Dominant Melody Enhancement in Cochlear Implants

Drew Cappotto[*], Wenye Xuan[†], Qinglin Meng[†*], Chaogang Zhang[#], and Jan Schnupp[*]
[*] Hearing Research Group, Biomedical Science Department, City University of Hong Kong, Hong Kong SAR of China.
E-mail: drew.cappotto@cityu.edu.hk, wschnupp@cityu.edu.hk Tel: +852-3442-7551
[†] Acoustics Lab, School of Physics and Optoelectronics, South China University of Technology, Guangzhou, China.
E-mail: mengqinglin@scut.edu.cn Tel: +86-20-87113191
[#]KuGou Inc., Guangzhou, China.
E-mail: zhangchaogang@kugou.net

***Abstract*— The ability of cochlear implant (CI) users to accurately perceive and enjoy listening to music has remained unsatisfactory in a significant subset of users. Basic elements of music such as pitch, melody, and timbre that are easily discerned by normal-hearing (NH) listeners are difficult for CI users to perceive, owing to factors such as limited resolution of the devices' electrode array, audio processing that relies on coarse separation of the frequency spectrum into a limited number of overlapping bands, and temporal envelope extraction that discards the temporal fine structure. Alternative signal processing methods have been explored to enhance music enjoyment in CI users, with varying results, and most are computationally complex or require reprogramming of the audio processing device. This paper explores a new pre-processing technique to enhance music enjoyment in CI listeners through extraction and enhancement of the music's dominant melody, a technique we refer to as DoME (Dominant Melody Enhancement). In the described pilot studies, we employed a DoME technique of adding a frequency-modulated sine wave at the fundamental frequency (F0) of the music's dominant melody, and conducted perceptual experiments on nine CI users to gauge the effect of DoME on their music enjoyment. Initial results suggest DoME could increase some CI users' enjoyment of music listening.**

I. Introduction

Despite serving as powerful tools to restore functional hearing in users with severe or profound sensorineural hearing loss, modern cochlear implants (CIs) face significant hurdles in accurately representing complex acoustic signals [1]. In particular, deficiencies in the representation of rich harmonic sounds and frequency contours prevent CIs from accurately processing elements of acoustic signals that are important for our perception of musical sounds [2] [3]. These deficiencies result from limitations in the two main components of a CI system - the electrode array that is surgically implanted into the user's cochlea to stimulate the auditory nerve, and the external sound-processing unit that converts acoustic sounds into electrical signals. Surgical and clinical factors can further limit the effectiveness of the CI in manners that can vary from patient to patient. These include the depth at which the electrode is placed into the cochlea, possible trauma to the cochlea or auditory nerve before or during the procedure, and other physiological or pathological differences between patients. Auditory nerve stimulation is also limited by the number of electrodes on a given array; the most advanced arrays available today can provide up to 24 electrodes within each cochlea, compared to the thousands of hair cells which stimulate the auditory nerve in normal hearing (NH) individuals [4], and electrical "crosstalk" between adjacent electrodes on the array limits the number of independent electrode channels that can be achieved.

Current signal processing methods are primarily focused on speech intelligibility and have been proven to be largely successful under ideal conditions [5], even so far as providing functionally normal levels of speech development to prelingually deaf children. At its most basic, the audio processing separates the frequency spectrum into bands corresponding to the number of active electrodes, each handling slightly overlapping frequency ranges. The temporal envelope of the incoming signal in each frequency band is estimated and a train of electrical pulses of corresponding amplitude is delivered to the corresponding electrode(s) in an interleaved sampling. This method works effectively for processing speech, owing to our reliance on broad spectral "formant" patterns in discriminating human vocalizations [6]. However, this stimulation strategy encodes very little detail of either harmonic structure cues or temporal fine structure cues for musical pitch and timbre.

Researchers in recent years have explored methods of enhancing music enjoyment in CIs, primarily employing one of two approaches:

1) Reprogramming the processing strategy of CIs to better represent musical acoustic cues. For example, amplitude-modulating envelopes according to the fundamental frequency (*F0*) contour in order to enhance pitch perception [7] [8], remapping channel frequency allocation around semitones to improve the representation of harmonic relationships [9] [10], and presenting bass and treble parts of music separately to each ear to improve music clarity by taking advantage of the binaural input in bilateral CI users and [11].

2) Reengineering the music itself according to the sound coding characteristics of CIs [12-16]. This can be attempted either in the preprocessing stage of the CI processor or during music production. Most studies employing this type of method report positive results. The general agreement is that percussion, rhythm, and vocal cues are most preferred, and that reducing the spectral and musical complexity is a feasible approach to improving music enjoyment in some users.

The idea of preprocessing music for CI users has recently been explored in the context of multi-track re-mixing

- using the recordings of individual instruments, or groups of instruments, that comprise a final stereo audio file and allowing the user to adjust relative levels of each element in the mix. A basic implementation of this technique is to provide listeners with control over the level of the vocal arrangement separate from everything else in the song (commonly referred to as vocal and instrumental "stems"). A recent study [15] provided such control to CI listeners and found strong preference for weighting the vocal stem louder than the instrumental. This is somewhat unsurprising, given CI ability to more accurately represent speech patterns. Several studies have also observed CI preference for music with upper harmonics reduced by a standard low-pass filter [14] or via more robust methods of reducing spectral complexity in audio recordings [16]. This is similarly unsurprising, given that the limited "effective bandwidth" of CIs prevents an accurate representation of complex audio signals. Related research has shown a detriment to the identification of melodic sequences in the presence of accompanying instruments playing in a similar frequency range [13], further supporting the concept that CI cohorts tend to prefer instrumentally and harmonically simple music.

While the above research has shown some degree of enhanced music enjoyment, the methods employed may not be practical or feasible for the average CI listener, owing to the technical complexity and computational load required to implement such methods. Rather than subtracting elements to reduce harmonic complexity, or deconstructing the music into elements assumed to best translate to CI listeners, this paper explores a new preprocessing method by extracting and enhancing the dominant melody (DoME) of typical music recordings. The approach was motivated by an attempt to work within one particular CI limitation - temporal resolution in CI users declines significantly above 300Hz [17]. This is within the F0 range of the average male and female spoken voice, and within the average melodic range (212Hz-1.4 kHz) of the musical excerpts chosen for this study. Because sung vocal lines in most genres of contemporary western popular music tend to carry the dominant melodic contour, the weighted preference for vocal stems in re-mixing experiments might be understood as a product of comparatively reduced musical complexity (via the amplitude reduction of musically complex instrumental stems), ease of ability to follow a melodic contour when less competing elements are present, and music where the dominant melody falls within the more temporally-sensitive F0 range.

In order to further investigate these concepts, a first implementation of DoME is proposed in this pilot study. It enhances the F0 component of dominant melodic contour by adding a frequency-modulated sine consisting of only the melody's F0. A series of re-mixing and preference studies were devised in which CI listeners were given control over the amplitude of either stems or full stereo mixes, paired with or without the frequency-modulated F0 melody. If our assessment is correct, we would expect to see a similar weighting of vocal over instrumental stems, and a preference for the added F0 melody resulting from an effective re-weighting of the frequency spectrum in favor of the dominant melody's F0, thus increasing temporal accuracy and reducing harmonic complexity of the dominant melody.

II. METHODS

A. *Preparation of Stimuli*

Stimuli was sourced from MedleyDB, a database of multi-track music recordings with detailed metadata, pitch, melody and instrument annotations developed primarily for music information retrieval (MIR) research [18]. Extraction of the dominant melody followed MedleyDB's second pitch annotation method, "Melody 2". MedleyDB employed a semi-automatic method of dominant melody extraction. A modified version of the YIN [19] pitch-tracking algorithm, p-YIN, was employed across each song's multi-track stems in order to establish activation levels defining the dominant melody. These computationally derived annotations were then manually checked/amended and cross-validated to establish final annotations. Melodies from the dataset were provided as time-stamped numerical values at 5.8ms frames. Melosynth, a tool provided by the MedleyDB authors, was used to generate continuous-phase sine waves interpolated between frames to create a continuous wave. At the onset and offset portions, a 10 ms ramp was used to fade in and fade out. All audio files used in the Experiment 1 were kept at their amplitudes as provided in in the database, with the F0 melody rendered at 0dB full-scale. Audio files used in Experiments 2 and 3 were loudness normalized to -20LUFS based on the EBU R128 standard.

B. *DoME: Dominant Melody Enhancement Algorithm*

The pilot implementation of DoME consisted of mixing the extracted F0 melody with original music recordings at either user-configured or predefined ratios. Fig.1 gives spectrogram (top) and electrodogram (bottom) visualizations of DoME, where a full-scale F0 melody and music recording are combined at unity gains. The original music excerpt is pictured on the left, with the DoME processed version pictured on the right. Increased energy across the F0 melodic contour can be observed in the DoME spectrogram, along with temporally corresponding pulses in the two lowest frequency channels shown on the electrodogram. For envelope-based strategies, these stronger durations may enhance the perception of rhythm; for some fine-structure enhanced strategies, the temporal information in low frequency channels may also be enhanced.

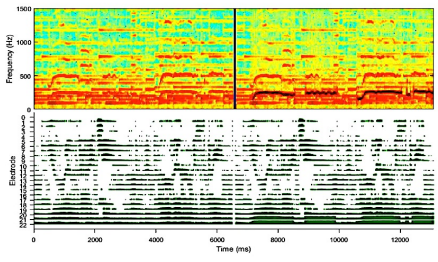

Fig.1 Spectrogram and electrodogram of pre/post DoME

Table I. CI user information

Subject	Gender	Age(yr)	CI Experience (yr)	CI Processor	Etiology	Speech Score	Experiment
C20	M	10	8	Right: Cochlear Freedom	Congenital	96.5%	1
C21	F	34	7	Right: Cochlear CP900	Drug-induced	93.5%	1, 2, & 3
C2	M	24	15	Left: Cochlear Kanso	Drug-induced	64.0%	1
C28	F	~30	10	Right: Cochlear Nucleus 5	Drug-induced	91.0%	1, 2, & 3
C25	F	38	6	Right: Cochlear CP802	Sudden deafness	92.0%	1
C23	F	29	3	Right: Nurotron NSP60B	Sudden deafness	88.0%	1
C30	F	23	1	Right: Cochlear Freedom	Sudden deafness	98.5%	2 & 3
C31	F	12	10	Left: Med-El OPUS 2	Congenital	94.5%	2 & 3
C35	M	15	11	Right: Cochlear SPrint	Congenital	99.0%	2 & 3

C. Music excerpt selection

As not all moments in a given piece of music contain a dominant melody, selection of the music excerpts was based on the shortest song section or musical phrase to be completed within a ~30-60 second clip with minimal interruption to the F0 melody. Songs were selected to represent a variety of different genres and musical styles and tempos, based on the metadata tagging provided in the database. As such, choruses from mid and up-tempo songs were preferred, in that they allowed for a higher number of F0 melodic tones with the minimal amount of time needed for a musical phrase to complete. In many excerpts, the dominant melody is shared across multiple instruments. Melody 2 annotations, described in the previous section, allowed us to better gauge the effects of DoME over an entire piece of music rather than for a specific instrument or vocal phrase.

D. Subjects and their CI processors

Nine CI subjects participated in this study (see Table 1.). All were compensated for their participation and all provided informed consent in accordance with the local institutional review board. The origin of deafness, type of CI, and years of CI experience varied widely between subjects. Six subjects participated in Experiment 1 and five participated in Experiments 2 and 3. Prior to the experiment, participants' speech recognition scores were measured using the Mandarin hearing in noise (MHINT) database.

E. Experiment 1: Multi-track scaling with a single excerpt

In our initial pilot study, six CI subjects participated in a remixing and informal preference experiment. One song excerpt was chosen from MedleyDB; participants used their default CI settings throughout the experiment, and stimuli were delivered via a pair of Bose Companion 20 loudspeakers placed at both sides of a laptop. Audio stems were separated into stereo audio files containing the vocal and instrumental groups as individual sub-mixes and, along with the synthesized F0 sine wave, were presented to participants on different tracks of a digital audio workstation (Adobe Audition) mixer interface. Participants were given individual control over the playback volume of each track via the mixer interface and instructed to adjust the values to their tastes. Labels were not provided to the participants as to what content was on each track of the mixer. Participants were asked to 1). Scale the volume of the vocal and instrumental track (without DoME) and 2). Scale the volume of all three tracks. The objective of Task 1 was to observe the relative preference of the vocal and instrumental tracks, while the goal of Task 2 was estimate the effects of DoME. Upon conclusion of the participants' level adjustment, the Investigator tested their response to set values in order to confirm their stated mix preference.

F. Experiment 2: Select a preferable Sine-volume with 17 excerpts

Five subjects participated in Experiments 2 and 3. Each of the 17 music excerpts and F0 melodies were presented with a Play button and a slider interface, allowing the participants to mix in desired the amount of F0 melody before moving on to the next song. The slider was not labeled, and participants were instructed to adjust the slider until the music sounded most pleasant to them. The avoid user preference, the functional direction of the slider was randomized to the left or right for each song, such that the direction required to increase the amplitude of the F0 melody was not the same for every trial. The default amplitude of the F0 melody was -6dB below -20LUFS at the middle position of the slider.

G. Experiment 3: A/B forced choice with 17 excerpts

Upon conclusion of Experiment 2, the same five subjects were given a forced A/B test, in which they were presented with two sound stimuli and told to choose between either option A or option B. Three F0 melody amplitude values were used for comparison against the original stereo mixes: 0dB (matched loudness at -20LUFS),

-6dB and -12dB below -20LUFS. Button positions for the unaltered mix and F0 melody alternated at random with each trial, as did the order in which the comparison F0 values were presented.

III. RESULTS

A. Experiment 1

For task 1 (i.e., weighting of instrumental and vocal stems), no apparent preference was found when compared to original mix. For task 2, five of the six subjects showed some preference (-21.1dB - -6.2dB) for the addition of the F0 melody (see Table II).

B. Experiment 2

Despite a large individual variance for each excerpt (see Table III), the median value for the addition of the F0 melody for each subject was from -9.6dB to -4.3dB below -20LUFS.

C. Experiment 3

In this A/B test (i.e., a two-alternative forced choice test) 17 excerpts, the total F0 melody preferences have been summed per-subject, as listed in Table IV. We compared observed choices against the null hypothesis that, if a subject were to choose randomly, they would identify the DoME result on half of the trials on average, and the number of DoME choices observed would follow a binomial distribution were n = 17 and p = 1/2. The probability of observing 12 or more DoME choices out of 17 by chance is then as small as 0.05. The probability of observing 5 or less DoME choices out of 17 by chance is also as small as 0.05. Consequently, subjects can be said to significantly prefer the DoME processed sound if their scores are at least 12 of the 17 trials (70.6%) and can be said significantly prefer the original signal if their scores are at most 5 of the 17 trials (29.4%).

The results showed that C28 significantly preferred the addition of the matched loudness (0dB) F0 melody in Experiment 3, but indicated a strong preference for the original signal in Experiment 1. C21 and C31 significantly preferred the addition of the -12dB F0 melody. C31 significantly preferred the original mix to both the -6dB and 0dB F0 melody. C30 significantly preferred original signal to the -6dB F0 melody. For other subject-DoME combination conditions, no significant preference was found.

D. Discussion, observation, and outlook

In this pilot study, we proposed a music pre-processing method to enhance the dominant melody in a piece of music, a technique we've dubbed DoME. The first implementation and its pilot experiment were presented above. In this implementation, the F0 component of the music's dominant melody was enhanced by adding a pitch-tracked frequency-modulated sine wave in parallel to the original signal. The pilot experiments showed some benefits for some patients. However, the results were inconsistent.

Limitations of this technique may result from the highly subjective perception of music enjoyability in CI users, the lack of normal hearing control group due to time limitations, and large variances among the hearing condition of the CI subjects. Different approaches in stimuli preparation between Experiments 1 and 2/3 also introduce complications when comparing results from each experiment. Many studies have shown that prelingually deafened early-implanted children, prelingually deafened late-implanted adult and postlingually deafened adult have very different perception ability and motivation both of which have significant effects on their preference reports [20].

Some observations during the experiments were conducted informally. For example, subjects C2 and C25 in Experiment 1 and C28 in Experiment 3 told the experimenter that they were very sure about their preference for the addition of the sine-wave melody, though empirical data was not collected to support those statements. C28 had recently attended weekly piano lessons, which may have had an effect on strong lack of preference for DoME in Experiment 1, though contrastingly showed significant preference for DoME in Experiment 3. All subjects were users of the Cochlear product line, which limited our scope and possible insights into the efficacy of DoME across the range of possible CI users. C31, a child using the Mel-El product showed significant lack of preference for the stronger F0 melody levels in Experiment 3 (-6dB and 0dB) but showed significant preference for or against the weaker F0 melody level (-12dB) compared with original music.

Despite the limitations and variables present in our initial pilot studies, DoME shows promise as a computationally light method of enhancing music enjoyment in some CI users. In the future, we plan to continue development of the DoME approach and refine our experimental designs for a set group of CI patients and cohort of normal hearing peers. Additional implementations of the DoME approach beyond the simple addition of the F0 melody will also be explored.

Table III Results of Experiment 2 (Unit: dB, 0dB = -20LUFS)

	C35	C28	C21	C30	C31
1	-19.2	0.0	-4.3	-19.8	0.0
2	-∞	0.0	-0.9	-5.4	-∞
3	-0.7	-13.7	-16.9	-5.8	-0.5
4	-0.6	-44.3	-16.3	-1.1	-11.5
5	-1.4	-17.7	-15.9	-5.9	-1.3
6	-2.9	-6.0	-1.3	-18.3	-16.4
7	-13.2	0.0	-7.8	-1.4	-18.1
8	-∞	-∞	-15.1	-1.4	-25.0
9	-17.7	0.0	-0.6	-6.0	-∞
10	-11.8	0.0	-2.7	-22.2	-3.2
11	0	-6.1	-3.1	-31.4	-7.1
12	-12.4	-5.9	-2.6	-0.6	-19.7
13	-9.6	-∞	-10.8	-21.4	-13.1
14	-2.1	-6.2	-2.8	-32.9	-6.1
15	-11.9	0.0	-11.2	-6.2	-4.0
16	-0.4	-∞	-2.6	-0.3	-4.4
17	-4.6	-6.1	-14.2	-0.2	-6.2
Median	-9.6	-6.1	-4.3	-5.9	-7.1

Table IV Results of Experiment 3

Subject	-12 dB	-6dB	0dB
C35	8	6	9
C28	8	11	**13***
C21	**14***	9	10
C30	11	**5**$^{\#}$	9
C31	**12***	**5**$^{\#}$	**4**$^{\#}$

IV. CONCLUSIONS

Dominant melody enhancement through a simple addition of a frequency-modulated sine wave tracked from the dominant melody's F0s may enhance some CI patients' enjoyment of music. Future music perception and enhancement research should pay careful attention to experiment design; a single group of CI patient types (i.e., prelingually deafened early-implanted children, prelingually deafened late-implanted adult, or postlingually deafened adult) is suggested to be the focus, rather than cohorts with a wide range of hearing loss etymology and years of CI experience.

ACKNOWLEDGMENT

This work is jointly supported by NSF of China (Grant No. 11704129 and 61771320), the Fundamental Research Funds for the Central Universities (SCUT), State Key Laboratory of Subtropical Building Science (SCUT, Grant No. 2018ZB23), and Shenzhen Science and Innovation Funds (JCYJ 20170302145906843). Qinglin Meng and Jan Schnupp are corresponding authors.

REFERENCES

[1] Svirsky, M. (2017). Cochlear implants and electronic hearing. *Physics Today*, 70(8), pp.52-58.

[2] Limb, C. and Roy, A. (2014). Technological, biological, and acoustical constraints to music perception in cochlear implant users. *Hearing Research*, 308, pp.13- 26.

[3] Zeng, F., Popper, A., Fay, R. and McDermott, H. (2011). *Auditory Prostheses.* New York, NY: Springer New York, pp.305-339.

[4] Zeng, F., et al. (2015). Development and evaluation of the Nurotron 26-electrode cochlear implant system. *Hearing Research*, 322, pp.188-199.

[5] Zeng, Fan-Gang, et al. "Cochlear implants: system design, integration, and evaluation." *IEEE reviews in biomedical engineering* 1 (2008): 115-142.

[6] Shannon, Robert V., et al. "Speech recognition with primarily temporal cues." *Science* 270.5234 (1995): 303-304.

[7] Laneau, Johan, Jan Wouters, and Marc Moonen. "Improved music perception with explicit pitch coding in cochlear implants." *Audiology and Neurotology* 11.1 (2006): 38-52.

[8] Milczynski, Matthias, Jan Wouters, and Astrid Van Wieringen. "Improved fundamental frequency coding in cochlear implant signal processing." *The Journal of the Acoustical Society of America* 125.4 (2009): 2260-2271.

[9] Kasturi, Kalyan, and Philipos C. Loizou. "Effect of filter spacing on melody recognition: acoustic and electric hearing." *The Journal of the Acoustical Society of America* 122.2 (2007): EL29-EL34.

[10] Omran, Sherif Abdellatif, et al. "Semitone frequency mapping to improve music representation for nucleus cochlear implants." *EURASIP Journal on Audio, Speech, and Music Processing* 2011.1 (2011): 2.

[11] Vannson, Nicolas, Hamish Innes-Brown, and Jeremy Marozeau. "Dichotic Listening Can Improve Perceived Clarity of Music in Cochlear Implant Users." *Trends in hearing* 19 (2015): 2331216515598971.

[12] Buyens, Wim, et al. "Music mixing preferences of cochlear implant recipients: A pilot study." *International journal of audiology* 53.5 (2014): 294-301.

[13] Kohlberg, Gavriel D., et al. "Music engineering as a novel strategy for enhancing music enjoyment in the cochlear implant recipient." *Behavioural neurology* 2015 (2015).

[14] Nemer, John S., et al. "Reduction of the harmonic series influences musical enjoyment with cochlear implants." *Otology & neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology* 38.1 (2017): 31.

[15] Pons, J., Janer, J., Rode, T. and Nogueira, W. (2016). Remixing music using source separation algorithms to improve the musical experience of cochlear implant users. *The Journal of the Acoustical Society of America*, 140(6), pp.4338-4349.

[16] Nagathil, A., Weihs, C., Neumann, K. and Martin, R. (2017). Spectral complexity reduction of music signals based on frequency-domain reduced-rank approximations: An evaluation with cochlear implant listeners. *The Journal of the Acoustical Society of America*, 142(3), pp.1219-1228.

[17] Zeng, F. (2002). Temporal pitch in electric hearing. *Hearing Research*, 174(1-2), pp.101-106.

[18] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research", in *15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[19] De Cheveigné, Alain, and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music."*The Journal of the Acoustical Society of America* 111.4 (2002): 1917-1930.

[20] Mitani, Chisato, et al. "Music recognition, music listening, and word recognition by deaf children with cochlear implants." *Ear and Hearing* 28.2 (2007): 29S-33S.