EUSIPCO 2013 1569739899

A HARMONIC/PERCUSSIVE SOUND SEPARATION BASED MUSIC PRE-PROCESSING SCHEME FOR COCHLEAR IMPLANT USERS

Wim Buyens[1 2 3], Bas van Dijk[1], Jan Wouters[2] and Marc Moonen[3]

[1]Cochlear Technology Centre Belgium, Schaliënhoevedreef 20 I, 2800 Mechelen, Belgium
[2]KULeuven, Department of Neurosciences (ExpORL), O&N 2, Herestraat 49 bus 721, 3000 Leuven
[3]KULeuven, Department of Electrical Engineering (ESAT-SCD) and iMinds Future Health Department, Kasteelpark Arenberg 10, 3001 Heverlee

ABSTRACT

Music perception and appreciation remain generally poor in cochlear implant (CI) users. Simple musical structures, a clear rhythm/beat and lyrics that are easy to follow are among the top factors to enhance music appreciation. A music pre-processing scheme for CI users is described in which vocals and percussion are elucidated using harmonic/percussive sound separation. The scheme is capable of modifying relative instrument level settings for improving music appreciation in CI users. The scheme is assessed with normal hearing subjects (N=5) using a pairwise comparison analysis and CI simulated music excerpts. All test subjects except one significantly preferred the music excerpts from the music pre-processing scheme.

Index Terms— sound separation, cochlear implant, music appreciation

1. INTRODUCTION

A cochlear implant (CI) is a medical device enabling severe to profoundly deaf people to perceive sounds by electrically stimulating the auditory nerve using an electrode array implanted in the cochlea [1]. Although CI users reach good speech understanding in quiet surroundings, music perception and appreciation generally remain poor [2]. Simple musical structures and a clear rhythm/beat were reported amongst the top factors that enhance musical appreciation for CI users [3]. A negative correlation was found between complexity and appreciation in CI users studied with pop, country and classical music [4]. Classical music was rated as more complex than pop and country music. Several plausible explanations were provided including the presence of simple musical structures and lyrics in pop and country music. Since CIs were mainly developed for transmitting speech sounds, the presence of lyrics might facilitate CI users to follow more easily the sequence of events in complex music. In addition, both pop and country music oftentimes contain a strong, simple beat which is more suitable for transmission through current-day CIs than the structural features of instrumental classical music.

The preference for clear vocals and a strong rhythm/beat in CI users was studied by modifying relative instrument level settings in pop music [5]. A significant difference in preference rating scores was found between normal hearing (NH) and CI subjects. For the pop songs provided, CI subjects preferred an audio mix with larger vocals-to-instruments ratio compared to normal hearing subjects. In addition, given an audio mix with clear vocals and attenuated instruments, CI subjects preferred the bass/drum track to be louder than the other instrument tracks. The relative instrument level settings were modified by altering the levels of the different, separately recorded, instrument tracks, which are, however, not widely available for most music. Therefore, a signal processing scheme is needed to modify the relative instrument level settings in complex music.

The paper is organized as follows. In section 2, the proposed music pre-processing scheme, which modifies the relative instrument level settings in complex music, is described in detail. In section 3 and 4, the music pre-processing scheme is assessed objectively and subjectively. Section 5 contains the conclusions.

2. MUSIC PRE-PROCESSING SCHEME

The music pre-processing scheme for CI users, which performs vocal and drum enhancement on complex music, is shown in Figure 1. The harmonic/percussive sound separation (HPSS) separates harmonic (H) and percussive (P) components by exploiting the "anisotropic smoothness" of these components in the spectrogram [6]. "Anisotropic smoothness" of sound is defined as partial differentials of the spectrogram in temporal or frequency direction: harmonic components are "smooth in temporal direction" because they are sustained and periodic; percussive

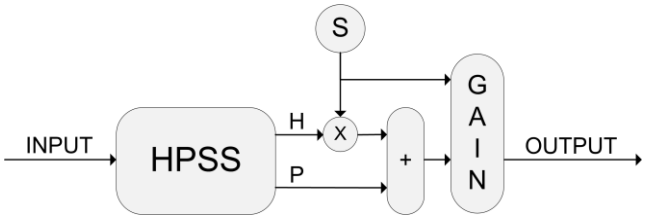


Figure 1: Music pre-processing scheme for CI using harmonic/percussive sound separation (HPSS)

components are "smooth in frequency direction" because they are instantaneous and aperiodic [7]. The original power spectrogram $W_{\tau,\omega} = STFT[w(t)]$ from input signal $w(t)$, in which indices τ and ω represent time and frequency, respectively, is decomposed into the harmonic component $H_{\tau,\omega}$ and the percussive component $P_{\tau,\omega}$. In the original paper [6], the L_2 norm of the power spectrogram gradients is examined for evaluating the anisotropic smoothness, that is, $H_{\tau,\omega}$ and $P_{\tau,\omega}$ are found by minimizing:

$$J(\boldsymbol{H},\boldsymbol{P}) = \frac{1}{2\sigma_H^2}\sum_{\tau,\omega}\left(H_{\tau-1,\omega} - H_{\tau,\omega}\right)^2 + \frac{1}{2\sigma_P^2}\sum_{\tau,\omega}\left(P_{\tau,\omega-1} - P_{\tau,\omega}\right)^2 \quad (1)$$

under the constraint of

$$H_{\tau,\omega}{}^2 + P_{\tau,\omega}{}^2 = W_{\tau,\omega}{}^2 \quad (2)$$

$$H_{\tau,\omega} \geq 0,\ P_{\tau,\omega} \geq 0 \quad (3)$$

where $\boldsymbol{H}$ and $\boldsymbol{P}$ are sets of $H_{\tau,\omega}$ and $P_{\tau,\omega}$, respectively, and σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness.

Several algorithms were explored to solve this optimization problem numerically, which were divided in two main approaches: power spectrogram estimation based on low-pass filtering and based on optimization [8]. The parameters of the different algorithms were tuned using the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) as performance criteria. The "optimization under hard mixing constraint", referred to as "HM2" in [8], was found to outperform the other algorithms. After estimating the power spectrograms $H_{\tau,\omega}$ and $P_{\tau,\omega}$, a time-frequency mask was considered before applying inverse STFT to estimate the corresponding waveforms $h(t)$ and $p(t)$. From the different considered types of time-frequency mask, the binary mask was most effective to improve SIR. Therefore, in the music pre-processing scheme of Figure 1, the "HM2" approach was selected for the HPSS together with the binary mask, resulting in the following iteration formulae:

$$H_{\tau,\omega}{}^2 \leftarrow \frac{\alpha_{\tau,\omega} W_{\tau,\omega}{}^2}{(\alpha_{\tau,\omega}+\beta_{\tau,\omega})} \quad (4)$$

$$P_{\tau,\omega}{}^2 \leftarrow \frac{\beta_{\tau,\omega} W_{\tau,\omega}{}^2}{(\alpha_{\tau,\omega}+\beta_{\tau,\omega})} \quad (5)$$

where

$$\alpha_{\tau,\omega} = (H_{\tau+1,\omega} + H_{\tau-1,\omega})^2 \quad (6)$$

$$\beta_{\tau,\omega} = \kappa^2 (P_{\tau,\omega+1} + P_{\tau,\omega-1})^2 \quad (7)$$

in which κ is a tunable parameter that was optimized to maximize SIR of vocals and drums. The binary mask, which was applied on the input power spectrogram to eventually separate H and P, is defined as:

$$BM_{\tau,\omega} = 1 \quad P_{\tau,\omega} > H_{\tau,\omega}$$
$$= 0 \quad P_{\tau,\omega} \leq H_{\tau,\omega} \quad (8)$$

Vocal tones which contain 4-8 Hz quasi-periodic vibrations of the fundamental frequencies (F0s) and which do not sustain for a long time are contrasted to chord tones which contain very few fluctuations and which are temporarily maintained stationary. The nature of the respective tones is called "temporal-variability" and "temporal-stability" [7]. Adjusting the time-frequency resolution of the short-time Fourier transform (STFT) in the HPSS calculations results in a different classification for the temporally-variable components. A STFT with long time window (100-500 ms) provides low temporal resolution and high frequency resolution. Consequently, sounds with long length and narrow bandwidth (temporally-stable sounds) appear "smoothly" in temporal direction (H-component) while sounds with moderately-or-very short length and moderately-or-very broad bandwidth (percussive and temporally-variable sounds) appear "smoothly" in frequency direction (P-component). Conversely, a STFT with short time window (30 ms) provides high temporal resolution and low frequency resolution resulting in the classification of temporally-variable sounds as well as temporally-stable sounds in the H-component.

In the music pre-processing scheme of Figure 1, the window length for the STFT in HPSS is 185 ms, resulting in the classification of temporally-variable components (such as vocal tones) as P-components. The obtained P-component with vocals and drums is added to the harmonic component (H), which is attenuated with an adjustable parameter 'S' ranging from -∞ to 0 dB. The output spectrogram after addition becomes:

$$W_{\tau,\omega}^{out} = P_{\tau,\omega} + S * H_{\tau,\omega} \quad (9)$$

With the attenuation parameter 'S' at 0 dB, the output signal of the HPSS remains unaltered compared to the input signal since the binary mask from equation (8) is applied on the original input spectrogram to separate H and P.

The final stage in the music pre-processing scheme applies a gain to the output signal as a function of the attenuation parameter 'S' to compensate for the decrease in output level due to the attenuated H-component.

The scheme presented in Figure 1 was implemented in Simulink® (Mathworks®) with a sampling frequency of 44.1 kHz.

3. OBJECTIVE TESTING

The music pre-processing scheme in Figure 1 was evaluated objectively with multi-track recordings of pop music [5]. A typical pop song consists of a vocal melody with piano, guitar, bass guitar and drums. The HPSS with a long time window separates vocals and drums (P) from the other instruments (H). The H-component attenuated with an adjustable parameter 'S' is added to the P-component, which results in the output signal. The remainder of the signal or "residual" signal is defined as the components of the original signal not present in the output signal. Consequently, the sum of output and residual signal corresponds to the input signal. The performance of the separation is analyzed with the energy ratio of the output signal and the residual signal for each track [6], and is calculated as

$$r_i^{output} = \frac{E_i^{output}}{E_i^{residual} + E_i^{output}} \quad (10)$$

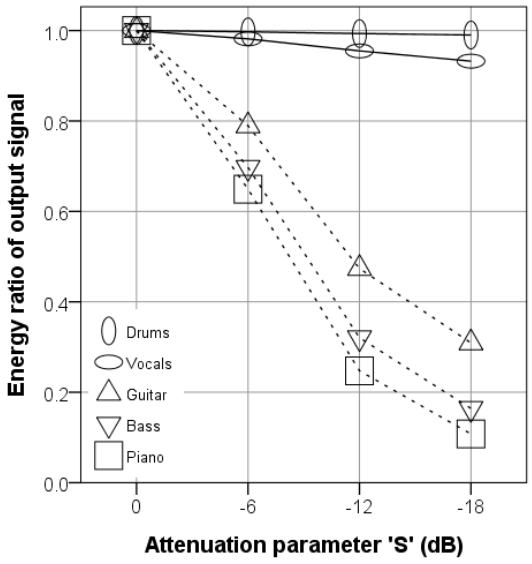
where

$$E_i^{output} = < f_i(t), output(t) >^2 \quad (11)$$

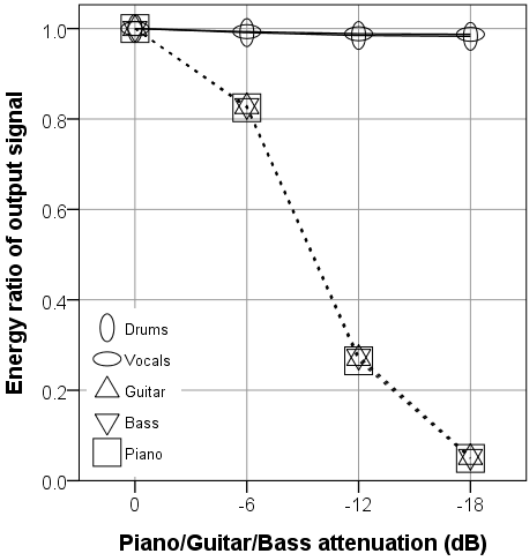
$$E_i^{residual} = < f_i(t), residual(t) >^2 \quad (12)$$

in which <> represents the cross correlation operation and $f_i(t)$ the signal of the i-th track.

The energy ratio of the output signal is shown in Figure 2a for the different tracks of the song 'The Dock of the Bay' by Otis Redding (Vocals, Piano, Guitar, Bass and Drums) [5]. The results are shown with attenuation parameter 'S' equal to 0 dB, -6 dB, -12 dB and -18 dB. With 'S' equal to 0 dB, the input signal and output signal are equal, resulting in a residual signal with zero amplitude and, consequently, an energy ratio of 'one' for each track. The vocal and drum track are preserved in the output signal regardless of the attenuation parameter, whereas the piano, guitar and bass guitar track shift towards the residual signal when decreasing the attenuation parameter. In Figure 2b, the energy ratio of the output signal is shown for the different tracks of the same song where the piano, guitar and bass track are attenuated by 0 dB, -6 dB, -12 dB and -18 dB relative to the vocal and drum track. The energy ratio is calculated with formula (10) in which the residual signal is



(a)



(b)

Figure 2: Energy ratio of output signal for vocal, piano, guitar, bass and drum track of "The dock of the Bay" by Otis Redding (a) for different values of the attenuation parameter 'S' in the music pre-processing scheme for CI and (b) for different Piano/Guitar/Bass attenuation relative to vocals and drums (0 dB).

obtained from the subtraction of the output signal from the input signal. The shape of the curves in Figure 2a and Figure 2b are comparable, which indicates that the music pre-processing scheme is capable of achieving the desired modifications in the relative instrument level settings [5]. The curves of the piano, guitar and bass track, which are by

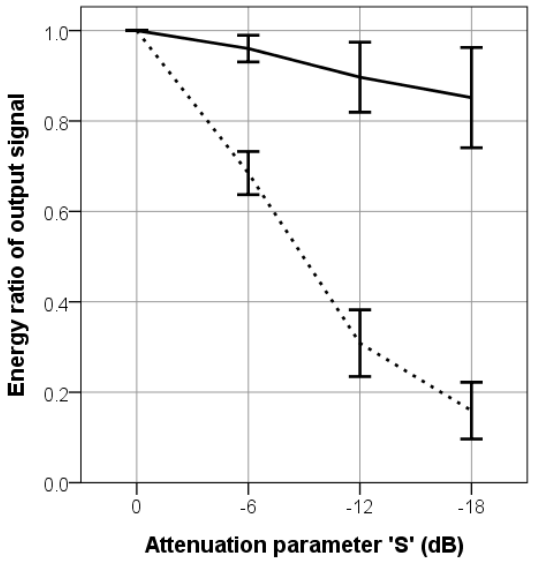


Figure 3: Energy ratio of output signal for the percussive and harmonic components from the songs studied in [5] for different values of 'S'. Percussive components (straight line) include vocal and drum tracks; harmonic components (dotted line) include piano, guitar and bass tracks.

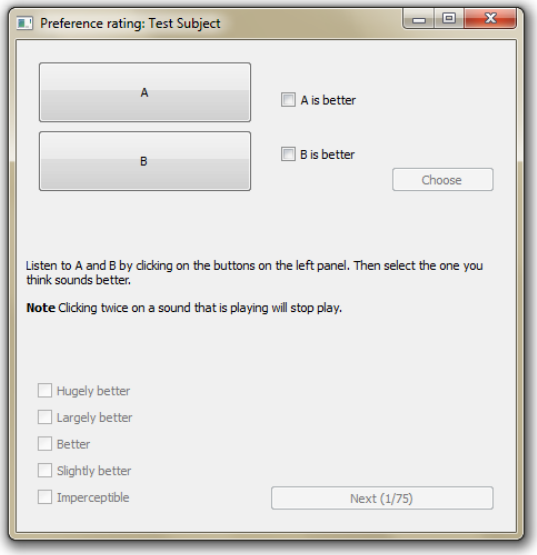


Figure 4: Graphical User Interface for pairwise comparison analysis with processed and unprocessed music excerpts.

definition equal in Figure 2b, show a certain spread in Figure 2a. Since in this song the piano is mainly playing sustained chords while the guitar is playing sharply, the guitar track appears to be more percussive than the piano track. Oftentimes, the opposite trend is observed since the piano is in fact a percussive instrument in which wired strings are struck by small hammers attached to the end of the keys.

Figure 3 shows the energy ratio of the output signal for the songs studied in [5] for different values of attenuation parameter 'S'. Energy ratios are calculated for each track in each song and are presented as harmonic and percussive components. The "harmonic" components (piano, guitar, bass) are attenuated in the output signal with decreasing attenuation parameter 'S', whereas the "percussive" components (vocals, drums) are mostly preserved in the output signal.

4. SUBJECTIVE TESTING

For the subjective evaluation of the music pre-processing scheme, a selection of pop/rock songs was used. The top twenty-five songs from the all-time greatest hits list of a popular radio station in Belgium (Joe FM) were gathered. Representative excerpts of the songs were selected with an average length of 24 seconds and were mixed down to MONO wave files with sampling rate of 44.1 kHz. The excerpts were processed with the music pre-processing scheme with an attenuation parameter 'S' equal to -18 dB.

Subject	Preference Processed (%)	95% Confidence Interval	Median Rating
S1	76	(66%-86%)	Better
S2	83	(74%-91%)	Slightly Better
S3	63	(51%-74%)	Better
S4	65	(54%-76%)	Slightly Better
S5	24	(14%-34%)	Slightly Better

Table I: Results of preference rating experiment, including preference for the processed songs, 95% confidence interval and median rating for the preferred condition.

A pairwise comparison was conducted with the processed and unprocessed condition for the 25 selected pop/rock songs. The pairs were randomly presented and repeated three times for each song. The music was presented to NH test subjects using a CI simulation with noise-band vocoder as used in [9]. In each of the 22 frequency bands, the extracted temporal envelope after the maxima selection process is used to modulate a pink noise signal, which has been bandpass filtered corresponding to the analysis channel. All the modulated channels are then summed to produce the vocoded stimuli. Finally, all stimuli are equalized in rms level. The excerpts were played through headphones (Beyerdynamic DT-770 pro) in a silent room. The test subjects were asked to select the condition that was most enjoyable and to quantify their preference with a rating score ranging from *Imperceptible*, *Slightly better*, *Better*, *Largely better* to *Hugely better* (Figure 4).

The pairwise comparison is presented in percentage indicating the preference for the processed songs. The additional preference rating score from *Imperceptible* to *Hugely Better* is presented as median for the preferred condition and indicates the strength of the preference.

A pilot experiment with five NH test subjects was performed. The test subjects had no self-reported hearing deficit and their age ranged from 26 to 66 years old (average = 37 years). The results are shown in Table I. Four test subjects significantly preferred the processed songs (Chi-square test: $p<0.05$), with a median rating of *Slightly Better* or *Better*. Test subject S5, who is a professional DJ, significantly preferred the unprocessed songs (Chi-square test: $p<0.05$) with a median rating of *Slightly Better*.

5. CONCLUSION

A music pre-processing scheme aimed at improving music perception in CI users has been described and evaluated. The scheme is capable of modifying the instrument level settings of music while preserving vocals and drums, which was found beneficial for the music appreciation in CI users. The scheme is evaluated subjectively using a preference rating experiment with NH test subjects and CI simulated music and showed significant preference for the processed music for all test subjects except one.

ACKNOWLEDGMENT

This work was supported by a Baekeland PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT090274) and Cochlear Technology Centre Belgium. The research work was also carried out in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC) and Concerted Research Action GOA-MaNet. The scientific responsibility is assumed by its authors.

REFERENCES

[1] P. Loizou, "Introduction to cochlear implants," *IEEE Signal Proc Mag.*, 15, pp. 101-130, 1998.

[2] H. McDermott, "Music perception with cochlear implants: A review", *Trends Amplif.*, 8, pp. 49-82, 2004.

[3] K. Gfeller, A. Christ, J. Knutson, S. Witt, K. Murray, and R. Tyler, "Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients", *J Am Acad Audiol.*, 11, pp. 390-406, 2000.

[4] K. Gfeller, A. Christ, J. Knutson, S. Witt, and M. Mehr, "The effects of familiarity and complexity on appraisal of complex songs by cochlear implant recipients and normal hearing adults", *J Music Ther.*, XL, pp. 78-113, 2003

[5] W. Buyens, B. van Dijk, M. Moonen, and J. Wouters, "Music mixing preferences of cochlear implant recipients", 2013 (to be published)

[6] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram", *Proc. EUSIPCO*, 2008

[7] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source", *Proc. ICASSP*, pp. 425-428, 2010

[8] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of various harmonic/percussive sound separation algorithms based on anisotropic continuity of spectrogram", *Proc. ICASSP*, pp. 465-468, 2012

[9] O.R. Qazi, B. van Dijk, M. Moonen, and J. Wouters, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility", *Hearing Res*., 299, pp. 79-87, 2013