is given for all channels considered in the results as shown in Section 6.7

In Figure 5.8 is shown a representation of these results for a word in the dataset.

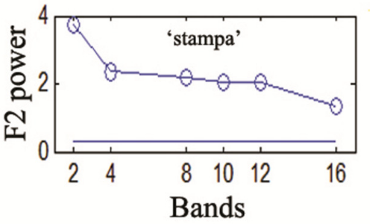


Figure 5.8: Power of the F2 of the first vowel /a/ of the word 'stampa'. The blue line represent the F2 power of the original signal, the blue circles represent the F2 power of the channel processed signals.

5.5.7 Consonant and vowel power

The values obtained from the computation of the power of plosives and fricative consonants and vowels in a word, are compared for all the processed signals by using their absolute value. The results of this analysis are shown in Section 0.

In Figure 5.9 is shown a representation of these results for a word in the dataset.

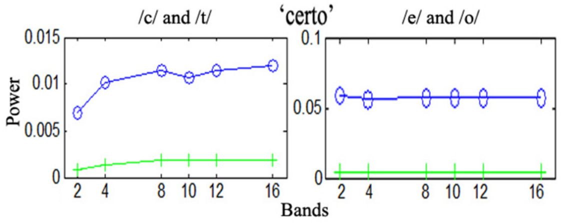


Figure 5.9: Consonant and vowel power for the plosive consonants /c/ and /t/ (left panel) and the vowels /e/ and /o/ (right panel) of the word 'certo'. The blue line represent the first occurrence, /c/and /e/, the green line represent the second occurrence /t/ and /o/.

5.5.8 CV-ratio

In the original signal, the two conditions of power ratio of adjacent consonants and vowels that can occur (the consonant power greater than the vowel power, and vice versa), requiring an analysis of the processed signals depending on the ratio in the original one. The ratio is then calculated as follows:

$$\mathrm{CVratio} = \begin{cases} \mathrm{ConsPWR/VowPWR} & \text{if } \mathrm{ConsPWR} > VowPWR \\ \mathrm{VowPWR/ConsPWR} & \text{otherwise} \end{cases}$$

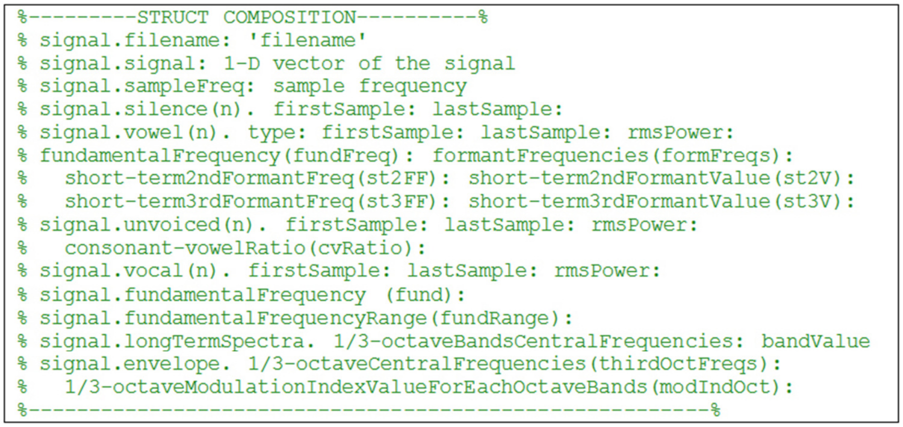
where ConsPWR is the consonant power and VowPWR is the vowel power. Having established this relationship, it is maintained in the computation of the CV-ratio for all the same CV couple in the processed signal.

5.6 Software

All the implementation regarding the Clear Feature Extractor, the Cochlear Implant Speech Generator, comparison of results and the charting is developed in MATLAB R2009a.

The analysis and segmentation of the speech signal is performed with Audacity ® 1.3.12-beta (Unicode),

The values and results of the features extraction by the CFE, obtained from the original signal and from the CISG processed version is stored in a MATLAB struct as the following:



```
%---------STRUCT COMPOSITION----------%
% signal.filename: 'filename'
% signal.signal: 1-D vector of the signal
% signal.sampleFreq: sample frequency
% signal.silence(n). firstSample: lastSample:
% signal.vowel(n). type: firstSample: lastSample: rmsPower:
% fundamentalFrequency(fundFreq): formantFrequencies(formFreqs):
%   short-term2ndFormantFreq(st2FF): short-term2ndFormantValue(st2V):
%   short-term3rdFormantFreq(st3FF): short-term3rdFormantValue(st3V):
% signal.unvoiced(n). firstSample: lastSample: rmsPower:
%   consonant-vowelRatio(cvRatio):
% signal.vocal(n). firstSample: lastSample: rmsPower:
% signal.fundamentalFrequency (fund):
% signal.fundamentalFrequencyRange(fundRange):
% signal.longTermSpectra. 1/3-octaveBandsCentralFrequencies: bandValue
% signal.envelope. 1/3-octaveCentralFrequencies(thirdOctFreqs):
%   1/3-octaveModulationIndexValueForEachOctaveBands(modIndOct):
%--------------------------------------------------------%
```

Chapter 6

Results

This chapter discusses all the results obtained from processing of signals from the dataset in their original version and processed by the Cochlear implant speech generator (CISG) as explained in Section 5.2.1. First the impact of the CISG on the signal is analyzed by comparing the processed versions with the original one. Following are regularly exposed the results obtained from the extraction of the characteristics of the original signal and of the processed signals, through the Clear Features Extractor, as described in Chapter 4.

6.1 Cochlear Implants Speech Generator (CISG)

The signal processed through the CISG for the Italian word 'conti', belonging to the dataset used, is shown as an example in Figure 6.1, in which are visible the original waveform (top panel) and its processed version with two channels (middle panel), with bands ranging from 188-1063 Hz and 1063-7928 Hz, and processed with twelve channels (bottom panel), with bands as described in Table 5.2.

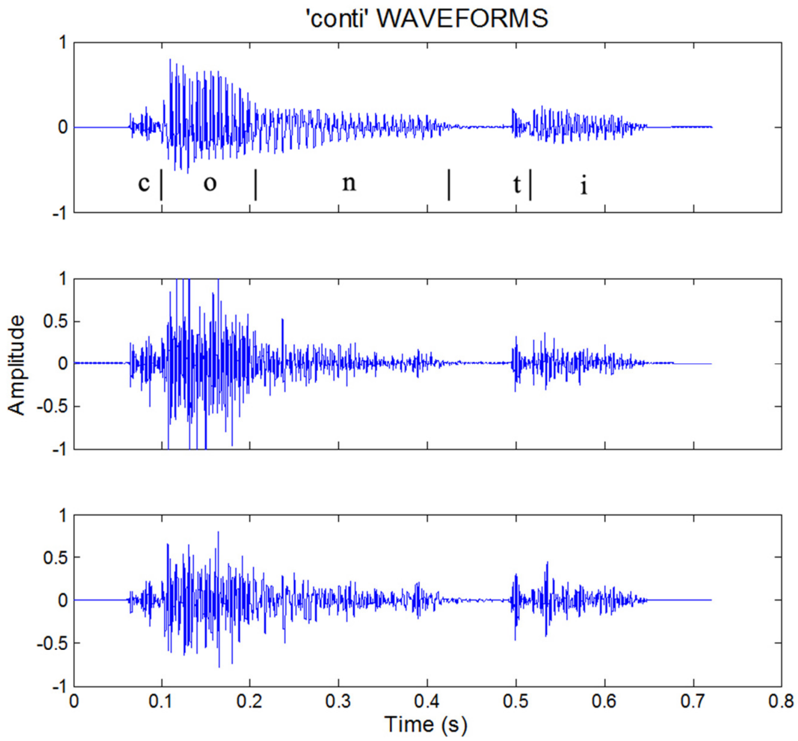


Figure 6.1: Waveforms of Italian word 'conti'. In the panel are visible the waveform of the original signal (top panel), the signal processed by the CISG with 2 channels (middle panel) and the signal processed by the CISG with 12 channels (bottom panel). Note the preservation of the envelope in both processed signals and the improvement in the approximation of the signal with the increase of the number of channels. It is also perceived a greater definition of the various phonemes forming the word.

In the time representation of the original signal various phonemes forming the word in question are identified. It is easy to distinguish the vocals, which have a clear periodicity, including the nasal consonant /n/ characterized by a lower intensity than the preceding vowel, and voiced palatal consonants /c/ and alveolar /t/, distinguished by a shorter duration and the absence of periodicity.

The frequency characteristics in processed signals are modified by the processing. A further analysis can be performed on the signal spectrum, as shown in Figure 6.2, in which the signal processed with two channel show a great difference with the original one, introducing various un-wanted frequency and show lower approximation with the original signal. The twelve band processed signal, instead, can replace the trend of the original signal, concentrating its distribution of frequency in the same range as the original one.

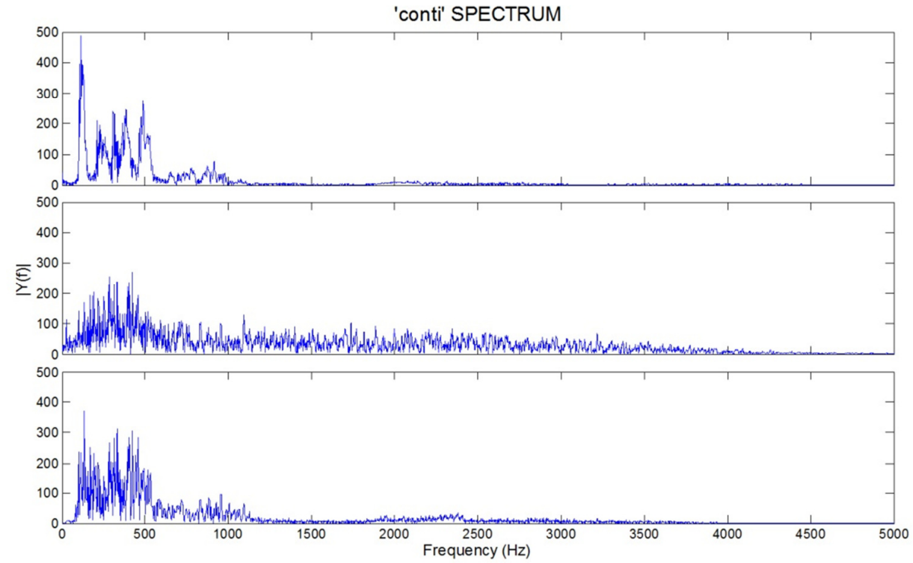


Figure 6.2: Spectrum of the Italian word 'conti' (top panel) and of its processed version with two channels (middle panel) and twelve channels (bottom panel), through the Cochlear Implant Speech Generator. It is evident how the twelve channel processed signal reproduce the original signal frequencies distribution, better than the two channel processed signal, that introduce unwanted frequencies especially in the range above 1000 Hz.

The processed signals are then used for the extraction of relevant features in the definition of clear speech, the results of which are described below.

6.2 Long-term spectra

The extraction of this parameter from the dataset of words in the frequency range between 0 and 5000 Hz are shown in Figure 6.3 and Figure 6.4 and shows the long-term power spectrum in absolute value in third-octave bands representation ranging from 62.5 to 5000 Hz.

For each word are then submitted to the representations of the original signal, the 2 channels processed signal (dotted red line) and 12 channels processed signal (green solid line), distinguishable as from legend.

For the original signal (solid blue) are recognizable a significant component of amplitude in the frequency range between 80 and 160 Hz and between 300 and 800 Hz. After a detectable decrease in up to 1000 Hz there is again an increase in the band between 1000 and 3000 Hz.

For the 2 channel processed signal (red dotted line) there is not significant variation, however, the maximum values are reached at about 2000 Hz for all the words.

For the 12 channels processed signal (green solid line) the trend of the amplitudes follows the original signal while maintaining lower values in the range between 80 and 160 Hz and higher values in the range from 1000 to 3000 Hz.

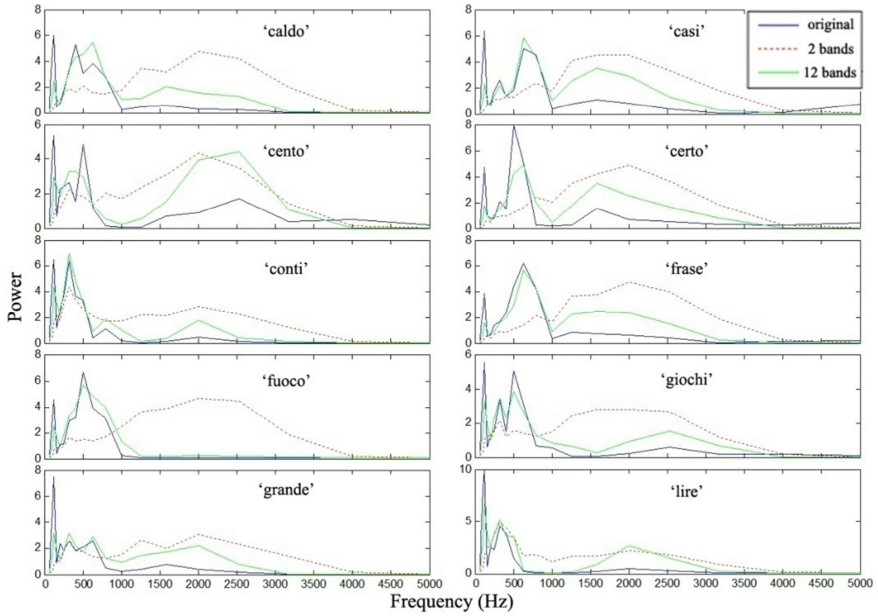


Figure 6.3: Representation of the extraction of long-term spectra for the first 10 words of the dataset. In each panel are represented the results by the Clear Features Extractor for the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with twelve channels (green

solid line). Bands which are concentrate the higher amplitude values are evident, i.e. the range between 80-160 Hz, 300-500Hz and 1000-3000Hz.

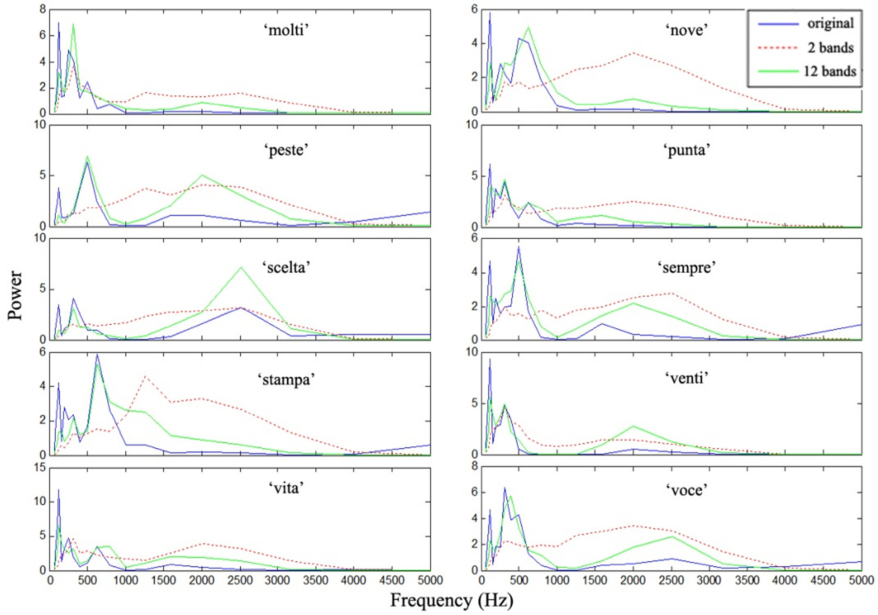


Figure 6.4: Representation of the extraction of long-term spectra for the second 10 words of the dataset. In each panel are represented the results by the Clear Features Extractor for the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with twelve channels (green solid line). Bands which are concentrate the higher amplitude values are evident, i.e. the range between 80-160 Hz, 300-500Hz and 1000-3000Hz.

A specific analysis is performed for the band between 1000 and 3000 Hz, shown in Figure 6.5 and Figure 6.6. This view represents the ratio of the absolute value of each frequency in the band and the sum of the absolute values of the band, for each signal independently, as explained in Section 5.5.1. From this representation is more evident the increase in the amplitude values in this band, and the differences between the absolute values of the different signals are leveled.

For the original signal (blue solid line) there is a distribution of peaks for different words, between 1500 and 2500 Hz

In the two channels processed signal (red dotted line) the distribution is fairly uniform, with no evidence of the presence of peaks at certain frequencies.

The twelve channels processed signal (green solid line) follows the same trend as the original signal, presenting some peak values at frequencies other than the original, as for the words 'frase',' grande', 'molti', 'peste', 'punta' and 'sempre'.

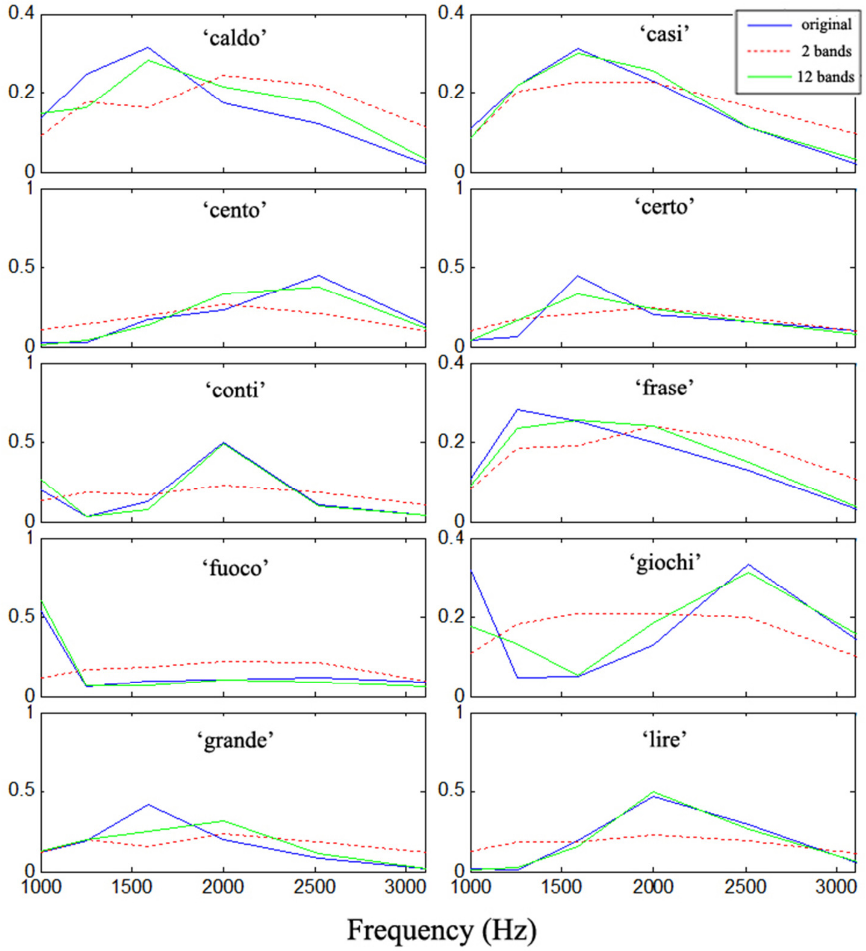


Figure 6.5: Representation of the extraction of long-term spectra for the first 10 words of the dataset in the range between 1000 and 3000 Hz. In each panel are represented the results from the Clear Features Extractor to the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with twelve channels (solid green) represented as the ratio between the amplitude and the sum of the individual amplitude values in the band. Note the more significant peaks of the original signal than the signals processed.

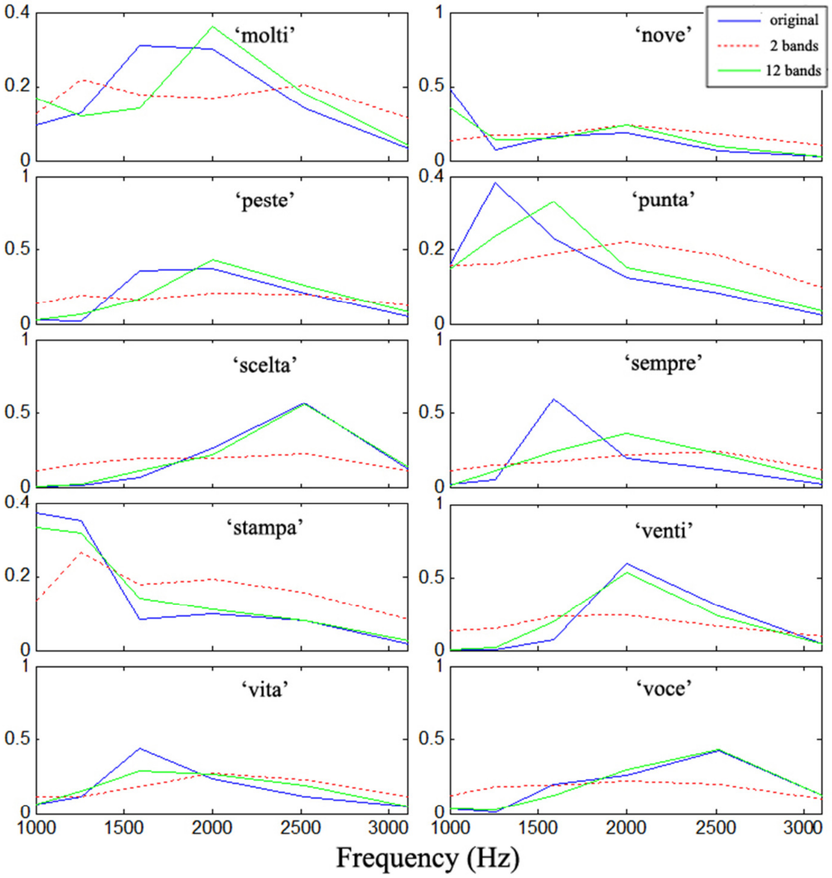


Figure 6.6: Representation of the extraction of long-term spectra for the second 10 words of the dataset in the range between 1000 and 3000 Hz. In each panel are represented the results from the Clear Features Extractor to the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with twelve channels (solid green) represented as the ratio between the amplitude and the sum of the individual amplitude values in the band. Note the more significant peaks of the original signal than the signals processed.

The overall analysis shows a greater intensity of the peak values of the original signal from the values of the signals processed when coincident. The processed signal to 12 channels is clearly more faithful to the original, compared to the signal with 2 channels.

Figure 6.7 shown in detail the analysis in the range 0 to 5000 Hz (top panel), and range from 1000 to 3000 Hz (bottom panel) in which are represented the normalized values, as explained above, for the word 'giochi'.

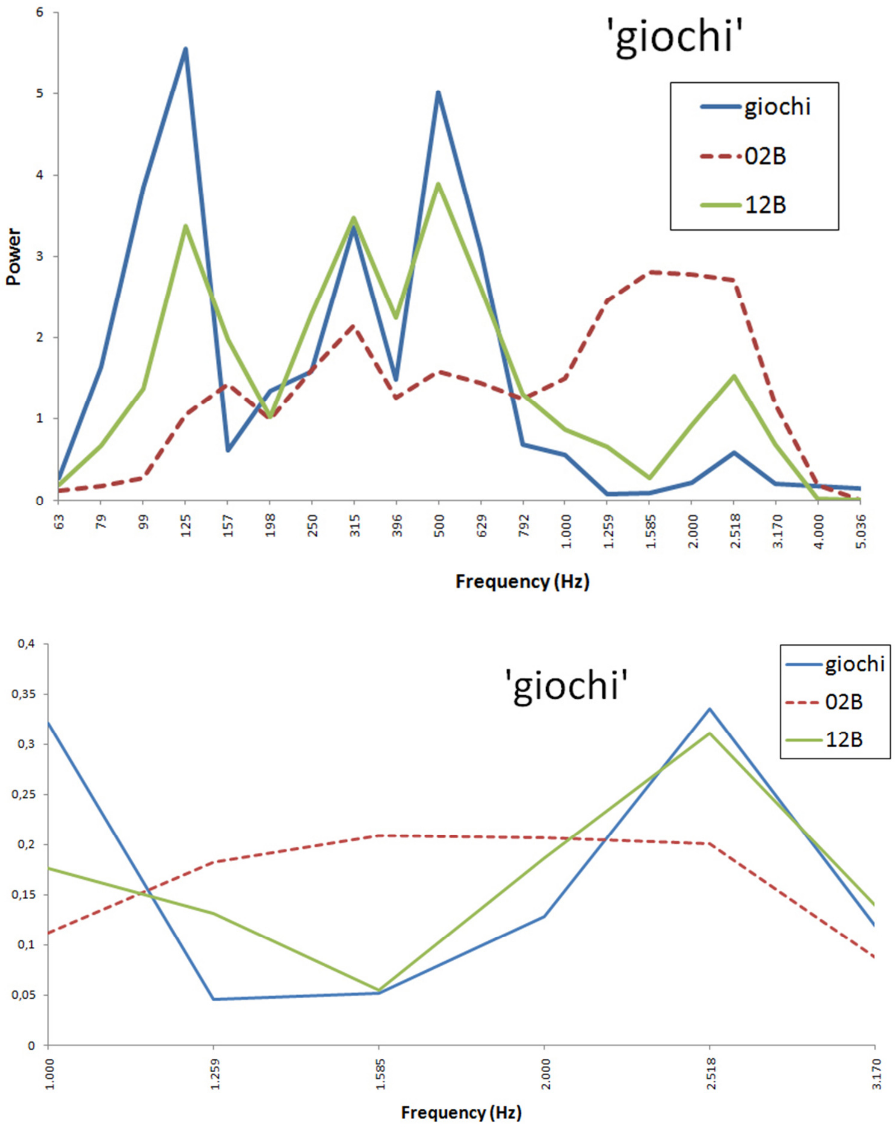


Figure 6.7: Long term spectra for the word 'giochi'. Top panel represent the spectra in the range 0 – 5000 Hz for the original signal (blue line), the 2 channel-processed signal (red dot line) and the 12 channel-processed signal (green line). Bottom panel is the representation of the ratio as in Figure 8.3.

In Figure 6.8 the analysis is given for the word 'punta'.

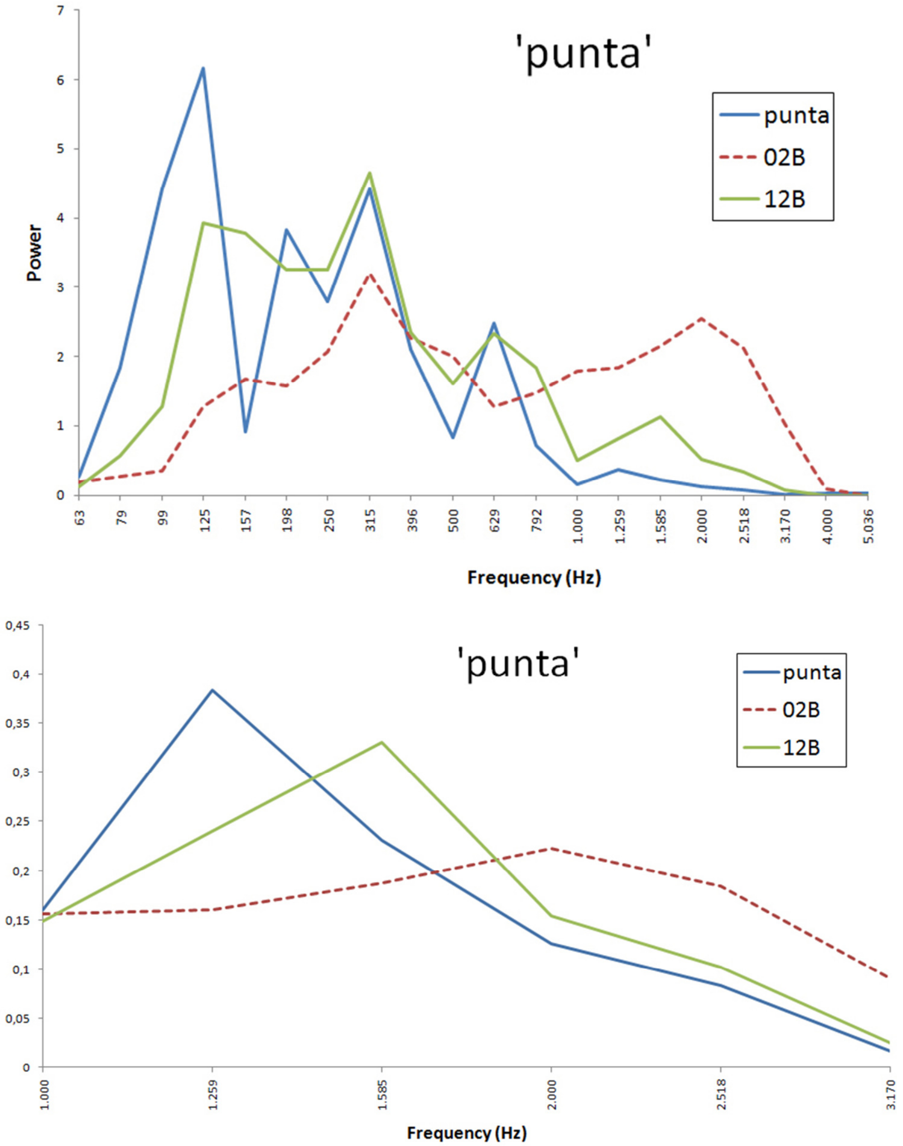


Figure 6.8: Long term spectra for the word 'punta'. Top panel represent the spectra in the range 0 – 5000 Hz for the original signal (blue line), the 2 channel-processed signal (red dot line) and the 12 channel-processed signal (green line). Bottom panel is the representation of the ratio as in Figure 8.3.

From these two examples is more evident the relevance of the original signal in the band 1000 - 3000 Hz analyzing values in proportion to the band, compared to absolute values.

6.3 Modulation index

The modulation index, obtained as explained in section 5.5.2 is shown for the entire data set of words in Figure 6.9 and Figure 6.10, in which the value is represented in the band between 0.4 and 20 Hz as third-octave bands.

The analysis of the original signal (blue solid line) shows a similar pattern for all the words, with peak index modulation concentrated around 1 Hz and between 2 to 4 Hz. In some cases, such as the words 'certo' and 'peste' peak between 2 and 4 Hz has higher values than the other frequencies, while for the word 'nove', 'frase', 'casi' and 'lire' its value is lower. From 4 Hz onwards, there were no further observed peaks on the modulation index.

The analysis of the processed signal to two channels (red dashed line) shows a similar pattern to the original signal, with similar values in an ever-defined major or minor.

The 12 channels processed signal (green solid line), however, follows the pattern of the original signal with values of modulation index higher than this reference.

In Figure 6.11 are shown more in detail the processing of the modulation index for two words, 'caldo', 'casi' (respectively top and bottom panel), from those that show the differences exposed previously.

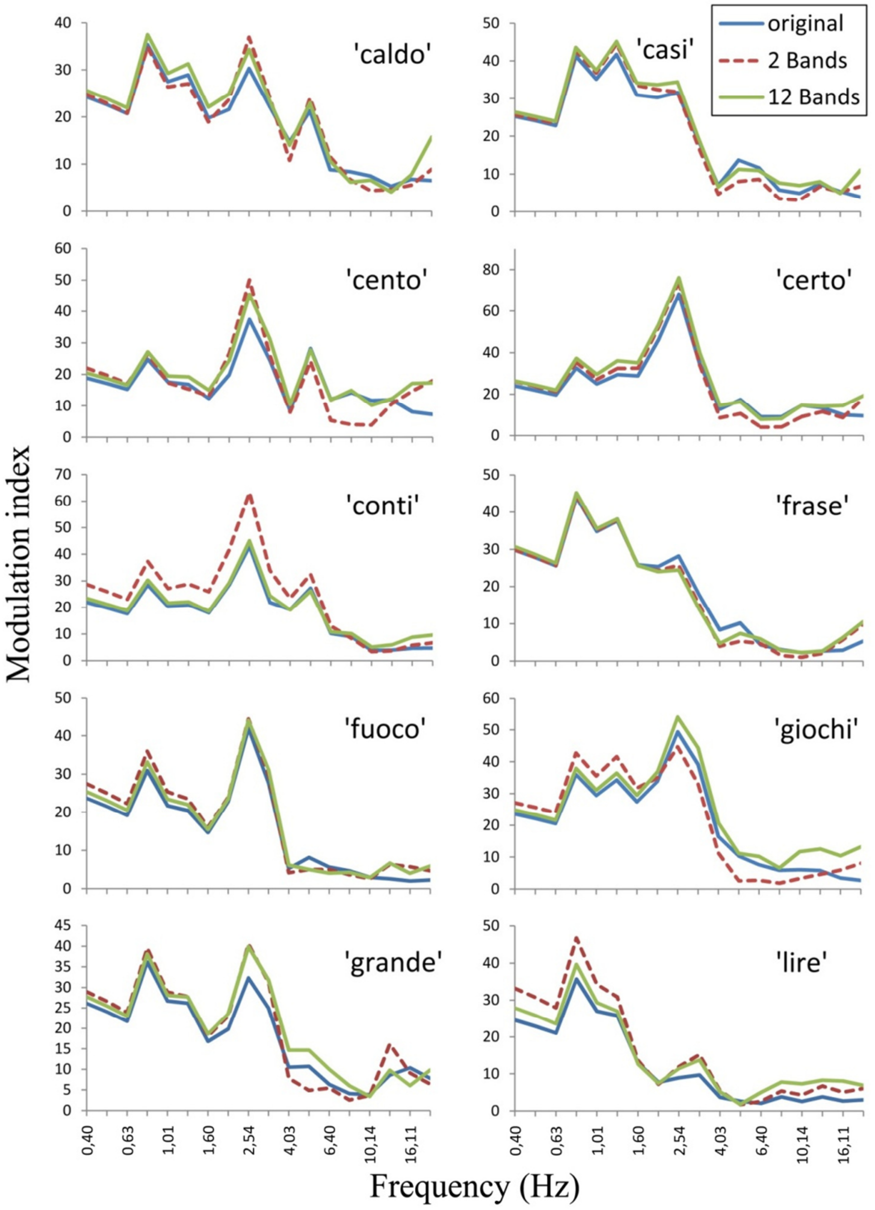


Figure 6.9: Representation of the modulation index for the first 10 words of the dataset in the range between 0 and 20 Hz. In each panel are represented the results from the CSE to the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with ten channels (solid green) represented as explained in Section 5.5.2

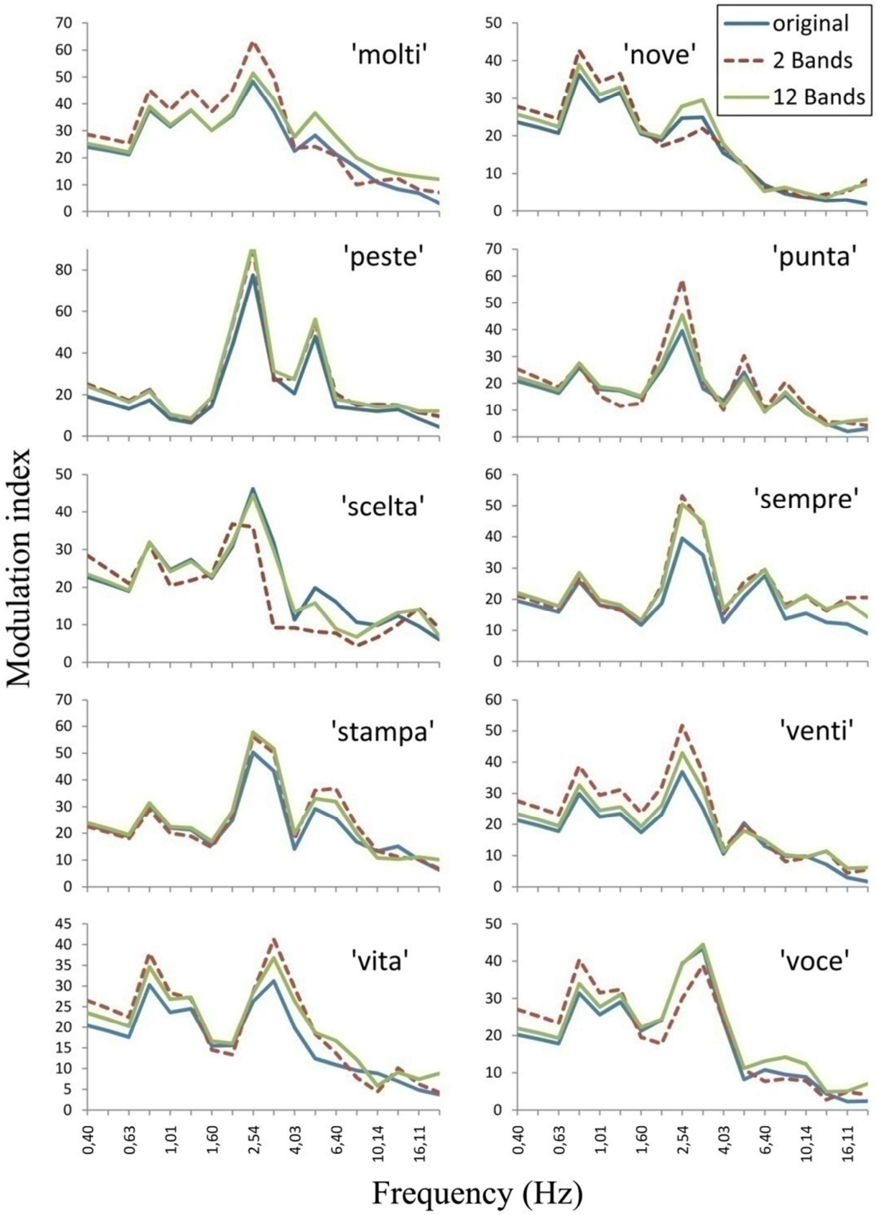


Figure 6.10: Representation of the modulation index for the second 10 words of the dataset in the range between 0 and 20 Hz. In each panel are represented the results from the CSE to the original signal (blue solid line), the processed signal with two channels (red dotted line) and the processed signal with ten channels (solid green) represented as explained in Section 5.5.2

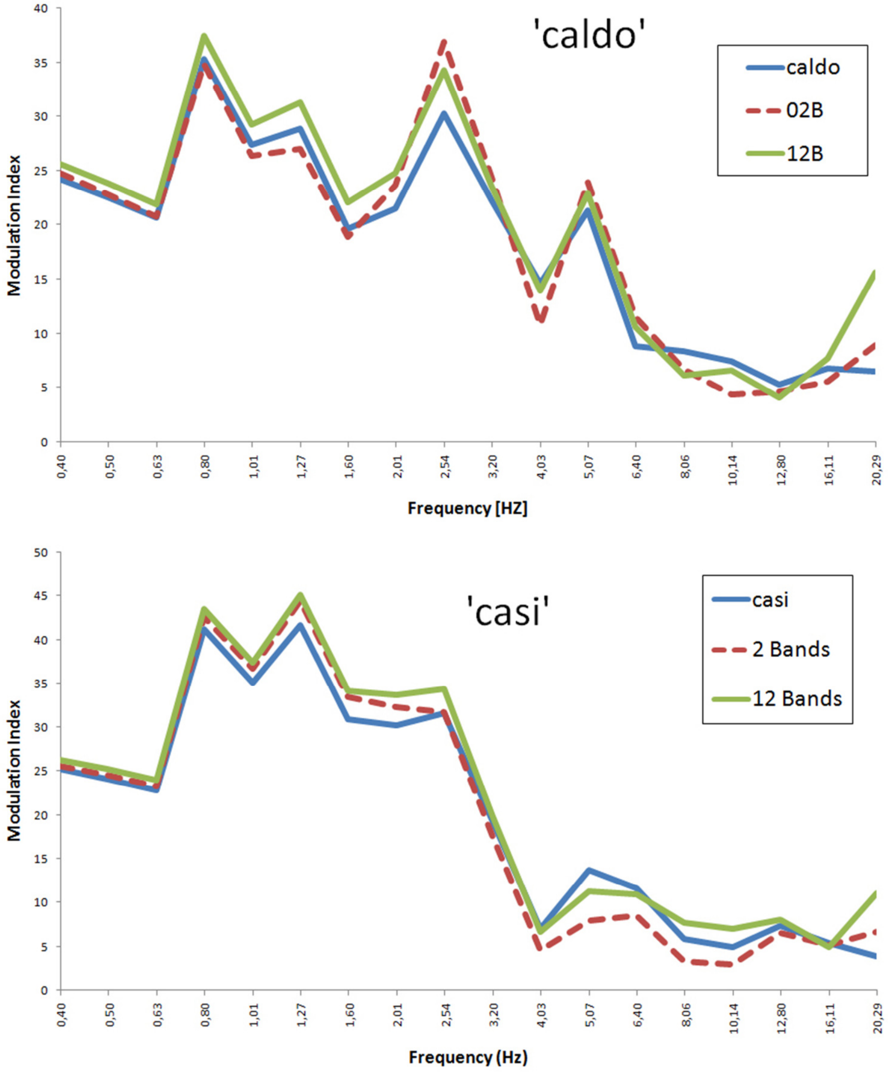


Figure 6.11: Modulation index for the words 'caldo' (top panel) and 'casi' (bottom panel). In the modulation index of the first word it is visible the peak between 2 and 4 Hz that is more pronounced. Also note the trend of the two channels processed signal, variable around the trend of the original signal and the twelve bands processed signal trend, always higher. From 4 Hz there are no additional peaks.

6.4 Fundamental frequency

The values obtained for the fundamental frequency are shown for each processed signal, referring to the fundamental of the original word. In Figure 6.12, this ratio is represented for each word in the dataset. The fundamental frequency value for each processed signal is represented by the blue circles, and the blue solid line defines the trend. The green solid line represents the reference value that is the fundamental frequency of the original signal estimated on the entire signal. We recognize common patterns representing a better representation of the fundamental frequency in the processed signals, with the increase of the number of channels. For some words, the improvement is only accentuated by the processed signal with 12 channels on, but the limit values are well approximated.

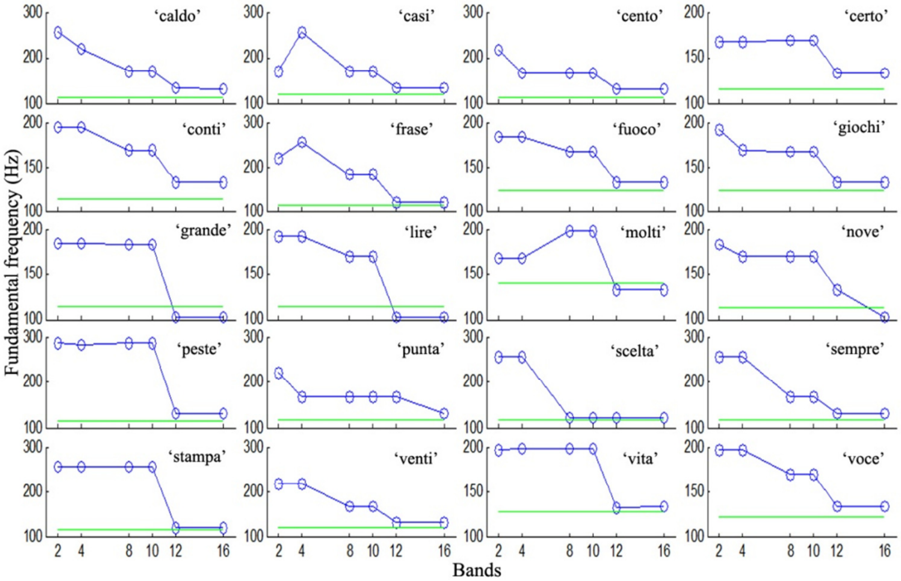


Figure 6.12: Representation of the ratio between the estimated fundamental in each processed signal for different channels and the estimation of the fundamental of the original signal. The improvement of the approximation as the number of channels of signal processing is evident.

6.5 Fundamental range

The frequency range of detection of the fundamental improves for many words in analysis. In Figure 6.13 the global trend of the frequency range are represented. The green line represents the range of the fundamental frequencies of the original signals of all the word in the dataset. A similar computation is performed for every set of processed signal with the same number of channels (e.g., for all the signal processed with two channels, representing all the word in the dataset, the range of the fundamental frequency are extracted). The results of these computations are represented with a blue circle line. In this representation is evident the best approximation to the original frequency range, given by the processed signal with highest number of channels.

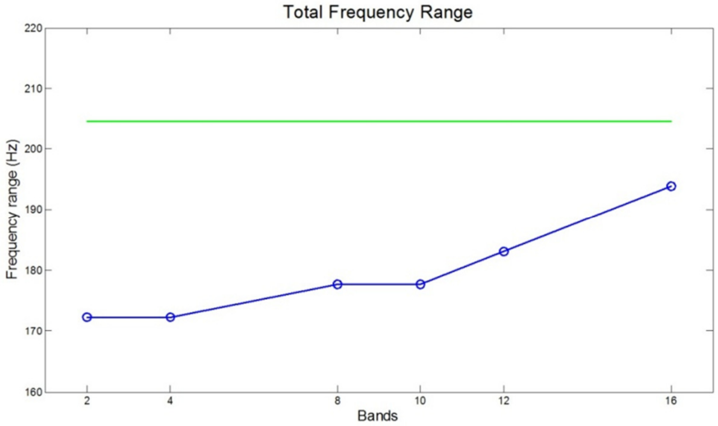


Figure 6.13: Representation of the range of deviation from the fundamental. In blue are represented the values of the processed signals, while green shows the range of the original signal.

6.6 Vowel space

The vowel space is calculated on the sentences of the dataset as they allow a better comparison for the greatest number of occurrences of the vowels. In Figure 6.14 shows the vowel space of the sentence 'franco è andato via di corsa' belonging to the dataset examined. The graph shows the average position of the occurrence of vowels in the sentence, to a Cartesian space represented by the first formant F1 and the second formant F2 recorded for each vowel.

As per legend, in blue shows the vowel space of the original signal. For the processing channels 2 and 4 the values obtained are not those of the original signal. From the 8 channel above, the processed signals gives a good estimation of the first two formants, approximating the original vowel space. As the increase of number of channels the values are significantly closer to the original signal.

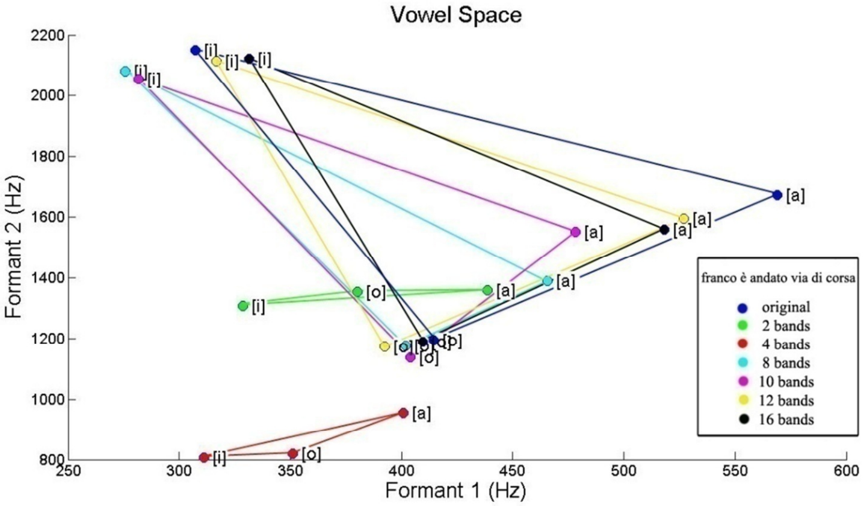


Figure 6.14: Vowel space of the sentence 'franco è andato via di corsa'. They represent the vowels of the sentence through the first and second formant estimation (respectively F1 and F2), mediated on the occurrences. The vowel space of the original signal is larger and is best approximated by the processed signals with greater number of channels.

As for the previous representation, in Figure 6.15 and Figure 6.16 are represented the vowel space of the other two sentences in the dataset, respectively 'abbiamo preparato una torta' and 'è rimasto solo al mondo'.

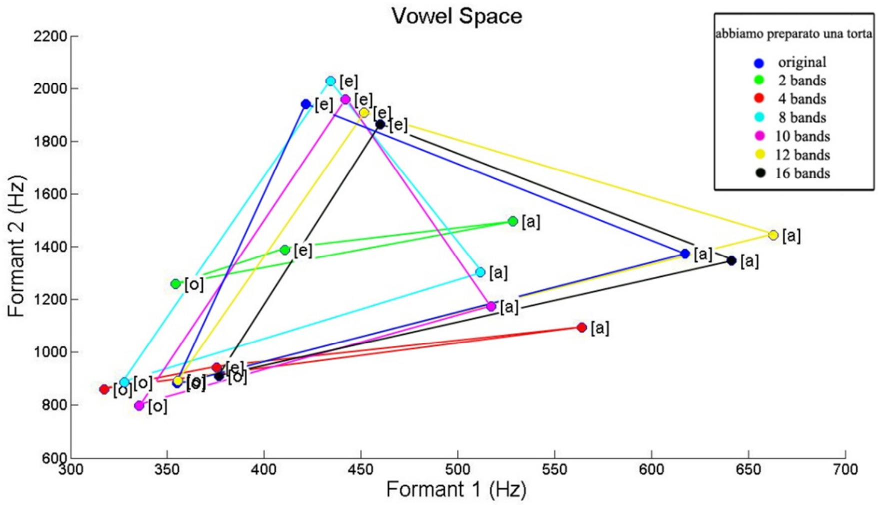


Figure 6.15: Vowel space of the sentence ‘abbiamo preparato una torta’.

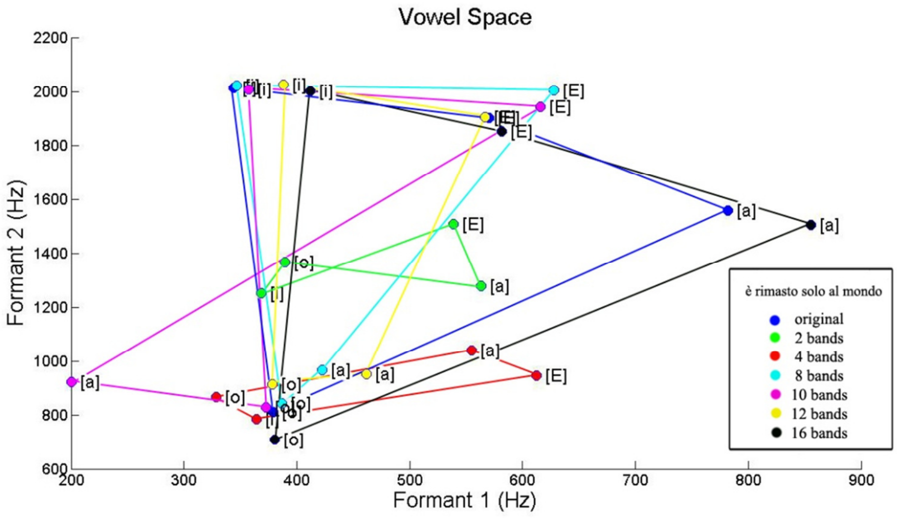


Figure 6.16: Vowel space of the sentence ‘è rimasto solo al mondo’. The error notice in the representation for the vowel ‘a’ even with an high number of bands channels is probably given by the single occurrence of the vowel ‘a’ (the other occurrence is part of a diphthong and then is not taken in account) and its position adjacent to a voiced sound like ‘m’ that make difficultly distinguishable the segmentation of the vowel.

As before, for the processing channels 2 and 4 the values obtained are not those of the original signal. As the increase of number of channels the values are significantly closer to the original signal. The error notice in Figure 6.16 for the representation for the vowel ‘a’ is probably given by the single occurrence of the vowel ‘a’ (the other occurrence is part of a diphthong and then is not taken in account) and its position adjacent to a voiced sound like ‘m’ that make difficultly distinguishable the segmentation of the vowel.

6.7 Second and third formant power

Regarding the power of the second and third formant of each vowel for the words in the dataset, the results are presented below as follows: in Figure 6.17 are represented the values of the second formant (F2) for the first vowel for each processed (circles), compared to the rms of F2 of the first vowel of the original signal (solid line). Similarly to Figure 6.18 are represented the values of F3 of the first vowel in each word with respect to the continuous line representing the rms of F3 of the first vowel of the original signal.

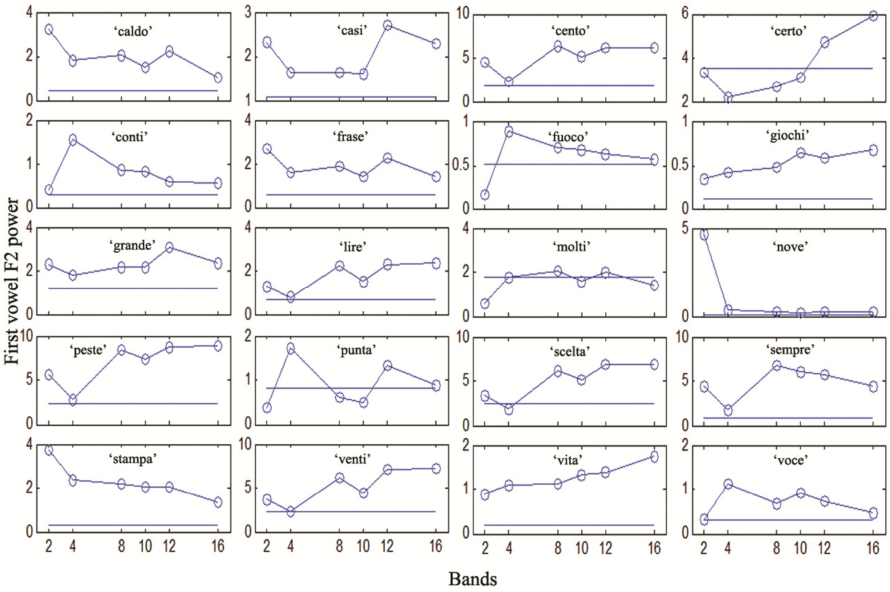


Figure 6.17: Power values of second formant of the first vowel of words in the dataset (circles). It is recognizable a pattern that approximates the value of the original signal (solid line).

The second formant of the first vowel to Figure 6.17 shows to approximate well the original value to most of the words under consideration. The exception are the words 'cento', 'peste', 'scelta', 'sempre', 'venti' and 'vita'.

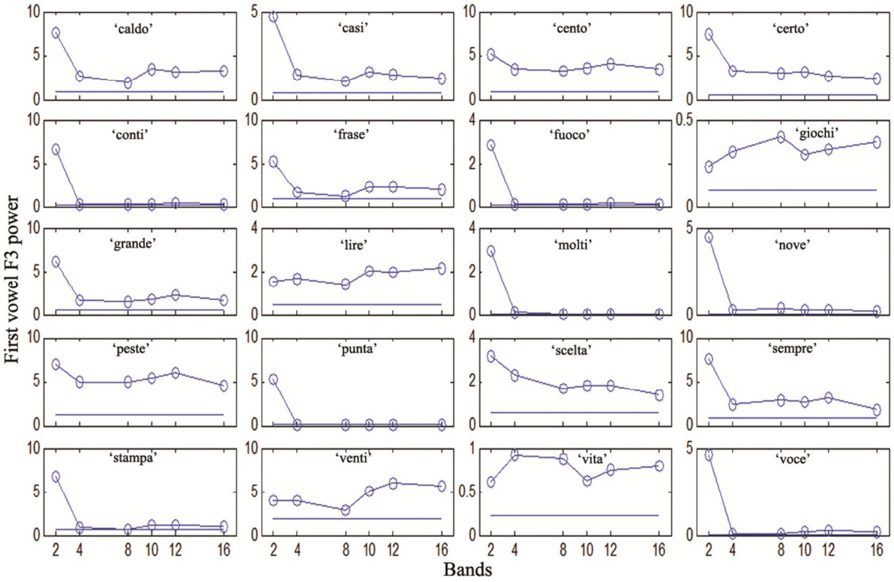


Figure 6.18: Power values of the third formant of the first vowel of processed signals (circles) referred to the power value of the third formant of the first vowel of the original signal (solid line).

For the third formant there is, instead, a close approximation of the values of power. Different trends are found only to 'peste' and 'venti' that show much higher values of the original signal. The improvement is most evident in signals processed with more channels.

In Figure 6.19 are represented the values of the second formant (F2) for second vowel signal for each processed signals (crosses), compared to the F2 of the second vowel of the original signal (solid line). Similarly to Figure 6.20 are represented the values of F3 of the second vowel in each word with respect to the continuous line that is the F3 of the second vowel of the original signal.

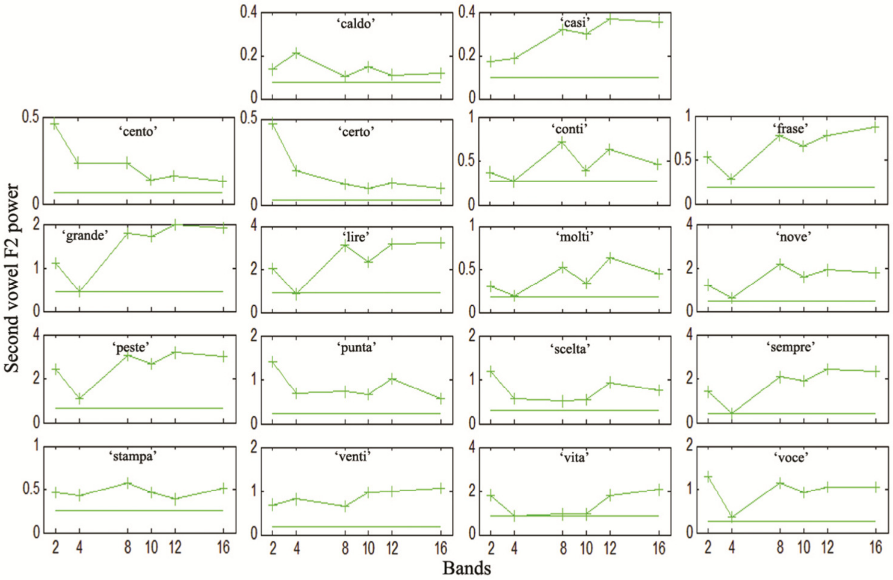


Figure 6.19: Power values of second formant of the second vowel of the words in the data set (crosses) compared to the value of second formant of the second vowel of the original (solid line), where exists.

For second formant of the second vowel there is a good approximation for every word in question.

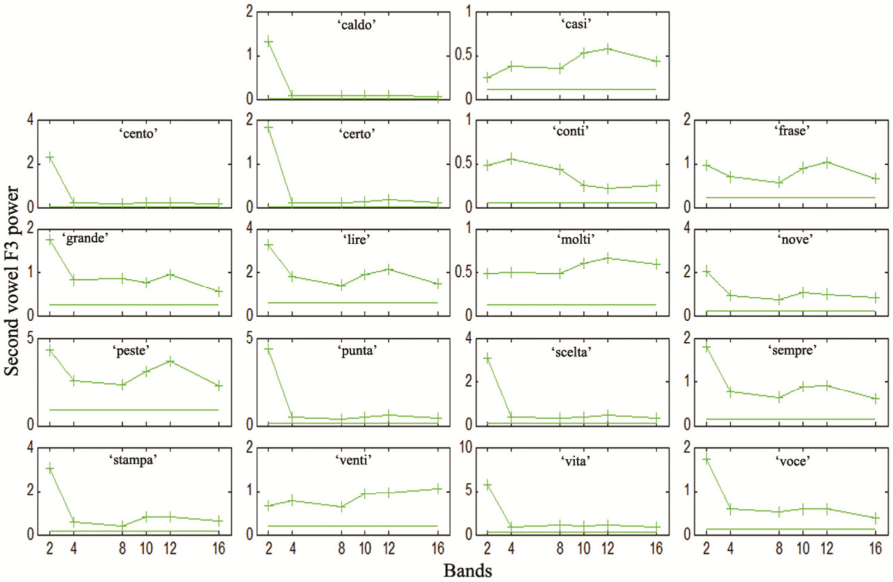


Figure 6.20: Power values of third formant of the second vowel of the words in the data set (crosses) compared to the value of third formant of the second vowel of the original (solid line), where exists..

Similarly the trend of the third formant of the second vowel approximates very well the value of the third formant of the second vowel of the original signal. The improvement is most evident in signals processed with more channels.

In comparison to the same word between two vowels, the second is to always have lower values, except in ‘lire’, ‘nove’, ‘voce’ and ‘punta’.

6.8 CV-ratio

The CV-ratio shown in Figure 6.21 shows the ratio between the power of plosive and fricatives consonants, and the adjacent vowels in the same syllable. This representation exposes the ratio in absolute value for each processed signal with several channels, for each consonant-vowel pairs of each word, in this way: the first pair is represented by the color blue with circle points, the second pair by the color green with cross points, the third pair by the color red with star points.

The words 'fuoco', 'grande', 'lire' and 'nove' are not shown in the figure because it does not contain plosive or fricative consonants (in the case of 'fuoco' are indistinguishable).

From the figure in question common trends of the word are not distinguishable, either between the words themselves. A reversal of the power ratings between consonants and vowels is noticeable in words like 'giochi', 'molti', 'voce' and 'casi' and in the second consonant-vowel pair of the word 'cento'. In these cases, the power ratio between consonant and vowel is inverted by signal processing (consonants that showed less power of vowel, shown higher values in the processed signals).

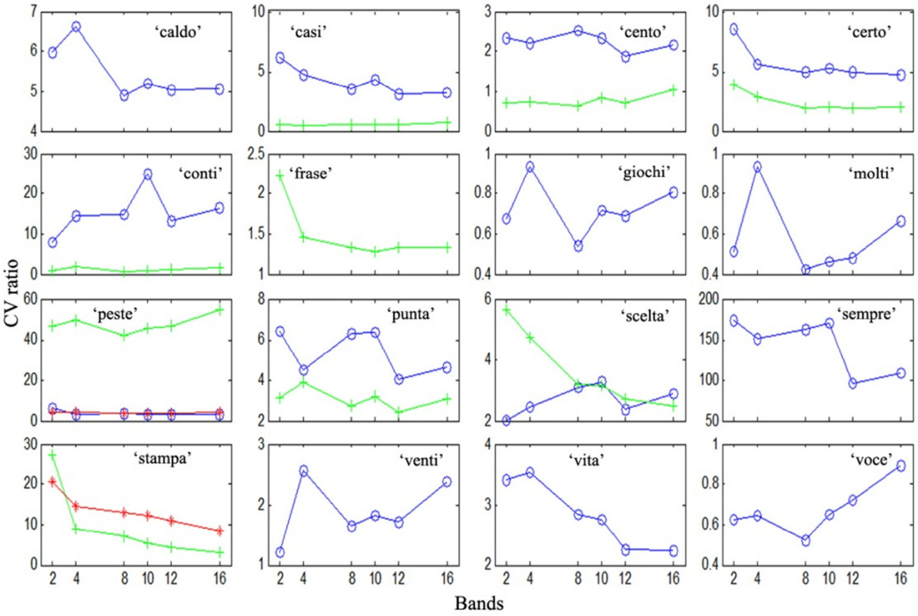


Figure 6.21: CV-ratio for each word in the dataset. The first pair of consonant-vowel of each word is represented by the color blue, the possible second pair by color green, and the possible third pair by the red. It is not an appreciable common pattern that defines an increase in the power ratio.

This analysis is helpless to distinguish some variation. Apparently, the CV-ratio does not give reliable indications on the influence of the number of channels. There are different trend, in words like ‘venti’ the CV-ratio increase with the number of channel increase, in words like ‘caldo’ did not. Then can be helpful a specific analysis of this two words, in the variation of the spectrum of their consonant/vowel couple.

In Figure 6.22 is represented spectrum of the pair /c/-/a/ in ‘caldo’, for the original signal (blue solid line), the two channel processed signal (red dotted line) and the twelve channel processed signal (green solid line). For the consonant /c/ the representation of the spectrum is limited between 0 – 5000 Hz, the approximation of the 12 channel signal spectrum to the original one is better than the one of 2 channel signal spectrum, but it shown a great number of spurious components. For the representation of vowel /a/ the spectrum is limited in the most relevant region between 0 – 2000 Hz, the 12 channel processed signal spectrum shown that this processing is able to reproduce the fundamental frequency (even if the first channel is limited to 188 Hz, above the fundamental frequency), and to reproduce the frequency distribution, with a great reduction of components above 1000 Hz.

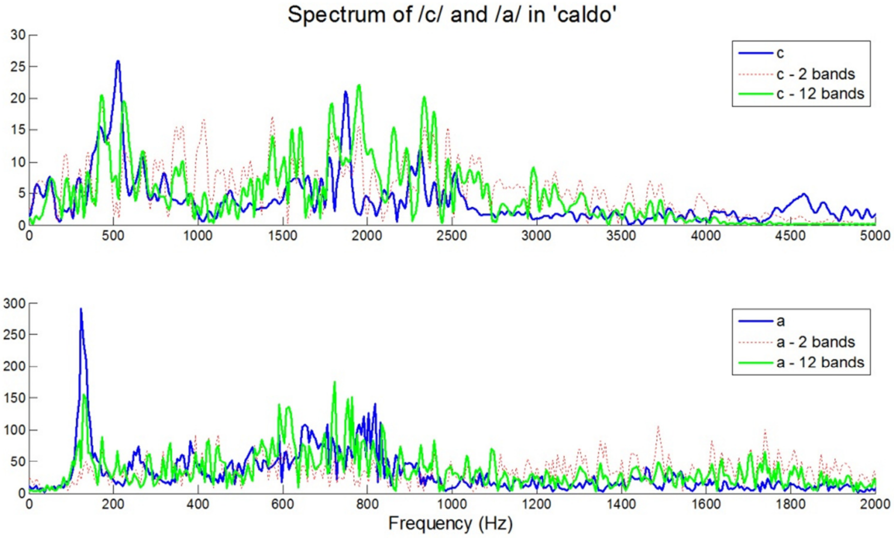


Figure 6.22: Spectrum of the consonant-vowel couple /c/-/a/ in ‘caldo’. It is represented the computation for the original signal (blue solid line), the two channel processed signal (red dotted line) and the twelve channel processed signal (green solid line).

In Figure 6.23 the spectrum of the couple /t/-/i/ in ‘venti’ is represented, for the original signal (blue solid line), the two channel processed signal (red dotted line) and the twelve channel processed signal (green solid line). For the consonant /v/ the representation of the spectrum is limited between 0 – 5000

Hz, The representation for the 12 channels signal show an increment in the frequency component in the same range of the increment in the original signal, but the spurious component is always present. For the representation of vowel /a/ the spectrum is limited between 0 – 2000 Hz, the 12 channel processed signal spectrum shown, as before, the same trend of the original signal, with a great reduction of components above 600 Hz.

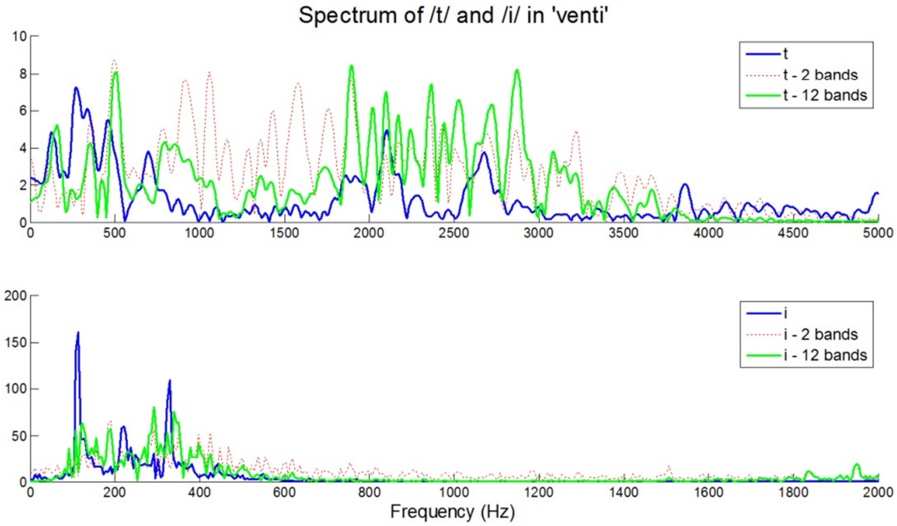


Figure 6.23: Spectrum of the consonant-vowel couple /t/-/i/ in 'venti'. It is represented the computation for the original signal (blue solid line), the two channel processed signal (red dotted line) and the twelve channel processed signal (green solid line).

An analysis of the power of consonants and vowels, represented respectively in Figure 6.24 and Figure 6.25, shows how the greatest variation of power takes place on consonants rather than vowels, which maintain stable values for each processed signal. The power of the consonants then appears to have a greater impact on CV-ratio.

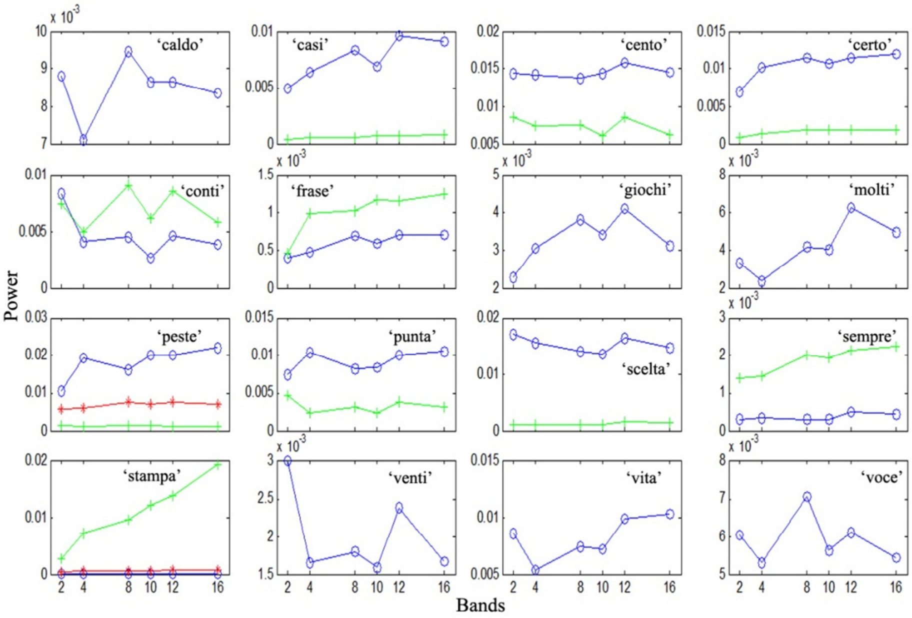


Figure 6.24: Plosive and fricative consonants power for each word in the dataset. The values of the first consonant of each word are represented in blue, green to second eventual, red for third.

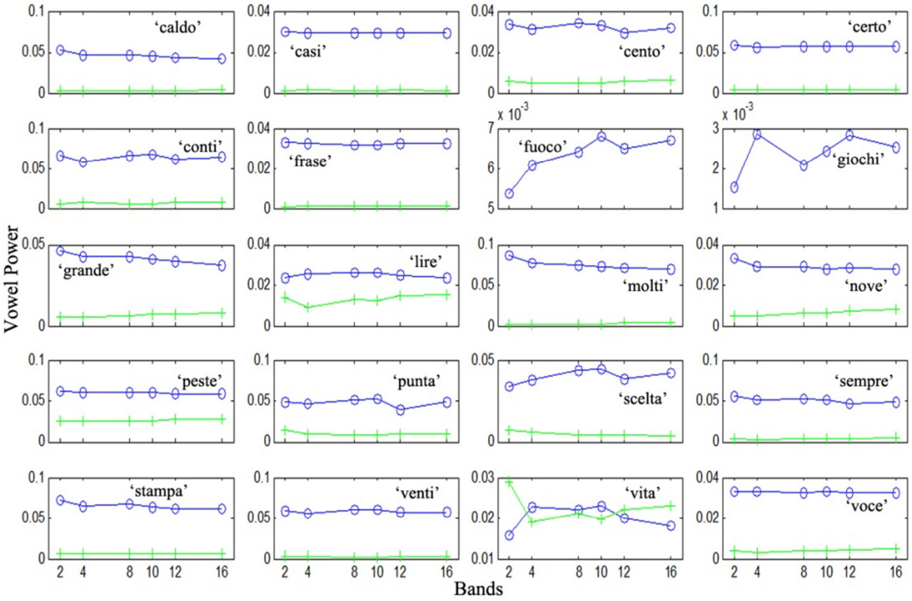


Figure 6.25: Vowel power for each word in the dataset. The values of the first consonant of each word are represented in blue, green to second eventual, red for third.

Chapter 7

Discussions and conclusions

7.1 General remarks

The CISG is the realization of an algorithm that allows the generation of signals that simulate the sound perception in cochlear implant patients. The steps of signal processing are designed so as to reproduce the basic features of a cochlear implant sound generation system, by acting on certain key parameters that determine system performance. In this process the variable parameter is the number of channels through which the simulator filters the input signal, and on which performs the necessary processing. So is tested the influence that the number of channels has on the processed signal and which of these calculations will be able to better preserve the essential characteristics to the preservation of the intelligibility of clear speech.

For each original signal, therefore, it is proposed a set of processed signals as described in Section 5.2.1, with number of channels between 2 and 16. These signals are compared with the original, both in its overall spectral characteristics both perceptual and for any selected feature of clear speech, describing the impact these have on the signal processing.

The processing performed by the CISG modifies the signal, unbalancing the power relationships between the various components of the signal. As seen in Figure 6.1, with the increasing of the number of channels for which the signal is processed, an improvement of the characteristics of the signal is obtained over the processing at least two channels. This improvement is noticeable both visually, as shown in Figure 6.1, both acoustically. In the two-band signal the distinction between phonemes is unclear and tends to create confusion especially in the recognition of voiced consonants, or in the perception of plosive consonants that come after them.

Comparing the listening of the various processed signals, can be perceived a sharper definition of phonemes forming the word as the increase of number of channels, allowing a clear distinction between the vowels and the following plosives consonants.

7.2 Long-term spectra

The long term spectra provides a measure of the energy of signal spectrum, viewing it in third-octave bands, to investigate the spectral distribution, identifying the most interesting bands in the comparison between the original clear signal and the processed signals. The goal is to verify that the spectral distribution of the processed signal deviates from the distribution of the original signal, and on what frequency bands that any deviation to be more present, verifying the results obtained by Krause and Braida(Krause & Braida, 2004) Greenberg and Arai (Greenberg & Arai, 2004) and Kain et al (Kain, Amano-Kusumoto, & Hosom, 2008), which detected an increase in the band between 1000 - 3000 Hz, related to an intelligibility increase of clear speech.

From the original and processed signals, then, has been obtained the spectral representation as described in Section, 5.5.1, and has been presented a direct comparison.

Analysis of the results shows a similar trend in the values of amplitude of the spectrum for all the dataset as shown in Figure 6.3 and Figure 6.4. This is probably due to the exposure of the words in question, shown in Section 5.1, which tends to be atonic and without accents. The distribution of frequencies is characteristic of the speaker itself.

After this initial analysis it has been possible to select the most relevant signals for comparison, showing the major and minor differences from the original signal. The selected signals are then the one processed with two channels, and the one with twelve channels.

The trend associated to the 2 channels processed signal shows a substantial loss of amplitude variation in the spectrum as represented in Figure 6.3 and Figure 6.4, associated with the use of white noise as the fine structure of the signal and the low modulation introduced by the two bands. This uniformity is also acoustically noticeable associating the perception of this signal with a lack of clarity and intense noise.

The 12 channels processed signal follows the trend of the original signal as shown to Figure 6.3 and Figure 6.4 detecting a loss at low frequencies due to the mode of processing that provides the value of cutoff frequency of the first channel, set at 188Hz, that is to affect the spectral amplitude at low frequencies. The use of the noise signal as the fine structure, amplifies the spectrum components between 1000 and 3000 Hz.

This result, seemingly at odds with the expectations of the experiment, has been analyzed in more detail comparing the relative energy spectrum in the band concerned, as explained in Section 5.5.1.

The analysis in the band 1000 - 3000 Hz signal in Figure 6.5 and Figure 6.6 shows a greater importance of the original signal respect to its processed version. This is to agreement with the results expected that verify an increase in the considered energy band and a greater importance in terms of perception, as shown by Greenberg and Arai (Greenberg & Arai, 2004) and confirmed by Kain et al (Kain, Amano-Kusumoto, & Hosom, 2008).

The results presented to Figure 6.7 and Figure 6.8 for the words considered, clarify the results reported making them more evident. Also visible in these results is the influence of signal processing on the band 1000 - 3000 Hz (top panels) for both processed signals, while the specific analysis on this band (bottom panels) shows the relative energy of the original signal is still greater than that of processed signals, clearly in the peaks of higher energy.

7.3 Modulation index

The modulation index is a measure of temporal envelope modulation of the signal and provides a weight to the frequencies that compose it. From the analysis of the modulation index is possible to verify the overall variation of a signal and the influence of the frequencies characterizing the envelope.

On the original signal has been expected a significant component in the frequency band around 3 to 4 Hz, as observed by Krause and Braida (Krause & Braida, 2004), while processed signals must follows the trend of the original signal, because the processing is done by modulating a noise signal with the temporal envelope extracted from the original signal.

The modulation index are then extracted from the original signal and from signals processed as described in Section 4.1.2 and are then compared. After a first analysis have been selected the two and twelve channels processed signals as the most relevant, and was presented a comparison between these and the original signal.

The analysis of the modulation index trend for the original signal shows peaks at frequencies in the range between 2 and 4 Hz. This result is consistent with the expected results, the different trend for some words, as for ‘lire’, ‘nove’ and casi’ in the example in Figure 6.9 and Figure 6.10, can be associated with the absence of plosive consonants in these words that can determine a less pronounced peak around 3 Hz. The accentuation of this trend may be due to the presence of a segment of silence before the plosive consonant, which characterizes the envelope of the word.

The values obtained for the processed signals are very similar to the original pattern, as shown in Figure 6.9 and Figure 6.10, mainly due to the method of construction of signals, processed as the modulation of the envelope of the original signal. For the processed signals with two channels has been noticed a greater deviation from the original signal than the signals processed with twelve channels. This is due to the lower definition of the envelopes to two processing channels, which do not allow to adequately approximate the original performance.

A similar pattern is therefore expected and confirms the goodness of the signal. The change in value compared to the original signal can be attributed to the use of noise that affects the average value of the envelopes calculated on different channels or to a greater extent in the same elaboration as the sum of the filtered signal over channels with adjacent bands, which will determine a overlap between the values, and therefore a global increase.

The analysis of two single words as in Figure 6.11 allows assessing more thoroughly the changes in the modulation index in the words with different trends, especially in the band of interest between 3 to 4 Hz. In the modulation index of the first word it is visible the more pronounced peak between 2 and 4 Hz. Is also noticeable the trend of the two channels processed signal, variable around the trend of the original signal and the twelve bands processed signal trend, always higher.

7.4 Fundamental frequency

An analysis on the fundamental frequency of each word is used to determine the pitch change between the different styles of speech. In this case, the fundamental frequency is estimated to test the ability of the algorithm of signal processing, to maintain this feature that should become decisive in the perception of speech. Studies by Picheny et al (Picheny, Durlach, & Braida, 1986)and Krause and Braida (Krause & Braida, 2004) reported an increase in both the average value of the fundamental frequency, and in the range of variation, as concerns the comparison clear/conversational speech. In this experiment, the objective is to detect a maintenance of the values of the original signal, in the processed signals.

The extraction of the fundamental is performed as explained in Section 4.2.1 and the comparison for each word processed, is assessed by reference to the original. In this analysis were taken into account the average value of the fundamental of the speaker needed to better assess the differences.

The result show a better approximation of the fundamental frequency of the signals processed, to the value of the original signal with increasing number of channels as shown in Figure 6.12. This trend is founded for every examined word and can be evaluated as a confirmation of the goodness of the CISG elaboration.

The simulation, indeed, is based on a type of cochlear implant that does not preserve the information about the fundamental frequency, and then the trend found is excellent.

Also notice the major improvement from the processed signal with twelve channels, that is another confirmation of the goodness of the processing, reflecting the implementation choices of the actual cochlear implant products.

7.5 Fundamental frequency range

As the average fundamental frequencies, the range will be provided by a relative difference of the change in pitch. Although the trend for this feature does not require verification of the studies conducted by Picheny et al (Picheny, Durlach, & Braida, 1986) and Krause and Braida (Krause & Braida, 2004) but wants to verify the capability of preserving the characteristics of the original signals , in the set of processed signals.

The range are calculated as the difference between the maximum and minimum fundamental frequency determined on all the vowels of all words, for each set of words processed with the same number of channels. The results of all sets are compared with the original result, as explained in Section 0.

Similar to what has been discussed for the fundamental, the result of a reduction in variation of the fundamental frequency in the signals processed respect to the value of the original signal with increasing number of channels as shown in Figure 6.13, is evaluated as a confirmation of the goodness the CISG.

The considerations about the type of process used, linked to the simulated cochlear implants, apply equally well to this parameter.

7.6 Vowel space

The vowel space is the representation of the average of the occurrences of each vowel, using the estimated value of the first two formants F1 and F2. In clear speech the hyper-articulation of words, characteristic of this style, influence the values of these two formants and determines a larger vowel space, as evidenced by the results obtained by Smiljianic and Bradlow (Smiljanic & Bradlow, 2005). The objective of this analysis is to detect an approximation of the vowel space of processed speech signals, to the one of the original signal.

In this processing are taken into account the sentences of the dataset for which the vowels are identified and for each one is performed the estimation of the two formants. For each vowel the averaged of the occurrences is computed and the results are represented by the size of F1 and F2. For each processed signal this calculation is repeated and comparison is made directly on the charts, as shown in Figure 6.14, Figure 6.15 and Figure 6.16, for the three sentences in question.

From Figure 6.14 it is evident that an increase in channels where the signal is processed by the CISG increases the definition of vowels and therefore better approximates the original vowel space. This, as defined by Ferguson and Kewley-Port (Ferguson & Kewley-Port, 2007) and Smiljianic and Bradlow (Smiljanic & Bradlow, 2009), it is more extended in comparison between clear and conversational speech.

Can be seen in Figure 6.15, the same way as before, the improvement of the approximation to the original signal with the increase in the number of processing channels. For the signals processed with 2 and 4 channels, the estimate of the formants is not reliable, because the creation of the signal through the filtering of white noise eliminates the vocals needed for the estimate. With the increase in the number of channels, the presence of periodic signal is restored, albeit partial, but sufficient to detect the formants at frequencies similar to the original

In order to processing the last sentence of the dataset considered is obtained results as in Figure 6.16 in which is still visible the improvement of the approximation to the original signal, with increasing numbers of channels, but also occurs in a fault recognition of a vowel. For the representation for the vowel ‘a’, indeed, there is a perfect recognition only for the elaboration with sixteen channels. This is probably given by the single occurrence of the vowel ‘a’ and its position adjacent to a voiced sound that make difficultly distinguishable the segmentation of the vowel.

This measure is taken as confirmation of the goodness of cochlear signal simulator that, for a number of processing channels of eight or more, preserves the characteristic properties of the vowel space.

7.7 Second and third formant power

The power of the second and third formant is a parameter derived from the analysis conducted on clear speech by Krause and Braida (Krause & Braida, 2004), Greenberg and Arai (Greenberg & Arai, 2004) and Kain et al (Kain, Amano-Kusumoto, & Hosom, 2008), related to the increase in energy observed in the spectrum of frequencies in the band between 1000 to 3000 Hz in this band do, indeed, the frequencies of the formants F2 and F3 of the vowels.

The analysis want to estimate how much the signal processing by simulating cochlear implant may influence the energy of these formants in particular. The use of the noise signal generation allows you to predict a change of energy, but line with the quality of the approximation of the signal, and thus with the increase in the number of channels, is expected to detect an approach of these values to those the original signal.

The values are then extracted for the original signal and the signals processed by each word, from the vowels that compose them. Then is checked the value of energy for these frequencies in the calculation of long-term spectra. The values thus obtained are compared as described in Section 5.5.6.

The power values of the second and third formants for the vowels of words processed with a high number of channels approximates the trends of the values of the original signals as shown in Figure 6.17, Figure 6.18, Figure 6.19 and Figure 6.20. The frequencies of F2 and F3 are generally between 1000 and 3000 Hz, were it is releaved an increase in the power spectrum, replacing the expectation. The deviation from the expected trend is probably given by the modulation with white noise in the processing of the simulated signal.

The CISG is shows therefore a method for the simulation of signal capable of maintaining the properties on the vowels in this band.

7.8 CV-ratio

The consonant-vowel ratio as the ratio of power between occlusive or fricative consonants and the vowel adjacent in a word, is the comparison between clear and conversational speech, a variation of the articulation of words results important in increasing intelligibility. The studies by Picheny et al (Picheny, Durlach, & Braida, 1986), in fact, show that in clear speech there is an increase in the ratio with respect to pre vocalic plosives. In the analysis performed is taken into account the variation of the ratio between the signals processed and the original signal.

Are presented results for both individual power values of vowels and consonants of each word, and their ratio, as explained inSection 5.5.7 e 5.5.8, with any further analysis where necessary.

The results obtained from the CV-ratio, as shown in Figure 6.21 show a relative increase in power ratio with rising of the bands in the processed signal. The major variations are noticed on the powers of the consonants, as shown in Figure 6.24, suggesting a greater influence on the processing on the same consonants. The power of vowels, in fact, remains fairly stable, as shown in Figure 6.25.

The different trends observed in the analyzed words do not allow you to make a specific assessment for all plosive or fricative consonants. Are distinguishable two main trends, such as those seen in Figure 6.21 for the word 'caldo', in which is not visible the desired trend with the increasing number of channels, and for the word 'venti' in which, however, the improvement is visible.

It was therefore considered necessary a deeper analysis, verifying the trends of the power spectrum for vowels and consonants that compose them. Has been analyzed the couple / c /-/ a / in the word 'caldo', visible in Figure 6.22, and the pair / t /-/ i / in 'venti' in Figure 6.23.

For both it has been noted that the trend of the spectrum of the two channels processed signal do not follow the original trend and present many spurious frequency components, both in vowels than for consonants. For the processed signal with 12 channels, however, the trend reflects that of the original signal, but the presence of spurious components is still there.

It is noticeable as the for vowels there is a most similar trend compared to the original, following the fundamental frequency peaks and the reduction at higher frequencies (1000 Hz for 'caldo', 600 Hz for 'venti').

The CISG, therefore, affected more consonants compared to the vowels, and cannot be detected a behavior similar to that followed by the studies reviewed.

7.9 Conclusions

In this thesis has been analyzed the influence of the speech signal processing of cochlear implants on the characteristics of clear speech defined as relevant for the preservation of intelligibility.

The development of digital hearing aids in recent years has been focused almost exclusively on increasing performance by acting on the elaboration of the signal according to the specific deficiencies in the auditory system of each patient. Progress has been made in the acquisition and signal processing, to improve the perception of frequency bands and in the definition of the optimal parameters to provide a better perception. Despite all developments, the patients with cochlear implants and hearing aids complain about a leak of naturalness of sound

The attention of researchers has been therefore begun to move on the analysis of such features of speech that are fundamental for the correct perception. In this area, still largely unexplored, fits this thesis, starting from a comprehensive analysis of the characteristics that differentiate the clear speech from conversational speech. Have been analyzed several studies that dealt separately the different characteristics and has been defined a set of those most relevant in the preservation of intelligibility.

Thus, from a set of signals (words and sentences) recorded by a speaker who produced clear speech, these features were extracted, and were compared with degraded versions processed through a simulation of a cochlear implant.

The implemented simulator allows varying the number of processing channels, resulting in a series of signals differently faithful to the original. From these were extracted the same features found on the original signal and it was possible to make a comparison, analyzing the influence that the number of processing channels, has on the characteristics under consideration.

The different analysis on the results have identified in the processed signal with 12 channels, the optimal solution, both for the maintenance of features, both for the acoustic response. This solution reflects the design choices of production of cochlear implants, usually set up on one number of 12 channels.

This observation comes from analysis of every single feature. The overall results are summarized it as follows:

- The analysis of *long-term spectra* shows greater fidelity to the original signal of the processed signal with 12 channels. It is noticeable an increase in power for the signals processed in the band 1000 - 3000 Hz which cannot be considered a carrier of an increase of intelligibility, as expressed by the results of Krause and Braida (Krause & Braida, 2004) and Greenberg and Arai (Greenberg & Arai, 2004), since the relative strength of band considered in the original signal is still higher compared to its processed versions;
- The *modulation index* is only slightly variable with the increase of processing channels; this is because the construction of processed signals requires a modulation with the envelope of the original signal. This feature is then of course maintained in a simulated cochlear implant processing;
- For the estimation of the *fundamental* and the computation of the *fundamental range*, is obtained an improvement the higher the number of processing bands. This result is even more significant considering that the simulated cochlear implant system does not provide for the transmission of information concerning the fundamental;
- Also for the *vowel space*, expressed as a space of representation of the formants F1 and F2 of vowels in the sentences under consideration,

there is a best approximation of the vowel space of the original signal, increasing the number of processing channels. This result is already visible with the number of channels above 8;

- The extraction of *second and third formant power* has some exceptions on the analysis of F2. Some words, indeed, differ from the common trend that approximates the value of the original signal, the higher the number of channels. This variation may be due, especially with low numbers of channels, to the use of white noise as modulated signal for the simulation. As seen for the long term spectra, in fact, there is an increase in the band 1000 - 3000 Hz for signals processed, which corresponds to the band which is usually present in the F2. For the power value of the F3, however, the trend is clearly better, values and is significantly improved as early as four channel processing.

A separate discussion is made for the CV-ratio. The analysis of the power of vowels and consonants shows how the signal processing significantly influence on the consonants. For the vowels, indeed, the values remain fairly consistent for each different process. For the consonants instead there is a considerable variability, apparently independent of the number of channels and also between different consonants of the same word. This affects on CV-ratio that does not show a common trend for all the words. A detailed analysis of the spectrum for those words with a desired performance, against an unwanted one, shows a more uniform trend to the original performance, for vowels, while there are significant spurious frequency components on the consonants. This can be justified by the use of white noise as test signal for the simulation, that have a greater influence in those segments of the signal similar to noise, as they are precisely the plosive and fricative consonants.

At the conclusion of the analysis performed can be established as the method of simulation of speech used in cochlear implants, used as a reference, preserve in a satisfactory way the characteristics of clear speech, considered as important for intelligibility. A process optimization parameter can be detected as the processed signal to 12 channels is proved optimal for processing. This value reflects, in fact, the implementation choices of cochlear implants on the market and the results obtained can serve as further confirmation of the goodness of the selected parameters.

With this analysis is then possible to improve signal processing in cochlear implants, focusing on increasing these characteristics, and have also laid the foundation for the development of a perception simulator of speech through cochlear implants that can be useful in the early stages study of the features modifications, verifying their impact on intelligibility could not carry out some tests of perception.

Chapter 8

Future works

This Thesis involved the study of a discussed aspect in literature but not uniformly or dedicated, as it was as a goal in this thesis. This is a first step towards creating a complete system that integrates the feature extraction, with their development, and strive for complete integration with hearing aids, and assessment tools for audiometric testing of intelligibility. For this reason, many scenarios are open for development, and application divided according to the following categorization.

Inclusion of additional features

The following features were not taken into account in implementation, but have nevertheless taken an analysis of the influence intelligibility. All these features can be implemented in the CFE studying the results of the previous elaboration. In this section are exposed the results of various articles that had studied them.

Modulation index peaks in the temporal envelope: defining the peaks as the part of amplitude envelope that exceed the target level (referred to long-term-rms level Leq), Drullman (Drullman, 1995) found that peaks up to 15dB below Leq can be "chopped off" without noticeable detrimental effect in intelligibility and that a 100% intelligibility is preserved between 19 dB below and 1 dB above the long-term-rms level (Leq). These results are related to the results on the modulation index peaks and can give a parameter for the envelope modulation.
Dynamic range fraction: Drullman (Drullman, 1995) found two different crossover levels: an acoustic crossover of on average 18dB below Leq that divides the temporal envelope into two equal peak and trough parts, a perceptual crossover of 6:9dB below Leq that yields equal intelligibility scores when removing either xdB below or above that level.

Number and duration variation: the number and the duration of pauses are most important to the increase in intelligibility but are even too characteristic of the speaking style. Picheny et al (Picheny, Durlach, & Braida, 1986) founds an increase in pause duration and length between clear and conversational speech (in clear speech there is a minimum duration of 10ms and an average of 120ms). This result is confirmed by Krause and Braida (Krause & Braida, 2004) that found a mean of 130ms of pause length in clear speech versus a mean of 42ms in conversational.

Spectral rate of change: defined as the root mean square of the first three formant slopes, the interaction between the spectral rate of change and the speaking style was studied by Wouters and Macon (Wouters & Macon, 2002). Results shown that this features increase their value in the clear speech and it is related even to the linguistic prominence, i.e. in stressed syllables, in accent words, in sentence-medial words and in clear speech (hyper-articulated speech).

Magnitude and phase of spectrum: referred to the study of Greenberg and Arai (Greenberg & Arai, 2004) that found that a desincronization of 25ms in some particular band of a test signal, have an deleterious impact on intelligibility of 10-20% (percentage of word recognition), while a jitter of 50ms have an impact of 40%

Inclusion of an automated classifier

A relevant necessary work for the definition of the analyzed features, concern the subdivision of the speech, in their different segment as vowel, plosive consonant, voiced consonant etc. An automatized approach can help in the speeding of the processing analysis. For this works can be possible to implement a consonant-vowel recognizer (as a Voiced-Unvoiced-Silence recognizer) to permit to distinguish between unvoiced consonants (plosive and fricative) and vowels. There are different approach to this problem that use an extraction of particular features, as energy of the signal, zero crossing rate of the signal, autocorrelation coefficient at unit sample delay, first predictor coefficient , energy of the prediction error in studies as the for example the one by Atal and Rabiner (Atal & Rabiner, 1976), or the MEL Frequency Cepstral Coefficient as for example in studies by Zolnay et al (Zolnay, and Schlueter, & and Ney, 2005) and the classification for voiced and unvoiced with Gaussian Mixture Model.

Influence of clear speech on brain responses

An interesting search field is on the brain responses to the acoustical stimuli. Several studies give results on the activation of particular brain zones to the acoustical stimuli, and an interesting research can focused on the different reaction to clear and conversational speech or to the modification of the features discussed in this thesis. As an example Ghitza and Greenberg (Ghitza & Greenberg, 2009) had studied the response at the variation of number of insertion and length of pauses in a set of sentences, while checking the brain responses and found that for particular values (optimal silence length of 20-120msec and optimal silence rate of 80 ms i.e. 8Hz) there is a greater increases in intelligibility. They relate this value to the synchronization of pauses length and duration to the brain fluctuation in theta range. Another example is the recent study proposed by Korczak and Stapells (Korczak & Stapells, 2010) that investigate the influence of articulatory features, as vowel-space contrast, place of articulation of stop consonants and voiced/voiceless distinctions on cortical event-related potential (ERPs). Other interesting studies is the one by Zhao et al (Zhao, Ravuri, & Morgan, 2010) on the spectro-temporal modulation features for speech recognition, inspired by cortical brain responses, or the one by Zaehle et al (Zaehle, Jancke, & Meyer, 2007) on the auditory cortex response to temporal features for speech/non-speech discrimination, or the one by Warren et al (Warren, Jennings, & Griffiths, 2005) on the analysis of the spectral envelope of sounds by the human brain.

All these works can suggest some interesting results to optimize the research on the importance of spectro-temporal features for the intelligibility, and then in the clear speech.

Influence of stress, accent and emotional prosody on clear speech

Another branch of research is the one that focus on the difference in stress accent and emotional prosody, and their influence on intelligibility. Starting ad example from the study of Banse et al (Banse & Scherer, 1996) can be recognizable different features that enhance the perception of speech and that can be extracted and implemented to obtain a greater intelligibility.

Enhancement of clear speech in speech processing (hearing aids /cochlear implants)

The simulation of the signal such as cochlear implant can be improved with the introduction of the next stages in the processing of the signal present on the actual cochlear implants, the first of a logarithmic compression on the envelopes of each channel, through a feature called Loudness Growth Function (LGF) that is an essential component of the CIS strategy (exposed in Section 2.2). This compression transforms acoustical amplitudes into electrical amplitudes, and it is necessary in the cochlear implants because the range in acoustic amplitudes in conversational speech is considerably larger than the implant's patience dynamic range (i.e., the range in electrical amplitudes between barely audible level and extremely loud level). Then the logarithmic compression fits the acoustical level to the patience's electrical dynamic range.

References

Atal, B., & Rabiner, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE transaction on acoustics, speech, and signal processing* , 201-212.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol , 70* (3), 614-636.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *J Acoust Soc Am , 97* (1), 585-592.

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am , 95* (5 Pt 1), 2670-2680.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am , 95* (2), 1053-1064.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *J Speech Lang Hear Res , 50* (5), 1241-1255.

Gabrielsson, A. (1998, June). The Effects of Different Frequency Responses on Sound Quality Judgments and Speech Intelligibility. *Journal of Speech and Hearing Research* , pp. 166-177.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica , 66* (1-2), 113-126.

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Transactions on Information and Systems , E87-D*, 1059-1070.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species--29 years later. *J Acoust Soc Am , 87* (6), 2592-2605.

Hamacher, V., Chalupper, J., & Eggers, J. (2005). Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *Eurasip Journal on Applied Signal Processing , 2005*, 2915-2929.

Houtgast, T., & Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am Volume , 77*, 1069-1077.

Kain, A., Amano-Kusumoto, A., & Hosom, J.-P. (2008). Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *J Acoust Soc Am , 124* (4), 2308-2319.

Klasmeyer, G. (1997, April 21-24). The perceptual importance of selected voice quality parameters. *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* , pp. 1615 - 1618 vol.3.

Korczak, P. A., & Stapells, D. R. (2010). Effects of various articulatory features of speech on cortical event-related potentials and behavioral measures of speech-sound processing. *Ear Hear , 31* (4), 491-504.

Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J Acoust Soc Am , 115* (1), 362-378.

Krause, J. C., & Braida, L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *J Acoust Soc Am , 125* (5), 3346-3357.

Liu, S., & Zeng, F.-G. (2006). Temporal properties in clear speech perception. *J Acoust Soc Am , 120* (1), 424-432.

Liu, S., Rio, E. D., Bradlow, A. R., & Zeng, F.-G. (2004). Clear speech perception in acoustic and electric hearing. *J Acoust Soc Am , 116* (4 Pt 1), 2374-2383.

Loizou, P. C. (2006). Speech processing in vocoder-centric cochlear implants. *Adv Otorhinolaryngol , 64*, 109-143.

O'Shaugnessy. (2008). *Springer Handbook of Speech Processing - Cap 11. Formant estimation and tracking.* (Y. H. Jacob, Ed.) Springer-Verlag Berlin Heidelberg.

Paglialonga, A., Tognola, G., Sibella, F., Parazzini, M., Ravazzani, P., Grandori, F., et al. (2008). Influence of cochlear implant-like operating conditions on wavelet speech processing. *Comput Biol Med , 38* (7), 799-804.

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J Acoust Soc Am , 95* (3), 1581-1592.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J Speech Hear Res , 28* (1), 96-103.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech. *J Speech Hear Res , 29* (4), 434-446.

Punch, J. L., & Beck, E. L. (1980). Low-frequency response of hearing aids and judgments of aided speech quality. *J Speech Hear Disord , 45* (3), 325-335.

Punch, J. L., & Beck, L. B. (1986). Relative effects of low-frequency amplification on syllable recognition and speech quality. *Ear Hear , 7* (2), 57-62.

Ricketts, T. A. (2009). Digital Hearing aids: current "state-of-the-art". *American Speech-Language-Hearing Association , Unknown*, 1-3.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci , 336* (1278), 367-373.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science , 270* (5234), 303-304.

Smiljanic, R. R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J Acoust Soc Am , 118* (3 Pt 1), 1677-1688.

Smiljanic, R., & Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Lang Linguist Compass , 3* (1), 236-264.

Tecca, J. E., & Goldstein, D. P. (1984). Effect of low-frequency hearing aid response on four measures of speech perception. *Ear Hear , 5* (1), 22-29.

Turrini, & Cutugno. (1993). Nuove parole bisillabiche per audiometria vocale in lingua italiana. *Acta Otorhinolaringoiatria Italiana , 13*, 63-77.

Uchanski, R. M. (2005). *The handbook of speech perception - Clear Speech.* (D. B. Pisoni, & R. E. Remez, Eds.) Blackwell.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *J Speech Hear Res , 39* (3), 494-509.

Warren, J. D., Jennings, A. R., & Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage , 24* (4), 1052-1057.

Wouters, J., & Macon, M. W. (2002). Effects of prosodic factors on spectral dynamics. I. Analysis. *J Acoust Soc Am , 111* (1 Pt 1), 417-427.

Xu, L., & Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: implications for auditory prostheses. *Hear Res , 242* (1-2), 132-140.

Zaehle, T., Jancke, L., & Meyer, M. (2007). Electrical brain imaging evidences left auditory cortex involvement in speech and non-speech discrimination based on temporal features. *Behav Brain Funct , 3*, 63.

Zhao, S. Y., Ravuri, S., & Morgan, N. (2010). Toward a many-stream framework of cortically-inspired spectro-temporal modulation features for automatic speech recognition. *Speech Comm. , Article in press*, Article in press.

Zolnay, A., and Schlueter, R., & and Ney, H. (2005). Acoustic Feature Combination for Robust Speech Recognition. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '05).*

Appendix A

Mathematical Background

A.1 Butterworth filters

A Butterworth filter is a signal processing filter designed to have as flat a frequency response as possible in the passband, until the cut-off frequency at -3dB with no ripples. Higher frequencies beyond the cut-off point rolls-off down to zero in the stop band at 6dB/octave (it has a 'quality factor', 'Q' of just 0.707). One main disadvantage of the Butterworth filter is that it achieves this pass band flatness at the expense of a wide transition band as the filter changes from the pass band to the stop band. It also has poor phase characteristics as well.

The general equation for a Butterworth filters frequency response is given as:

$$H(j\omega) = \frac{1}{\sqrt{1+\varepsilon^2(\frac{\omega}{\omega_p})^{2n}}}$$

Where: n represents the filter order, ω is equal to $2\pi f$ and ε is the maximum pass band gain, (A_{max}). If A_{max} is defined at a frequency equal to the cut-off -3dB corner point (fc), ε will then be equal to one and therefore ε^2 will also be one. However, if you now wish to define A_{max} at a different voltage gain value, for example 1dB, or 1.1220 (1dB = 20logA_{max}) then the new value of epsilon, ε is found by:

$$H_1 = \frac{H_0}{\sqrt{1+\varepsilon^2}}$$

Where: H_0 is the Maximum Pass band Gain, $\mathrm{A_{max}}$ and H_1 is the Minimum Pass band Gain.
Taking the transpose of this last equation, results:

$$\frac{H_0}{H_1} = 0.122 = \sqrt{1+\varepsilon^2} \text{ gives } \varepsilon = 0.5088$$

The Frequency Response of a filter can be defined mathematically by its Transfer Function with the standard Voltage Transfer Function $H(j\omega)$ written as:

$$H(j\omega) = \frac{V_{out}(j\omega)}{V_{in}(j\omega)} \quad or \quad H(s) = \frac{V_{out}}{V_{in}} = \frac{1}{(s-s_1)(s-s_2)\cdots(s-s_n)}$$

Where: V_{out} is the output signal voltage, V_{in} is the input signal voltage, j is equal to the square root of -1 ($\sqrt{-1}$) and ω is the radian frequency ($2\pi f$). The second equation represent a transfer function for a generic low pass filter in which ($j\omega$) can also be written as (s) to denote the S-domain.

A.2 Hilbert Transform

In signal processing the Hilbert transform is a linear operator which takes a function, $g(t)$, and produces a function, $H(g)(t)$, with the same domain. In Fourier analysis provides a concrete means for realizing the conjugate of a given function or Fourier series. In harmonic analysis, it is an example of a singular integral operator, and of a Fourier multiplier. It is also important in the field of signal processing because it is used to derive the analytic representation of a signal $g(t)$,

The Hilbert transform of the function $g(t)$, is defined by

$$H(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(t)}{x-t}\,dt$$

Because of the possible singularity at $x = t,$ the integral is to be considered as a Cauchy principal value. Consider the following limit

$$\lim_{z\to 0+}\left[\int_{x+z}^{\infty} \frac{g(t)}{x-t}dt + \int_{-\infty}^{x-z} \frac{g(t)}{x-t}dt + \right]$$

When this limit exist it is called the Cauchy principal value around $x = t$ of the integral and is written as

$$\int_{-\infty}^{\infty} \frac{g(t)}{x-t}dt$$

So when the Hilbert transform exist, it is written as presented at first equation. Other forms for $H(x)$ can be obtained by change of variable.

$$H(x) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(x-t)}{x}\,dt$$

$$H(x) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{g(x+t)}{t}\,dt$$

This equation shown that Hilbert transform is a convolution:

$$H(x) = \frac{1}{\pi\, x} \times g(x)$$

From the convolution theorem we can say how the spectrum of $H(x)$ is related to that of $g(t)$. Applying the Fourier transform to both sides of last equation, we obtain:

$$F\big(H(x)\big) = F\left(\frac{1}{\pi\, x}\right) F(g(x))$$

where F is the Fourier transform.

The Fourier transform of $\frac{1}{\pi x}$ is $-i\, sgn$ of frequency, which is equal to $-i$ for positive frequency and $+i$ for negative frequency. Hence Hilbert transform is equivalent to an interesting kind of filter, in which the amplitudes of the spectral components are left unchanged, but their phases are altered by $\pi/2$, positively or negatively according to the sign of frequency.

The Hilbert transform presents interesting properties for causal functions. A function $g(t)$ is called causal function if is zero when *t<0*, because in many physical systems the fact that effects cannot precede cause places this constraint on certain functions describing the system.

One of the interesting properties implied by causality is that the imaginary part of Fourier transform is completely determined by knowledge of its real part and vice-versa.

If the causal function $g(t)$ contains no singularities at the origin, then $F(h(t)) = R(f) + iX(f)$ is its Fourier transform, and $R(f)$ and $X(f)$ are the Hilbert transform that satisfy the equations:

$$X(f) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{R(y)}{\omega - y} dy \quad \text{and} \quad R(f) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{X(y)}{\omega - y} dy$$

A.3 Linear Predictive Coding

Linear prediction (LP) is widely used in speech applications (spectral estimation, recognition, compression, modeling, etc.). This is due to the fact that the speech production process is well modeled with LP. Indeed, it is well recognized that a speech signal can be written in the following form

$$x(k) = \sum_{l=1}^{L} a_l \; x(k-l) + \; G\,u(k) \qquad \text{(A. 1)}$$

where k is the time index, L represents the number of coefficients in the model (the order of the predictor), $a_l,\ l = 1, \cdots, L$, are defined as the linear prediction coefficients, G is the gain of the system, and $u(k)$ is the excitation signal, which can be either a quasiperiodic train of impulses (for vowel signal) or a random noise source (for noisy consonants, e.g. plosive /t/) and also a combination of both signals (for voiced consonants, e.g. fricatives such as /v/ and /z/). The parameters, a_l, determine the spectral characteristics of the particular sound for each of the two types of excitation.

Using the z-transform we can rewrite the Equation (A. 1) in the frequency domain. If $H(z)$ is the transfer function of the system, we have:

$$\boldsymbol{H(z)} = \frac{\boldsymbol{G}}{\boldsymbol{1-\sum_{l=1}^{L} a_l \; z^{-l}}} = \frac{\boldsymbol{G}}{\boldsymbol{A(z)}} \qquad \text{(A. 2)}$$

which is an all-pole transfer function. This filter $[H(z)]$ is a good model of the human vocal tract.

Consider a stationary random signal $x(k)$ and assuming that it is real, stationary, and zero mean, we can use the LPC to predict the value of the sample $x(k)$ from its past values, i.e. $x(k-1)$, $x(k-2)$, etc.. (Forward Linear Prediction). We define the forward prediction error as,

$$\begin{aligned} e_{f,L} &= x(k) - \hat{x}(k) \\ &= x(k) - \sum_{l=1}^{L} a_{L,l} \; x(k-l) \\ &= \boldsymbol{x(k) - a_L^T \, x(k-l)} \qquad \text{(A. 3)} \end{aligned}$$

where the superscript 'T' denotes transposition, $\hat{x}(k)$ is the predicted sample, $\boldsymbol{a_L} = \left[a_{L,1} \; a_{L,1} \; \cdots \; a_{L,L}\right]^T$ is the forward predictor of length L and $\boldsymbol{x}(k-1) = [x(k-1)\, x(k-2) \cdots x(k-L)]^T$ is a vector containing the L most recent samples starting with and including $x(k-1)$.

We would like to find the optimal Wiener predictor. For that, we seek to minimize the mean-square error (MSE):

$$\boldsymbol{J_f}\left(\boldsymbol{a_L}\right) = \boldsymbol{E}\left\{\boldsymbol{e}_{\boldsymbol{f},\boldsymbol{L}}^{\boldsymbol{2}}(\boldsymbol{k})\right\} \qquad \text{(A. 4)}$$

where $E\{\cdot\}$ denotes mathematical expectation. Taking the gradient of $\boldsymbol{J_f}\left(\boldsymbol{a_L}\right)$ with respect to $\boldsymbol{a_L}$ and equating to $0_{L\times 1}$ (a vector of length L containing only zeroes), we easily find the Wiener–Hopf equations:

$$\boldsymbol{R_L a_{o,L}} = \boldsymbol{r_{f,L}} \qquad \text{(A. 5)}$$

where the subscript 'o' in $\boldsymbol{a_{o,L}}$stands for optimal,

$$\begin{aligned} \boldsymbol{R}_L &= E\{\boldsymbol{x}(k-1)\boldsymbol{x}^T(k-1)\} \\ &= E\{\boldsymbol{x}(k)\boldsymbol{x}^T(k)\} \\ &= \begin{pmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & r(0) & \cdots & r(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(L-1) & r(L-2) & \cdots & r(0) \end{pmatrix} \end{aligned}$$

the correlation matrix, and

$$\begin{aligned} \boldsymbol{r}_{f,L} &= E\{\boldsymbol{x}(k-1)x(k)\} \\ &= [r(1)\ r(2)\ \cdots\ r(L)]^T \end{aligned}$$

is the correlation vector. The matrix $\boldsymbol{R_L}$ has a Toeplitz structure (i. e., all the entries along the diagonals are the same); assuming that it is nonsingular, we deduce the optimal forward predictor:

$$\boldsymbol{a}_{o,L} = \boldsymbol{R}_L{}^{\boldsymbol{-1}}\,\boldsymbol{r}_{f,L} \qquad \text{(A. 6)}$$

Expanding $\boldsymbol{e}_{\boldsymbol{f},\boldsymbol{L}}^{\boldsymbol{2}}$ in (A. 4) and using (A. 5) shows that the minimum mean-square error (MMSE),

$$\begin{aligned} \boldsymbol{J}_{f,min} &= \boldsymbol{J}_f(\boldsymbol{a}_{o,L}) \\ &= r(0) - \boldsymbol{r}_{f,L}^T\boldsymbol{a}_{o,L} = E_{f,L} \end{aligned} \qquad \text{(A. 7)}$$

This is also called the forward prediction-error power. Define the augmented correlation matrix:

$$\boldsymbol{R}_{L+1} = \begin{pmatrix} r(0) & \boldsymbol{r}_{f,L}^T \\ \boldsymbol{r}_{f,L} & \boldsymbol{R}_L \end{pmatrix} \qquad \text{(A.8)}$$

equations (A. 7) and (A. 4) may be combined in a convenient way:

$$\boldsymbol{R}_{L+1}\begin{pmatrix} 1 \\ -\boldsymbol{a}_{o,L} \end{pmatrix} = \begin{pmatrix} E_{f,L} \\ \boldsymbol{0}_{L\times 1} \end{pmatrix} \qquad \text{(A. 9)}$$

We refer to (A. 9) as the augmentedWiener–Hopf equations of a forward predictor of order $\boldsymbol{L}$. From (A. 8) we derive that,

$$det(\boldsymbol{R}_{L+1}) \; = \; E_{f,L} \, det(\boldsymbol{R}_L \,) \qquad \text{(A. 10)}$$

where 'det' stands for determinant.

Let us now write the forward prediction errors for the optimal predictors of orders L and $L - i$:

$$e_{f,o,L}(k) = x(k) - \textstyle\sum_{l=1}^{L} a_{0,L,l} x(k-l) \qquad \text{(A. 11)}$$

$$e_{f,o,L-i}(k) = x(k) - \textstyle\sum_{l=1}^{L-i} a_{0,L-i,l} x(k-l) \qquad \text{(A. 12)}$$

From the principle of orthogonality (A. 6) we know that:

$$E\{e_{f,o,L}(k)\boldsymbol{x}(k-1)\} = \; \boldsymbol{0}_{L\times 1} \qquad \text{(A. 13)}$$

For $1 \; \leq \; i \; \leq \; L$, we can verify by using (A. 13), that:

$$E\{e_{f,o,L}(k)e_{f,o,L-i}(k-i)\} = 0 \qquad \text{(A. 14)}$$

As a result,

$$\lim_{L\to\infty} E\{\, e_{f,o,L}(k)e_{f,o,L-i}(k-i)\}$$

$$= \; E\big\{e_{f,o}(k)e_{f,o}(k-i)\big\} = 0.$$

This indicates that the signal $e_{f,o}(k)$ is a white noise. So the optimal forward predictor has this important property of being able to whiten a stationary random process, provided that the order of the predictor is high enough.