P4 REPORT:

Evaluation

Out of The Blox

Fang He

Manasvi Lalwani

Regis Verdin

Table of Contents

Introduction……..……..……..... 3

Wristband Evaluation…….. 3

Techniques, Tasks and Users…….…3

Evaluation Design Rationale……..….4

Description of Results…….....6

Design Implications……..…….12

Potential Design Improvements……15

Computer Interface Evaluation ……15

Techniques, Tasks and Users……15

Evaluation Design Rationale…….17

Description of Results……..…….18

Design Implications……..…….19

Potential Design Improvements……...……19

Future Work and Conclusions…….……...……...20

Introduction

BumbleBeat is a system that helps deaf and hard of hearing people dance along to music. We previously prototyped two parts of this system: a wristband that gives vibrotactile feedback to indicate musical rhythm, and a series of screenshots of a computer interface for converting song files into a series of vibrations that the wristband could play. Our group has performed expert reviews and user evaluations of these two prototypes. We analyzed and reviewed the data from our testing to decide what future iterations of the prototyping stage would involve.

Wristband Evaluation

Techniques, Tasks and Users

Techniques and Tasks:

Timer Test:

To test whether users can feel the music and whether our wristband can help deaf people detect the beat precisely, we designed a timer test. For this test, users will be asked to wear wristband and record time that they have felt vibration. And after that, we will have a list of time that shows when the subjects have detected the vibration generated by wristband, and can compare it to benchmark. The benchmark is the accurate beat of music.

The detailed steps are as follows. First, test subjects will be asked to wear the wristband prototype, and they can adjust it so that it is comfortable. Secondly, we will start the song track (silent) and the wristband will vibrate. Users will be asked to press space key as soon as possible after they have felt the vibration of wristband and the time

they press key will be recorded by MaxMSP which is a visual programming language for media. After that, we will check the deviation of pressing the space key with benchmark and see whether our prototype has enhanced people's feeling of music by vibration and what aspects of our prototype should be ameliorated.

Survey:

After the timer test, there will be a survey to test users' feedback concerning usability of the wristband prototype. There will be 7 likert scale questions, 4 free-response questions and 1 yes/no questions. And after this survey, test subjects are free to propose any problems they found during process of test and give us suggestions to complete our prototype.

Users:

We have recruited 5 students from MS-HCI to participate in the timer test and post survey. Regretfully, we were not able to recruit deaf people to participate in our test. So we recruit people with normal hearing to use our prototype and give us their feedback. To diminish the natural hearing ability difference between hearing subjects and deaf subjects, we don't give our subjects any hearing indication of music during the whole process of test, and the only information they can get from the music is the vibrations that are conveyed by the wristband.

Evaluation Design Rationale

Timer test:

The basic rationale for timer test is that test subjects will be given the access to press space key when the wristband prototype has started vibrate. So we can acquire a list of time that test subjects have felt vibrate of wristband representing music beat. And after that, we can compare the time list created by each test subjects with benchmark and

calculate deviation time for each key pressing. By taking subjects' response time into account, we can know the difference between subjects data and benchmark. Also, the average deviation time of key pressing will be calculated after the test and ratio that a test subject has successfully pressed space key after he or she has felt vibration of wristband will be counted for each test.

Survey:

The post survey after timer test is a chance for us to know about subjects real feeling after using the prototype and figure out potential usability problem. Survey questions are listed as following:

7 point likert scale questions:

1. The product was enjoyable to use.
2. I would want to wear this product while dancing.
3. I would feel comfortable wearing this product in public
4. The vibration was easy to feel.
5. I would use this product often.
6. The product is easy to use.
7. The product was comfortable to wear.

The 7 point likert scale is mainly used for test subjects' attitude toward a specific aspect when they use the wristband prototype. And we can specifically improve one part of our design based on the information we get from likert scale.

4 free-response questions:

1. What did you like about the device?
2. What did you dislike about the device?
3. How did the band hinder your movement? (Answer only if you answered yes to the previous Question)
4. Do you have any other comments or suggestions?

The 4 free-response questions give test subjects a chance to freely express their general attitude toward our design and they can propose problems that they found and improvement suggestions for this prototype.

1 yes/no question:

Did the band hinder your movement in any way?

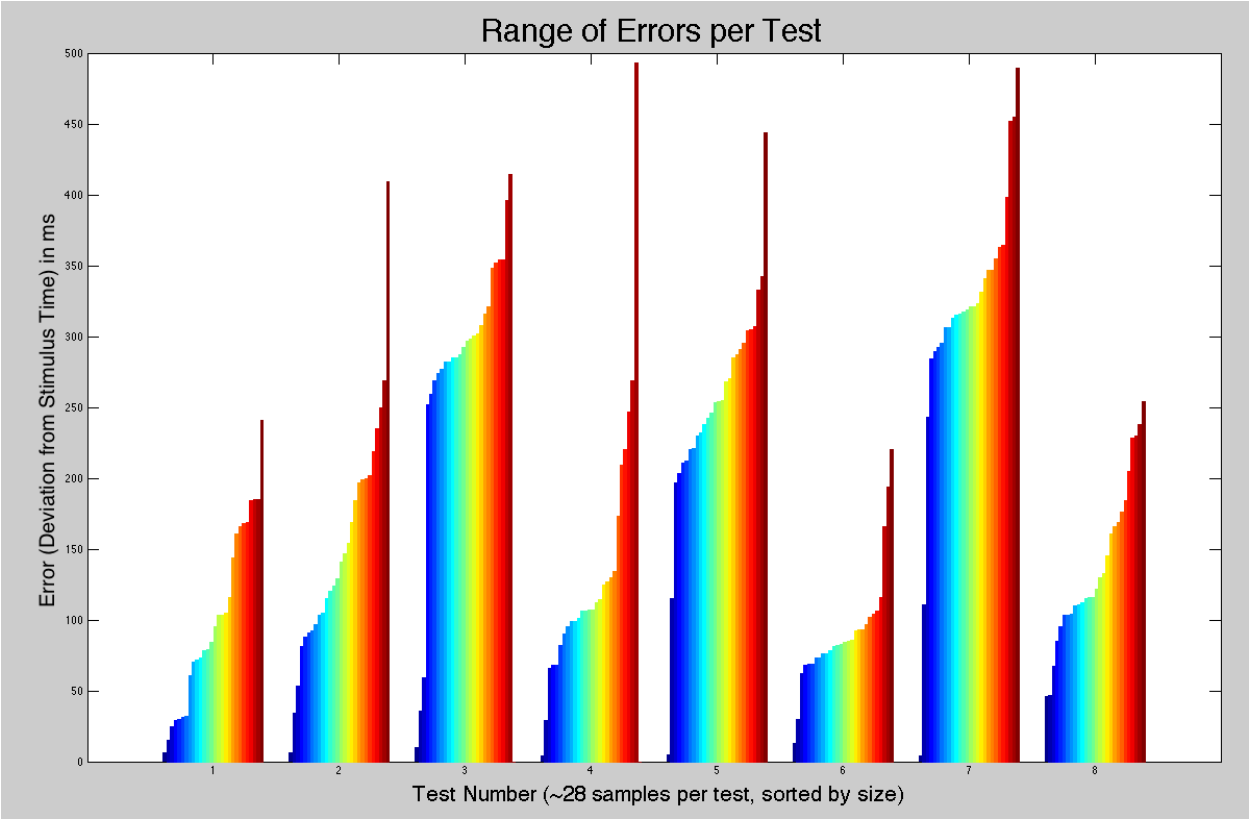
For this question, we specifically want to know whether the wearable prototype is comfortable to wear and will it impact activity of human in real environment. We think this is an essential part for all wearable product, especially for wearable product like ours which intend to help users finish a specific task (dance) easier. Therefore, if test subject think this prototype does hinder movement, we may have to improve our prototype from aspects such as wearable comfortability, wear place of prototype, size of prototype and so forth.

Description of Results: Timer Test

Our timer-based wristband test mainly sought to find if the feedback from the wristband was easy and fast enough to perceive for cueing the user to move in time. Our findings were fairly positive, indicating that the users were, for the most part, able to react to the vibrations within an acceptable amount of time. We defined this amount of time as half of the shortest time-interval between vibrations. Our reason for this definition: if the user were to react that late it would signal a large subdivision of the beat, which would create a rhythmically confusing dance movement if unintentional.

Our results came in the form of two lists of time values: the actual vibration time of the wristband, and the time the subject hit the spacebar. The start times of these lists were synchronized, and so

subtracting the actual times from the subject times yielded the subject's error for each beat. This is presented below:

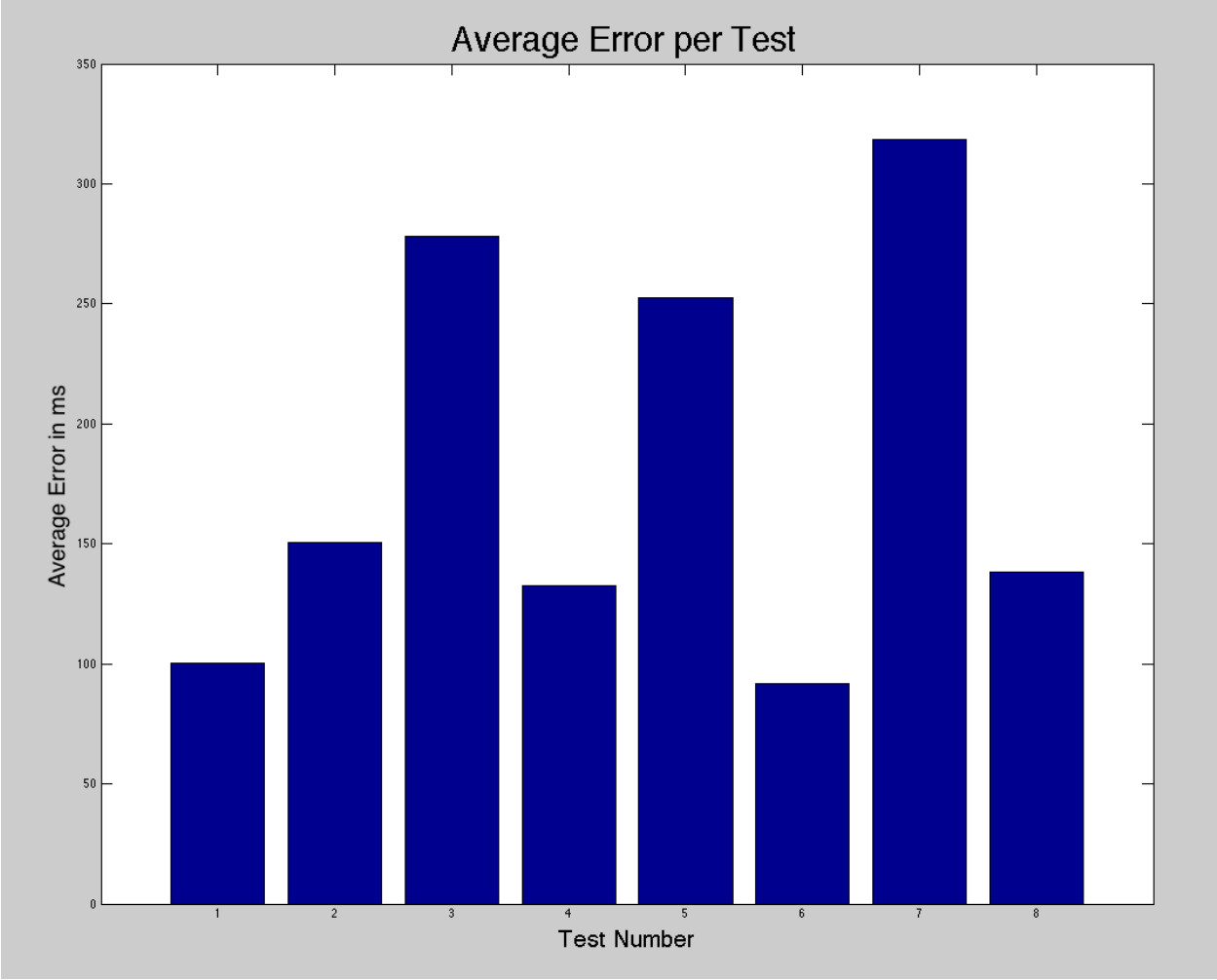


Here, we have 8 tests (2 per subject), and approximately 28 beats in each test that the subject had to identify. The height of each bar shows the subject's delay in ms, sorted from low to high for each test.

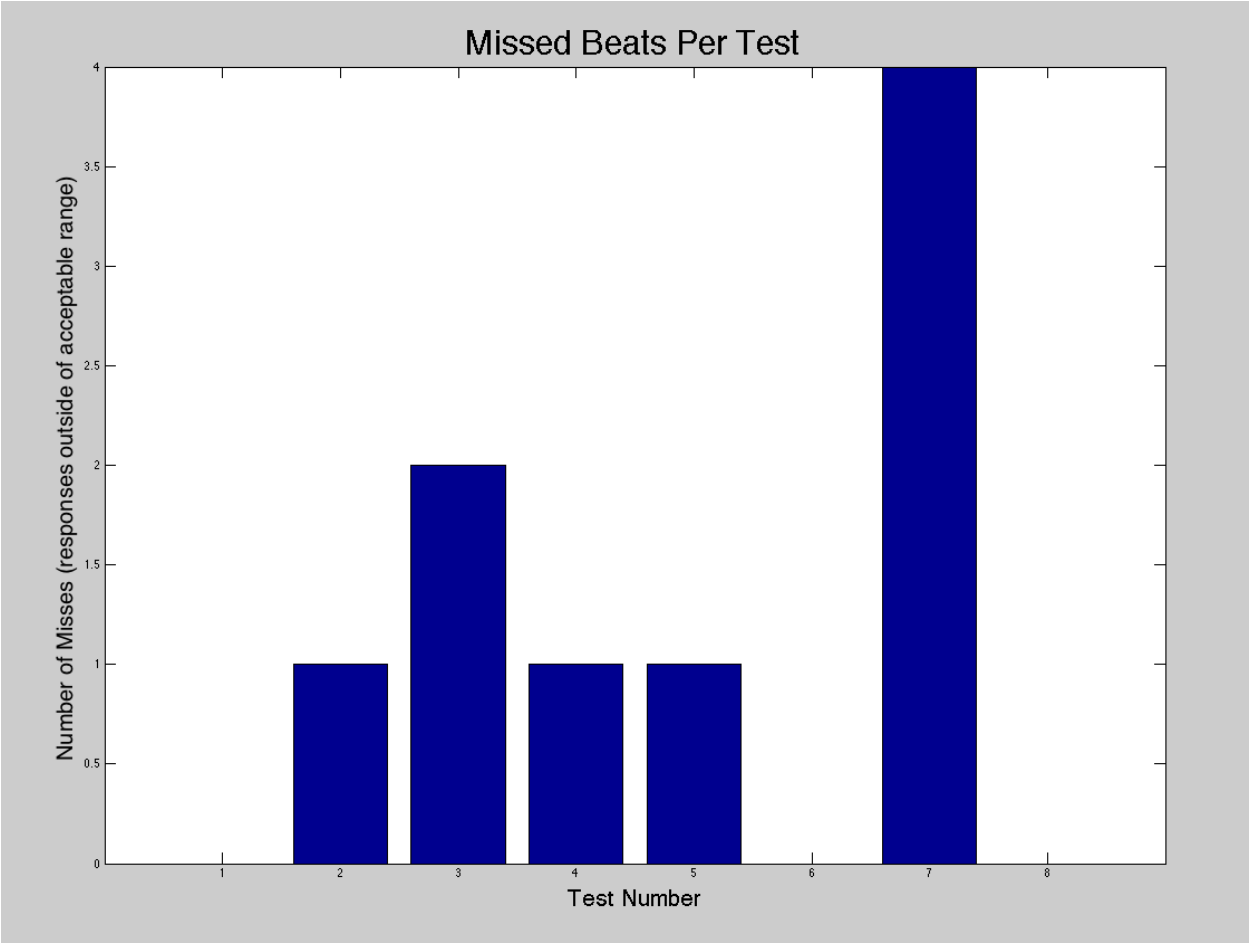
The minimum delay time was 4ms in test 7 (subject 4), which seems remarkably low for a reaction time. One possible reason for this super-human reaction time is that the test induced a sense of pulse in the user, and they were able to predict when the next beat would occur. This is a fairly positive finding, as we hoped that we could induce a sense of rhythm in this way with the wristband. This is an important feature for the products usability. Most of the experiments had several low delay-times like this, which is encouraging considering we varied

the rhythm after several beats. They were able to get a feel for the pulse very quickly.

The maximum delay time was 493ms in test 4 (subject 2), and we see spikes like this in most of the other tests. The distribution of values within tests is fairly condensed towards their mean. There are much fewer spikes and low values, and these middle values seem to indicate a more realistic reaction time without the subject using prediction. The range of subject errors is also of interest, with a mean of 358.75 between the within-tests ranges. Taking the mean of each test's values, we get the following graph:



We were also interested in how many of the subjects' responses were in an acceptable range of error. Part of this was for data analysis: we needed a threshold within we could unambiguously say that the subject had hit the spacebar with the intention of corresponding to a certain vibration. The other reason was that we felt that responses that were too late could not be considered positive results. The chart below shows the number of "missed" responses per test.



The subjects did fairly well here, with an average of 1.125 misses per test, out of 28 total beats in each test. This shows that they were able to hit the spacebar in an acceptable amount of time 96% of the time.

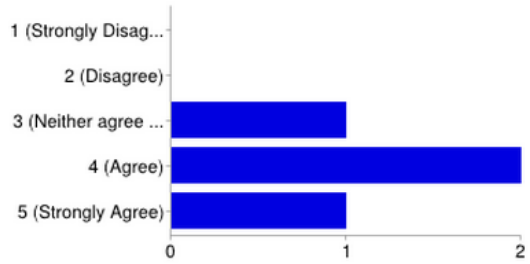
Another issue we thought of is if a subject were to hit the spacebar more than once in between two vibrations. Thankfully this didn't occur,

but if it had we would have probably removed any hits after the first (in between the vibrations).

Description of Results: Survey

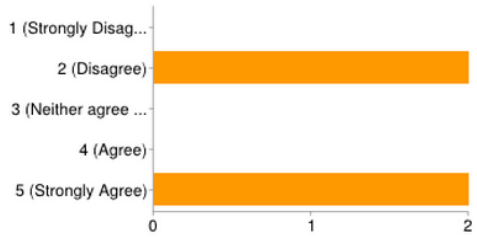
Our survey consisted of 7 Likert-scale questions, 4 free-response questions, and one yes/no question. Results of the Likert questions were as follows:

1. The product was enjoyable to use. [Rate the following statements based on how much you agree with them.]



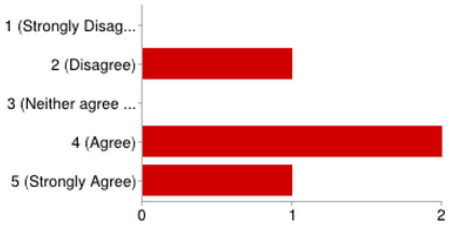
1 (Strongly Disagree)	0	0%
2 (Disagree)	0	0%
3 (Neither agree nor disagree)	1	25%
4 (Agree)	2	50%
5 (Strongly Agree)	1	25%

2. I would want to wear this product while dancing. [Rate the following statements based on how much you agree with them.]



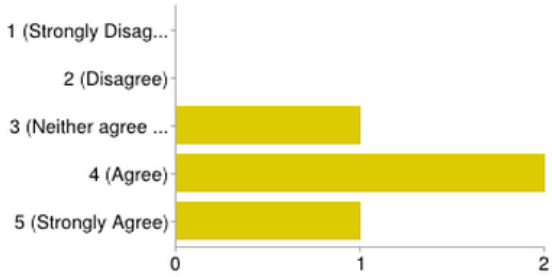
1 (Strongly Disagree)	0	0%
2 (Disagree)	2	50%
3 (Neither agree nor disagree)	0	0%
4 (Agree)	0	0%
5 (Strongly Agree)	2	50%

3. I would feel comfortable wearing this product in public. [Rate the following statements based on how much you agree with them.]



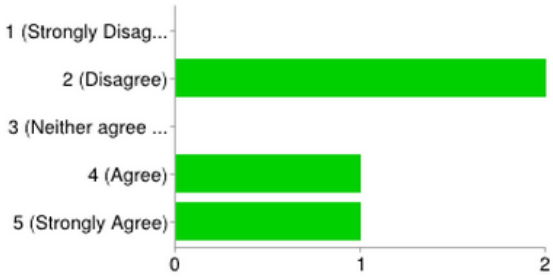
1 (Strongly Disagree)	0	0%
2 (Disagree)	1	25%
3 (Neither agree nor disagree)	0	0%
4 (Agree)	2	50%
5 (Strongly Agree)	1	25%

4. The vibration was easy to feel. [Rate the following statements based on how much you agree with them.]



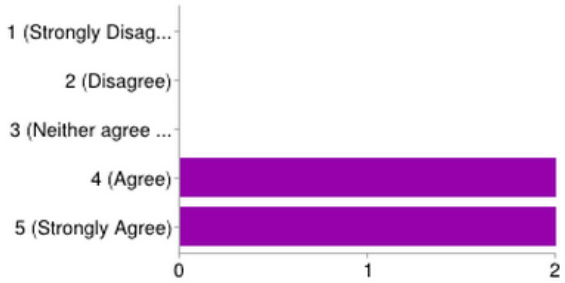
1 (Strongly Disagree)	0	0%
2 (Disagree)	0	0%
3 (Neither agree nor disagree)	1	25%
4 (Agree)	2	50%
5 (Strongly Agree)	1	25%

5. I would use this product often. [Rate the following statements based on how much you agree with them.]



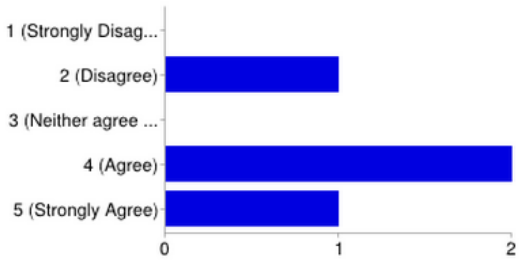
1 (Strongly Disagree)	0	0%
2 (Disagree)	2	50%
3 (Neither agree nor disagree)	0	0%
4 (Agree)	1	25%
5 (Strongly Agree)	1	25%

6. The product is easy to use. [Rate the following statements based on how much you agree with them.]



1 (Strongly Disagree)	0	0%
2 (Disagree)	0	0%
3 (Neither agree nor disagree)	0	0%
4 (Agree)	2	50%
5 (Strongly Agree)	2	50%

7. The product was comfortable to wear. [Rate the following statements based on how much you agree with them.]



1 (Strongly Disagree)	0	0%
2 (Disagree)	1	25%
3 (Neither agree nor disagree)	0	0%
4 (Agree)	2	50%
5 (Strongly Agree)	1	25%

The Likert-scale questions leaned towards the agreeing end, which is ideal as our statements were framed in a positive way about the product. Our most positive results were from the following prompts “The product was enjoyable to use”, “The vibration was easy to feel”, and

"The product is easy to use". The second question was especially important to use, as the prototype was largely made to test this feature. The main functionality of the product depends upon the feedback being easy to perceive. While these results were fairly positive, we suspect that if we hadn't proxied our users with fellow students, they would have been a bit more negative.

The free-response questions mostly confirmed the likert-scale ratings, and gave us additional interesting feedback. The aesthetic appeal of our design was criticized, which we expected as we didn't focus on that element for this prototype. The wristband was also too tight for some users, which we should have found a solution for prior to the evaluation (knowing that we would probably have users with different sized hands and wrists). One of the most interesting points was that it was difficult for the user every time the rhythm changed, and they felt their dancing would be negatively affected by this. The issue was that the feedback happens at the moment when the dancer needs to respond, and not slightly before. One respondent suggested giving a "heads up" before the music is about to change, which could be accomplished with the lights.

Many of the other suggestions made were aspects of the wristband that we had planned to implement at a later iteration of prototyping. These include: stronger vibration (more motors), an interface on the wristband for switching songs, and not requiring being plugged into a computer (we hadn't yet installed a battery on the wristband, so it was bus-powered).

Design Implications from Evaluation Results

Based on the timer-test, we concluded that the reaction time was acceptable when using vibrotactile feedback, but that it is difficult to follow when the rhythm changes. This is largely the result of our test song, and for future evaluations we would adjust the song so that it would contain more typical rhythmic features.

We could decrease the reaction time in several ways. One is simply by adding more motors to increase the overall intensity of stimulus. Another is to look carefully into the type of motors, as the one we used was slow to begin vibrating. It had to "ramp up", and the time in which it did so was very noticeable (around 200ms). Perhaps there is another kind of vibration motor or actuator that can deliver a more discrete pulse, rather than a continuous vibration that we simply truncate.

Based on the open-ended questions in our post-test questionnaire we were able to get user-feedback on the design of the band. We categorized the feedback into positive and negative feedback as follows:

Success of the design:

> Product is light and comfortable.
>
> Feedback from the band is easy to perceive, without being too annoying, and easy to understand i.e. low learning curve for the most part.

Failures of the design:

> The aesthetics of the band could be improved upon.
>
> The wrist band felt tight.
>
> The band could not convey when the beat was about to change.
>
> It was a little difficult to feel the vibrations.
>
> The band restricted my movement because it was connected to a laptop.

Additional feedback:

> A possible improvement would be to vary the strength of the vibration with the beat. If it is going to be faster, the vibrations could be weaker and short spanned and if it is going to be slower, they could be stronger and long spanned. A very high fidelity

prototype would probably allow users to select the song they like from a playlist.

It might be worthwhile exploring if the vibrating motor should be place on the bottom of the arm/hand (as opposed to close to the underside of the wrist).

On observing the feedback above one can see that while some of our users found the band comfortable, others found it too tight.

The implication of this is that we should have a band that is able to adapt to different wrist sizes. This could be like a watch or something else altogether.

More than one user reported that while they could feel the vibrations, the change in the rhythm of the vibrations was currently not conveyed by the band.

Implication- The above information would have to be encoded in some way in the next iteration of the prototype. There are musical characteristics in a song which are able to convey this and our design would have to provide a similar cue in some manner.

While some users were able to perceive the vibrations of the band easily, some others would have liked for it to be stronger.

This implies that the next iteration of our band would have to provide a way to adjust the strength of the vibration according to individual preference.

One of the users reported that the band restricted their movement by virtue of being connected to the laptop. This was because the prototype lacked an onboard battery (which we were unable to procure). A future iteration should include an onboard battery so the band can operate without the need to be plugged in.

Potential Design Improvements

Future Work:

An onboard display for users to be able to see the song on the band and to be able to navigate back and forth between them.

A possible way to convey information about beat changes.

A watch like form factor that can adapt to a wider variety of wrist sizes.

Means of controlling the intensity of the vibration provided by the band.

Consideration of other means of stimulus, such as an actuator that taps the wrist, or a vibration motor that takes less time to reach its maximum vibration level.

It might be useful to conduct our study with the actual target population instead of proxying it to be able to receive more reliable feedback on its usefulness and social acceptability.

It might be useful to test the band by asking users to dance aided by the band. It will lend validity to the usefulness of our product(or not, as the case maybe).

Computer Interface Evaluation

Techniques, Tasks, and Users

Overview of Techniques:

We employed the heuristic evaluation usability method to evaluate the Bumble Beat software interface. The test was conducted based on heuristics provided by Jakob Nielsen.
We gathered 4 MS-HCI students posing as expert evaluators and asked them to perform certain tasks and uncover flaws in the design of our interface based on the following pre-defined nine heuristics set forth by Jakob Nielsen:

1. Use Aesthetic and Minimalist Design
2. Effective Menu/Command Structure
3. Use Simple and Natural Language
4. Minimize User's Memory Load
5. Be Consistent
6. Provide Feedback
7. Provide Clearly Marked Exits
8. Deal with Errors in a Positive Manner
9. Provide Help

Tasks and methodology:

Our evaluators, most of whom were HCI students, were gathered in the same room at the same time. They were given an overview of the system and our target users (deaf and hard of hearing people.) They were then asked to perform the following tasks:

Import songs to the 90's Music Playlist
Convert the songs to extract beats in the music and upload the band readable beat information file to the band.
Check if songs had been uploaded to My Bumble Beat 1

At the conclusion of their tasks they were asked to report their problems on a common document. We then went over each Evaluator's problems and put them in one table (overlapping problems were only reported once). The experts then discussed the problems and reached consensus on the severity rating of each and the ease of fixing each. The following scale was used for this process:

Severity Ranking

0-Violates a heuristic but doesn't seem to be a usability problem.
1-Superficial usability problem: may be easily overcome by user or occurs extremely
infrequently. Does not need to be fixed for next release unless extra time is available.
2-Minor usability problem: may occur more frequently or be more difficult to overcome.
Fixing this should be given low priority for next release.
3-Major usability problem: occurs frequently and persistently or users may be unable or unaware of how to fix the problem.
Important to fix, so should be given high priority.
4-Usability catastrophe: Seriously impairs use of product and cannot be overcome by users. Imperative to fix this before product can be released.

Ease of fixing Ranking:

0-Problem would be extremely easy to fix. Could be completed by one team member before next release.
1-Problem would be easy to fix. Involves specific interface elements and solution is clear.
2-Problem would require some effort to fix. Involves multiple aspects of the interface or
would require team of developers to implement changes before next release or solution is not clear.
3-Usability problem would be difficult to fix. Requires concentrated development effort to finish before next release, involves multiple aspects of interface. Solution may not be immediately obvious or may be disputed.

Evaluation Design Rationale:

Our prototype of the computer interface was in the form of a series of screen mockups at different states of the program. Certain areas of each screen were made clickable, and linked to other screen shots. The

effect was that the user could be asked to perform a task, and the correct path could be clicked through. We chose a heuristic evaluation with semi-expert reviewers because these types of evaluation are often useful early in the design process. It seemed appropriate to enlist others with HCI expertise to look over our interface, and identify any glaring issues before bringing it to our deaf and hard of hearing users. We used severity rankings based on Jakob Nielsen's use of these in his description of heuristic evaluation.

Description of Results:

Problem	Severity Ranking	Ease of Fixing Ranking	Heuristic Number	Broad Heuristic
Icons are not clear, labels needed as well(reported for the Import Convert Upload buttons at the top-center of the interface)	2	0	2, 3, 4	Effective Menu/Command Structure Use Simple and Natural Language Minimize user's Memory Load
Feedback needs to be provided when icon is clicked (reported for the Import Convert Upload buttons at the top-center of the interface)	2	0	6	Provide Feedback
Icon sizes are not consistent (reported for the Import Convert Upload buttons at the top-center of the interface)	1	0	5,2	Effective Menu/Command Structure Be Consistent
Show systems status at the top about convert and upload	2	1	6	Provide Feedback
Shows songs on device only if they have been uploaded	1	0	6	Provide Feedback

Visibility of my device panel could be improved	1	0	2	Effective Menu/Command Structure

The evaluators also noted that overall the interface looked minimalistic and aesthetic. They liked that it bore familiarity to iTunes.

Design Implications from Evaluation Results

Based on our evaluation the we derived a list of 3 superficial usability problems and 3 minor usability problems. All of these were reported as easy to fix.
Superficial problems:

Icon sizes are not consistent
Does not show songs on device only if they have been uploaded
Visibility of my device panel could be improved

3 minor problems are:

Icons are not clear, labels needed as well
Feedback needs to be provided when icon is clicked
Does not show overall system status about convert and upload

Potential Design Improvements

With these problems identified in the evaluation, we had a clear list of changes to make in future iterations of our design, some of which we made prior to our poster session.

Consistency of icon size is more important than we anticipated, and helps build a sense of trust in the user about the functionality of the system. It was a mistake in our design that occurred between screenshots, and was a simple fix involving resizing.
Not showing songs on the device falls under the category of providing the user with feedback to their actions. We corrected this by showing all of the tracks in the device immediately, and showing their individual upload statuses on the track.
For the visibility of the "my device" panel, we discussed highlighting this in some way, either spatially or by adding a prominent border to the panel. Currently it is in the bottom left corner of the screen, but perhaps it could be moved up with the playlists, as in the iTunes interface. The inconsistency with iTunes (which our prototype otherwise resembles) could be contributing to this confusion.
We added labels to the three main icons (import, convert, upload) for our poster, to improve clarity about their functionality.

Future Work and Conclusions

Future Work

For the future work, we will consider the following aspect that we learned from conducting our evaluations.

First of all, we should continue testing users mental model for using this kind of prototype which applies synaesthesia method, that is using tactile to replace auditory senses to guide deaf people finishing a task. We found there's many components and aspects of music that haven't been embodied in our prototype design. For example, users of wristband do not only need to know the beat of music, they may also need to know the intensity of music, the stress of each bar in a song, the variance of music emotion and so on. All these factors can likely not be expressed by the vibration of wristband alone; other display methods may need to be used. As deaf people have auditory sensing problem, we should figure out methods other than hearing to convey above

information to them. In our preliminary design, we have tried to use mood light to show the emotion of music. This may be a practical way of conveying music emotion. The focus of researching users mental models here is try to explore connections between different types of information and display forms which can be quickly accepted and understood by deaf people.

Secondly, if it is possible, we will conduct a comparison study researching the difference of in feeling music between deaf people after they have used Bumblebeat wristband and hearing people. Unfortunately we were unable to have deaf people test our prototype in our project this semester, so we had to ask hearing people to use our prototype and give feedback. Although we have tried our best to diminish the deviation, the difference between deaf people and hearing people can not be ignored, and our study likely reflects the preferences of the hearing slightly more than the deaf.

What is more, as our ultimate goal is to establish a system which can instruct deaf people how to dance, the wristband can only be seen as the first part of our design which focuses on helping deaf people understand music. The other part, as we see it, should instruct deaf people on the basic components of choreography including pace and gesture. For our preliminary design, we considered using google glass to display arrows of direction and other information which can help deaf people with the physical motions of dancing or designing a device like dancing carpet which already has instructions on it (and users would dance in a specific area above that device). In summary, in order to help deaf people dancing, our prototype needs at least two components: a wristband to help deaf people feel music and a device in charge of guiding basic choreography for deaf people. For our continuing work, we will focus on finding an appropriate device form that can guide the user in following choreographed dance steps.