

# Adaptive Pitch Transposition: Smart Auditory Spectral Shifts in Cochlear Implants

Kevin Struwe, Ralf Salomon

Institute of Applied Microelectronics and Computer Engineering, University of Rostock, Rostock, Germany  
Email: kevin.struwe@uni-rostock.de, ralf.salomon@uni-rostock.de

**How to cite this paper:** Struwe, K. and Salomon, R. (2017) Adaptive Pitch Transposition: Smart Auditory Spectral Shifts in Cochlear Implants. *Engineering*, 9, 739-754.  
<https://doi.org/10.4236/eng.2017.99045>

**Received:** July 20, 2017

**Accepted:** September 16, 2017

**Published:** September 19, 2017

Copyright © 2017 by authors and Scientific Research Publishing Inc.  
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).  
<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

Patients with severe hearing loss have the option to get a cochlear implant device to regain their hearing. Yet, the implantation process is not always optimal, which in some cases results in a shallow insertion depth or an accidental insertion into the wrong cochlear duct. As a consequence, the patients' pitch discrimination ability is suboptimal, leading to an even more decreased vowel identification, which is vital for speech recognition. This paper presents a technical approach to solve this problem: the adaptive pitch transposition module modifies the frequency content in a fashion so that the pitch is fixed to an optimal value. To determine this value, a patient-individual best pitch is determined experimentally by evaluating speech recognition at different pitches. This best pitch is subsequently called the comfort pitch. As a result of the considerations a technical implementation is presented in principle. A system comprised of pitch detection, pitch transposition and an arbitrary chosen comfort pitch is described in depth. It has been implemented prototypically in Matlab/Octave and tested with an example audio file. The system itself is designed as a preprocessing stage preceding cochlear implant processing.

## Keywords

Cochlear Implants, Pitch Transposition, Speech Processing

---

## 1. Introduction

Hearing is one of our five senses, and losing it constitutes a severe impairment. This loss might happen simply due to age, gene defects, or the excessive exposure to loud noise. Those patients are normally helped by wearing hearing aid devices.

On the functional level, deafness might be caused by the destruction of the

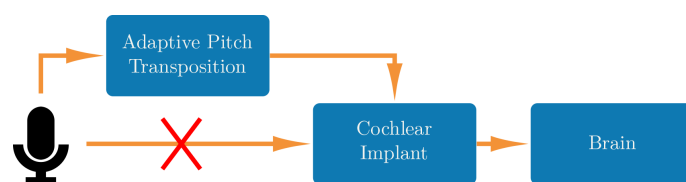
hair cells within the cochlea. The cochlea is a central organ deep inside the ear. Its very function is the transformation of mechanical waves, *i.e.*, the sound, into electrical signals, *i.e.*, the action potentials of the nerve cells. The hearing nerve maps these action potentials onto the auditory cortex, which is responsible for further processing.

Often, the term “destruction of the hair cells” does not refer to the entire destruction of those cells but to destruction of the hairs. This kind of deafness is usually called “sensorineural hearing loss”. Without these hairs, the cells cannot sense the mechanical oscillations within the cochlea, and can thus not invoke any action potential to send toward the auditory cortex. Those patients receive help from cochlear implants. A cochlear implant bypasses the processing chain from the earpiece up to the cochlea. It employs a microphone, a speech processor, and a small number of electrodes, with which it stimulates a certain area of hair cells. For every incoming sound sample, the speech processor creates separate frequency filtered signals, which are turned into pulse train signals, that stimulate the hair cells directly. Section 1 provides the background on both the hearing sense and cochlear implants as far as required for the understanding of this paper.

Even with properly implanted and operating cochlear implants, not all patients enjoy sufficient speech recognition. In addition to the small number of electrodes and limited brain plasticity, the cochlea itself might be the origin of further problems: due to its construction and possible additional ossifications within it, the electrodes cannot be inserted deep enough; some of the (usually) twelve available electrodes remain outside the cochlea. Consequently, a very small number of potential frequencies remains available, which does not necessarily map the frequency spectrum of the incoming sound. This problem is described in more detail in Section 3.

As an alleviation, this paper proposes to insert a signal preprocessor between the microphone and the actual speech processor (see, also, **Figure 1**). This newly introduced module is called adaptive pitch transposition, or APT for short, and has been introduced in [1]. Its very purpose is to perform a speech transposition on the fly such that the cochlear implant patient can achieve optimal comprehension through usage of his device.

As is already suggested by its name, *pitch* is the central concept of the adaptive pitch transposition module. Pitch is a concept that is rather hard to understand



**Figure 1.** The sound processing chain for cochlear implants is broken up, and a new module, “Adaptive Pitch Transposition” is introduced. It lies between the sound capturing with the microphone and the actual speech processor of the cochlear implant.

as it is not a frequency that is contained in the sound sample, but rather a neurophysiological phenomenon that reflects the inherent periodicity of the sensed sound. Therefore, Section 4 provides a brief description of pitch and its occurrence in speech.

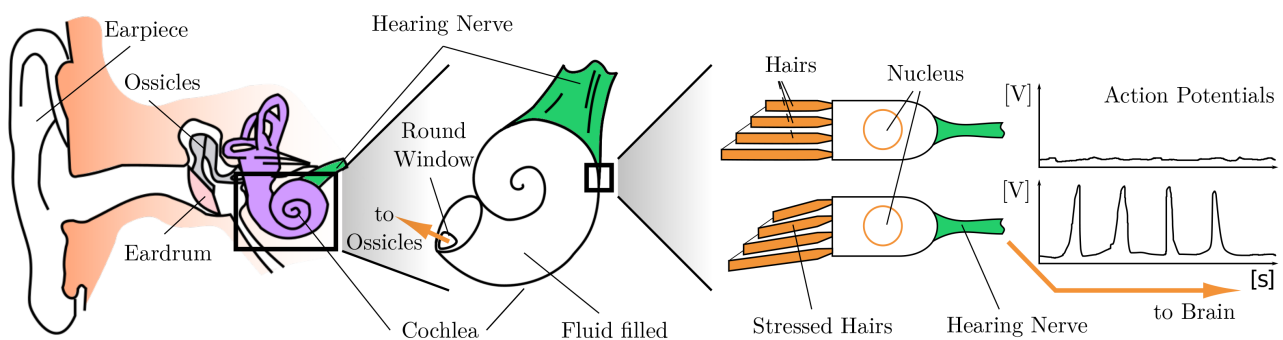
Depending on the actual situation of the implant, every cochlear implant patient has a very specific frequency area that provides the most comfortable and best speech recognition. This personal frequency area is called the comfort pitch range and does not necessarily match the pitch of an incoming speech sample. This frequency mismatch is a main reason for limited speech recognition. As a remedy, Section 8 proposes the *adaptive pitch transposition* module. Its very purpose is to transpose, *i.e.*, shift and scale, the frequency spectrum of all incoming speech samples to one pitch in the comfort pitch area. To this end, the adaptive pitch transposition module consists of a pitch detection unit and a pitch transposition unit. To achieve the adaptivity of the module, a test to acquire the comfort pitch has been developed as well.

The adaptive pitch transposition module has been prototypically implemented in both Matlab® and C++. These prototypes are described in Section 1, and allow for the offline evaluation of recorded sound samples as well as the online demonstration of its effects.

The preliminary practical experiences clearly indicate that the chosen approach functions. These experiences, however, also indicate room for improvements and several limitations. The main drawback of the current approach is that it heavily hinders speaker separation. Section 11 discusses these aspects and concludes this paper.

## 2. Background: The Ear and Cochlear Implants

This section provides some background information about the ear, its cochlea, and cochlear implants as far as relevant for the understanding of this paper. This description is illustrated by **Figure 2**, which shows the structure of the ear including the most important parts for hearing.



**Figure 2.** The function of a healthy human auditory system. An arriving sound wave travels from the earpiece (pinna) through the eardrum (tympanic membrane) to the ossicles (malleus, incus, and stapes). From there, the stapes relays the wave to the oval window (fenestra vestibuli) and sets the cochlear fluid into vibration (longitudinal waves). This in turn, stresses the hair cells' hairs, which releases an action potential through the hearing nerve to the brain. The spatial location of the hair cells codes the frequency component of the incoming sound wave.

## 2.1. The Ear

From a medical point of view, the ear consists of three parts: the outer ear (earpiece (*pinna*) and eardrum (*Tympanic membrane*)), the middle ear (the ossicles: *malleus*, *incus*, and *stapes*), and the inner ear (cochlea and hearing nerve). With respect to this paper, the cochlea is the most important part, since it converts incoming mechanical signals into electrical signals. This is accomplished by its hair cells: The entire functional chain consisting of the outer and middle ear inject mechanical sound waves into the liquid of the cochlea. Due to the strong coupling, these oscillations excite the hair cells' hairs. Depending on the actual excitation, the hair cells release certain action potentials, which are subsequently processed by the hearing nerve and the auditory cortex. This process is shown in **Figure 2**.

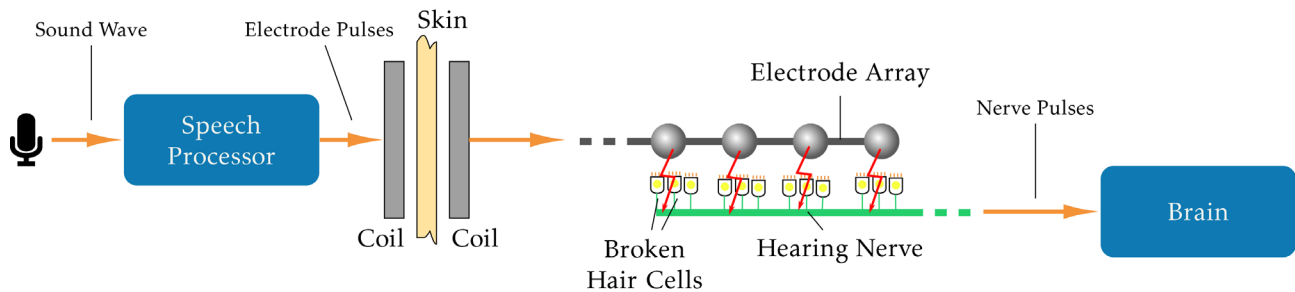
All incoming sound waves are reflected at the cochlea's end. This leads to a superposition of several sources. Due to its construction and the damping inside its tubes, the cochlea performs a frequency-to-spatial transformation: Every single hair cell is particularly sensible to one frequency of the perceived sound signal, with higher frequencies being detected at the cochlea's entry (also known as *base*, *i.e.*, near the oval window (*Fenestra vestibuli*)) and lower frequencies being detected at the cochlea's end (*apex*).

Due to age, diseases, inheritance, and too loud sound sources, the hair cells, particularly their hairs, might be damaged or even broken. As a result, the hair cells' hairs cannot stimulate the hair cells' nerves. In the long run, this leads to partial or even total loss of the hearing sense, even though the hair cells' bodies are still functioning. This is known as sensorineural hearing loss, which is the most common form of hearing loss. Other forms are conductive hearing loss, where the mechanical forwarding mechanism is broken (e.g., if the middle ear is damaged), or a mixed form of sensorineural and conductive hearing loss. With conductive hearing loss, cochlear implants are not useful but other devices are available.

## 2.2. Cochlear Implants

Cochlear implants, also called CIs, have been developed since the 1950's. Their very purpose is to stimulate the nerve cells electrically. That is, a small number  $n \approx 12$  of electrical electrodes substitute the mechanical excitation of the nerve cells with electrical signals.

A cochlear implant is not just a bunch of electrodes. Rather, it is a complete processing chain as shown in **Figure 3**. In the very beginning, a microphone captures incoming sound stimuli. On that data, the cochlear implant's processor splits the audio signal through several bandpass filters with different frequency bands. The number of bands corresponds to the number of electrodes. After that, the contour (envelope) of each band is calculated. The resulting signals are then superposed with a pulse train signal, to form the correct amplitudes to steer the electrodes. These are then transmitted via a transmitting coil on top of the skin



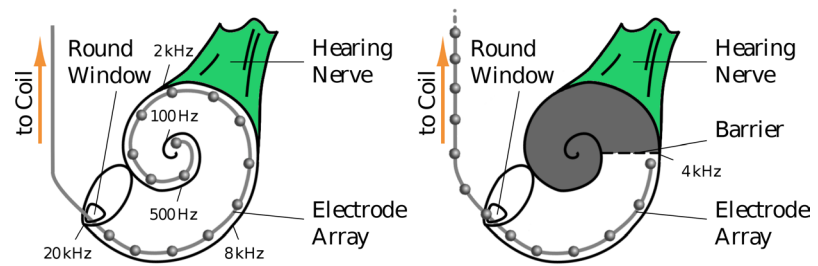
**Figure 3.** The simplified function of a cochlear implant. An arriving sound wave is captured by a microphone. A speech processor proceeds to process the sound into as much signals as there are active electrodes. These signals may consist of pulses with different durations and heights instead of discrete equidistant values like in digitized sound waves. The pulse trains are then transmitted via a transmitting coil to the receiving coil inside the head. From there, the pulse signals are relayed to the electrodes in the electrode array, where they stimulate the broken hair cells. This invokes the action potentials of the cells, and result in the sensation of perception in the brain.

to an implanted receiver coil. From there, the signals are then relayed to the electrodes to invoke the hearing sensation.

In the mode of operation, the cochlear implant processes the acoustic signals, decomposes them into independent frequency bands, and uses those to stimulate the nerve cells via the electrodes. From that description, it should be clear that cochlear implants require repetitive recalibration of all electrodes and frequency bands, since the transition resistance of the electrodes or even the characteristics of the hair cells themselves might change (age) over time.

### 3. Main Problem: Low Insertion Depth

Even with such sophisticated cochlear implants as mentioned, not all patients can understand speech to a sufficient degree. One of the major problems is a reduced insertion depth due to a partly ossified cochlea. The upper part of **Figure 4** displays the ideal insertion depth of the electrode array. The figure shows that with all electrodes ideally inserted into the cochlea, the patient can perceive the entire frequency, spectrum e.g., 20 Hz to 20 kHz because the electrode maps this entire range. Unfortunately, even optimal insertion of electrodes rarely results in stimulation beyond the 1 kHz nerve location [2]. However, these faded out frequencies may be mapped to the lower electrodes later on, so patients can at least perceive these frequencies. Because of this, and cochlear implant's coarse frequency representation, patients may have up to 24-times worse pitch perception than a hearing control group [3], even with the current best insertion depth. Rather, ossification within the cochlea prohibit the electrode array to be inserted to its full extent (lower part of **Figure 4**). In the displayed example, too low an insertion depth has two consequences: 1) the patient cannot perceive any frequency lower than 4 kHz, and 2) only seven (out of twelve) electrodes remain effective. In addition, all signal frequencies are mapped onto higher frequencies in the cochlea, which is, however, compensated during calibration.



**Figure 4.** Ideal and problematic cochlear implant electrode insertion. The left-hand picture displays the ideal insertion of a cochlear implant electrode array. The array penetrates the cochlea deeply and covers the entire range from base (at 20 kHz) to apex (at 20 Hz). The right-hand picture displays a partly ossified cochlea in which it is not possible to insert the electrode array to a full or even its mediocre extent. The topmost electrode barely reaches the hair cells at 4 kHz.

However, research aiming at enhancing pitch perception was recently done by Laneau *et al.* [4], which developed F0mod. This system modulates the electrode signal with the fundamental frequency, which has been shown to be beneficial to speech recognition for frequencies up to 1000 Hz. Francart *et al.* [5] have shown that pitch related tasks can benefit from the F0mod approach.

## 4. Pitch

Pitch is the perceived “height” of a sound and constitutes a neurophysiological phenomenon. It is *not* a frequency in the traditional sense. Instead, it is a perceptive phenomenon in which the human brain extracts harmonic information from a signal, and turns it into the illusion of being a frequency. This harmonic information is measurable and is called the *fundamental frequency*. The fundamental frequency is detectable in a sound signal, whereas the pitch sensation is only created in the brain. In other words: the “fundamental frequency” refers to the incoming sound, whereas “pitch” refers to the physiological phenomenon in the auditory system.

This section explains the concept behind pitch, and shows how to estimate it. In addition, it outlines the role of pitch in speech and describes some pitfalls in dealing with pitch in digital speech processing systems. This section also briefly describes the transposition of the incoming sound signal, since the adaptive pitch transposition module aims at a perfect match between the frequency at which the patient has the best speech comprehension and the fundamental frequency of the incoming sound.

### 4.1. The Neurophysiological Phenomenon

Pitch is an attribute of sound, which only manifests itself in the brain of the listener. The perception of pitch is and has been subject to intensive research.

Two theories have emerged: The *place coding* and the *temporal coding* theory [6]. Place coding argues that the location of the most excited hair cells define the perceived pitch. However, this only holds for high frequencies and is not applicable to lower frequencies, as the frequency discrimination characteristics

in that area do not allow for such fine frequency resolution humans display. For lower frequencies, the temporal coding theory can be used. Temporal coding states that the location of the excitation is almost irrelevant. However, the firing rate (rate of excitation) of the hair cells allows to infer a pitch, *i.e.*, the higher the rate, the higher the pitch. The temporal coding theory is yet limited in frequency as the nerve firing rate of neurons is naturally limited to 300 - 500 times per second and is thus not capable of encoding 20.000 Hz.

## 4.2. Pitch in Sound

A distinction should be made, between sounds, that have a perceivable pitch and sounds that do not. A pitched sound has an inherent periodicity (a fundamental frequency). This periodicity usually appears as distinct, easily observable peaks in the frequency spectrum. Non-pitched sounds, however, are all forms of noise or noise-like, such as a flowing river or rustling leaves. With those having no detectable periodicity, they also have no distinct fundamental frequency. Thus, using the theories of pitch perception, those sounds have no perceivable pitch.

## 4.3. Pitch in Speech

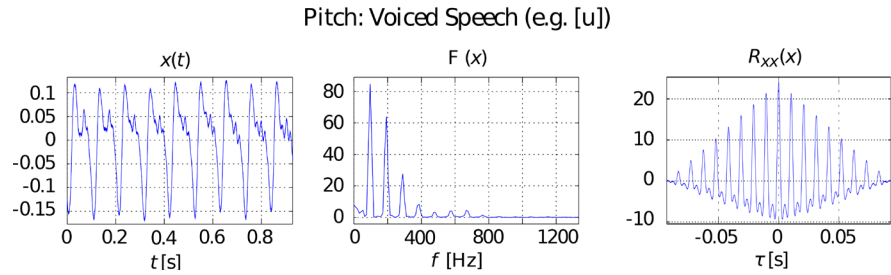
The human body can create sounds that have a perceivable pitch. What is usually called the voice, is actually the vocal folds vibrating, and giving the exhaled air a periodic oscillation characteristic. That is, by using the vocal folds, humans create speech with a perceivable pitch, which is commonly referred to as *voiced* speech. However, if the vocal folds are kept still, the air does not vibrate, and thus the created sound does not have a fundamental frequency. These tones, which are usually only shaped by the current shape of the mouth, are called *unvoiced*. As a third category, also silence is considered as sound, but it does not play a role in the current discussion.

In speech, voiced sounds are called vowels, and unvoiced sounds are usually known as consonants (or whispered speech). Examples for these sounds are [ʃ] like in fish, or [θ] like in teeth. **Figure 5** and **Figure 6** show signals, spectra, and autocorrelation examples (see Section 6) with and without a fundamental frequency, where these characteristics can be easily seen. Pitch is usually perceived as lying between 50 Hz and 500 Hz for speech. This is the frequency range for vowels, where pitch perception plays a vital role [7].

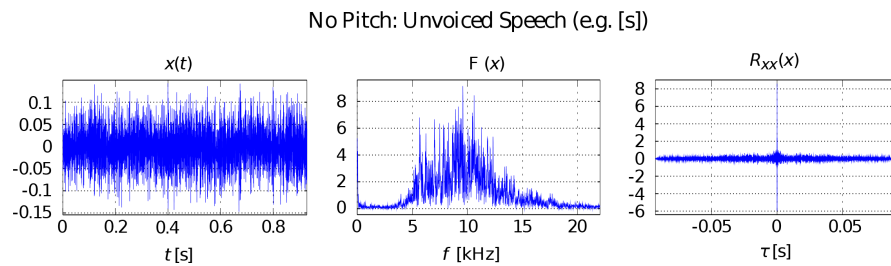
On a side note, the decreasing peaks on the vowel spectrum, constitute another attribute of sounds, namely the timbre, which is defined through the shape which is formed by the single peaks.

## 4.4. The Estimation of Pitch

Pitch cannot be determined directly. Rather, it is normally estimated via the fundamental frequency  $f_0$  of the considered sound sample. The fundamental frequency  $f_0$  is the lowest in the spectrum that forms a set of harmonic frequencies  $f_n$ , which are integer multiples of the fundamental frequency. Frequency  $f_0$  itself is also a harmonic. Any harmonic sound is composed of



**Figure 5.** A waveform and spectrum example of voiced speech. The left-hand picture depicts a part of the waveform signal, which is a periodic sound, and thus should have a pitch. The frequency spectrum in the central diagram displays the expected distinct peaks. The right-hand diagram shows the waveform autocorrelation result. The interesting region of this plot is between 2 ms and 20 ms (between 50 Hz and 500 Hz respectively), from where a pitch of around 95 Hz has been estimated.



**Figure 6.** A waveform and spectrum example of unvoiced speech. The left-hand of the figure depicts a part of the waveform signal. The noise like structure of the signal is evident at this representation. In the central picture, the frequency spectrum of that part is shown. The frequency distribution looks quite random. The right-hand picture shows the autocorrelation result when applied to the waveform signal. Only the lag  $\tau = 0$  shows a significant value, therefore it can be assumed, that no pitch is present.

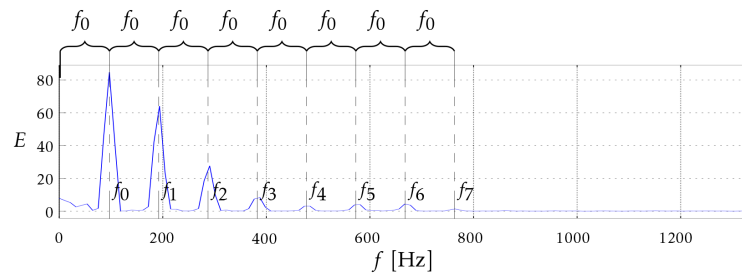
one or more harmonics and is thus periodic with respect to  $f_0$ . In case a sound has no fundamental frequency, it cannot be considered periodic (or harmonic, respectively). **Figure 7** depicts the relationship between the fundamental frequency  $f_0$  and its harmonics by the example of a pure [u:] sound.

As already shown, pitch can be estimated by evaluating the distance between the individual peaks. Due to several side effects, it is usually not possible to just use the largest peak of the spectrum as an  $f_0$  estimate. Thus an alternative approach to measure the periodicity of a signal is required. The autocorrelation function (ACF,  $R_{xx}$ ) is one of the widely used algorithms for this very purpose. Autocorrelation slides a portion of the data over itself with an offset lag  $\tau$  to determine a similarity measure. A maximum in this function indicates a harmonic frequency determined by the corresponding lag  $\tau$ . The equation for the autocorrelation function reads as follows:

$$R_{xx_n}(\tau) = \sum_{n=-N}^N x(n)x(n-\tau) \quad (1)$$

The application of the autocorrelation function to the original signal (*not* the spectrum) yields a good pitch estimate. However, this is only true for signals that





**Figure 7.** In the relationship of the fundamental frequency  $f_0$  to its harmonics  $f_n$  pitch can be seen as a representative of  $f_0$ . The fundamental frequency  $f_0$  is the lowest in the set, while each harmonic  $f_n$  is an integer multiple  $f_0$  of  $f_0$ .

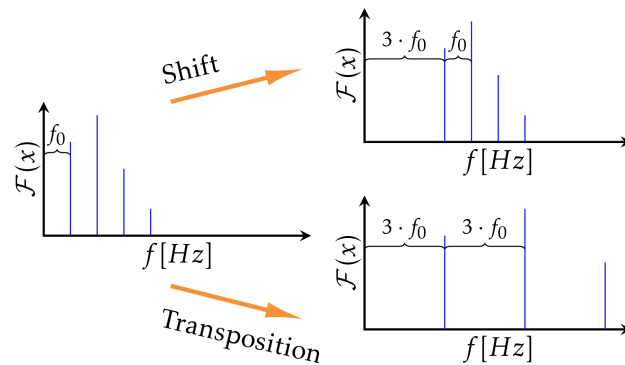
have a harmonic structure. For unvoiced signals, the autocorrelation results appear to be pseudo-random. Therefore, the results have to be verified. This verification process is called voice-activity-detection (VAD) or voiced-unvoiced-decision (V/UV) for speech audio signals, and is used in speech recognition software or other speech-related tasks. Voice-activity-detection usually classifies a given signal part into voiced, unvoiced, and silence. Indicators for voiced speech are high signal energy combined with a low zero crossing rate, which indicates the number of sign changes in a digitized signal. A vice versa constellation, *i.e.*, low energy and a high number of zero crossings, is a good indicator for noise. However, a variety of different algorithms of different complexities and results are offered by the pertinent literature [8].

#### 4.5. Pitch Shifting

Pitch shifting or pitch transposition is the act of changing the fundamental frequency  $f_0$  to change the perceived pitch of a signal. Since pitch is an elaborate concept, this is not a trivial task. Usually pitch shifting is referred to as pitch transposition, since a shift means moving everything in the signal by an absolute value. Transposition however, considers the relationships of the harmonics in the spectrum and keeps those intact. In other words: pitch transposition is harmonic conservant, where pitch shifting is not. **Figure 8** illustrates this difference.

Transposing the fundamental frequency of a signal is usually a two-step method. The first step consists of changing the signal length, without affecting the pitch. Then, the second step simply consists of altering the replay speed. This step can be imagined as a player for vinyl records, which is intentionally set to the wrong speed. In case of a higher-than-intended speed, both the record's pitch increases and the duration shortens. In the opposite case, the pitch decreases, whereas the playtime increases.

In terms of digital signal processing, the two supplied steps are called 1) time stretching and 2) resampling. First, the signal is stretched in time with the original pitch maintained. Then the data is resampled to the original signal



**Figure 8.** This figure shows the difference between shifting and transposing the fundamental frequency  $f_0$  of a signal. Both the operations are applied with a factor of 3. Shifting the signal's  $f_0$  results in a usually disharmonic outcome, whereas transposing yields a well sounding result.

length. A signal stretched by the factor of two, resampled to the original length, will thus have a pitch at double frequency.

For this task, the pertinent literature offers several algorithms. The PSOLA algorithm [9] is a well-known example in the time domain, whereas the phase vocoder [10] is dominant in the frequency domain.

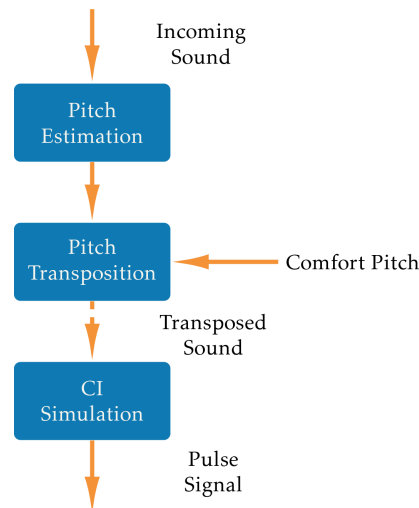
## 5. Solution: Adaptive Pitch Transposition

This section describes both, the idea and functionality of the adaptive pitch transposition module. From a top-level perspective, adaptive pitch transposition is a module that is inserted between the microphone and the cochlear implant's speech processor. Thus, this module changes all incoming utterances, which are then processed by the implant as usual. After a brief overview of the general idea, the single parts of the concept are outlined. These are the concept of *comfort pitch*, which constitutes the voice pitch of best comprehension, the actual pitch detection, and the final pitch transposition.

### 5.1. Overview

Adaptive pitch transposition, or APT for short, is a method that shifts (or transposes) the pitch of an arbitrary signal to a fixed, predetermined value. The resulting signal will thus have only a single pitch for each of the voiced signal parts. Based on the former explanations, this will most likely help cochlear implant patients in their speech comprehension, since it accounts for both the low insertion depth as well as their low brain plasticity. The required functional units are depicted in **Figure 9**.

The APT concept works as follows: It assumes that a particular patient has a certain voice pitch that yield the best recognition rates. This particular pitch is called *comfort pitch*. Then, APT “smartly” transposes all incoming *voiced* sound samples to that pitch. The “mart” part is to measure the current value of the



**Figure 9.** The functional units of the adaptive pitch transposition.

pitch and adjust it adaptively to the patient’s comfort pitch. This approach requires the following processing stages: 1) determination of the current pitch, 2) determination of the required shift amount, and 3) shift/transposition of all incoming signals to the target comfort pitch. The transposed signals may then be forwarded to an arbitrary cochlear implant speech processing algorithm.

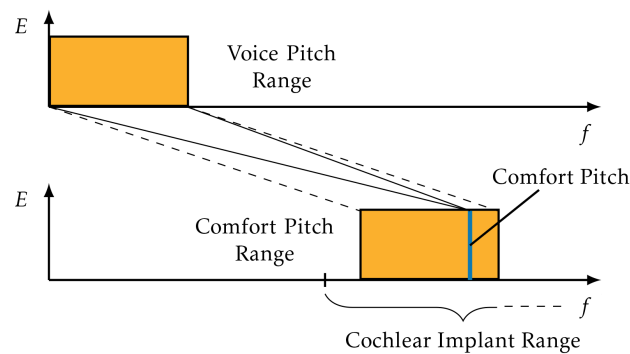
## 5.2. Pitch Detection

For performing the actual pitch transposition, the pitch of the current signal has to be determined (estimated). This is especially true, since usually the signal pitch is varying (due to intonation), which leads to varying conversion factors, because the signal has always to be transposed to the very same target pitch. In other words: not the transposition factor is constant, but the target pitch is. This varying conversion factor is called the transposition coefficient  $k$ , which in turn defines how much shift has to be applied.

Pitch detection itself is executed as explained in Section 6. In addition, Voice activity detection is required, since unvoiced speech does not require any further processing.

## 5.3. Comfort Pitch

The comfort pitch is the frequency of best comprehension. For humans with an intact auditory system, the comfort pitch may not have a distinct value, but rather consist of a certain range of frequencies. Given the brains accustoming to the human voice, this frequency most likely lies within the frequency range of the human voice. Cochlear implant patients, however, suffer from the shortcomings explained above: Their comfort pitch might be a single value, which presumably lies within the range of the cochlear implant’s electrode frequency range. The relationship of these *different* pitches, is depicted in **Figure 10**.



**Figure 10.** Difference of voice pitch range and comfort pitch range. The two graphs represent different pitch ranges in the frequency domain. The voice range of possible voice pitches is spatially disjunct from the range, a cochlear implant can address. Therefore the pitch has to be transposed into the cochlear implant range.

This paper defines the comfort pitch  $f_c$  as the one among a given set of pitches with which a particular listener has the best speech recognition rates. This definition directly defines the means of how to estimate the comfort pitch.

The comfort pitch will be determined by a simple test. This test will be comprised of voice samples of different pitches, uttering random phrases from a fixed set of sample phrases. A sample speaker pitch configuration might be as shown in [Table 1](#).

In this table, every speaker was recorded with a certain set of phrases, which are then presented to the patient in a randomized manner. Subsequently, the patient communicates the content of the phrase according to what he or she understood. Pitch-score pairs are then ranked for best speech recognition rates on the patient's side. The voice pitch with the best score will be the new comfort pitch for optimal recognition when using the system. From a measure of two best scores, the lowest one will be picked, since it results in lower computational costs during the subsequent pitch transposition. Through this test, the comfort pitch adapts to the patient.

The comfort pitch itself is a robust feature, since it is determined with an already implanted cochlear implant, and just depends on the nerve ends in the cochlea, which should not deteriorate in an unusual manner. The speech recognition rates themselves will probably rise even further, since the principle of the comfort pitch relies on the brain's adaptivity to the equal sounding stimuli. It is adapted to the existing structures, and may benefit from the remaining brain plasticity.

#### 5.4. Pitch Transposition

As the name already suggests, the pitch transposition is the vital part of the APT module. The source audio signal is changed to a new pitch within this block (second block in the overview on [Figure 9](#)). This change happens relative to the current signal pitch. Therefore a transposition coefficient  $k$  is required.

**Table 1.** A sample speaker pitch configuration.

Speaker	M1	M2	M3	F1	F2	F3	F4
Pitch	95	110	140	200	240	290	320

The transposition coefficient  $k = f_0/f_c$  is determined by using the comfort pitch  $f_c$  and the fundamental frequency  $f_0$  (see, also, Section. 6 and the previous Section.). In APT, the pitch transposition utilizes  $k$  to calculate the relative change in pitch. This ensures an always steady pitch at the output. The output signal is a high definition broadband audio signal with almost the same characteristics as the input signal has, except for the altered pitch. **Figure 11** illustrates the intermediate steps mentioned in Section 4 through the resulting signals. However, the resulting spectrum in the bottom graph cannot yet be generated by the algorithm, since the pitch detection algorithm is yet partly unreliable.

## 6. Implementation

For test purposes, the APT algorithm was implemented in both Matlab and C++. Both programs share the same basic structure and algorithms. The implementations focus on proof-of-concept and maximal flexibility in order to allow further experimentation. This section provides a brief overview of both implementations.

### 6.1. Overview

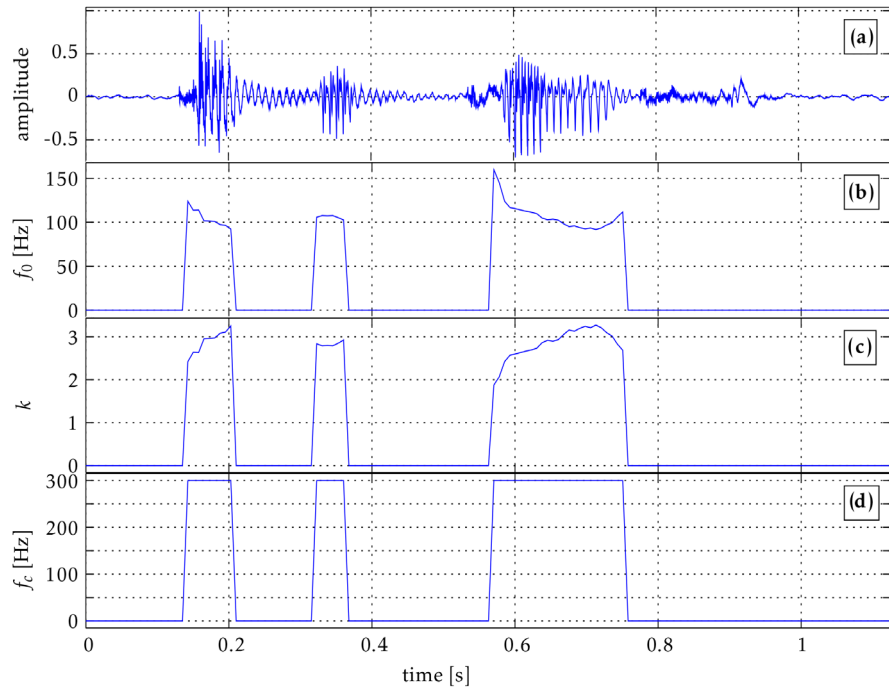
The implementation and testing has been done using an Intel® Core™ i7 Q720 CPU at 1.6 GHz. The input of the programs may be any file in the Waveform Audio File Format (WAV). The C++ implementation also allows for live microphone input with simultaneous output (best suited for headphone use). The comfort pitch is yet assumed to be a fixed value, and is to be set as a parameter.

After obtaining the audio data, it is segmented into overlapping frames. Each of the following steps is executed per frame of size  $N$  (e.g. with  $N = 1024$ ). The output signal, has the modified fundamental frequency  $f_c$ .

The given hardware is under considerable load when running the program. Therefore, it can be concluded that a microprocessor solution would have to be highly optimized, or alternatively, fundamental improved algorithms have to be developed.

### 6.2. Pitch Detection

A simple autocorrelation has been used for the pitch detection, as already described in Section 6. The resulting signal is then parsed for its maximum with respect to the lag  $\tau$ , equivalent to the pitch range of 50 - 500 Hz. As already explained, the results are not reliable, since not every frame has a pitch. Therefore a verification function evaluating both the zero crossing rate (ZCR)



**Figure 11.** Exemplary processing of the German phrase 'assistenz'. The detected pitch is rough in its structure; after alteration it has just one single value. (a) is the waveform of the phrase; (b) is the pitch detected with the autocorrelation function; (c) is the calculated transposition coefficient for an  $f_c = 300$  Hz; (d) is the desired resulting pitch fixed to 300 Hz (for this example).

and the energy ( $E$ ) has been used. Equation (2) calculates a validity measure with  $p > 0$  indicating a voiced frame. This equation achieves good results for pre-normalized signals, but not for normalized frames.

$$p_n = \frac{E_n + 1}{ZCR_n + 1} - 1 \quad ZCR, E \in [0, \dots, 1] \quad (2)$$

If the voice activity detection indicates a voiced frame, APT uses the  $f_0$  estimated by the autocorrelation algorithm.

However, even with voice activity detection, not every frame is classified correctly. Because of that, the voice activity results of some of the recently calculated frames are taken into account to determine the result of the current frame. Using this, a meaningful pitch trajectory in time can be established (since, if the last frame was pitched, the next will most likely also be).

The real-world application made clear that the pitch detection algorithm (especially the voice activity detection part) is the bottleneck of the implementation. Wrongly detected pitches result in jumps in the created pitch trajectory. This destroys the quasi steady character of the pitched regions, and leads to worse comprehension, even for hearing people.

### 6.3. Comfort Pitch

Since no patient data is available yet, the comfort pitch is included as an arbitrary value in the programs, which might be set later according to the

determined  $f_c$  of selected patients. Currently it is used as an indicator whether or not the algorithm works. This evaluation is done manually by listening tests.

The transposition coefficient for frame  $n$  is calculated according to

$$k_n = \frac{f_{0n}}{f_c} \quad (3)$$

#### 6.4. Pitch Transposition

Both implementations use a modified phase vocoder technique based on [11] for the time stretching step in the pitch transposition (see Section 7). A phase vocoder modifies the spectrum of a frame in such way that two adjacent frames can be combined to form a signal of arbitrary length while maintaining the current pitch of the incoming frame. It uses spectral information to alleviate frequency errors made in the Fourier transform step, to make estimates of *true frequencies*. These can be used to generate the resulting time stretched signal by generating new synthetic spectra and combining the result of the inverse transformation. Further explanations to these methods are given in [10] and [12].

Knowing the transposition coefficient  $k$ , it is possible to prolong the signal using such a phase vocoder. As a last step, the signal has to be resampled. This means, basically, to transform the signal back from digital to analogue, and to sample it again, using a different sampling frequency. Digital signal processing systems however, already provide filtering and interpolation.

After resampling, the signal has a changed pitch and the original length, and is now ready to be fed into the actual speech processing of a cochlear implant.

### 7. Discussion

Not every cochlear implant patient has a sufficient speech comprehension to get through his or her everyday life. Due to ossifications in the cochlea, the implant's electrodes do not reach deep enough into the cochlea to provide the full frequency spectrum to the patient. To alleviate this problem, this paper has presented the concept of adaptive pitch transposition. This is possible due to considerations of how words can be distinguished on a coarse level through vowels alone, which is utilized through shifting the determining information in the audio signal into the perceptive range of the cochlear implant patient.

One major drawback is, however, the loss of speaker differentiation, since every speaker will sound alike after the modification. The only indicators left will be the timbre and visual cues (e.g., moving mouth). Nevertheless, in prospect of regaining the ability to understand speech, this seems negligible.

On the other hand, through the ever same stimuli, the brain will adapt to the comfort pitch over time. Under the assumption a comfort pitch exists for every patient, the brain is already accustomed to being stimulated in that particular area of the cochlea. This enables the brain to learn to differentiate these stimuli even better, and thus, to use the remaining brain plasticity.

As for the introduced parameter called comfort pitch, it is yet to be investigated,

whether a single best pitch exists for cochlear implant patients (as is for normal hearing). The comfort pitch is not likely to change after it has been determined, since it depends on the remaining nerves and is fitted to the established structures in the brain.

The extent to which the concept together with the proposed system enhance the speech recognition rates in patients is subject of future research. It will be directed to the evaluation and testing with hearing people as well as cochlear implant users. Furthermore, performance and quality increase of the developed implementation are required, in order to allow for real time usage and testing. The current implementations are good foundations to do so.

## Acknowledgements

The authors gratefully thank the welisa graduate school for its support. Part of this research was funded by the German Research Foundation (DFG) grant number GRK 1505.

## References

- [1] Struwe, K. (2017) APT: Enhanced Speech Comprehension Through Adaptive Pitch Transposition in Cochlear Implants. In: Giokas, K., Bokor, L.-Z. and Hopfgartner, F., Eds., *eHealth 360°: International Summit on eHealth*, Budapest, 14-16 June, 2016, Revised Selected Papers. Springer International Publishing, 224-228.
- [2] Shannon, R.V., et al. (2004) Speech Perception with Cochlear Implants. In: *Cochlear Implants: Auditory Protheses and Electric Hearing*, Springer, 334-376.
- [3] Zeng, F.-G., Tang, Q. and Lu, T. (2014) Abnormal Pitch Perception Produced by Cochlear Implant Stimulation. *PLoS One*, **9**, e88662.
- [4] Laneau, J., Wouters, J. and Moonen, M. (2006) Improved Music Perception with Explicit Pitch Coding in Cochlear Implants. *Audiology and Neurotology*, **11**, 38-52.
- [5] Francart, T., Osses, A. and Wouters, J. (2015) Speech Perception with F0mod, a Cochlear Implant Pitch Coding Strategy. *International Journal of Audiology*, **54**, 424-432.
- [6] De Cheveigne, A. (2005) Pitch Perception Models. In: *Pitch*, Springer, 169-233.
- [7] Patterson, R.D., Gaudrain, E. and Walters, T.C. (2010) The Perception of Family and Register in Musical Tones. In: *Music Perception*, Springer, 13-50.
- [8] Mak, M.-W. and Yu, H.-B. (2014) A Study of Voice Activity Detection Techniques for NIST Speaker Recognition Evaluations. *Computer Speech & Language*, **28**, 295-313.
- [9] Charpentier, F.J. and Stella, M.G. (1986) Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP86*, **11**, 2015-2018.
- [10] Flanagan, J.L. and Golden, R.M. (1966) Phase Vocoder. *Bell System Technical Journal*, **45**, 1493-1509.
- [11] Ellis, D.P.W. (2002) A Phase Vocoder in Matlab. <http://www.ee.columbia.edu/ln/rosa/matlab/pvoc/>
- [12] Laroche, J. and Dolson, M. (1999) Improved Phase Vocoder Time-Scale Modification of Audio. *IEEE Transactions on Speech and Audio Processing*, **7**, 323-332.



**Submit or recommend next manuscript to SCIRP and we will provide best service for you:**

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.

A wide selection of journals (inclusive of 9 subjects, more than 200 journals)

Providing 24-hour high-quality service

User-friendly online submission system

Fair and swift peer-review system

Efficient typesetting and proofreading procedure

Display of the result of downloads and visits, as well as the number of cited articles

Maximum dissemination of your research work

Submit your manuscript at: <http://papersubmission.scirp.org/>

Or contact [eng@scirp.org](mailto:eng@scirp.org)