Designing a Transcription Game

A Thesis Presented by

Beatrice Liem

To Applied
Mathematics

in partial fulfillment of the honors requirements for
the degree of Bachelor of Arts Harvard College

Cambridge, Massachusetts
Friday, April 1, 2011

Abstract

In this thesis, we develop a method to generate audio file transcripts with the high accuracy of professional transcription and the low costs of computerized transcription. We apply the principles of \Games With A Purpose" to this problem, creating a solution in the form of a game. In designing the Transcription Game, we overcome the obstacle of being unable to determine which transcripts are accurate, creating an incentive structure that results in a Perfect Bayesian Equilibrium in which all players enter the most accurate transcripts possible. We test our hypothesis that an iterative implementation of this game, in which players improve upon others' entries, performs better than a parallel implementation, in which players enter transcripts without seeing others' entries. Empirical results comparing the accuracy, efficiency, and enjoyability of the two versions support our initial hypothesis. Ultimately, while this game was not an instant hit, it provides solid groundwork for future development. The 96.6% accuracy of the transcripts obtained through the iterative process clearly demonstrates the potential of the methods implemented herein and hints at the possibility of someday being able to use a similar system on a larger scale for audio file transcription.

Acknowledgments

This thesis would not have been possible without the help and support of Yiling Chen and Haoqi Zhang. Both have provided invaluable guidance in developing thesis topics, challenging my ideas, critiquing my analysis, and suggesting possible solutions to the myriad problems I encountered. In addition, they have provided much-needed moral support and encouragement throughout this process, transforming my thesis-writing experience from an academic exercise into an exciting adventure. To them, I offer my wholehearted thanks.

AIn addition, I thank everyone who took the time to play the Transcription Game, particularly those who offered feedback on their experience. In particular, I thank my sister, Victoria Liem, for her enthusiastic participation, as well as for her understanding during the most intense portion of my thesis writing. I thank Colin Fogarty for his extensive LT E Xassistance, as well as for the time he spent discussing ideas with me. Finally, I thank my family and friends for their support.

Contents

Chapter 1

Introduction

1.1 Thesis Motivation and Goals

There is a widespread need for transcription services converting audio files into written text for various purposes: meeting minutes, court reports, medical records, interviews, videos, speeches, and so on. Written text is easier to store and search than audio files, and apart from this, there are many circumstances one could imagine for needing to transcribe human speech: those who are deaf still need to listen to certain audio files; people with limited ability to type, such as those who are paralyzed or suffer from Carpal Tunnel Syndrome, still need to draft documents; and often, when we're in public places, we like to watch subtitled videos rather than listening to sound on our speakers.

Speech recognition was first introduced to the world in 1952 with the construction of a device that recognized single spoken digits [1]. Following soon after was the IBM Shoebox, a 1961 computer that recognized 16 spoken words in addition to the ten digits [2]. In 1987, Kurgweil Applied Intelligence released the first large-vocabulary speech recognition software, which recognized 20,000 words, uttered one at a time [3]. In 1990, Dragon Dictate launched the first continuous-word speech system for PCs [4]. Since then, speech recognition software has been under continual development.

Currently, speech recognition software uses Hidden Markov Models, analyzing the frequency of words to decipher both pre-recorded speech and words spoken in real time [5]. Many real-time speech recognition programs, such as Dragon Dictate and Dragon NaturallySpeaking, allow users to train computers to recognize their speech patterns. Though the accuracy levels of these programs depend on factors such as the clarity of a user’s speech and how much training the computer has undergone, the latest versions of these softwares advertise accuracies of up to 99% [6]. While this seems impressive, we found during the course of this thesis that these programs fail to reach such high levels when transcribing pre-recorded audio or unfamiliar voices; they often encourage users to \speak more slowly or clearly."

Once we move to programs that do not allow for individual user training, accuracy levels plummet. Though there are a number of platforms for so-called \computerized" speech-to-text transcription, such as Google Voice’s voicemail-to-text program [7], their accuracies lie in the range of 78%-86% [8]. Other programs designed for the purpose of transcribing pre-recorded audio, such as Adobe Soundbooth Pro, only reach accuracy levels of approximately 50% [9]. Transcripts of clips fed into Adobe Soundbooth for this thesis were nearly unrecognizable at times, as is shown in Appendix A.

Professional transcription serves as a more accurate alternative to computer transcription software. Though not typically used by the general public, professional transcription is a large industry; medical transcription alone is a $10-$25 billion industry in the United States each year, projected by the Bureau of Labor Statistics to grow 11% by 2018 [10]. Typical transcription firms market their services online, guaranteeing accuracies as high as 99%, for fees as \low" as $1 per minute of transcribed text [11]|a sum that adds up quickly. Upon closer examination, however, these charges are not necessarily as outrageous as they seem. If we assume that audio files are played in real time, the minimum time it takes to transcribe an audio recording is simply equal to the length of the recording. Other factors, such as limitations in typing speed, unclear recordings, uncommon words, and so on, increase this time. Thus, a $60-per-audio-hour fee may actually be a $30-per-labor-hour fee in the case of a recording that takes twice as long to transcribe. While this seems marginally cheaper, it is still quite costly to transcribe large volumes of audio material.

Clearly there is a need for a solution that bridges the gap between low-cost, low-accuracy computerized transcription and high-cost, high-accuracy professional transcription. In this thesis, we propose a solution in the form of a game played by online users motivated by their own competitive natures, a chance to win a slight financial reward, a willingness to help others, and their own enjoyment of the game. By having these players listen to audio clips and either transcribe what they hear or correct previous players’ transcripts, we are able to form transcripts that are 96.6%

accurate.

1.2 An Overview of the Transcription Game

1.2.1 Game Design Considerations and Goals

Two formats that lend themselves nicely to providing low-cost human labor are Amazon Mechanical Turk (MTurk), a task-oriented, paid \job-like" platform, and online game implementation, whose minimal cost stems from users freely volunteering their time. We choose the latter because it offers flexibility in implementation, allows players to build responses iteratively in a very quick manner, can be scaled more easily and at lower costs, and provides entertainment value. Furthermore, the social aspect of the game provides both a competitive and cooperative environment that encourages players to enter more accurate transcripts in larger volumes. (Cooperation is established because though players compete in a game, they must also cooperate to maximize the number of points they earn.) This decision is further discussed in Chapter 2.

The Transcription Game, as I call my solution, takes advantage of a phenomenon known as \crowd surfing"|namely using the efforts of the general public to solve a problem. In designing this game, we adopt similar methods to those presented in \Games With A Purpose," a set of games designed by Carnegie Mellon Professor Luis von Ahn and his colleagues that present real-world problems as games that people play for entertainment [12]. These games, such as the \ESP Game / Google Image Labeler," \Tag-a-Tune," and \Foldit" label images, tag tunes, teach computers to fold proteins, and accomplish a myriad of similar useful tasks by taking advantage of the fact that people are willing to spend time playing enjoyable games. Similarly, in the Transcription Game, we use the output obtained through game play to deliver transcriptions with accuracies that beat those of computerized transcription and are comparable to professional results. Rather than simply hoping, however, that users will submit accurate results that eventually end up matching others' transcripts for a given audio file, we must establish a game structure that encourages players to submit accurate outputs that build iteratively on previous players' work. It is our hope, ultimately, to design a game that is both fun to play and useful in that it ultimately results in the accurate transcription of a given audio clip.

Our goals for this thesis are as follows:

1. Design and analyze a game that people can play to transcribe audio clips for a low cost and a high accuracy comparable to professional transcription. This involves ensuring that incentive structures are properly aligned so that people are motivated to enter the most accurate transcript possible. To do so, we must first overcome the limitations imposed by the fact that we cannot determine what is and isn’t accurate when we do not yet have an accurate transcript of an audio clip.

2. Determine whether a parallel or iterative implementation of the game is more effective. The former refers to a situation in which players work independently to submit accurate transcripts; the latter involves improving on previous players’ transcripts to produce an accurate final product. Based on the results of Little et al.’s 2010 paper, \Exploring Iterative and Parallel Human Processes," which finds that iterative processes are more efficient and slightly more accurate than parallel ones [13], we hypothesize that iterative processes will be more successful, as players can combine their understandings of different portions of a clip to produce an accurate final result.

We evaluate the two implementations of the game, as well as the success of the game overall, in terms of the following:

1. The accuracy of the transcripts. This factor is important in demonstrating the viability of the Transcription Game as a new approach to transcription. To measure accuracy, we use a standard metric known as the \Word Error Rate (WER)," a measure of word edit distance, to compute Word Accuracy [14]. In many cases, because it is easier to compute, we also use the Levenshtein edit distance, a similar metric that operates on a character level [15], to calculate what we call Character Accuracy. Both measures will be explained later in this paper.

2. The efficiency of the game. This is measured in two ways: effort efficiency (how accurate the \best" transcript is after niterations) and time efficiency (how long it takes to complete a single iteration of the transcription process). Though this is not as crucial as the overall accuracy of the result, it is a factor that we consider heavily in the design of the game.

3. The degree of enjoyability the game offers. This will be measured quantitatively by

the amount of time a user spends playing the game and qualitatively by users' comments on the game. This factor is instrumental in evaluating whether or not the Transcription Game is successful as a form of entertainment.

While it is difficult to attain ideal levels of accuracy, efficiency, and enjoyability, we attempt to maximize the game's achievements in these three categories. In terms of accuracy, this means producing transcripts that are more accurate than computerized transcripts and comparable to professional transcripts; in terms of efficiency, this means converging upon an accurate result in a timely manner; in terms of enjoyability, this means designing a game that people play multiple times and are interested in returning to in the future. We compare how well the parallel and iterative implementations of the game meet each of these standards and hypothesize that the iterative implementation will produce better results. The success of the game as a whole will be measured according to these standards.

1.2.2 Game Structure and Empirical Results

The original, parallel, and iterative implementations of the Transcription Game operate fairly similarly on the player's end. In all instances, players are asked to listen to a randomly-selected audio clip and transcribe its contents. This transcription is then submitted for scoring purposes, and the process repeats itself. Differences between the two implementations lie in the fact that the original and iterative forms show players up to two transcripts entered by previous players and ask players to edit these transcripts. Players in the parallel implementation do not see other players' transcripts and must simply enter their own guesses from scratch.

The results of these experiments will be detailed in Chapter 5; however, they are worth mentioning briefly. The parallel and iterative implementations generated final transcripts with overall Word Accuracies of 93.6% and 96.6% respectively. We found that while these accuracy levels were fairly comparable for the two processes, there was greater variance in the accuracies of transcripts in the parallel process, and we saw slight evidence that the iterative process was marginally more accurate. Effort efficiency appeared to be slightly higher for the parallel process when only a few iterations had occurred, but higher for the iterative process as the number of iterations increased, indicating that the iterative process performs better after a larger number of iterations. Time efficiency and enjoyability were markedly higher for the iterative process, though there was room to improve on overall enjoyability. We conclude from these results that the iterative process is more promising that the parallel one, though future efforts should be targeted towards increasing both the degree of improvement from one transcript to the next and the enjoyability of the game.

* * * * *

In this thesis, we provide a way to obtain audio file transcripts with the high accuracy levels of professional transcription and the low costs of computerized transcription. We apply the principles of \Games With A Purpose" to this problem, creating a unique new approach to transcription in the form of a game. In designing the Transcription Game, we overcome the obstacle of being unable to determine which transcripts are accurate, to create a game structure that results in a Perfect Bayesian Equilibrium in which all players enter the most accurate transcripts possible. We test our hypothesis that an iterative implementation of this game performs better than a parallel implementation in terms of accuracy, efficiency, and enjoyability. Our results slightly favor the iterative process, providing support for our hypothesis. Ultimately, while this game was not an instant hit, it provides the groundwork for future improvements. The 96.6% accuracy of the transcripts obtained through the iterative process clearly demonstrates the potential of the methods implemented herein and hints at the possibility of someday being able use to a similar system on a larger scale for audio file transcription.

Chapter 2

Related Work

In this chapter we further discuss the current state of transcription, as well as two methods of obtaining large amounts of human labor for extremely low costs. In particular, we describe a method called \Games With a Purpose," which presents tasks as games that people play for fun.

2.1 Professional versus Computerized Transcription

A 2003 experiment conducted by Al-Aynati et al., pathologists at St. Joseph's Healthcare in Hamilton, Ontario, Canada, compared the accuracies and costs of professional and computerized transcriptions of 206 pathology reports recorded by the primary experimenter [16]. Professional transcriptionists consisted of four professionals employed by St. Joseph's Healthcare, whose experience ranged from two to 29 years. The software used for computerized transcription was IBM Via-Voice Pro Version 8 with pathology vocabulary support, and the computer that was used had previously been trained to recognize the primary experimenter's voice.

The 206 pathology reports were both given to professional transcriptionists and played to computer transcription software. There were a total of 23,458 words in these reports, averaging 114 words per report. Experimenters found that professional transcripts had a mean accuracy of 99.6% (range, 99.4%-99.8%), while computerized transcripts had a mean accuracy of 93.6% (range, 87.4%-96.0%). Because of the lower accuracy of computerized transcription, the extra time needed to edit these transcripts averaged twice as long as the time needed to edit professional transcripts. Nevertheless, there were significant financial savings resulting from the use of computerized transcription. Because of the trade-off between accuracy and cost, and because of the additional time it took to correct computerized transcripts, experimenters did not fully recommend one system over the other. They concluded that while they would not encourage the use of computerized transcription services in hospitals that already have widescale professional transcription services available, it could be of use to hospitals with a shortage of these services, particularly as transcription software improves over time.

This experiment, along with others like it, makes an increasingly compelling case to overcome our challenge: to combine the high accuracy levels of professional transcription with the low costs of computerized transcription. We explore two potential methods of accomplishing this.

2.2 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk), named after a 19th century chess-playing automaton called \The Turk,'' provides a way to obtain human labor for low costs [17]. MTurk, an Internet website, employs the idea of crowdsourcing, a phenomenon in which tasks traditionally performed by employees or contractors are outsourced to the general public to reduce costs [18]. On the Internet, this essentially allows anyone to contribute to a project. Well-known examples of crowdsourcing include Wikipedia, an open encyclopedia that anyone can contribute to; the Netflix Prize, an open competition to develop the best algorithm to predict user ratings for a film based on previous film ratings; and reCAPTCHA, which displays distorted text that can be read by humans but not bots for the dual purpose of digitizing books and verifying that it is indeed a human who is trying to access a certain website [19].

Through MTurk, requesters can post jobs for others to do, listing amounts that they are willing to pay for a job to be completed. These amounts are typically appallingly low by U.S. standards, paying an average of $0.01 to $0.10 in many cases, for work that may span 5 minutes to an hour [20]. Many of the \Turkers" who take on these jobs, however, are either those who aren’t really working for the money, or those in developing nations, (with a large concentration in India) to whom these \low'' amounts are reasonable by local standards. Studies have shown that the median reservation wage (the lowest rate at which a worker is willing to accept a job), is $1.38

per hour [21], and this low pay means that companies or individuals pay only a fraction of the cost that they otherwise would [22]. Amazon.com profits from hosting the MTurk website through a 10% surcharge on all transactions.

Despite these low costs, however, requesters are often worried about the quality of the output they receive. In general, tasks are not restricted to Turkers of certain qualifications, though requesters may stipulate that a set of standards be met, and they can reject work that is improperly completed [23]. In the context of audio transcription, we note that most Turkers are not professional transcriptionists and thus lack their skills and constant practice; they should not be expected to produce the same level of results. Still, those willing to overlook this can post an audio clip, pay some nominal amount, and get back a reasonably accurate transcription of an audio file.

For those who will settle for this reasonably accurate transcript and who are willing to go through it and correct errors, such a solution might be appropriate. However, due to the fact that only one person will be listening and transcribing the audio clip, there is much more room for error than if multiple people listened to a clip. Requesters looking for higher quality transcriptions should post a second job, upload the audio file and the transcript from the first Turker, and advertise for someone to correct the transcript or transcribe it anew.

At the end of my thesis-writing experience, I was made aware of an online transcription website called CastingWords.com, which implements the strategy just described using the MTurk platform. Though the accuracies of CastingWords' transcripts are not advertised online, CastingWords is described by the Wall Street Journal as \[t]he most accurate and detailed of all our services [i.e. among the five it reviewed]" [24]. Still, this accuracy comes at a cost: it was, according to the article, the second most expensive service [24], charging rates of $0.75-$2.50 per minute of text transcribed [25]. CastingWords' algorithm consists of posting tasks on MTurk for varying compensations (i.e. $0.18 for a 3.5-minute clip with a maximum bonus of $0.72), having Turkers transcribe clips or modify existing transcripts, and having other Turkers grade these transcriptions before reposting them for others to correct [26, 25]. Though CastingWords eschews using professional transcriptionists in favor of using Turkers to transcribe clips, its increased accuracy still bears a high cost (some of which is likely due to Turkers' higher-than-average pay). As such, it does not fit the requirements of our initial goals.

Employing a game-like structure with a centralized database is much more versatile than using the MTurk platform and provides the following additional benefits:

1. A centralized game structure offers an easy way to allow multiple people to process a clip iteratively, with very little downtime between one person's transcript being submitted and the next person being able to see it for correction.

2. The incentive structure of the game encourages players to cooperate with one another to generate matching transcripts, as players are awarded points based on how similar their entries are to others'. Additionally, four players from two independent groups must agree on the same transcript, so the cooperative theme imposed here makes it more likely that an accurate transcript will be produced in a timely manner. Though a similar incentive structure could be arranged using MTurk, the social atmosphere fostered by a game format makes this aspect of cooperation much more prominent.

3. Playing a game is always somewhat competitive, and this competition encourages players to play longer and more often to maximize the number of points they receive. This allows them to place higher on the leaderboard than their competitors, and it increases their chances of winning a prize|in this case, a $25 Amazon.com gift card. A similar monetary incentive is, of course, provided through MTurk, but players uncertain of their chances of winning a prize may choose to play longer than Turkers whose rewards are known and fixed.

4. Though there is a cost in providing a prize for the Transcription Game, if the site becomes popularized, the cost of obtaining accurate audio transcripts is likely lower for a game implementation than for MTurk, where each worker must be paid for his efforts. This would require that the volume of transcripts processed per dollar spent in the Transcription Game exceeds the ratio that can be obtained on MTurk. Additionally, if the site were to be popularized, it would be targeted towards the public as a whole, rather than those simply looking to work, providing greater scalability.

One can argue that the above benefits can all be achieved by posting a fun version of the transcription tasks to MTurk and setting up incentives using money rather than points. In this case, it seems that the two forums are nearly equivalent, and one simply ends up paying people to complete enjoyable tasks. The social aspect of an online game website, however, produces both cooperative and competitive elements that are less prominent in a MTurk task, and it is these aspects that we hope will encourage players to enter accurate transcripts in larger volumes. For these reasons, as well as the practical ease-of-implementation argument, we choose to design a Transcription Game.

2.3 Games With A Purpose (GWAPs)

Our game is designed in the spirit of \Games With A Purpose" (GWAPs), an idea coined by Carnegie Mellon professor Luis von Ahn. Von Ahn states that according to the Entertainment Software Association, 200 million hours are spent daily on playing computer and video games, and that by the age of 21, the average American has spent over 10,000 hours (equivalent to five years of working a 40-hour per week job) playing these games [27]. The development of GWAPs serves simply as a means by which to redirect this energy towards a seemingly more productive end. Von Ahn and his colleagues focus their efforts on creating games that contain tasks that are difficult for computers but easy for humans. \When you play a game at Gwap," says the GWAP.com website (http://www.gwap.com), \you aren't just having fun. You're helping the world become a better place. By playing our games, you're training computers to solve problems for humans all over the world" [12]. Still, when designing a GWAP, von Ahn and his colleagues keep in mind that what they are building is not simply an interface that people interact with because of their altruistic feelings and desires to advance computing; it is a game, designed for entertainment and enjoyable to play [27].

There are a multitude of different types of GWAPs, such as the \ESP Game," \Tag-a-Tune," and \Foldit," which ask players to label images, tag tunes, and teach computers to fold proteins. We will examine what is perhaps the simplest and most popular of the GWAPs, the \ESP Game," later acquired by Google and renamed the \Google Image Labeler." This game is set up to help tag images|again, a task easy for humans but difficult for computers. Each player is randomly matched with an unknown partner, and both must enter image labels for a picture that they both see in order to get points. Points are awarded for matching labels, so the two players must work cooperatively to try to come up with the same labels based solely on the picture presented to them. To increase their chances of matching, players enter as many labels as they can think of.

Sometimes, however, certain labels are \taboo" and can't be used|this usually occurs with more commonly used labels|encouraging players to think more creatively and to find more complex ways to describe the images they see. The game has been extremely popular, with some players playing more than 40 hours per week [28]. Clearly, this is an example in which von Ahn is able to advance computing by providing image tags to search engines while simultaneously entertaining the masses.

2.3.1 Different Types of GWAPs

The ESP Game and other GWAPs can be divided into three categories: output-agreement games, inversion-problem games, and input-agreement games [27]. Output-agreement games are those in which players are randomly paired and expected to generate matching outputs based on some common input that both are given. Players are not allowed to communicate or see each others' answers. Because of this lack of communication, the best way for them to match their outputs is to base their outputs off of the input that they are given. Output-agreement games may not necessarily have a single correct answer, as in the case of the ESP Game, for example. As a result, players are incentivized to produce a series of common outputs, and because these outputs come from independent sources, the probability that something is correct when both parties have entered the same thing is increased.

Inversion-problem games are those in which players are once again randomly paired and go back and forth in a manner similar to the guessing game \20 Questions." One player, the \describer," is given some input and from it, generates an output that is sent to the other player, the \guesser," who is then expected to guess the original input. With each false guess, the describer must provide an additional clue, and the game continues until the guesser successfully guesses the initial input. This type of game structure is effective for eliciting facts about some input, and it promotes accuracy from both the describer and the guesser, as the describer must provide helpful clues, and the guesser must guess correctly in order to end the game. Once a round ends, players switch roles so that they remain equally engaged in the game.

Finally, input-agreement games are those in which players are randomly paired and given two inputs. They provide each other with outputs to determine if their inputs are the same, so they are incentivized to provide accurate outputs. Additionally, because players could simply guess randomly, scoring is set up so that the number of points they earn increases with the number of consecutive correct guesses they make. Thus, they are incentivized to guess correctly to earn higher scores.

Notice that while these games may be presented differently, they all follow the same cooperative two-player format. Von Ahn writes, however, that both single-player and multi-player modes can also be implemented [27]. For example, a single-player game can be advantageous in cases where there are uneven numbers of players, or where the number of players at a given time is not large enough to run a game (i.e. when the game is just starting out). Prerecording a game may be beneficial to allow a single player to play against this prerecorded set of actions, simulating the existence of another player. This method, however, may be more difficult to implement in some games than in others. On the other end of the spectrum, having multiple players may be advantageous in that it allows people to compete directly|for example, having two of three or more players match labels in the ESP Game. This, however, changes the game's atmosphere from cooperative one to competitive, which may be more enjoyable in some instances, but may compromise the accuracy of the game. Additionally, using multiple players may not be as efficient in terms of effort, particularly if only two are needed to obtain the same result and additional players can be grouped together to produce another set of outputs.

2.3.2 Evaluating a GWAP

Regardless of the format of the games, von Ahn uses the same criteria to assess their success: accuracy, design efficiency, and enjoyability [27].

The accuracy of the output is clearly paramount, as unreliable or inaccurate output is not useful. The ideal game is one that is extremely enjoyable and employs incentive structures that encourage players to give accurate responses. To measure accuracy is easier than measuring enjoyability: von Ahn hires people to perform the same work as people playing the game, and he determines whether the outputs from the two parties are similar. To increase the likelihood of obtaining accurate outputs among the unpaid players, however, von Ahn employs the following techniques [27]:

Random Matching. As mentioned above, randomly matching individuals makes it less likely that they can collaborate to manipulate the system.

Player Testing. Problems with known correct outputs may be used to test players. If their outputs do not match the predetermined correct outputs, their inputs may be disregarded.

Repetition. Outputs are not considered to be correct until a number of pairs have entered the same thing. This decreases the probability that an incorrect answer is marked as correct simply because two people have agreed on it.

Taboo Outputs. In the case of the ESP Game, for example, many outputs may be considered correct. To distinguish them from other incorrect labels that may appear with high frequency, certain correct words are made \taboo" to increase the frequency of other correct words. This allows von Ahn and his colleagues to determine with greater accuracy whether something is correct.

In addition to accuracy, design efficiency in the game is important. Von Ahn measures this as the average number of problem instances solved, or in the case of the ESP Game, the number of labels generated per human-hour. This statistic is averaged over all of the people playing a game. In the Transcription Game, we assess efficiency using two metrics: the amount of time people spend to generate a single output and the accuracy of the \best" transcript after a given number of outputs have been generated.

Finally, the enjoyability of a game is important because it determines the volume of useful output the game will generate and ultimately, how successful a game truly is. The more enjoyable a game is, the more likely it is that individuals will play this game, perhaps switching from another game or introducing other players to the game. To measure enjoyability is difficult, but one way of estimating this is to look at the average length of time people spend playing or how often they return. For increased enjoyability of a game, von Ahn employs the following techniques [27]:

Timed Response. This increases the level of difficulty in many instances, keeping the game from becoming boring and repetitive. It is important, however, that limiting time does not compromise accuracy.

Scorekeeping. Assigning points allows users to keep track of how well they are performing and provides them with a goal: to increase their total score.

Player Skill Levels. Introducing a ranking system among players motivates them to increase their scores by playing more to accumulating more points.

High-Score Lists. Similarly, introducing a leaderboard allows players to compete with each other and motivates them to play more to increase their scores.

Randomness. Inputs should be randomly assigned and players randomly paired to prevent players from scheming together or getting bored with the same partner or same input in the game.

As was mentioned in Chapter 1, we consider these three criteria in the design and assessment of the Transcription Game, particular in our comparison of the parallel and iterative implementations.

2.4 Parallel versus Iterative Processes

A fundamental design decision in this game was choosing to implement it primarily as an iterative process. Intuitively, it seems that an iterative process should deliver more accurate results, as players can fill in or correct portions of transcripts that others miss. This process is also expected to be more time-efficient, as players can simply fix others' mistakes rather than re-entering transcripts. Finally, because the iterative process is more interactive (in the sense that one corrects another player's result), it seems that the iterative process should be more fun than a parallel process in which players work independently.

In addition to these intuitions, however, we further base our decision on previous work comparing parallel and iterative processes. Little et al.'s 2010 paper \Exploring Iterative and Parallel Human Processes" compares the two processes under three circumstances|writing, brainstorming, and deciphering blurry text|and concludes that iteration increases the average quality of responses in each of these processes, with a statistically significant result in writing and brainstorming [13]. Little and his colleagues find that despite this result, in brainstorming and text deciphering, the best approach is not necessarily clear, as both tasks benefit from the larger variety of responses generated in the parallel process. We focus on the case of text deciphering, as it best mirrors the transcription issue we are addressing.

In the deciphering experiment, researchers created blurred images of 12 sentences and posted them on MTurk for Turkers to decipher. This is similar to the Transcription Game, in which players listen to an audio file, have to decipher what they hear, and write down their best guess. Experimenters ran the task as both an iterative and a parallel process, presented to two separate groups of Turkers. They hypothesized that the iterative process would result in a greater likelihood of deciphering the text, as people would be able to build on each others' guesses.

Experimenters found that on average, results from the iterative process were 65% accurate, compared to 62% in the parallel process, though this difference was not statistically significant. Examining the accuracy of each process after niterations, they found that after four iterations, the iterative process produced clearly superior results; this difference was greatest after eight iterations. These differences, however, were never statistically significant. Finally, researchers noted that there was a case in which the parallel process produced nearly perfect results, but the iterative case produced results that were only 30% accurate, as people built off of others' incorrect responses and were unable to think creatively to decipher a word after seeing previous guesses. Still, overall, there was a statistically significant difference in the amount of time people spent deciphering the text, with those in the iterative process taking nearly 20% less time than those in the parallel process. Researchers concluded that the iterative process could perhaps be improved to attain significantly higher accuracies if completely incorrect guesses were hidden to avoid future Turkers from being led off track.

The results we find in the Transcription Game resemble Little et al.'s findings, with the iterative process producing final transcripts that are slightly more accurate than those in the parallel process. We find that the accuracy of the iterative process is lower than that of the parallel process when only a few iterations have occurred, but that the former exceeds the latter as the number of iterations increases, though the difference is not statistically significant. Additionally, we find that the time efficiency of the iterative process is higher than for the parallel process|a statistically significant result. Finally, we find cases in which the iterative process was unable to converge upon an accurate result, but the parallel process did, indicating that as Little suggests, players can be misled by others' mistakes.

2.5 The Transcription Game

The Transcription Game developed in this thesis uses a variant on the output-agreement game structure described above. In this game, we seek to develop a similar cooperative environment that harnesses the input of multiple humans to transcribe an audio clip, and to test our hypothesis that iterative methods produce higher levels of accuracy, efficiency, and enjoyability than parallel ones. Unlike existing GWAPs, however, the Transcription Game is implemented as a single-player game for ease of implementation and testing.

In this chapter, we have compared existing methods of transcription, discussed various forums in which to implement a new approach to transcription, and settled upon the GWAP framework. What remains, therefore, is to structure a game that provides a way to transcribe audio clips with high accuracy and low costs. To do so, we must overcome the impediment of not knowing whether a transcript is accurate or not to design an incentive structure that properly motivates people to enter the most accurate results possible. Our focus in this thesis is mainly on the design of an iterative game, which conveniently offers a solution to the incentive structure problem: each person is correcting the previous person's input, theoretically improving the final output from one iteration to the next, and each person's results are compared to hidden results of another group, providing an incentive for them to produce the most accurate transcript possible. The design decisions of the game are discussed fully in the next chapter.

Chapter 3

Designing the Transcription Game

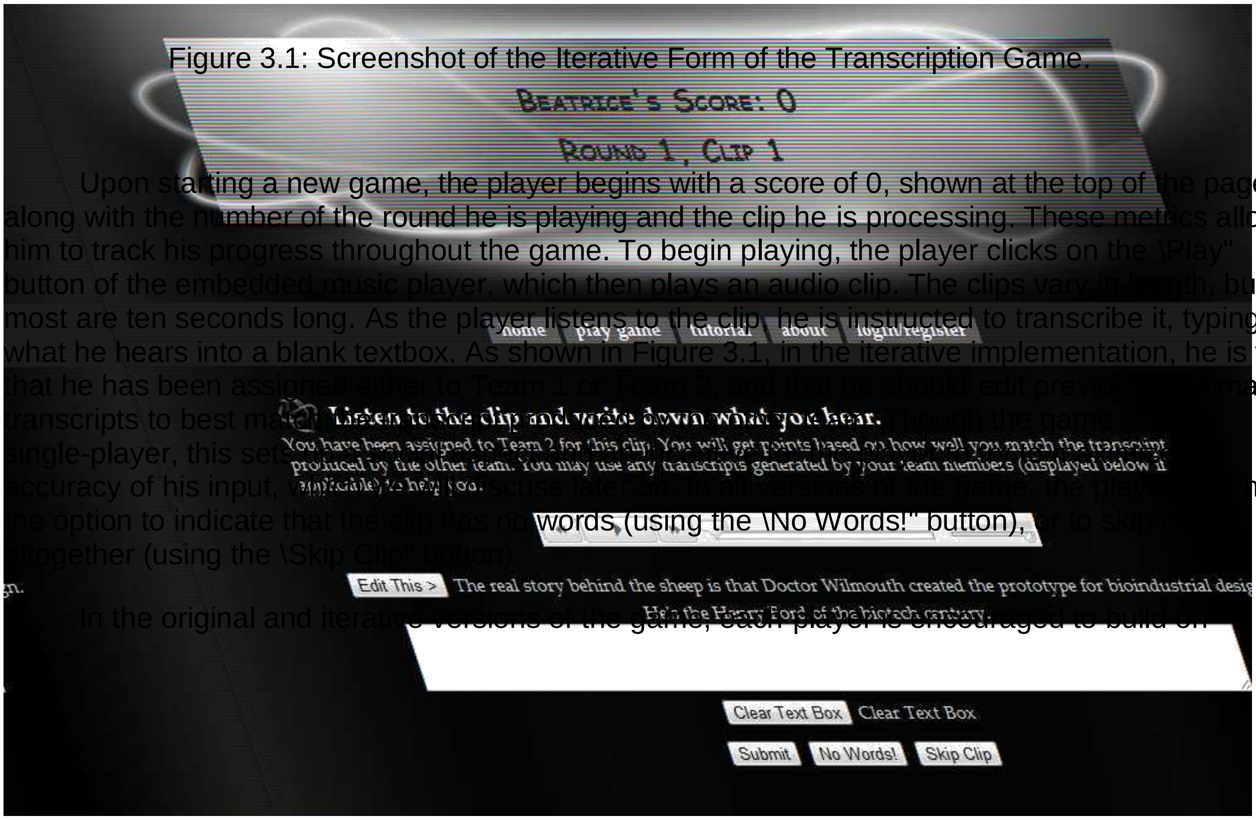
The Transcription Game takes the form of a single-player online game hosted on the Harvard Computing Services website and supported by a SQL database. The website was coded in PHP/SQL and hosted at http://hcs.harvard.edu/~ bliemthesis/. While the details of the code will not be discussed, we review and explain key design decisions, and we describe the original implementation (sometimes called the \original iterative implementation"), the parallel implementation, and the iterative implementation (sometimes called the \final iterative implementation") of the game. Except where noted, the contents of this chapter are applicable to all versions of the Transcription Game.

3.1 The Player's Experience

The first time a player goes to the Transcription Game website, he is asked to register with a username, password, and email, if he wants to be eligible for a drawing to win an Amazon.com gift card. Upon registering, he is assigned a user ID, and his information is stored in a SQL database. From then on, the player can simply log in using his username and password.

Once logged in, the player is redirected to the home page, where a leaderboard displays the usernames and scores of players who have obtained the highest \total score" (across all games) and the highest \high score" (in a single game). The player can then either click on a link to a tutorial that teaches him how to play the game, or on a button to begin playing directly. Figure 3.1 is a screenshot of the iterative form of the game.

Figure 3.1: Screenshot of the Iterative Form of the Transcription Game.



Listen to the clip and write down what you hear.

You have been assigned to Team 2 for this clip. You will get points based on how well you match the transcript produced by the other team. You may use any transcripts generated by your team members (displayed below if applicable) to help you.

gn.

Edit This > The real story behind the sheep is that Doctor Wilmouth created the prototype for bioindustrial desig He's the Henry Ford of the biotech century.

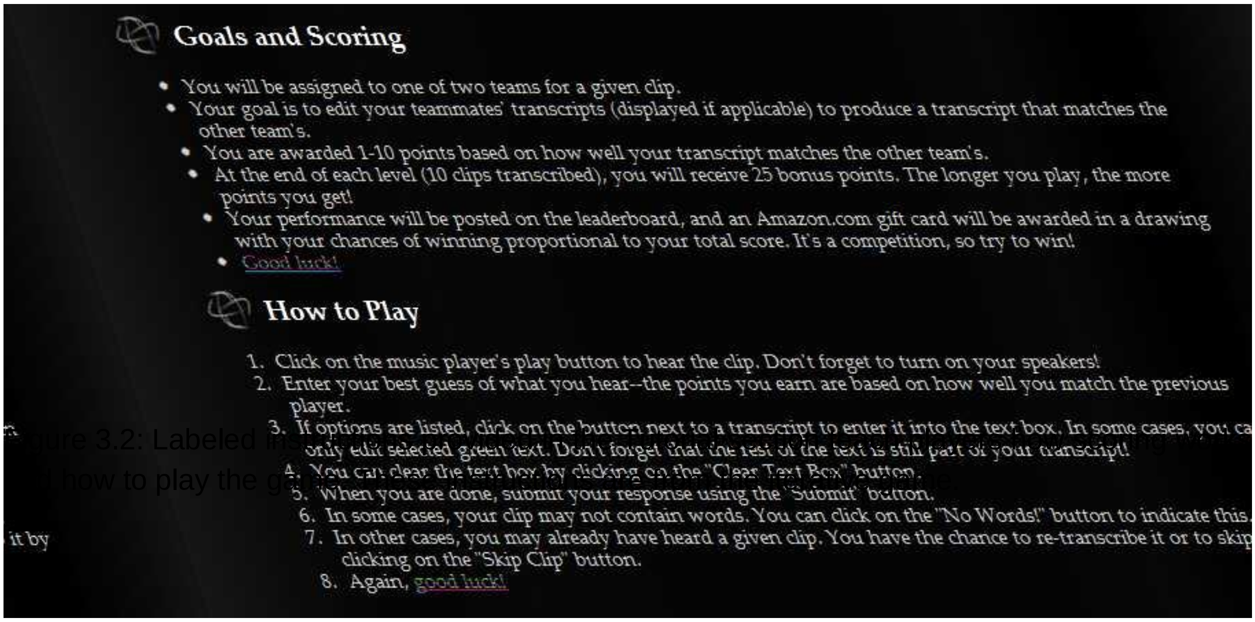
Clear Text Box Clear Text Box

Submit No Words! Skip Clip

Upon starting a new game, the player begins with a score of 0, shown at the top of the page along with the number of the round he is playing and the clip he is processing. These metrics allow him to track his progress throughout the game. To begin playing, the player clicks on the \Play" button of the embedded music player, which then plays an audio clip. The clips vary in length, but most are ten seconds long. As the player listens to the clip, he is instructed to transcribe it, typing what he hears into a blank textbox. As shown in Figure 3.1, in the iterative implementation, he is told that he has been assigned either to Team 1 or Team 2, and that he should edit previous teammates' transcripts to best mat[illegible]. Though the game is [illegible] single-player, this sets [illegible] accuracy of his input, [illegible] discuss later on. In all versions of the game, the player [illegible] has the option to indicate that the clip has no words (using the \No Words!" button), or to skip [illegible] altogether (using the \Skip Clip" button).

In the original and iterative versions of the game, each player is encouraged to build on previous players' responses. We include up to two previous responses on the page for the iterative implementation (four in the original version), and the player can click on a button to add these transcripts to the textbox. He can then modify the text until he is satisfied that what he hears matches what he has written. Players in the parallel version of the game do not have this option and must enter their own transcripts from scratch. All players can use the \Clear Textbox" button to start over, or the \Submit" button to submit their transcripts and move on to the next clip. After processing ten clips, the round ends, players are awarded 25 bonus points (50 in the original implementation), and a new round begins. Players can play indefinitely, until they have processed all of the clips in the database.

Figure 3.2 contains a screenshot of the instructions from the Tutorial page of the iterative game. This tutorial details goals and scoring rules for the game, as well as instructions on how to play.



Goals and Scoring

- You will be assigned to one of two teams for a given clip.
- Your goal is to edit your teammates' transcripts (displayed if applicable) to produce a transcript that matches the other team's.
- You are awarded 1-10 points based on how well your transcript matches the other team's.
- At the end of each level (10 clips transcribed), you will receive 25 bonus points. The longer you play, the more points you get!
- Your performance will be posted on the leaderboard, and an Amazon.com gift card will be awarded in a drawing with your chances of winning proportional to your total score. It's a competition, so try to win!
- Good luck!

How to Play

1. Click on the music player's play button to hear the clip. Don't forget to turn on your speakers!
2. Enter your best guess of what you hear--the points you earn are based on how well you match the previous player.
3. If options are listed, click on the button next to a transcript to enter it into the text box. In some cases, you ca only edit selected green text. Don't forget that the rest of the text is still part of your transcript!
4. You can clear the text box by clicking on the "Clear Text Box" button.
5. When you are done, submit your response using the "Submit" button.
6. In some cases, your clip may not contain words. You can click on the "No Words!" button to indicate this.
7. In other cases, you may already have heard a given clip. You have the chance to re-transcribe it or to skip clicking on the "Skip Clip" button.
8. Again, good luck!

it by

gure 3.2: Labeled ins

how to play the ga

3.2 General Design Decisions

Ideal game design and executable game design are not always the same. There were times in which certain decisions that significantly altered the game's structure had to be made in order to accommodate the limitations of implementation. Ultimately, however, game design was guided by Luis von Ahn's paper, \Designing Games With A Purpose," and decisions were made based on how well they were thought to maximize the accuracy, efficiency, and enjoyability of the Transcription Game.

3.2.1 Basic Game Structure

The basic structure of the game was one of the most important features to specify before even considering how to implement the parallel and iterative forms. This involved determining the number of players, the type of game (output-agreement, inversion-problem, or input-agreement), and basic game features, such whether or not to limit the time players had to enter transcripts. A major challenge in designing the game was to determine whether a transcript was right or wrong without being able to compare it to the correct transcript.

A Single-Player Game

My first consideration in game design was to determine the number of players the game should involve. Initially, I considered implementing a multi-player game, where perhaps two players would enter their own transcripts, and a third player would judge whose transcript was more correct. However, this structure appeared to have a few problems. Firstly, it is difficult to incentivize a judge to choose the correct answer, when he could simply choose the wrong one, potentially for the same number of points. Having multiple judges could perhaps correct this problem if judges' points were awarded based on whether their decisions matched, but if both transcripts are incorrect, it may be difficult to judge which is more correct. Secondly, the use of three or more players in a single round seems inefficient, as only one output comes out of the efforts of the group. Finally, the implementation of an online game with three players seemed to be a nontrivial task, and the volume of players that would be required for such a structure was likely more than could be supported by a game still in its testing stages.

I then considered a simple two-player version of the game that used principles drawn from the ESP Game. In this formulation, two players would listen to the same audio clip, transcribe it independently, and be awarded points based on how well their transcriptions matched. This formulation, however, presented similar limitations to the multi-player game. While efficiency would automatically be increased from the multi-player game (getting one output from two people is more efficient than getting one input from multiple people), the difficulty of implementing a two-player game online was a significant obstacle, and again, the volume of players might not be enough to support this structure. As a result, I decided to implement a single-player game.

A single-player game has advantages in that it can generate one unit of output per player submission, is easier to implement, and does not require that the site is frequented by a large volume of players present at the same time. Part of what makes a game fun, however, is the social interaction aspect, which is greatly decreased in the single-player game. While players presumably understood that they were playing on their own, the iterative game revives the social aspect by telling players that they have been assigned to a team, and that they should improve teammates' transcripts to match the opposite team's entry. This social interaction was missing in the parallel form of the game.

An Output-Agreement Game

After deciding upon a single-player game, the next step was to determine whether it should be structured as an output-agreement, inversion-problem, or input-agreement game. The challenge here was to find a way to implement these structures within the single-player framework.

Recall that an output-agreement game is one in which players match their submissions, based on some common input. This seemed to be the most natural way to approach the game: give players a clip, let them transcribe it, and compare the results to previous submissions. The inversion-problem game, in which players go back and forth with one person sending clues and the other responding with guesses (as in \20 Questions"), seemed not to work for a single-player game. Finally, the input-agreement game, in which players are provided with two different inputs and must create outputs to determine if the inputs are the same, also seemed not to make sense in this context, as it was unclear how this sort of structure would meet the goals of the Transcription Game. Thus, I settled on the output-agreement game structure, encouraging players to listen to clips and enter accurate transcripts to increase their probabilities of matching previous players' transcripts.

Basic Game Features

We turn to von Ahn's \Designing Games With a Purpose" for inspiration for methods used to increase the accuracy, efficiency, and enjoyability of a game. From von Ahn's list of common techniques, we choose to use repetition to determine whether an output is accurate, requiring multiple players to converge upon the same result before we deem a clip's transcription correct. Additionally, we keep track of scores and set up a leaderboard, though we do not award players ranks based on their scores. We assign players to clips randomly, permitting them to skip clips, but allowing them to see clips only once in the parallel and iterative implementations. We modify some of von Ahn's suggestions for the game, such as player testing. In the iterative form of the game, we reject submissions that are more than 50% in edit distance away from previous transcripts, as we want to minimize the chances that players will be misled in their transcriptions, per Little et al.'s suggestions [13]. Finally, there are cases in which we consciously decide not to use certain techniques that von Ahn suggests, such as timed response, as we feel that this will decrease the quality of a player's output.

Interestingly enough, players' feedback suggested that such a timer be added in, and that other features that were also decided against for simplicity's sake (such as a flashing scoreboard when players did particularly well) be added in future versions of the game. Players' feedback indicated that such features, as well as other interactive comments (\other players transcribed this clip more accurately!") would have increased the enjoyability of the game.

3.2.2 Rewarding Players in Real Time

After outlining a basic game structure, the next step was to design an effective reward system for each game. For the output-agreement structure to work, it is important to incentivize players to collaborate with one another, for it is only through matching users' transcripts multiple times that we can be increasingly sure of obtaining accurate results. When playing games, however, people tend to be competitive, particularly when something is at stake|an Amazon.com gift card or a leaderboard position, for example. Thus, it was important to ensure that players had no incentive to enter incorrect results for the purpose of decreasing others' scores, and that instead, they would be motivated to produce the most accurate transcripts possible.

As mentioned previously, we are theoretically limited by our lack of knowledge of what is and isn't an accurate transcript. This presents a significant hurdle in determining how to score clips, as we would ideally prefer to award points based on accuracy, with the highest number of points going to those who produced the most accurate transcripts. Once all transcripts have been collected for a single clip, this is fairly easy to do; examining each of the transcripts gives us an idea of what the best (most accurate) one might be, and we can award points based on similarity to the best transcript. However, while we are still collecting transcripts for a given clip, we can only compare players' entries to earlier responses. (Such a comparison applies primarily to the original and iterative processes, as in the parallel process, there is no real sense of transcript arrival order.) Because transcripts should theoretically be improving over time, a player who enters the first accurate transcript may not receive the highest score, as his entry will not match previous ones exactly. Thus, because we do not know what the best transcript is in the middle of the transcription process, this lack of information gives rise to situations in which the best transcripts may not always receive the highest scores.

The most obvious way to address this issue is to award points after collecting all transcripts. Delaying the awarding of points allows us to judge the accuracy of an earlier player's transcript by comparing it to a transcript entered by a later player, or even having later players vote on the accuracy of the earlier transcript. However, as was the problem with the multi-player game setup, it is difficult to check whether later players are correctly assessing the accuracy of earlier players' inputs, and furthermore, it is uncertain when additional players will come along to judge the transcript. Delaying the awarding of points means that players are not given feedback about their outputs until an undetermined later time, which may be after they have left the game. This detracts from the enjoyability of the game and greatly decreases one's motivation to play. Being able to see one's points increase immediately after submitting a transcript provides instantaneous feedback and a sense that one is moving in a certain direction (i.e. further up the leaderboard); thus, it engages a player more directly. We choose, therefore, to accept potentially awarding players more or fewer points than they deserve based on the accuracy of their transcripts, and to provide them with instant gratification by comparing their transcripts to those of previous players.

Now that we have established the basic structure of the game, let us take a look at the specific forms of each of the three implementations.

3.3 Game Implementation

3.3.1 The Original Implementation

The original implementation of the game was an early version of the iterative form, in which players could see and edit the four most popular entries (listed in order of popularity as long as two or more players have entered the same transcript). Players were awarded points half based on the similarity of their entries to these four transcripts (weighted by frequency of appearance), and half based on the similarity of their entries to that of the previous player. In an ideal world, if people were to play for some common good, this naive version could be very efficient, with people improving on previous entries and arriving at some accurate result after a small number of iterations. A gametheoretic analysis of this implementation is presented in Chapter 4, and we find, unfortunately, that the Perfect Bayesian Equilibrium of this game is one in which all players enter trivial results rather than the most accurate transcript. Though empirical results did not support this analysis (players behaved irrationally and entered fairly accurate transcript), we modify the original implementation to produce the iterative implementation discussed later on.

3.3.2 The Parallel Implementation

[1]The next implementation of the game was a parallel version, in which people could not see what others wrote. Though this seems simple to design, the supposed time independence of each submission created problems: it is unclear how each transcript should be scored if scoring is to occur in real time. Scoring players based on how well their submissions match others' implies some ordering of transcript arrivals, so this was decided against, and instructions on the game play page simply read, \Transcribe the Clip!"Points were awarded by random assignment. Because the parallel implementation was considered secondary to the main game, incentive

structures were not elaborately analyzed prior to execution. If players had been explicitly informed

[1] Players who chose to lo ok in the Tutorial p ortion of the website would have seen leftover instructions from the original implementation, telling them that they would b e scored based on how similar their entries were to previous players' transcripts. This misstatement was an unintentional mistake and will not b e rep eated in the future.

of the random scoring system, their optimal strategy would have been to enter a trivial transcript, as this costs the least effort and has no negative impact on their expected score. We hope that players instead simply followed the \Transcribe the Clip!" instructions that they were given. Empirical results showed that players tended to enter transcripts that resembled the contents of the clip, indicating that they were likely unaware of the scoring system and followed the instructions that they saw. If we were to decide to use a parallel version of the game in the future, however, further thought would have to be given as to how to score transcripts.

When implementing the parallel version of the game, it was difficult to know when to call a transcript accurate|do we simply wait to obtain a certain number of matching transcripts, or should we find the transcript with the highest frequency after a certain number of entries have been submitted? Both options leave open the possibility that if two different transcripts of the same audio clip are both very common, the order of transcript arrivals is important, and an incorrect result could be agreed upon fairly easily. For the purposes of this experiment, we simply collected responses for each clip, not restricting the number of transcripts we received, and we decided that a clip converged when the same transcript was generated twice. At any point in the process, if pressed to choose the best transcript, we randomly selected one of the two that were most similar to one another. The results of this strategy are further discussed in Chapter 5.

3.3.3 The Iterative Implementation

An improvement on the original implementation addresses the problems of checking the first person's output and aligning incentives such that people are motivated to improve upon existing transcriptions. As we shall see in Chapter 4, the original implementation did not set up incentives such that players entered their best guess of the most accurate transcript. Players knew what they had to match, and it was simply a matter of submitting something similar to what previous players had written, rather than entering an accurate transcript.

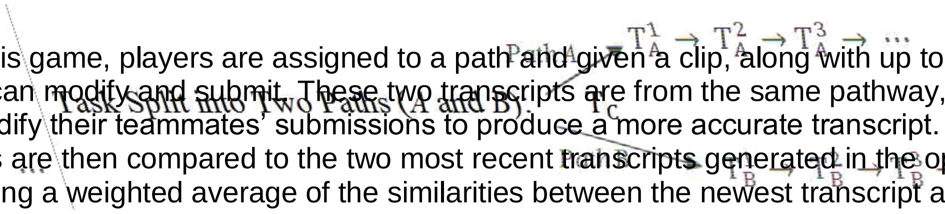
The (final) iterative implementation solves this problem by dividing players into two groups and thus separating what the player must improve from what he is scored against. This \dual pathway structure" thus allows us to employ an iterative process by using one group's transcripts as a basis against which to compare the other group's submissions. It also somewhat revives the social aspect of the game lost when we decided upon a single-player implementation, in that people are nominally assigned to teams and therefore are said to be part of a smaller group working towards a common goal.

The Dual Pathway Structure

The dual pathway structure introduces the idea of having two independent paths of evolving transcripts that can be compared to one another. In this setup, we nd that because players are unaware of what they are being compared against, there exists a Perfect Bayesian Equilibrium in which everyone produces the most accurate transcript possible. This result will be further analyzed in Chapter 4.

Figure 3.3 shows a diagram of the dual pathway structure, with the two paths Aand B displayed in green and blue respectively. We use Tdenote the transcript obtained when clips are processed via computer (in our analysis, using Adobe Soundbooth CS4 on High Quality in American English). Let Tn iC(i2fA;Bg;n= 1;2;:::) denote the nth transcript produced in pathway i. In this gure, the transcripts are listed from left to right in the order that they are generated. We assume independence between the two pathways, as subjects from one pathway can presumably interact with those from the other pathway only through means not accessible via the game. Thus, if the two pathways evolve along similar lines, this is likely a matter of chance.

. Figure 3.3: Players are alternately assigned to one of two di erent pathways A(in green) or B(in blue). They modify previous transcripts from their own pathway, and receive points based on how well their transcripts match the two most recent entries in the hidden opposite pathway.


Path A
$T_A^1 \rightarrow T_A^2 \rightarrow T_A^3 \rightarrow \cdots$
Task Split into Two Paths (A and B).
T_C
Path B
$T_B^1 \rightarrow T_B^2 \rightarrow T_B^3 \rightarrow$
...

In this game, players are assigned to a path and given a clip, along with up to two transcripts that they can modify and submit. These two transcripts are from the same pathway, and players are told to modify their teammates' submissions to produce a more accurate transcript. Players' transcripts are then compared to the two most recent transcripts generated in the opposite path and scored using a weighted average of the similarities between the newest transcript and the two used for comparison. We decide to compare a transcript against two other transcripts to account for the fact that players may mistakenly change a correct portion of a transcript, and we have no way of measuring this. To account for Little et al.'s findings that incorrect transcripts mislead future players [13], and to ensure that future transcripts are not wrongfully penalized due to others' mistakes, we ignore players' submissions for future display and comparison purposes if they differ by more than 50% from the transcripts they are being compared against. Because the earliest transcripts are being compared to a computerized transcript that at least vaguely resembles the contents of the clip, we felt that this would ensure that users did not simply enter irrelevant submissions [8].

On the back-end, implementation is fairly straightforward. First, the clip is passed through Adobe Soundbooth to generate T_C, the computerized transcript. The first player to transcribe the clip is assigned to path A. He produces a transcript T_{1A}, which is then compared to $T_C$and scored accordingly. The next player is assigned to path B, and she produces a transcript T_{1B}, which is compared to $T_C$and $T_{1A}$and scored accordingly. The third player is then assigned to path A. He sees $T_{1A}$and modifies it to produce a transcript T_{2A}, which is then compared to $T_C$and $T_{1B}$and scored accordingly. Subsequent players are assigned alternately to paths Aand B, and they are shown the last two transcripts in the same path (as is available). This means that players assigned to path A never see transcripts from path B, unless the two paths converge, so they are completely unaware of what they are being scored against. No one ever sees T_C, and we shall see in Chapter 4 that this serves as a mechanism to ensure that the first two players have an incentive to enter a correct transcript rather than simply entering gibberish. To summarize our procedure mathematically:

(where n= k+1 2

n2 A

n2 B and/or $T_{n1\,B}$

n2 B

n B k 2

n A

n1 A

n2 B

n1 A

For odd k, Player kis assigned to produce transcript T_n A (where n=). ~~He~~ sees transcripts Tand $T_{n1\,A}$if availab
compared to transcripts Tand $T_{n1\,B}$(or to t
deemed too far off).

For even k, Player kis assigned to produce transcript T). He sees transcripts Tand $T_{n1\,B}$if available, and produ
compared to transcripts Tand $T_{n\,A}$(or to the last two pe
Thave been deemed too far off).

Note once again that because of the way this reward system is designed, the first accurate transcript will not necessarily be rewarded the maximum number of points, as it differs from the inaccurate one to which it is compared. This issue has been discussed previously, and we accept it as a sacrifice that we must make for the ability to award players points in real time.

Implementing the Reward System

In implementing a point-based reward system, we need a way to measure the similarity between a player’s output and that of previous players. One method commonly used to measure the similarity of two strings is to calculate the Levenshtein distance between the strings [15]. The Levenshtein distance measures the minimum number of edits (insertion, deletion, or substitution of a single character) needed to convert one string into another, also known as the edit distance. For example, SATURDAY and SUNDAY are a distance of 3 apart. To see this, we perform two deletions (of the first \A" and the \T") and a substitution (of the \R" for a \N") for a total of 3 edits: SATURDAY ! STURDAY !SURDAY !SUNDAY. While there are various other ways to measure the edit distance between strings, the Levenshtein distance measure is a very simple one and part of the PHP environment, so we use it for the sake of simplicity. (For the comparison of final transcripts, we turn to a more standard industry measure, known as Word Accuracy, which uses a similar calculation on a word level. In our analysis, accuracies calculated using both methods seem to produce very similar metrics.) Levenshtein distance treats uppercase and lowercase letters as different characters, and it does not ignore capitalization; thus, before comparing any two transcripts, we remove extra spacing, strip out punctuation, and convert the results to uppercase.

Once we have decided how to determine the similarity of two strings, we compare the newly entered transcript against the two most recent eligible transcripts from the opposite path. Because the later one should theoretically be more accurate, we weight it slightly more, with of the score coming from a comparison against the earlier transcript, and 1 of it coming from a comparison against the later transcript. (For two transcripts, 0:5; for one, = 1, meaning that we are effectively using the same transcript twice. In our experiment, = 0:4.)

Let us write down the score calculation algorithm in mathematical language. Let $L(T_i;T)$ be the Levenshtein distance between two transcripts T_i and T_j. Let T_{kj} be the transcript submitted by the kth player, while T_1 is the most recent eligible transcript submitted by a player on the opposite path and T_2 is the next most recent eligible transcript submitted by a player on the op posite path. The ultimate score awarded to the current player ($Score_k$) is calculated according to the following equations:

LD_k = ( )L(T_2;T_k) + (1)L(T_1;T_k) (3.1)

$Length_k$ = ( )(length(T

1;(nuo(110

1

2))+(1)(length(T

=min

Note that we have chosen to normalize the score between 0
minimum award of 1 point. When the transcripts match very

The calculated score is also used to determine whethe
ones that preceded it. We use the fact that computerized tra
accurate to arbitrarily decide upon a 50% tolerance, as we s
changes of more than 50% from one iteration to the next alo
thus deemed \eligible" for future comparison if $Score_k$ 5. All

0 points are awarded for blank entries. While it is argua
and rewarded for listening to blank clips and verifying tha
was abused when introduced in the original implementat
rewarded for this work. Additionally, to prevent users fro
gathering bonus points at the end of each round, markin
one's count of the number of clips processed.

1-10 points are awarded for all other entries, depending
two most recent eligible entries on the opposite path. Th
Levenshtein distance as detailed above.

25 points are awarded at the end of each round, which
creates an incentive to continue to play the game. The b
people have an incentive to continue to play the game, b
skipping through a round merely to collect the bonus poi

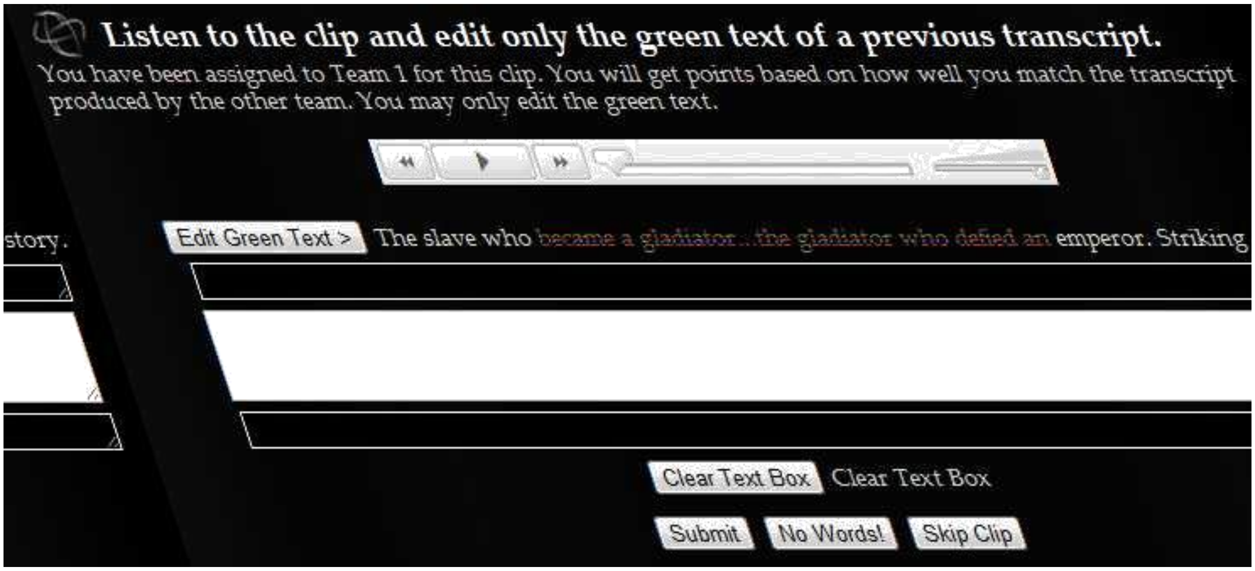
We will show in Chapter 4 that this results in a Perfect Bayesian Equilibrium in which all players enter the most accurate transcripts possible.

3.4 Preparing and Assigning Audio Clips

We have discussed the structure of the game; now let's take a look at its contents. If this project were to be implemented on a larger scale, audio clips would consist of submissions from parties who requested transcripts, or perhaps of certain transcripts that could later be used to train voice recognition software. However, given that the game is purely in its experimental stages, we will discuss the audio clips included here. Currently, clips are obtained from http://www.americanrhetoric.com/, a website that hosts audio clips and their corresponding transcripts. These clips consisted of audio segments from movies, speeches, and the like. Clips ranged in clarity; content matter; the degree to which they used uncommon words, proper nouns, and slang; and length. They were passed through Adobe Soundbooth CS4 (transcribed on High Quality, using American English) to produce a transcript for comparison. A list of clips used, alongside their Adobe Soundbooth transcriptions, are presented in Appendix A.

To make the game playable, audio clips were strictly divided into segments of ten seconds each (i.e. time t= 0 seconds to t= 10 seconds, t= 10 to t= 20, :::). Clips whose lengths were not multiples of ten seconds were simply broken down into ten-second clips with a shorter clip at the end; for example, a 12-second clip would be divided into t= 0 to t= 10 and t= 10 to t= 12. These \short clips" (i.e. anything ten seconds or fewer in length) were immediately available to all players for transcription.

To account for the fact that certain words would be spliced in half, we created 20-second \long clips" that consisted of two adjacent short clips combined (i.e. time t= 0 seconds to t= 20 seconds, t= 10 to t= 30, :::). These clips were not available for transcription at the beginning of the experiment; however, when a transcript had been agreed upon for each of the short clips that made up this longer clip, the longer clip became available. As a result, rather than being asked to generate a transcript for the entire 20-second clip, users were shown the final transcripts for each of the two short clips, but only allowed to change the middle portion of the transcript. This allowed us to combine the final transcripts easily later on, while still addressing the fact that certain words would be interrupted when clips were cut into pieces. These longer clips were presented only in the final version of the iterative game, and they were treated like the short clip for scoring purposes. Figure 3.4 shows a cropped screenshot of what users see when transcribing the twenty-second clip.



Listen to the clip and edit only the green text of a previous transcript.

You have been assigned to Team 1 for this clip. You will get points based on how well you match the transcript produced by the other team. You may only edit the green text.

story. The slave who became a gladiator...the gladiator who defied an emperor. Striking

Figure 3.4: Longer Clip Presented for Transcription. This 20-second clip consisted of two tensecond segments, the first of which was transcribed as \The slave who became a gladiator..." and the second of which was transcribed as \the gladiator who defied an emperor. Striking story." Users clicked on the \Edit Green Text" button to add this transcript into the textboxes below, with the light blue text being added to the top and bottom black textboxes, and the green text being added to the white textbox. Users could only edit the text in the white box, and the contents of the three boxes were combined to create a joint transcript.

Audio clips were randomly assigned to players, and player were only allowed to process a given clip once. Besides this, there were no restrictions on which clips players could see, but in the future, if there were sensitive clips that needed to be transcribed, rules could be built in to allow players to see certain clips only if they had not already seen others.

3.5 Creating a Final Transcript

In the iterative implementation, pieces were ultimately assembled to form a final transcription, though because the experiment had to conclude, this assembly was done in some cases before all clips had been fully processed. In theory, by the time transcripts are said to be finished, multiple players should have agreed upon a transcript, increasing the chances that it will be accurate. This is done for both the short clip and the long clip, so in effect, the transcript has undergone two comprehensive screening processes before being finalized.

As mentioned above, in the processing of the longer clip, we prevent players from changing the first and last parts of the joined transcript. This allows us to match the clips up, and it ensures that the middle portions of the clips can be joined together easily. To illustrate this, let's say we want to combine three short transcripts that are initially transcribed as abcx, yfgh, and ijkl. We join these together into two longer clips, providing starting transcripts abcxyfgh and yfghijkl , where underlined text represents what players can change. If, for instance, the word between the first two clips were cut off, there could be a mistake in which x should actually be d and y should actually be e. Players can correct the first long transcript to abcdefgh but cannot fix the error in the yfghij kl version, so they leave the latter alone. Still, when the transcripts are joined, changes are taken into account, to produce a final transcript abcdefghijkl. Repeating this process for the various clips thus enables us to create a larger final transcript without requiring users to listen to an entire clip. As such, we are able to take the smaller pieces produced by a number of players and create a larger transcript without forcing any one player to listen to the entire clip.

* * * * *

In this chapter, we have discussed the various design decisions that went into creating this game, described the original, parallel, and iterative implementations that were eventually tested live, and illustrated how final transcripts were generated. Additionally, for the iterative implementation, we have provided extensive detail regarding the reward structure. In the next chapter, we will analyze the original and iterative implementations of the game to show that the latter leads to a Perfect Bayesian Equilibrium in which actual game play results in all players exerting high effort to try to enter the most accurate transcripts possible.

Chapter 4

A Game-Theoretic Analysis

Having delineated the structures of each of the games, we will solve for the Perfect Bayesian Equilibrium in the original and iterative implementations. These equilibria represent strategies for each player such that each cannot improve his score by deviating from his current strategy, given that other players do not deviate from their strategies. We find that in the original implementation, it is a Perfect Bayesian Equilibrium for all players to enter trivial transcripts, and we find that in the final iterative implementation, it is a Perfect Bayesian Equilibrium for all players to copy previous players' entries when they are unrelated to the clip, to enter the most accurate transcript possible when they observe relevant transcripts, and to mark transcripts as correct when they believe them to be perfect. However, because the first player's dominant strategy in this game is to enter the most accurate transcript possible, we only see the latter two situations, and game play results in all players entering the most accurate transcripts possible.

Clearly ex-post, when players know what all of the other players have entered, these strategies may not be ideal, particularly in the iterative implementation. In this case, a player's best strategy is simply to copy the transcript to which he is being compared. This strategy minimizes the edit distance and maximizes the player's score, but needless to say, does not provide for a very accurate transcript. Luckily, the game is sequential, and players do not have full information about the contents of others' transcripts, so this is not a possible action. Thus, we focus on an ex-ante game-theoretic analysis.

4.1 Defining A Bayesian Game

For this analysis, rather than considering the entire Transcription Game to be a game-theoretic game, we consider a \game" as the actions of a set of players as they pertain to a specific clip. In particular, we define the following:

Players: There is a set of players P= f1;2;:::;Ngwho transcribe a specific clip. We assume that these players are all rational.

Types: Players are of certain types = f ; g; that is, they vary personally in some manner. We define two types differentiated on ability: \High" types ( ) who are more likely to produce

accurate transcripts given a set amount of effort, and \Low" types ( ), who are not as likely to produce accurate transcripts, even given the same amount of effort. By this definition,

> . We know that people are the \High" type with probability pand \Low" type with

probability 1 p.

Actions: There is a set of actions A= fA1;A2;A3gdescribing the possible actions that a player can take during his turn. These actions, available to both High and Low types, are as follows:

Action 1 (A_1): Exert zero effort (e_1= 0). This is equivalent to entering a trivial (and incorrect) transcript, or to copying what the previous player has entered. Cases in which players see a transcript that they believe to be correct and submit it without modification also fall into this category.

Action 2 (A_2): Exert low effort (e_2= e >0). This is equivalent to writing something other than your best guess of what the clip actually says. Entering an approximate guess about what you hear, or intentionally entering an incorrect transcript would fit into this category, so choosing this outcome is not always undesirable.

Action 3 (A_3): Exert high effort (e_3= e>e>0). Though you may not necessarily be correct, you enter your best guess of the most accurate transcript, possibly modifying others' transcripts along the way.

Strategies: Separate from actions, there are strategy sets Sthat each player can adopt, defining a player's actions in response to each possible combination of actions undertaken by previous players.

Information Sets: An information set His the set of possible moves that a player believes may have occurred by a given point in the game, based on the observed actions of previous players. Information sets evolve as game play progresses.

Beliefs: Players decide on a given strategy and action based on their beliefs, |their best guess about the true nature of unknown circumstances such as other players' types and actions, based on their current information sets. Players look at preceding players' actions to update their prior beliefs according to Bayes' Rule.

Payoffs: There is a payoff, or utility, given as the difference between the reward that the kth player earns from a given transcript and the effort he has put in to create that transcript. A player's reward is a measure of how closely his transcript matches those of previous players (here denoted by k); thus, it is a function of his action, his type (ability), and previous players' actions and types. Note that we assume that effort is fixed for each task, regardless of who is performing it. We define the utility that the kth player (type $_k$) receives from taking Action ias follows:

$$u_k(A_ij\ _k) = r_k(A_ij\ _k;A_k;\ _k)\ e_i \qquad (4.1)$$

We note that in each game, game play occurs occurs sequentially, with a single player playing at a time, and with each player's outcome affected only by the actions of previous players. Because players have incomplete information about previous players' types, characteristics, and actions, we call this a \Bayesian game."

Our goal now is to solve for the Perfect Bayesian Equilibrium in the original and final iterative implementations of the Transcription Game. A Perfect Bayesian Equilibrium is a strategy profile s , coupled with a set of beliefs such that

is optimal given his beliefs (based on his information set h_i

$$_i\ ) = \text{argmax}\ _iE\ _{i(xjh_i)}u_i(\ _i\ ;\ _ijh_i;\ _i;\ _i)\ 8i2P \qquad (4.2)$$

$_i(h_i$

1. A player's strategy $_i$ and the set of nodes xin said information set) and his opponents' strategies:

iwith some probability j(aj2. Player i's beliefs about Player j's type is always updated (based on the fact that player j takes action Aj j)) according to Bayes' Rule when applicable:

i(jjaj) = p(j) j(aj) ~ j2 jp(~ j) jj(ajj ~ j) (4.3)

In a Perfect Bayesian Equilibrium, players have no incentive to deviate from their optimal strategies i, given that other players play according to their optimal strategies i.

4.2 Modeling the Set of All Transcripts

We begin by examining the notion of accuracy, measured by the distance between transcripts. To do so, we create a model for the distribution of possible transcripts. Let Tdenote the set of all possible transcripts for a given clip, and let T be the single accurate transcript for said clip. All other Ti2T(Ti) belong to the set of incorrect transcripts that vary both in their degree of correctness and in the nature of the errors they contain. Geometrically, consider an n-dimensional space in which all points are denoted by transcripts. We assume that T lies in center of all these transcripts, as mistakes can be made at any part of a transcript, and these mistakes can be of all different natures. Thus, we consider all inaccurate transcripts to be some sort of variation on the most accurate transcript. The transcript space is illustrated in two-dimensional form in Figure 4.1.

To measure the accuracy of a given transcript, we can draw a vector ~ Dfrom T to each incorrect transcript. (In Figure 4.1, Tiis one such example.) D, the magnitude of ~ D, is determined

by the gravity of the mistake in the transcript (measured as the Levenshtein edit distance, in a sense), while^ D, the direction of~ D, can be thought of as the nature of the mistake such that two

different transcripts can both be a measure Daway from T but still differ in the mistakes they contain (different^ D). Our assumption that T lies in the center of all the transcripts comes from the assumption that mistakes of varying severity and type are made. (Indeed, empirical data shows that the correct transcript tends to take elements from all of the other transcripts, implying that the most accurate transcript lies somewhere between all of the incorrect versions.) This concept is illustrated in Figure 4.1.

In both the original and iterative implementations of the game, players are scored not based

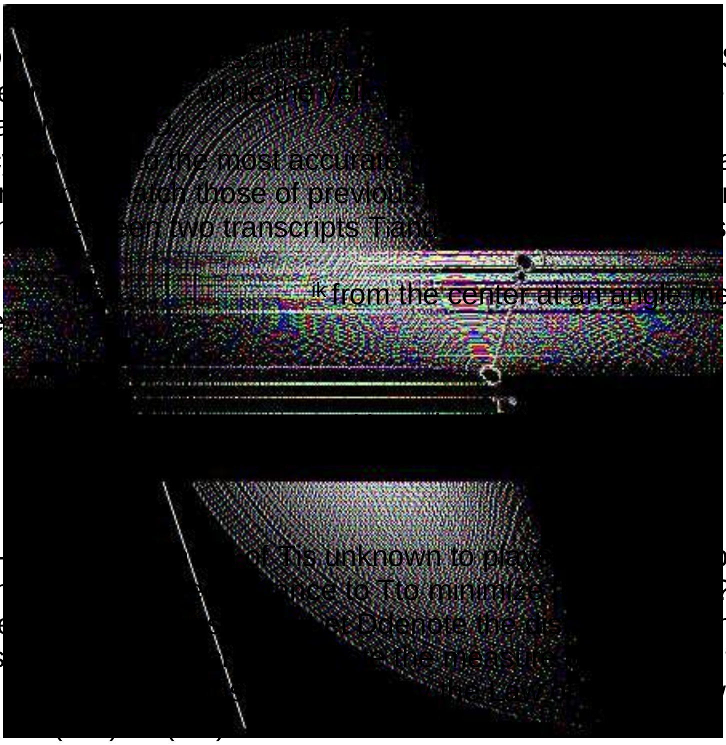


Figure 4.1: A Two-D Space. The blue dot
denotes the accurate ct transcript, located a
distance Daway, in a
on absolute accurac the most accurate ack this), but based on
how closely their tra those of previous nterested in
measuring the dista two transcripts Tiand , consider the case
where T_i

is located a distance ^{ik} from the center at an angle measure $_i$ ^{k} ^{k} , and $_k$
$_i$

above the positive x- of Tis unknown to play e producing transcript
Tand wants to minim nce to Tto minimize ? We begin by defining
the distance, Dbetwe Ddenote the di he kth player's
transcript to the mos the measu formed between the
vector pointing from ve have
$= D_i^2 + D_k^2 \; 2D_iD_k$

Because distance is never negative, we can square this quantity to make it integrable when we sum over all possible values of i. Expectations sum linearly, and the probabilities of D_{ik} and D^2_{ik}

) that minimizes $E[D^2_{ik}]$ also minimizes $E[D_{ik}$

$\frac{1}{2}$ $(D^2_i + D^2_k \; 2D_iD_k \quad \cos(_i \; _k)) \; d \; _i$ (4.5)

$E[D^2_{ik}] = Z_0^2$ $= D^2_i$

$= D^2_i$

occurring are identical, so the $(D_k; \; _k$]. We find this solution:

1 $(D_iD_k \quad \sin(_i \; _k)j \; _{i=2} \; _{i=0})$ (4.6)

(4.7)

$+ D^2_k$

$+ D^2_k$

Setting D_k = 0 minimizes the above expression. Because we have not restricted $D_i$in any way, this analysis applies to all situations, regardless of whether one knows the distance of the transcript one is being compared against. Though this is a two-dimensional proof, the same result should extend to higher dimensions by symmetry.

We show, therefore, that if we do not know the manner in which a transcript is inaccurate, then regardless of how inaccurate it is, we can minimize the expected distance from this clip by entering the most accurate transcript. This shows that players who are unaware of the transcripts that they are being compared against should maximize their expected reward by entering the most accurate transcript. This conclusion will be used in the analysis of the iterative implementation.

4.3 Effort, Ability, and Accuracy

While we have shown that players who are unaware of the contents of the transcripts they are being compared against should enter the most accurate transcript, they may not always be able to do so. We make the realistic and often-observed assumption that the accuracy of a player's transcript is affected by two factors: the player's effort and the player's ability. The former consists of a player's conscious decision to exert zero, Low, or High effort, as reflected by his action set; the latter is assigned by some force of nature and cannot be controlled by the player.

We assume that as players exert higher levels of effort, they generally produce more accurate transcripts, and we assume that High types have a greater probability of producing accurate transcripts than Low types. Formally, we assume that D, the distance between a transcript Tand T_i, has a probability density function f(D) that is monotonically decreasing and more concave for higher effort/ability than for lower effort/ability. In Figure 4.2, we see that this results in stochastic dominance, where $F_{High}(D)$ $F_{Low}(D)$ for all D(where High and Low are used both for effort and ability), and there exists some value of Dsuch that $F_{High}(D)$ >F(D). This means that as people exert more effort, their probability of producing an accurate transcript increases, and their expected distance from T_{Low}decreases. Similarly, High types have a greater probability of producing accurate transcripts, and their expected accuracy is higher. Note that this model accounts for the reality that even when players exert High effort or are High types, there is still a possibility that they will not be entirely correct. We do not try to compare ability and effort; we only note that people with High effort / High ability tend to do better than those with Low effort / Low ability, and those with High effort / Low ability or Low effort / High ability do somewhere in between. It is therefore hard to distinguish High types exerting Low effort from Low types exerting High effort.

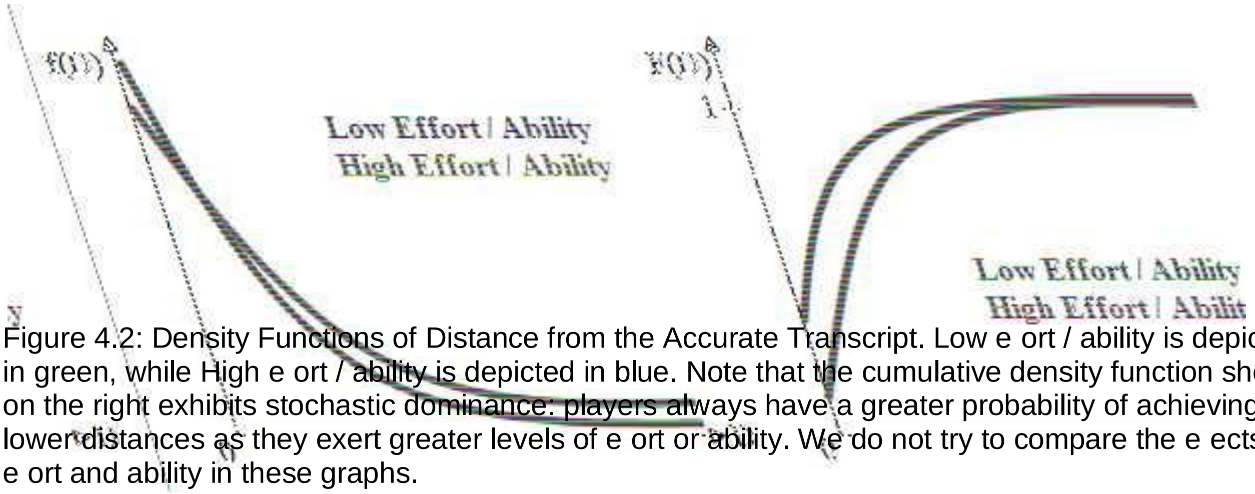


Figure 4.2: Density Functions of Distance from the Accurate Transcript. Low effort / ability is depicted in green, while High effort / ability is depicted in blue. Note that the cumulative density function shown on the right exhibits stochastic dominance: players always have a greater probability of achieving lower distances as they exert greater levels of effort or ability. We do not try to compare the effects of effort and ability in these graphs.

As a final note, because we tie increased reward to increased effort, but our utility function considers effort as a cost, there is the question of whether the gain in reward is perhaps negated by the cost of effort, meaning that certain players may be less inclined to exert high effort, even though it gives a better reward. Because we are designing the point system for this game, we can set rewards high enough such that when players are unaware of the transcripts they are being scored against,

the possible gain in utility a player receives from exerting extra effort eclipses the cost of putting in this effort, regardless of the player's type: $E[r_k(A_3 j\ _k;A_k;\ _k]e_3>E[r_k(A_2 j\ _k;A_k;\ _k]e>$
$E[r_k(A_1 j\ _k;A_k;\ _k]e_1 8\ _k 2f$; g. In future versions of this game, we could modify the reward system to make it convex (perhaps by squaring it) such that there is a clear benefit to putting in the extra effort. Empirical results backed by survey data show that players did not tend to sacrifice higher rewards simply because of the cost of effort. It seems that by the time players commit to playing the Transcription Game, they do not mind the slight increase in effort that it takes to enter an accurate transcript rather than an \approximately correct" one.

4.4 Analyzing the Original Implementation

Recall that in the original implementation of the game, players were given the four most popular transcripts (ordered by frequency) and scored half based on these transcripts and half based on the most recent transcript. Let T_1, T_2, T_3, and T_4denote the four most popular transcripts, and n, n_2, n_3, and $n_4$denote the frequencies at which they appeared respectively. Let T_{k1}be the transcript of the kth player and T_{k1}be the previous transcript. Recall that $L(T_i;T_j)$ is the Levenshtein distance between two transcripts. Scoring for the kth transcript is as follows:

Score$_k$ = 8 >> <> > :5 for k= 1max n2;10 $_{1\ 2}$ $L(T_{k1};T_k)$ + $_{4\ i=1} n_i L(T_i;T)$ $_4$ $_{i=1} n_{ik}$ o for k>1 (4.8)

To simplify our analysis of the kth player's actions, we group players 1 through k2 into a single hypothetical player. Rather than considering these k2 players, we consider our scoring function to be based only on the similarity of a player's transcript to the immediately previous player and to a group of more distantly previous players represented by our hypothetical player. We presume that transcripts that are displayed are similar in nature (at least two people must have entered the same transcript for it to be displayed), particularly as time goes on and more transcripts are entered, so this grouping generally makes sense. We will see that this generalization does not affect the outcome of our results.

Let us examine the behavior of the first player who listens to a given clip (k= 1). This player receives 5 points regardless of what he enters. His utilities are thus as follows: $u_1(A_1j\ ) = 5$, $u_1(A_2j\ _1)$ = 5 e, $u_1(A_3j\ _1)$ = 5 e, with $u_1(A_1j\ _1)$ >$u_1(A_2j\ _1)$ >$u_1(A_3j\ _{11})$ regardless of his type $_1$. Thus, his strategy is simply to take Action $A_1$and enter a trivial transcript (i.e. \a"), as this requires the least effort.

Now consider the case of the kth player (k > 1). Call this player \Player C," with the (k1)th player called \Player B" and the 1 through k2 previous players grouped as \Player A." (For k= 2, Players A and B are identical.) Players A and B are each either High types (with probability p) or Low types (with probability 1p). To start a game, Player A, who is of some type $_A$2f ; g, takes some unobservable action A_A. Player B, of type $_B$2f ; g, cannot observe this

action, but can use the transcripts he does see to update his beliefs about previous players (all but one of whom are grouped into what we call Player A). Player B takes another unobservable action accordingly. (Player B is congruent with Player C, just one time step back, so we will not analyze his incentives specifically.) Player C sees only the transcripts that are displayed (results of Player A and possibly, though not necessarily, of Player B as well); he does not explicitly know the actions of Players A and B. Judging from the transcripts he sees, however, he can tell if previous players entered trivial transcripts or non-trivial ones that appear to resemble the clip. Let us examine each of these situations separately.

$_A$In the case in which he sees a trivial transcript, Player C knows that Player A must have entered something trivial: A= A_{1B}. This does not give him any real insight into Player A's type, as both High and Low types could rationally choose to enter a trivial transcript as we shall soon see. Player B's actions are a little less obvious: if previous transcripts were displayed, they were likely trivial as well, so Player B would have entered something trivial. Otherwise, if nothing was displayed, Player B would have played as the first player who listens to a given clip does, entering something trivial as well. In both instances, A= A_1. Again, this gives little insight into Player B's type.

Despite lacking knowledge on previous players' types, Player C can still make a well-informed decision on which action to take. Players' types and actions simply give us information on what types of transcripts they will produce, and because Player C already sees these transcripts, previous players' types and actions are somewhat redundant. To maximize his reward, Player C must simply enter the most similar transcript to previous players' entries. Because he already sees these entries, all he must do is to copy them. Given this analysis, Player C concludes that $E[r_k(A_1 j\ {}_k;A_A;\ {}_A;A_B;\ {}_B)] > E[r_k(A_2 j\ {}_k;A_A;\ {}_A;A_B;\ {}_B)]$ and $E[r_k(A_1 j\ {}_k;A_A;\ {}_A;A_B;\ )] > E[r_k(A_3 j\ {}_k;A_A;\ {}_A;A_B;\ {}_{BB})]$, as copying existing trivial transcripts (which, as mentioned previously, are listed by frequency) means that one will be more likely to match previous players than if one entered something di erent. He also knows that $e_1<e_2<e_3$, so $E[u_k(A_1 j\ {}_k)] > E[u_k(A_2 j\ )]$ and $E[u_k(A_1 j\ {}_k)] > E[u_k(A_3 j\ {}_k)]$ 8 ${}_{kk}$2f ; g. Player C will thus choose to take Action 1 when he

$= A_1$

–

entered something nontrivial: A_A

^{B} ${}_B)]$ and $E[r_k(A_1$ $j\ {}_k;A_A;$ ${}_A;A_B$

${}_k(A_1 j\ {}_k;A_A;\ {}_A;A_B;\ {}_B)] > E[r_k(A_2 j\ {}_k;A_A;\ {}_A;A_B$

= A_1

elieves he observes previous players taking Action 1: A_C

; $_B$)] >

.bIn the case in which he sees a nontrivial transcr have

2fA_2;A_3g. Judging simply by how accurate Player C be can update his belief slightly: if he believes that the tra prior belief of the probability that Player A is a High typ more accurate transcripts, given a fixed effort level; if he accurate, he increases his prior belief of the probability are less likely to produce extremely accurate transcript

Once again, Player B's actions are not as obviou they were likely nontrivial as well, so Player B could ha entries, taken Action 2 and entered something vaguely best guess of the most accurate transcript. Because A greatest expected reward, we conclude that this is Pla which nothing is displayed, Player B would have playe clip does, entering something trivial. In both instances, transcript, but we can conclude that the rational course entries (A), so B's transcript should be similar (or ident Given this analysis and the fact that Player C cannot s transcript he entered, he should maximize his expecte players, whose transcripts are listed by frequency. Ent displayed transcripts decreases his similarity to these t transcript, assuming that Player B is rational. Thus, Ac reward than Actions 2 and 3: E[r;

$_k(A_1 j\ _k)] > E[u_k(A_2 j\ _k)]$ and $E[u_k(A_1 j\ _k)] > E[u_k(A_3 j\ _k)]$ 8 $_k$

$= A_1$

44

C

$[r_k(A_3 j\ _k; A_A;\ _A; A_B;\ _B)]$. Furthermore, effort costs are lowest for Action 1 (e_1

$< e_2$ $< e_3$

), so $E[u_2 f$ E; g. Player C wi

to take Action 1 even when he
Actions 2 or 3: A. (Because th
responses are to play Action 1
taken as long as all players ar

Our Perfect Bayesian Ec
strategy: to take Action 1 (ente
other players' actions. The first
fact, does worse do to the hig
strategy either, as deviation do
regardless of others' strategie
information about other player

We find in the original im
they are being compared agai
previous analysis, where we s
previous transcripts should ma
transcripts, here we find that p
Regardless of one's type, prev
we find that a player's domina
one's expected reward, given
effort. Thus, players' strategies
Under these circumstances, re
This early implementation suffe
matching previous players' en
transcript. Because the first pl
the unhappy Perfect Bayesian
transcription such as \a" and r
experiment, this turned out no
enter accurate transcripts. Thi
clear from

this analysis that for the final version of the iterative game in this paper, we must find a way to uncouple what a player is trying to improve from what he is being scored against.

4.5 Analyzing the Iterative Implementation

We begin our analysis of the iterative implementation by recalling the setup of the game. In the final iterative implementation, players are divided into two pathways in which they see responses generated by the last two players on the same pathway and are scored against the last two responses generated in the opposite pathway. Transcripts more than 50% in edit distance away from the opposite path are disregarded so as to avoid misleading or misscoring future players. While the idea is for players to improve on existing transcripts in the same path, seeing a transcript does not necessarily mean that the transcript is incorrect or that it does not match the opposite path, as four players (two in each path) must consecutively agree on the same transcript before it is deemed accurate.

The first two players see nothing, and their results are compared to the computerized transcript (and, in the case of the second player, to the first player's transcript). Subsequent players are then shown the transcripts of the last two players on the same path, and their entries are compared to the unseen transcripts of the last two players on the opposite path. For scoring (but not display) purposes, we treat the seed transcript as the 1 or 2 \players" preceding the earliest players. Let the kth player's transcript be denoted T, with the previous two players' transcripts being denoted $T_{k1}$and T_{k2k}respectively. Recall that a player's score, a function of the transcript he enters, is calculated as follows:

$Score_k(T_k)$ = min 1;round 10 1 ( )L(T_{k2};T_k) + (1)L(T_{k1};T) ( )(length(T_{k2k})) + (1)(length(T_{k1})) (4.9)

and T_j

k1

A player is able only to control T_k, his own transcript; T_{k2} $_{k1}$;T_k), the weighted

$_{k2}$;T_k

and is awweight such that 0:5, putting g

here L(T_i;T_j) is the Levenshtein distance between two transcripts T_i more accurate) eligible transcript on the o

and Tare fixed. Thus, he can maximize h

Levenshtein distances between his transcript and the previous two. His reward function, r(A), increases as the aforementioned weighted distance decreases. We have demonstrated previously that when he does not know $T_{k2}$and T_{k1i}, he should maximize his expected reward by trying to make his transcript as accurate as possible.

Let us analyze this in the game-theoretic framework we have set up. Again, we simplify matters by combining players. Let Player A be the two previous players on the same path, while Player B is made up of the two previous players on the opposite path, to whom Player C, our current player, is compared. Again, these groupings make sense because we expect the transcripts of the two players in each group A and B to be fairly similar, as players build on previous results.

We examine the behavior of the rst two players to listen to a clip. These players see nothing and know that they are being compared to some unknown computerized transcript, and possibly to another player's entry as well. They are unable to update their prior beliefs of the type of player (or the ability of the computer) that generated these transcripts, so they have no idea how accurate they are or what potential mistakes might be. We have shown that the more accurate a player's transcript is, the greater his expected reward. Thus, E[r_k(A_3j $_k$;$A_{Computer}$;] > E[r_k(A_2j $_k$;$A_{Computer}$; $_{Computer}$] >E[r_k(A_1j $_k$;$A_{Computer}$; $_{ComputerComputer}$]. We previously established that the extra e ort one exerts to take Action 3 is worth the potential extra reward, so E[u_1(A_3j)] > E[u_1(A_2j $_1$)] >E[u_1(A_1j $_{11}$)] 8 2f ; g, and both players choose Action 3, entering the most accurate transcripts possible. Now consider the case of the kth player (k>2). Again, we call this player \Player C," with

$_A$the previous two players on the opposite path called \Player B" and the previous two players on the same path called \Player A." Players A and B are each either High types (with probability p) or Low types (with probability 1p). To start a game, Player A, who is of some type 2f ; g,

takes some observable action A_A, and his transcript is displayed for Player C. Player B, of type $_B$2f ; g, cannot observe this action, but rationally makes a decision as well, based on what he

observes as the previous transcripts in his own path. (Player B is congruent with Player C, just one time step back, so we will not analyze his incentives speci cally.) Player C does not observe Player B's actions and sees only Player A's transcript; he does not explicitly know the actions of Players A and B. Judging from the transcripts he sees, however, he can draw a few conclusions.

If Player C sees a transcript vastly di erent from what he hears, he should be alarmed|Player

A must have entered something trivial or completely irrelevant (A_A 2fA_1;A_2
2fA_2;
A_3
B

$2fA_2;A_3$

g),

or he is simply very, very bad at this game (unlikely). This allows Player C to update his belief of Player A's type, and he increases his guess at the probability that A is a Low type. From this transcript, Player C can use the idea that transcripts are ignored if they are more than 50% in edit distance away from other transcripts to conclude that Player B's transcript must be somewhat similar to Player A's. Furthermore, Player C can use this to update his belief of Player B's type, and he increases his guess of the probability that B is a Low type. Because neither transcript resembles the contents of the clip, however, Player C is at a loss for what to do. In this case, the player takes Action 1 and copies Player A's transcript, as he knows it must be somewhat similar to B's, and changing the transcript dramatically will likely give him worse results. Thus, if players observe previous players playing Action 1 or Action 2 to an extreme (i.e. entering a wildly incorrect transcript), they should play Action 1.

Fortunately for everyone, the only way the above circumstance could occur is if people slowly evolved the transcript into an unrelated statement. Because this evolution would have to take place over two independent pathways, it is a very low-probability event, given that we assume that players on the two pathways do not communicate. Furthermore, because the first players on each path should rationally enter the most accurate transcripts, if we observe extremely inaccurate transcripts, then we can conclude that there must have been a point at which previous players chose Actions 1 or 2. We will now show, however, that when player observe that previous players have entered relevant transcripts, they have no incentive to play anything but Action 3. Thus, if players act rationally, we should never end up in a situation in which completely irrelevant transcripts are displayed.

When players see transcripts that resemble the contents of a given clip, we arrive at a more satisfactory conclusion: that they should exert High effort and enter accurate transcripts. Player C knows that Player A likely took Action 2 or Action 3 (Ag); the more accurate the transcript appears, the more likely it is that Player A took Action 3 or is a High type. Player C updates his beliefs accordingly. The accuracy of the existing clip does not give Player C very much information on the accuracy of Player B's transcript; it only tells him that the two transcripts should be fairly similar and that Player B likely took Action 2 or 3 as well (Ag). Still, Player C may be able to deduce that the accuracies of Player A's and Player B's transcripts should be comparable, given that they come at roughly the same points in the transcription process and have gone through approximately the same number of iterations. Player C knows that the two transcripts are unlikely to be identical (for clips are removed after two players on each path agree on the same transcript), but perhaps the accuracies are on par with one another. He can thus update his prior belief of Player B's type, increasing his guess at the probability that B is a High type.

We have previously shown that having an idea of the accuracy of a given transcript does not give a player any insight into what types of mistakes are made in that transcript; thus, despite this knowledge, we cannot match the transcript exactly. We found that in cases where a player knows that another transcript is off but does not know the nature of the mistake made, he maximizes his reward by correcting it: $E[r_k(A_3 j\ _k;A_A;\ _A;A_B;\ _B)] >E[r_k(A_1 j\ _k;A_A;\ _A;A_B;\ )]$ and $E[r_k(A_3 j\ _k;A_A;\ _A;A_B;\ _B)] >E[r_k(A_2 j\ _k;A_A;\ _A;A_B;\ _B)]$, so $E[u_k(A_3 j\ _k)] >E[u_k(A_{1B} j\ )]$ and $E[u_k(A_3 j\ _k)] >E[u_k(A_2 j\ _k)]$ 8 $_k$2f ; g, and Player C will choose to take Action 3: A_{Ck}= A_3. Finally, in cases in which the player believes Player A's transcript to be perfect, Player C
can conclude that Player A must have played Action 3. Still, Player C lacks insight into Player B's actions and thus, should enter the most accurate transcript possible. This can be done in two ways, by taking Actions 1 or 3, which are equal in reward: $E[r_k(A_1 j\ _k;A_A;\ _A;A_B;\ )] =$ $E[r_k(A_3 j\ _k;A_A;\ _A;A_B;\ _B)] >E[r_k(A_2 j\ _k;A_A;\ _A;A_B;\ _{BB})]$. However, because Action 1 takes less effort (e_1<e_3), $E[u_k(A_1 j\ _k)] >E[u_k(A_3 j\ _k)] >E[u_k(A_2 j\ _k)]$ 8 $_k$2f ; g, and Player C will

= A_1

hoose to take Action 1: A_C

, leaving the correct transcript untouched when he believes it to be accurate. Thus, w
when Player C observes previous players entering slightly incorrect transcripts by pl
(in a more conservative manner) or Action 3, he should play Action 3 (make it as ac
possible); when he observes a player taking Action 3 and entering perfect transcript
take Action 1 (leave it alone).

To verify this Perfect Bayesian Equilibrium, consider a deviation in each player's str
the first player deviates by entering something trivial or inaccurate, his expected rew
than if he takes Action 3 and enters the most accurate response. Thus, he has no in
deviate, and similar logic applies to the second player. Assuming that everyone else
the strategy delineated above, it does not make sense for the kth player to deviate i
If he sees an irrelevant transcript, taking Action 1 will guarantee him a certain level o
whereas

changing it would pose a risk of getting fewer points. If he sees a relevant but imperfect transcript, taking Action 1 gives a very low score, and taking Action 2 likely gives a lower score than Action 3. Finally, if he sees a perfect transcript, taking Action 2 rather than Action 1 decreases his reward, and taking Action 3 increases the amount of effort he spends to get the same number of points. Thus, we have shown that there is never an incentive for players to deviate from this strategy regardless of their types or of their beliefs of others' types. Their actual actions during game play depend only on their beliefs of the actions (but not the types) of previous players, which is hinted at in the transcripts they see.

All in all, we have shown that it is a Perfect Bayesian Equilibrium for the first player on each path to take Action 3, and for subsequent players to take Action 1 if they observe previous players entering irrelevant transcripts (Action 1 and an extreme form of Action 2), Action 3 if they observe previous players entering related transcripts (a conservative form of Action 2 or Action 3), and Action 1 if they observe previous players entering perfect transcripts (Action 3). The first set of circumstances leads to them updating their beliefs and increasing their guesses at the probability that previous players are Low types; the second and third set leads to them updating their beliefs and increasing their guesses at the probability that previous players are High types. Note, however, that despite our discussion of player types, knowing whether the previous player is a High or Low type does not matter for the purpose of deciding one's actions, as this information only gives a player insight into the accuracy of the previous transcript, rather that the exact nature of its contents. We conclude that player types are immaterial in this case, as all players share the same strategy. If all players act according to these strategies, real game play should result in all players playing Action 3 and entering or correcting transcripts most of the time, only playing Action 1 and marking a transcript as correct when they believe it is perfect. This results in players delivering the most accurate transcripts possible.

This analysis shows that the dual pathway structure of the iterative game leads to a Perfect Bayesian Equilibrium that encourages people to exert high effort to maximize the accuracy of their transcripts and to stop modifying transcripts when they are correct. The difference between the iterative implementation and the original one is that the former separates what people see from what they are compared against. Players' ignorance of the natures of the transcripts they are being compared to means that they maximize their expected reward by entering the most accurate submissions possible. As a result, we are confident that rational players will continue to modify the transcripts they are given to make them as accurate as possible, and that they will stop making changes when they come across a perfect transcript. Results from our implementation of the iterative game, as well as a post-experimental survey, showed that players generally did as we predicted, entering transcripts to the best of their abilities.

* * * * *

In this chapter, we have shown that while the original implementation of the game results in an unsatisfactory Perfect Bayesian Equilibrium in which players exert no effort and simply enter trivial transcripts, a switch to the dual pathway structure employed in the iterative implementation leads to a Perfect Bayesian Equilibrium in which players exert high effort to try to enter the most accurate transcripts possible, marking transcripts as correct when they believe them to be perfect. We now turn to the empirical results from our game implementation experiments to assess the accuracy, efficiency, and enjoyability of the parallel and iterative forms of our game.

Chapter 5

Empirical Work

Three separate versions of the game were implemented from February 10 to March 14, 2011 through the website http://hcs.harvard.edu/ bliemthesis. A link to the website was distributed widely to friends and classmates and sent across House lists with each new version of the game, primarily targeting the Harvard undergraduate population. As incentive, we provided the chance to win a $25 Amazon.com gift card, given to a person chosen at random, with each person's chances directly proportional to the number of total points accumulated during game play. To be eligible, players had to enter their email addresses when they registered. In total, 180 players registered, 147 of whom actually played a game. A total of 1740 valid (non-blank) transcripts were collected over the course of 168 games and more than 22 hours of game play across all three implementations.

5.1 Overview of Experimental Results

Experiment 1, the original implementation of the game, allowed players to see and edit up to four previous players' entries, with players scored according to how close players' transcripts were to previous entries. Experiment 1, which ran from February 10-25, 2011, consisted of ten audio files that were divided into 22 ten-second clips. It produced 80 games and 883 transcripts (an average of 11.0 per game), but the results of this experiment will not be analyzed here, due to the misaligned incentive structure of this game, as was demonstrated in Chapter 4. It is important to note, however, that a cursory glance at the transcripts produced in Experiment 1 reveals that players tend not to simply enter the most trivial transcripts possible, contradicting the results of our theoretical analysis. This implies that players are not strictly rational, and that there may be some other force at hand|for example, the fact that players tend to be more inclined to help with a senior thesis project than to try to game the system.

Experiment 2, the parallel implementation of the game, did not allow players to see what others were entering, and players' entries were scored randomly. Experiment 2, which ran from February 26 to March 6, 2011, consisted of ten audio files that were divided into 20 ten-second segments. Longer clips that spanned the gap between these short clips were not created for this experiment. Overall, Experiment 2 produced 40 games and 308 transcripts (an average of 7.7 per game). If we define convergence on a transcript to occur when two players randomly arrive upon the same transcript (ignoring capitalization and punctuation), we find that after an average of 15.4 transcripts per clip were submitted 17 of the 20 clips (85%) had converged. Across all transcripts, the total Word Accuracy was 93.6%.

Experiment 3, the iterative implementation of the game, employed the dual pathway structure detailed in the previous chapter, allowing players to see what others on the same pathway had entered, and scoring their entries based on the opposite pathway. Experiment 3, which ran from March 7-14, 2011, added 10 new audio files to the ones used in Experiment 2, for a total of 44 shorter ten-second clips and 25 longer 20-second clips that spanned these shorter clips, to be used for the purpose of addressing words that were cut off when clips were split. The experiment concluded before all of these clips were processed; thus, 17 of the longer clips were never processed. To preserve the integrity of our results, players who had processed certain clips in Experiment 2 were no longer allowed to process these same clips. Overall, Experiment 3 produced 48 games and 549 transcripts (an average of 11.4 per game). Defining convergence on a transcript to occur when four players (two from each path) agreed on a given transcript (again ignoring punctuation and capitalization), we found that 27 of them had converged. It is important to note, however, that the longer clips only become eligible for transcription after the transcripts corresponding to the two components of the longer clip converge, so we cannot compare this convergence rate to that of Experiment 2. Overall, the Word Accuracy in the iterative process was 96.6%.

Accurate transcripts for audio files used in this game can be found in Appendix A, alongside Adobe Soundbooth transcripts. Final transcripts from the parallel and iterative processes are also displayed alongside their Word Accuracy measures where available.

5.2 Comparing the Parallel and Iterative Processes

To test our hypothesis that the iterative process is a better prototype for this game than the parallel one, we must compare the accuracy, efficiency, and enjoyability of the game under both implementations. We find many of the same results as were found in Little et al.'s 2010 paper \Exploring Iterative and Parallel Human Computation Processes" discussed in Chapter 2.

5.2.1 Accuracy

To compare our results to industry figures concerning transcription accuracy, we used Word Accuracy, which is measured as a percentage and calculated on a word basis as follows:

NumberOfWordsInAccurateTranscript

Because Word Accuracy is calculated on a word basis, however, it is harder to implement when comparing many transcripts. ~~We used a variation on this, which we call Character~~ Accuracy. This metric computes accuracy using the Levenshtein distance as follows:

LevenshteinDistance = $\text{Insertions}_{\text{Char}} + \text{Deletions}_{\text{Char}} + \text{Substitutions}_{\text{Char}}$

WordAccuracy= 1 $\text{Insertions}_{\text{Word}} + \text{Deletions}_{\text{Word}} + \text{Substitutions}_{\text{Word}}$

CharAccuracy = 1

LevenshteinDistanceNumberofCharactersInAccurateTranscript

We find that in all cases tested, Character Accuracy and Word Accuracy were comparable.

Comparison to Accurate Transcripts

As previously discussed, we compared the final transcripts obtained from each process to the correct transcript of the clips (as given by the source of the clip) using Word Accuracy measures. As another standard for comparison, we transcribed the soundtracks using both Google Voice and Adobe Soundbooth, confirming that results were not nearly as accurate. As a result, no quantitative comparison was made, but Adobe Soundbooth transcripts are included in Appendix A.

The parallel process yielded an overall Word Accuracy of 93.6%, which was calculated by looking at all transcripts submitted for a given clip and finding the two most similar results. In cases of ties, the single version of the transcript with the highest frequency of occurrences was selected. These transcripts were aggregated to produce the overall Word Accuracy. It is important to note, in this case, that 6 of the 17 clips that converged (35%) ended up converging on multiple different transcripts, indicating that if we were to stop collecting transcripts when two matched, results would not necessarily be correct and would certainly vary based on the order in which transcripts were entered. Using the most popular transcript for each of these clips in our calculation for Word Accuracy is a best-case-scenario analysis of the accuracy. This method produced a 100% Word Accuracy for each of the 17 clips that converged, and 31%, 95%, and 96% Word Accuracies for the three that did not.

In the iterative process, the accuracy of the clips that had converged was 97.4%, compared to an average of 95.5% for those that had not. Combining all clips to produce final transcripts (and using the most recently entered transcripts for clips that had not yet converged), we found the overall Word Accuracy to be 96.6%. Given more time, however, this accuracy would likely have increased, as in many instances, errors came not in the middle of transcripts, but across breaking points between clips.

This analysis shows that overall, the iterative process seems to produce slightly more accurate results, particularly once we take into account the fact that it is easier for the parallel process to converge upon an incorrect transcript. It is important to note, however, that there are instances in which the parallel process produces matches with 100% accuracy, but the iterative process is unable to converge upon an accurate result. This indicates that users are, at times, misled by others' mistakes as suggested by Little et al. [13]. On the whole however, both processes appear to produce results that are somewhat comparable to what we find with professional transcription, and they exceed the accuracies of computerized transcription for non-trained voices.

Accuracy Distribution

To get a better idea of the average quality of the output from each process, we compare the distributions of the Character Accuracies of the transcripts. Figure 5.1 shows that the two processes appear to be very comparable at first glance.

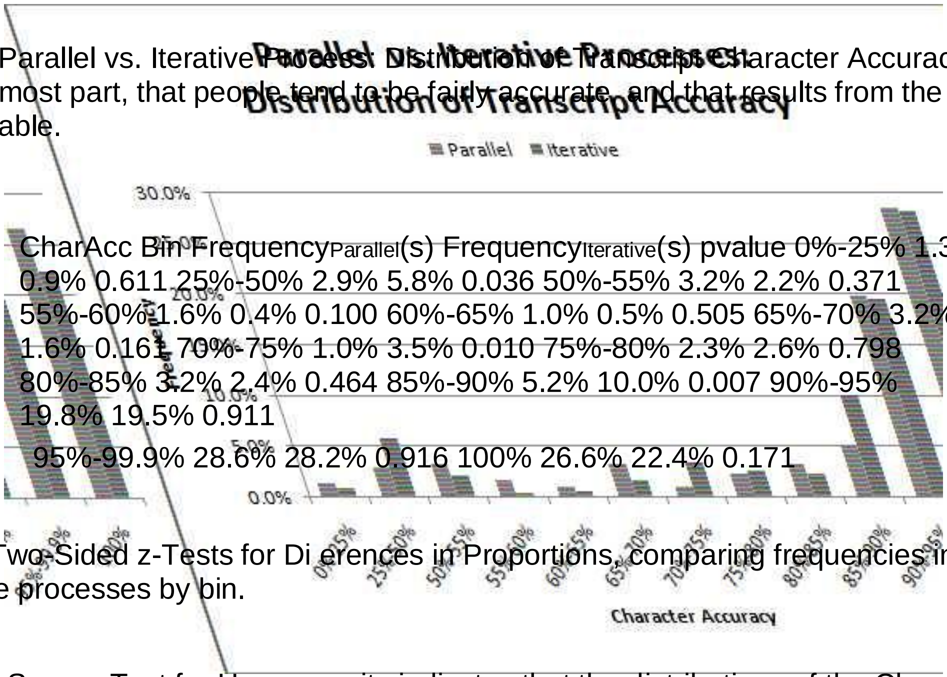


Figure 5.1: Parallel vs. Iterative Process: Distribution of Transcript Character Accuracy. We can see that for the most part, that people tend to be fairly accurate, and that results from the two processes are comparable.

CharAcc Bin	Frequency$_{Parallel}$(s)	Frequency$_{Iterative}$(s)	pvalue
0%-25%	1.3%	0.9%	0.611
25%-50%	2.9%	5.8%	0.036
50%-55%	3.2%	2.2%	0.371
55%-60%	1.6%	0.4%	0.100
60%-65%	1.0%	0.5%	0.505
65%-70%	3.2%	1.6%	0.161
70%-75%	1.0%	3.5%	0.010
75%-80%	2.3%	2.6%	0.798
80%-85%	3.2%	2.4%	0.464
85%-90%	5.2%	10.0%	0.007
90%-95%	19.8%	19.5%	0.911
95%-99.9%	28.6%	28.2%	0.916
100%	26.6%	22.4%	0.171

Table 5.1: Two-Sided z-Tests for Differences in Proportions, comparing frequencies in the parallel and iterative processes by bin.

A Chi-Square Test for Homogeneity indicates that the distributions of the Character Accuracies of transcripts in the parallel and iterative implementations are not the same (p-value=0.022), so we turn to multiple two-sided z-tests for differences in proportions to determine whether the differences for each bin are significant. If we break the data into two bins by various cutoffs (i.e. Below / Above 80% Character Accuracy), we find that there is no statistically significant difference between the accuracies of the two processes (p-value=0.795 for an 80% cutoff, p-value=0.986 for a 90% cutoff, etc.). Conducting similar tests by the bins shown in Figure 5.1, we find from a two-sided z-test that only three bins have statistically significant differences at the 5% level, as is shown in Table 5.1. Though we observe that the iterative process' performance in the 100% bracket is lower than that of the parallel process, this result is not statistically significant, and furthermore, it is biased, as additional transcripts were not created in the iterative process once a clip had converged.

To see what happens in a worst-case scenario, we select the three least accurate clips for each process (of those mutual to both experiments), which gives us a total of five clips, as one performed poorly under both conditions. Figure 5.2 shows the distribution of accuracies for these clips. We find once again that on average, the parallel process produces more low-accuracy results, though

Parallel vs. Iterative Processes:
Distribution of Transcript Accuracy for Worst Results
Parallel
Iterative
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Percent Total
Character Accuracy

Figure 5.2: Parallel vs. Iterative Process: Distribution of Transcript Accuracy for Worst Results. Data from the three worst clips for each process (five total) are shown. While differences between the two processes a [illegible] e process tend to be conce [illegible] transcripts that are very far o [illegible]

this time, it is unclear whether the iterative process gives more high-accuracy results. Combining the last two bins to create a 95%-100% bin, we find that the parallel process produces results in this range 31.4% of the time, compared to 30.8% for the iterative process. It is comforting to see, however, that the average quality of results in the iterative process is higher than in the parallel process, which suggests that people tend to correct transcripts that they find to be very far off.

Though this particular analysis shows no great difference between the accuracies of the transcripts produced in the parallel and iterative processes, our previous analysis of the final transcripts from each process suggests that while accuracy levels between the two processes are comparable, the iterative process appears to be slightly better.

5.2.2 Efficiency

To compare the efficiencies of the parallel and iterative processes, we examine both the time efficiency (amount of time it takes to enter a transcript) and the effort efficiency (average accuracy of the \best" transcript after a given number of iterations) in the two processes.

Time Efficiency

While the number of iterations that are necessary to converge upon a final result depends both on one’s definition of convergence (in terms of the number of people who must agree on a certain clip) and on the clip itself, we can analyze the amount of time a player spends producing a transcript for an individual clip. There are cases in which players may load a page, get distracted, and then return to the game to enter a transcript after a few minutes; these instances greatly increase the average amount of time we record as having been spent on a clip. As a result, we disregard the rare instances in which players spent more than two minutes on a clip.

On average, players appear to spend an average of 38.5 seconds during the parallel process, with a median of 32.5 seconds, compared to 36.6 and 28.0 seconds for the iterative process respectively. Combined, none of these distances are statistically significant (p-value=0.147, df=304; one-sided, unpaired, unequal variance). Running a t-test only on mean transcription times for the 20 clips that were in both the parallel and the iterative process, we find that the mean transcription time overall is 39.5 seconds in the parallel process, compared to 33.1 seconds in the iterative process (p-value= 0.0150, df=38; one-sided, paired, equal variance). Conducting a series of t-tests on the

Clip Name	Mean$_{Parallel}$(s)	Mean$_{Iterative}$(s)	pvalue	df
afewgoodmensymploce	49.7	51.0	0.469	11
barbarajordanscesisonomaton1	35.8	28.5	0.157	13
barbarajordanscesisonomaton2	38.2	23.2	0.041	15
barbarajordanscesisonomaton3	19.6	14.3	0.093	13
billysundayepistrophe	57.7	62.1	0.450	13
gladiatoranadiplosis1	32.4	27.0	0.118	52
jamesbibleanadiplosis1	50.2	69.5	0.002	23
jamesbibleanadiplosis2	36.0	19.4	0.012	24
jeffbridgesparadox1	50.5	68.2	0.208	18
jeffbridgesparadox2	34.9	25.2	0.098	14
johnfkennedyparallelism1	42.0	29.2	0.012	23
johnfkennedyparallelism2	50.0	32.4	0.006	25
johnfkennedyparallelism3	40.1	41.0	0.471	15
johnfkennedyparallelism4	57.0	26.2	0.001	16
johnfkennedyparallelism5	31.6	20.8	0.017	16
rockyhorrorpictureshowexpletive	32.3	18.0	0.013	14
stingscesisonomaton1	51.7	43.9	0.238	25
stingscesisonomaton2	27.6	35.8	0.138	18
stingscesisonomaton3	31.7	15.5	0.054	14
topgunassonance	21.8	11.5	0.028	15

Table 5.3: t-Tests on Individual Transcription Times by Clip for Clips in Parallel and Iterative Processes. One-Sided, Unpaired, Unequal Variance.

individual transcription times for each clip, we see that 9 of these clips yield significant results at the 5% level, with 8 of these indicating a lower average transcription time under the iterative process. Details of these tests are shown in Table 5.3.

Though we conclude what we expect to find|namely that in most cases, the iterative process takes less time to execute|this is important because it provides a possible explanation for the lower level of accuracy found in the parallel process. It is possible that because it takes more effort to transcribe a clip than to correct it, players are less willing to do so and stop halfway, transcribing only the beginning of a clip.

Accuracy after nIterations

More important than time efficiency, however, is some measure of effort efficiency. This can be interpreted either as how many transcripts it takes to converge upon an accurate result, or as how accurate a transcript is after niterations. While we can calculate metrics for the former definition, pitfalls arise if we converge upon an incorrect result or if we do not converge at all. Instead, we choose to measure accuracy of the \best" transcript generated after niterations such that we can derive some measure of accuracy per unit of effort exerted.

In the iterative process, we define the best transcript as the most recent one (as long as it has not been rejected for being wildly different from the opposite path). Looking back once we have the most accurate transcript, we may see that this algorithm does not always produce the most accurate result, but we generally do not know this until after the fact, as we assume that transcripts are improving iteratively. If a transcript on one path is modified from its previous form, its distance from the transcript to which the previous form was compared could theoretically increase. Still, this does not necessarily mean that the newest transcript is less accurate than before; it could have solved a problem common to the other two transcripts.

Our analysis for the iterative process excludes transcripts once they have converged, as including these clips would create an upward bias due to the greater accuracy of converged clips. Additionally, because it seems artificial to project accuracies for clips that never attained more than a certain number of iterations, we exclude these from our analysis after their last transcripts have been submitted. Retaining the rest of the data, we calculate Character Accuracy for each clip-group combination at each iteration. We see in Figure 5.3 that accuracy tends to increase with each iteration, though because sample sizes drop off as the number of iterations increases, this result is not always smooth.

In the parallel process, we need to define not only what the \best" transcript is, but also what an iteration is. As before, it is easy to choose the best transcript after collecting them all; however, at any point in the transcript collection process, your best guess is to take the two most similar transcripts. Note that this definition means that even if there are two inaccurate but perfectly matched entries, this could potentially be considered the best transcript. This phenomenon accounts for the sharp drop in the average accuracy of the transcripts for iteration 5 (Figure 5.3).

To simulate the time independence of the submissions of the transcripts in the parallel process, we randomly generated permutations of the transcript arrivals for each clip. We then looked at the first nclips and calculated the distance between all ntranscripts available at the nth step of the game, found the two most similar ones (breaking ties by random selection for ease of calculation), and calculated the average accuracy of these clips. As before, our analysis excluded transcripts once

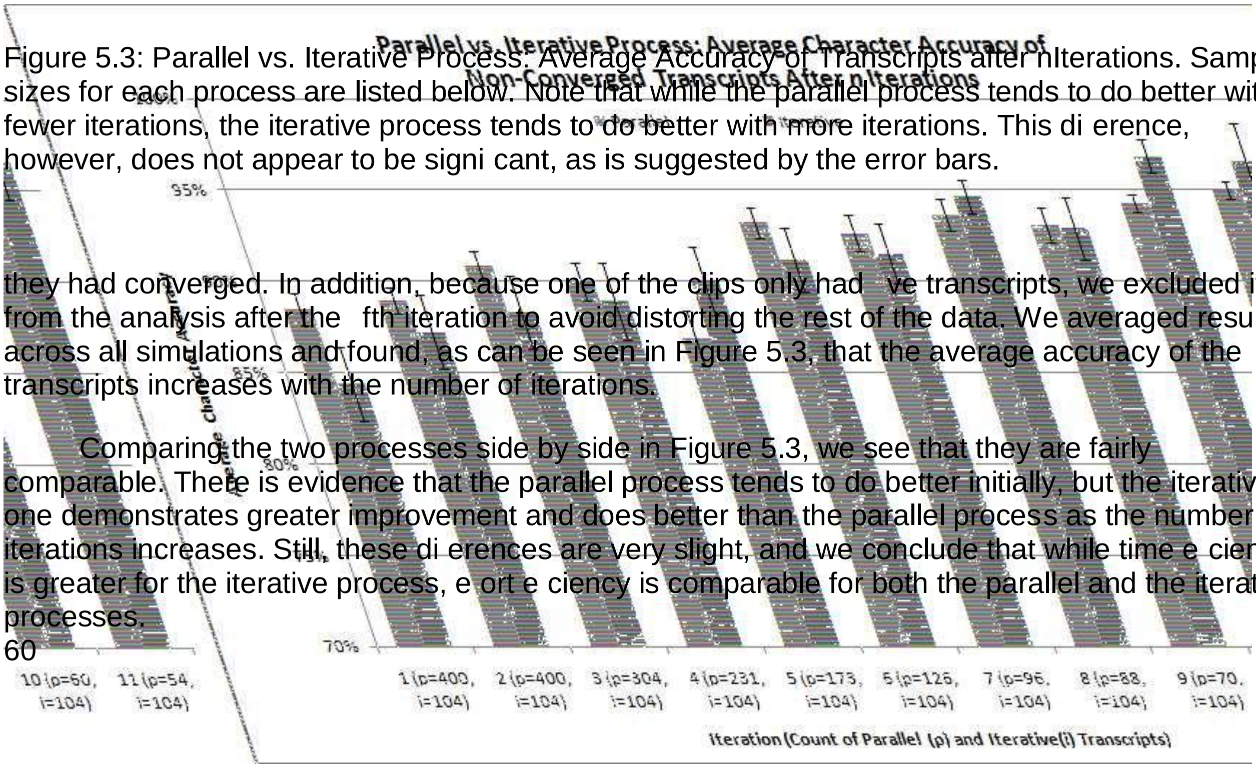


Figure 5.3: Parallel vs. Iterative Process: Average Accuracy of Transcripts after n Iterations. Sample sizes for each process are listed below. Note that while the parallel process tends to do better with fewer iterations, the iterative process tends to do better with more iterations. This difference, however, does not appear to be significant, as is suggested by the error bars.

they had converged. In addition, because one of the clips only had five transcripts, we excluded it from the analysis after the fifth iteration to avoid distorting the rest of the data. We averaged results across all simulations and found, as can be seen in Figure 5.3, that the average accuracy of the transcripts increases with the number of iterations.

Comparing the two processes side by side in Figure 5.3, we see that they are fairly comparable. There is evidence that the parallel process tends to do better initially, but the iterative one demonstrates greater improvement and does better than the parallel process as the number of iterations increases. Still, these differences are very slight, and we conclude that while time efficiency is greater for the iterative process, effort efficiency is comparable for both the parallel and the iterative processes.

5.2.3 Enjoyability

Looking first at the amount of time players spent playing each type of game, it seems that the iterative version of the game was more popular than the parallel one. Comments from a postexperimental survey answered by 17 players indicated that people found correcting clips more fun than transcribing them anew, and empirical results strongly supported this. Players spent an average of 6.0 minutes playing the parallel version, compared to an average of 9.6 minutes playing the iterative version. A t-test of game play times for the two games shows that this difference is significant (p-value=0.034, df=81; one-sided, unpaired, unequal variance).

To measure enjoyability of the game overall, we turn to survey results, with the caveat that because the survey was completed only by a few participants, we cannot generalize results to the population at large. Still, those who answered appeared to be fairly neutral with respect to the game: players had mixed feelings about whether they would play more if they had more time, and while some called the game \addictive," others said that they got bored of the game after playing for a while. It seems from these comments that while some enjoyed it, there is certainly room for improvement. Players suggested changes such as including more clips from interesting movies, adding a timer, making the game more interactive by revamping the score board, giving players feedback about their performance in terms of transcription speed and accuracy compared to others, and so on. While many of these features were previously decided against, they are certainly factors for reconsideration.

We conclude therefore, that while accuracy and efficiency were comparable for the parallel and iterative processes (with the iterative process perhaps slightly winning out), enjoyability measures overwhelmingly favored the iterative process. Because of this, the iterative process appears to be more promising than the parallel one, though there is room for improvement in all three aforementioned categories.

5.3 Players’ Strategies in the Iterative Process

Because we tend to favor the iterative process, let’s take a closer look at the nuances in player strategies and the nature of the errors that were made.

On average, players modified incorrect transcripts 63% of the time, increasing character accuracy 63% of the time by 16.1%, and decreasing it by 10.7% the rest of the time, for an average increase of 6.3% in character accuracy across all changes. Thus, players' edits did not always improve character accuracy. Still, character accuracy is based on Levenshtein distance, and changes in these distance may not necessarily be good indicators of accuracy in cases in which transcripts are very far off. For example, consider a garbled clip that says, \Kangaroos make great pets." One person may hear \Kangaroos migrate west," while the next hears \Kangaroos, unlike rats." Neither is particularly good, but the former has an edit distance of 9, while the latter has an edit distance of 10. Thus, if the second listener edited the transcript of the first, this change would have increased the edit distance and decreased the character accuracy without necessarily making the transcript much better or much worse off.

Turning to survey results for more insight into players' strategies, we find that most people claim to have always entered the most accurate transcript possible. Two of the 17 respondents admitted that when they had trouble understanding the words, they did not try as hard, but the rest asserted that they still tried to enter their best guess of what was accurate. People rated their efforts between 3 and 5 on a scale of 1 (no effort) to 5 (highest effort). While there may be sample bias in these survey results if more engaged players are the ones answering the survey, it seems that of the players who responded, most made a concerted effort to improve the accuracy of existing transcripts. Some even went so far as to transcribe beyond the length of the transcript, guessing words that were partially truncated or completing turns of phrase based on what they had heard. These results support our theoretical analysis, as we observe that players tend to exert High effort.

Despite these efforts, however, people made a few of the same errors repeatedly. Of these, some, such as misspelling names and switching words from plural to singular, would likely be made even by extremely accurate computer transcription software. Others errors, however, were uniquely human, with people writing down what they thought the audio file should say rather than what it actually said: they fixed subject/verb agreement errors, substituted a/the for one another, and inserted words such as \that." Additionally, they made changes to capture the tone or mood of certain clips, using slang such as \gonna" rather than \going to" in a clip from Pirates of the Caribbean or adding onomatopoeia such as \heh" into clips that contained laughter. (This last point arguably should not even be considered to be an error, depending on what one considers to be suitable material for transcription.) These types of mistakes decrease the probability that two independently evolving dual pathways would converge if players fail to correct these mistakes, and they further provide motivation for us to encourage players to correct existing transcripts rather than simply transcribe them.

5.4 An Evaluation of the Transcription Game

As discussed above, empirical evidence suggests that both the parallel and the iterative processes produce accurate results with comparable effort efficiency, though there is slight evidence that the iterative process performs marginally better in both categories. Players preferred playing the iterative process to the parallel process, volunteering this information as a comment on the game overall. We conclude, therefore, that future efforts should be focused on improving the iterative process.

As it stands, accuracy in the iterative process was fairly high, and would likely have been higher if players had had the opportunity to complete the processing of the clips. What should be improved, therefore, is the effort efficiency of the game, or the rate at which transcripts are improved. Rather than being instructed to correct mistakes in existing transcripts, players were asked to enter correct versions, with the option to add existing transcripts to a textbox for editing. The emphasis, it seems, should be on maximizing improvement (and thereby accuracy), rather than trying to maximize accuracy directly.

Finally, it seems that for a game that was designed to be enjoyable, actual levels of enjoyability are lackluster. Players' suggestions to make the game more interactive or social should certainly be taken into account in future instances of this game.

Chapter 6

Discussion and Conclusion

At the beginning of this thesis, we established two goals: 1) to design and analyze a game that people can play to transcribe audio clips for a low cost and a high accuracy comparable to professional transcription, and 2) to determine whether a parallel or iterative implementation of the game is more effective in terms of accuracy, efficiency, and enjoyability. We succeeded in both of these endeavors, creating a game that produces results that are 96.6% accurate, and we found support for our hypothesis that the iterative implementation of the game was more successful than the parallel one. From our experiments, we also concluded that future work should focus on improving the iterative process, possibly by setting improvement, rather than accuracy, as the main goal. In addition to this, further efforts should be made to increase the level of enjoyability of the game, perhaps by making it more interactive or social. We conclude this thesis by discussing these two areas of future work.

6.1 Awarding Players for Improvement

As we have seen, there is strong evidence that enhancing the overall accuracy and efficiency of the game relies on targeting improvement by encouraging players to correct transcripts. Errors seen in our experiments tended to be those that were made when players wrote down what they thought they heard, rather than what they actually heard. These errors included plural-versus-singular mistakes and subconsciously correcting subject-verb agreement errors in audio clips. If players had focused on correcting these errors, it is likely that they could have caught these mistakes and generated transcripts that were more faithful to reality. An emphasis on correcting mistakes in previous transcripts could be created by changing the

reward system from one that awards players points based on accuracy to one that awards them points based on the number of changes they make and whether these changes bring the transcript closer to or further from transcripts in the opposite path. Such a change ensures that players do not achieve some baseline score simply by leaving a transcript alone; they must exert the same level of effort as the previous player in order to get the same number of points. This institutes a greater degree of fairness in the game, where the nth player does not automatically receive a higher score than the first player simply by coming later in the transcription process. Furthermore, it motivates players to maximize the number of points they receive per unit of effort exerted, which helps our measure of efficiency.

To implement such a system, however, we need to address the question of how to measure improvement. If both pathways are making the same mistake at a given point in time, and a player on one pathway corrects this mistake, this is considered a deviation from the opposite pathway and is thus penalized in scoring. Though the player does not know this while entering the transcript, we should explore ways to avoid marking such changes as inaccurate.

Additionally, this scoring system raises the question of how to measure honest effort when a player enters a transcript. Because once again, we do not know what is accurate until we arrive upon a converged transcript, it is difficult to separate those who are simply injecting garbage into the transcription process from those who are genuinely making changes to improve it. Once again, we could use a second player to vote on the correctness of a transcript, but this brings up the problem of delayed gratification and how to align incentives for the second player.

Finally, we want to ensure that no matter how we measure honest effort and improvement, this measure focuses not only on the quantity of changes made, but also on the quality. In cases where there are many easy-to-correct mistakes in a transcript, along with a single difficult-to-correct mistake, we need to devise a fair system that rewards people more for correcting the difficult mistake than for correcting an easy mistake. While this could plausibly be done by counting the number of iterations that have passed before something is changed, we do not want to open the door for malicious players to make changes to a part of the transcript that people have deemed to be correct simply because such a change is perceived as a fixing a difficult-to-correct mistake and rewarded accordingly. Thus, while there is a clear need for us to consider ways in which to reward players for corrections and improvements, a solution is not immediately obvious and will be a large consideration for future work.

6.2 Increasing the Level of Enjoyability

The other area for improvement lies in levels of enjoyability: we need to make this game more fun. It seems, from survey results, that people playing this game craved increased levels of interaction, whether this was with the game itself or with other players. In the spirit of promoting efficiency, we could include a timer and ask players to make as many corrections to a given transcript as they can before time runs out. In this way, we do not compromise the integrity of our results, as we are no longer striving for 100% accuracy in each iteration, but rather for a marked improvement in accuracy levels which we hope will eventually translate into 100% accuracy. Such a timer would also allow players to feel more engaged, due to increased competitive pressures as they race against a clock.

In addition to this, we could provide players with feedback about how their performance compares to others', letting them know how high the scores they achieved for each clip are relative to the scores that others obtained. By tapping into peoples' competitive spirits, we would be able to foster a greater sense of social interaction and motivate people to perform better.

Finally, it is possible that we could choose to move away from the single-player game entirely, returning to the two-player format common to most of the GWAP games. Though this would be a radical departure from the current version of the Transcription Game, such a change would certainly increase the level of social interaction and could allow us to solve some of the problems brought up in the previous section. Having two players in a game could give us a way to check whether players agree on which parts of various transcripts are incorrect and the manners in which they should be fixed, but we would have to be careful to avoid a situation in which both players mark everything as incorrect and change it to a blank transcript. The implementation of a two-player game thus requires further thought, but promises to increase social interaction and with it, enjoyability.

* * * * *

In conclusion, while the Transcription Game was not an instant success, it provides solid groundwork for future improvements. The iterative form of the Transcription Game provides a dual pathway structure that helps us overcome the obstacle of being unable to determine which transcripts are accurate. Furthermore, its current incentive structure successfully results in a Perfect Bayesian Equilibrium in which all players enter the most accurate transcripts possible, marking transcripts as correct when they believe them to be perfect. Our tests of both the parallel and iterative forms of this game confirms our hypothesis that an iterative implementation is more effective than a parallel implementation in terms of accuracy, efficiency, and enjoyability. Future efforts will be focused on emphasizing improvements rather than accuracy, as well as increasing the degree of enjoyability the game offers. We assert, therefore, that this thesis offers a unique low-cost, high-accuracy approach to transcription that, though somewhat unrefined, demonstrates potential to eventually become a widespread method by which transcription can be executed.

Appendix A

Transcripts of Clips and Word Accuracy

Iterative Process: Overall Word Accuracy = 96.6%.
Parallel Process: Overall Word Accuracy = 93.6%.

Notes: Transcripts from the iterative and parallel processes are provided whenever possible. Adobe Soundbooth and accurate transcripts are also provided. Clips containing a \P" before the final number indicate that they are longer 20-second clips. For transcripts obtained through the iterative process, the latest version is recorded in cases of non-convergence. For transcripts obtained through the parallel process, the two versions that are closest to one another are provided. Word Accuracy (WAcc) is calculated as an average for these clips.

afewgoodmensymploce Accurate Transcript You don't want the truth because deep down in places you don't talk about at parties, you want me on that wall { you need me on that wall.

Iterative Result (100% WAcc)	You don't want the truth because deep down in places you don't talk about at parties, you want me on that wall, you need me on that wall.
Parallel Result (100% WAcc)	you don't want the truth because deep down in places you don't talk about at parties, you want me on that wall, you need me on that wall!!!

Adobe Soundbooth because you can't use is don't talk about the party's you what Leon will he be on the wall

barbarajordanscesisonomaton1 Accurate Transcript Let there be no illusions about the difficulty of forming this kind of a national community. It's tough, Iterative Result(88.9% WAcc)Let there be no illusions... about the difficulties of forming this kind of national community. It's tough...

Parallel Result (100% WAcc)	let there be no illusions, about the difficulty of forming this kind of a national community, its tough!

Adobe Soundbooth go to the NATO peace no you don't shoot we are going to find the great national community it's top

barbarajordanscesisonomaton2

Accurate Transcript difficult, not easy. But a spirit of harmony will survive in America only if each of us remember-

Iterative Result (100% WAcc) Difficult, not easy, but a spirit of harmony will survive in America, only if each of us remember

Parallel Result (100% WAcc) difficult, not easy, but a spirit of harmony will survive in America, only if each of us remember

Adobe Soundbooth it all not to say the radar I mean what's the bottom up we gave each of us read numbers

that we share a common destinybarbarajordanscesisonomaton3 Accurate Transcript -s that we share a common destiny Iterative Result (100% WAcc)

Parallel Result (100% WAcc) that we share a common destiny

Adobe Soundbooth we share a common destiny

billysundayepistrophe

Accurate Transcript Booze sold to a preacher or a high school girl has the same effect as when it's sold to an automobile thief, or a horse thief.

Iterative Result(69.2% WAcc) whose soul to a preacher or a high school girl has the same effect as my soul to an automobile piece or a heart's beat.

Parallel Result (31.0% WAcc) ... where I high school girl has the same effect ... / a highschool girl has the same effect as my soul to a

Adobe Soundbooth the real quick Hi Lou Saban said for more King gladiatoranadiplosis1

They call for you. The general became a slave.Accurate Transcript They call for you: The general who became a slave; Iterative Result(90.0% WAcc)

Parallel Result (100% WAcc) they call for you, the general who became a slave

Adobe Soundbooth the he said s gladiatoranadiplosis2

Accurate Transcript the slave who became a gladiator; the gladiator who defied an Emperor. Striking story.

Iterative Result(85.7% WAcc) Slave who became a gladiator, gladiator who defied an emperor striking story.

Adobe Soundbooth the story jamesbibleanadiplosis1

Accurate Transcript But every man is tempted when he is drawn away of his own lust and enticed. Then, when lust has conceived, it bringeth forth sin.

Iterative Result(96.0% WAcc) But every man is tempted when he is drawn away of his own lust, and enticed. Then, when lust hath conceived, it bringeth forth sin.

Parallel Result (100% WAcc) but every man is tempted when he is drawn away of his own lust and enticed, then when lust hath conceived, it bringeth forth sin

Adobe Soundbooth you're going to win it all and she is going for him jamesbibleanadiplosis2

Accurate Transcript And sin, when it is finished, bringeth forth death.

Iterative Result (100% WAcc)	and sin, when it is finished, bringeth forth death
Parallel Result (100% WAcc)	And sin, when it is finished, bringeth forth death.
Adobe Soundbooth	continued for that
je bridgesparadox1	
Accurate Transcript	There was a little boy who didn’t know if he wanted to be born. His mommy didn’t know if she wanted him to be born either. They lived in a cabin, in the wood-
Iterative Result (100% WAcc)	There was a little boy, who didn’t know if he wanted to be born. His mommy didn’t know if she wanted him to be born either. They lived in a cabin in the wood
Parallel Result (100% WAcc)	There was a little boy, who didn’t know if he wanted to be born. His mommy, didn’t know if she wanted him to be born either. They lived in a cabin, in the woods.
Adobe Soundbooth	more you didn’t know if he wanted to do more this month no one is going to warn you that they lived in the cabin and awards
je bridgesparadox2	
Accurate Transcript	-s, on an island, in a lake, and there was no one else around. And in the cabin { there was a door in the floor.
Iterative Result (100% WAcc)	On an island in a lake and there was no one else around and in the cabin there was a door in the floor.
Parallel Result (100% WAcc)	On an island, in a lake, and there was no one else around, and in the cabin, there was a door in the floor.
Adobe Soundbooth	on an island in the lake there was no one else around the cabin there was a door and walk
johnfkennedyparallelism1	
Accurate Transcript	For the Ireland of 1963, one of the youngest of nations and the oldest of civilizations, has discovered that the achievement of na-
Iterative Result(95.7% WAcc)	For the island of 1963, one of the youngest of nations and the oldest of civilizations has discovered that the achievement of na-
Parallel Result (100% WAcc)	For the island of 1963, one of the youngest of nations, and the oldest of civilizations, has discovered that the achievement of
Adobe Soundbooth	for the island of nineteen sixty three one of the youngest termination and the oldest civilization then discovered that the achievement of nation or
johnfkennedyparallelism2	
Accurate Transcript	-tionhood is not an end but a beginning. In the years since independence, you have undergone a new and peaceful revolution,
Iterative Result (100% WAcc)	Nationhood is not an end, but a beginning. In the years since independence, you have undergone a new and peaceful revolution.
Parallel Result (95% WAcc)	-ationhood is not an end but a beginning. In the years since independence, you have undergone a new and peaceful revolution / Nationhood is not an end but a beginning. In the years since independence, you have undergone a new and peaceful revolution.

Adobe Soundbooth it's not an end but a beginning in the year since independent you're about to go on a new and

johnfkennedyparallelism3 Accurate Transcript An economic and industrial revolution, transforming the face of this land while still holding to the old spiritual and cultural values. Iterative Result (100% WAcc)An economic and industrial revolution transforming the face of this land, while still holding to the old spiritual and cultural values

Parallel Result (100% WAcc) an economic and industrial revolution, transforming the face of this land, while still holding to the old spiritual and cultural values.

Adobe Soundbooth peaceful revolution an economic and industrial revolution transforming the face of this land while still holding daily or spiritual and cultural value

johnfkennedyparallelism4 Accurate Transcript You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your faIterative Result (100% WAcc)You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your fa-

Parallel Result (100% WAcc) you have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your

Adobe Soundbooth you have a long night your economy punish your honor diversified here with us they have brought your tray electrified your mom

johnfkennedyparallelism5 Accurate Transcript -rms, accelerated your rate of growth, and improved the living standards of your people. Iterative Result (100% WAcc)Accelerated your rate of growth and improved the living standard of your people

Parallel Result (100% WAcc) Accelerated your rate of growth, and improved the living standard of your people.

Adobe Soundbooth accelerating a rate of growth and improve the living standard of your people i would like, if i may, to take you, on a strange journeyrockyhorrorpictureshowexpletive Accurate Transcript I would like, if I may, to take you on a strange journey. Iterative Result (100% WAcc)

Parallel Result (100% WAcc) I would like, if I may, to take you on a strange journey.

Adobe Soundbooth I would lie and they did two JQ a straight actor

stingscesisonomaton1

Accurate Transcript But four years ago Jimmy Swaggart said this about me. He said, \This here song by The Police, 'Murder by Numbers', was written

Iterative Result(95.7% WAcc) But four years ago Jimmy Swagger said this about me. He said \This here song by the Police, 'Murder by Numbers,' was written."

Parallel Result (96.0% WAcc) four years ago, Jimmy Swaggart said this about me, he said: \this here song by the police, Murder by Numbers, was written" / Well four years ago, Jimmy Swaggart said this about me. He said, "This here song by The Police, 'Murder by Numbers,' was written..."

Adobe Soundbooth the following stingscesisonomaton2
by Satan, performed by the sons of satan, beelzebubAccurate Transcript by Satan, performed by the Sons of Satan { Beelzebub, Iterative Result (100% WAcc)

Parallel Result (100% WAcc) by satan, performed by the sons of satan. Beelzebub

Adobe Soundbooth a it was stingscesisonomaton3
Lucifer, the horned one.Accurate Transcript Lucifer, The Horned One. Iterative Result (100% WAcc)

Parallel Result (100% WAcc) Lucifer! The horned one.

Adobe Soundbooth hard and topgunassonance
Accurate Transcript I feel the need, the need for speed. Iterative Result (100% WAcc)I feel the need. The need for speed.

Parallel Result (100% WAcc) I feel the need, the need, for speed.

Adobe Soundbooth I'm sure many they could do
barbarajordanscesisonomatonP2
Accurate Transcript di cult, not easy. But a spirit of harmony will survive in America only if each of us remembers that we share a common destiny

Iterative Result (100% WAcc) Di cult, not easy, but a spirit of harmony will survive in America, only if each of us remembers that we share a common destiny

Adobe Soundbooth it all not to say the radar I mean what's the bottom up we gave each of us read numbers we share a common destiny

je bridgesparadoxP1 Accurate Transcript There was a little boy who didn't know if he wanted to be born. His mommy didn't know if she wanted him to be born either. They lived in a cabin, in the woods, on an island, in a lake, and there was no one else around. And in the cabin { there was a door in the oor.

Iterative Result (100% WAcc) There was a little boy who didn't know if he wanted to be born. His mommy didn't know if she wanted him to be born either. They lived in a cabin in the woods on an island in a lake and there was no one else around. And in the cabin, there was a door in the oor.

Adobe Soundbooth more you didn't know if he wanted to do more this month no one is going to warn you that they lived in the cabin and awards on an island in the lake there was no one else around the cabin there was a door and walk

johnfkennedyparallelismP3

Accurate Transcript An economic and industrial revolution, transforming the face of this land while still holding to the old spiritual and cultural values. You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your fa-

Iterative Result (100% WAcc) An economic and industrial revolution transforming the face of this land, while still holding to the old spiritual and cultural values. You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your fa-

Adobe Soundbooth peaceful revolution an economic and industrial revolution transforming the face of this land while still holding daily or spiritual and cultural value you have a long night your economy punish your honor diversified here with us they have brought your tray electrified your mom

johnfkennedyparallelismP4 Accurate Transcript You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your farms, accelerated your rate of growth, and improved the living standards of your people.

Iterative Result (100% WAcc) You have modernized your economy, harnessed your rivers, diversified your industry, liberalized your trade, electrified your farms, accelerated your rate of growth, and improved the living standard of your people

Adobe Soundbooth you have a long night your economy punish your honor diversified here with us they have brought your tray electrified your mom accelerating a rate of growth and improve the living standard of your people

Patience Iago, patience.aladdindiacope1 Accurate Transcript Patience, Iago, patience. Iterative Result (100% WAcc)

Adobe Soundbooth a student of the time

bazluhrmannanalogy1

Accurate Transcript Don't worry about the future; or worry { but know that worrying is as effective as trying to solve an algebra equation by chewing bubble gum.

Iterative Result (100% WAcc) Don't worry about the future. Or worry, but know that worrying is as effective as trying to solve an algebra equation by chewing bubble gum.

Adobe Soundbooth D don't worry about anyone well you know why it is trying to slow down to the nation by

bladerunnersententia1 Accurate Transcript We're not computers, Sabastian, we're physical. I think, Sabastian, therefore, I am. Iterative Result(96.2% WAcc)We're not computers, Sebastian, we're physical. I think, Sebastian, therefore I am.

Adobe Soundbooth welcome to this action yes

gonewiththewindepizeuxis1

Rick, Rick! Rick, if you go, where shall I go? What shall I do?Accurate Transcript Rhett, Rhett, Rhett! If you go, where shall I go? What shall I do? Iterative Result (100% WAcc)

Adobe Soundbooth I the the the the the U

jeremyrifkinallusion1

Accurate Transcript And finally you're all familiar with Dr. Wilmut's cloned sheep. We actually missed the real story behind this. Were so interested in talking about when this will happen with humans. (And, by the

Iterative Result (100% WAcc)	and finally you're all familiar with Doctor Wilmouth's cloned sheep. We actually missed the real story behind this. We're so interested in talking about when this will happen with humans, and by the w-

Adobe Soundbooth you're all familiar with the overwhelming reaction to the outside we're interested in how wonderful it would have been either way

jeremyrifkinallusion2 Accurate Transcript way, if we haven't already done it somewhere, the cloning of a human being is likely anytime. It's no longer a theoretical issue; it's just a question of who's going to do it.)

Iterative Result (100% WAcc)	If we haven't already done it somewhere, the cloning of a human being is likely any time. It's no longer a theoretical issue, it's just a question of who's going to do it.

Adobe Soundbooth we have already gotten somewhere calling it human being is likely to be the rapidly due to the question to you

jeremyrifkinallusion3 Accurate Transcript The real story behind the sheep is that Dr. Wilmut created the prototype for bioindustrial design. He's the Henry Ford of the Biotech Century.

Iterative Result (100% WAcc)	The real story behind the sheep is that Dr. Wilmut created the prototype for bioindustrial design. He's the Henry Ford of the Biotech Century.

Adobe Soundbooth the real story behind that she did not do well and created a prototype even Henry Ford

jeremyrifkinallusion4 Accurate Transcript It is now possible to replicate in countless numbers exact copies of an original living creature with the same kind of qualIterative Result (100% WAcc)It is now possible, to replicate in countless numbers exact copies of an original living creature, with the same kind of qual-

Adobe Soundbooth it is possible you ran a good solid number yes when you think readers quality

jeremyrifkinallusion5 Accurate Transcript -ity controls and engineering standards we did using mass production and assembly line factory work with inert materials. That's what's so important about this an-

Iterative Result (100% WAcc)	quality controls and engineering standards we did using mass production and assembly line factory work with inert materials. That's what's so important about this an-

Adobe Soundbooth control and you didn’t need it using their production I’m actually working there the interior four

jeremyrifkinallusion6 Accurate Transcript -imal. We moved from the industrial age to the bioindustrial age. Iterative Result (100% WAcc)...animal. We moved from the industrial age to the bioindustrial age.

Adobe Soundbooth and we will need to do it mickjaggerasyndeton1

Accurate Transcript And we thought we were totally unique animals. I mean there was no one like us. And then we heard there was a group from Liver-

Iterative Result (100% WAcc) and we thought that we were totally unique. Animals, I mean, there was no one like us. And then we heard there was a group from Liver-

Adobe Soundbooth and re lled every year it slightly you need I don’t I mean there was an on line outs and every head it was recruited from Liverpool

They had long hair, scru y clothes, but they had a re-mickjaggerasyndeton2 Accurate Transcript -pool. They had long hair, scru y clothes, but they had a recIterative Result (100% WAcc)

Adobe Soundbooth they had long hand it’s nothing to lose that directly mickjaggerasyndeton3

contractAccurate Transcript -ord contract.
Iterative Result (100% WAcc)

Adobe Soundbooth contradicts oprahwinfreyepizeuxis1

Accurate Transcript So, they sent me to a salon where they gave me a perm, and after a few days all my hair fell out and I had to shave my head. And then they really did-

Iterative Result (100% WAcc) so, they sent me to a salon, where they gave me a perm, and after a few days all my hair fell out and I had to shave my head and then they really d-

Adobe Soundbooth stalled the US alone where they gave me a current accurate few days all my hair fell out of my head shaved my head and then you really do

oprahwinfreyepizeuxis2 Accurate Transcript -n’t like the way I looked, cause now I am black and bald and sitting

on TV. Not a pretty picture. But even worIterative Result(95.8% WAcc)Didn’t like the way I look. ’Cause now I am black and bald and sitting on TV, eh heh, not a pretty picture. But even worse

Adobe Soundbooth like the way I look now I am black and gold and sitting on TV not a pretty picture

oprahwinfreyepizeuxis3 Accurate Transcript -se than being bald, I really hated, hated, hated being sent to report

on other people’s tragedies as a part of my daily duty. Iterative Result (100% WAcc)Worse than being bald, I really hated, hated, hated being sent to report on other people’s tragedies as a part of my daily duty. No

Adobe Soundbooth but even once in a bean ball I really hated hated hated being sent to report on other people's tragedies as a part of my daily duty

oprahwinfreyepizeuxis4 Accurate Transcript Knowing that I was just expected to observe, when everything in
my instinct told me that I should be doing something. I should be lending a hand.

Iterative Result (100% WAcc) knowing that I was just expected to observe when everything in my instinct told me that I should be doing something, I should be lending a hand

Adobe Soundbooth knowing that that was expected to observe when everything in my instinct told me that I should be doing something I should be lending a hand

pirateseuphemismos1 Accurate Transcript We're going to steal the ship? That ship? Commandeer. We're going
to commandeer that ship. Nautical term. Iterative Result(88.2% WAcc)We're gonna steal the ship? That ship? Commandeer. We're gonna commandeer that ship. Nautical term.

Adobe Soundbooth G the question during that time s
tonyblairdiacope1
Accurate Transcript The people everywhere, not just here in Britain, everywhere { they kept faith with Princess Diana.

Iterative Result (100% WAcc) The people everywhere, not just here in Britain - everywhere, they kept faith with Princess Diana.

Adobe Soundbooth Gupta Yes Kerry Gri n of the draft but you're great you're going to try it

My name is Robert Neville. I'm a survivor living in New York Citywillsmithiamlegendepanalepsis1
Accurate Transcript My name is Robert Neville. I'm a survivor living in New York City. Iterative Result (100% WAcc)

Adobe Soundbooth going on GA The assists the New York City
willsmithiamlegendepanalepsis2
Accurate Transcript I am broadcasting on all AM frequencies. If you are out there, if anyone

Iterative Result (100% WAcc) I am broadcasting on all a.m. frequencies. If you are out there, if anyone

Adobe Soundbooth all they have three wins we're up there this anymore
willsmithiamlegendepanalepsis3
Accurate Transcript is out there, I can provide food, I can provide shelter, I can provide security

Iterative Result(93.3% WAcc) It's out there, I can provide food, I can provide shelter, I can provide security.

Adobe Soundbooth he's out there and I do the shelter but to provide security in the limb
willsmithiamlegendepanalepsis4

If there's anybody out there.Accurate Transcript { if there's
anybody out there. Iterative Result (100% WAcc)

Adobe Soundbooth	there’s anybody out there and
willsmithiamlegendepanalepsisP3	
Accurate Transcript	is out there, I can provide food, I can provide shelter, I can provide security { if there’s anybody out there.
Iterative Result(95.0% WAcc)	he’s out there. I can provide food. I can provide shelter. I can provide security. If there’s anybody out there
Adobe Soundbooth	he’s out there and I do the shelter but to provide security in the limb there’s anybody out there and

Works Cited

[1] B. Juang and L. Rabiner, \Automatic Speech Recognition{A Brief History of the Technology Development," Encyclopedia of Language and Linguistics, Elsevier, 2005.

[2] \IBM Shoebox." http://www- 03.ibm.com/ibm/history/exhibits/specialprod1/ specialprod1_7.html, March 2011.

[3] \Speech recognition." http://en.wikipedia.org/wiki/Speech_recognition, March 2011. [4] \History of speech & voice recognition and transcription software." http://www.

dragon- medical- transcription.com/history_speech_recognition.html, March 2011. [5] S. Melniko , S. Quigley, and M. Russell, \Implementing a hidden Markov model speech recognition system in programmable logic," in Field-Programmable Logic and Applications, pp. 81{ 90, Springer, 2001.

[6] M. Muchmore, \Dragon NaturallySpeaking 10." http://www.pcmag.com/article2/0, 2817, 2327354, 00.asp, August 2008.

[7] G. Williams, \5 easy speech-to-text solutions." http://chronicle.com/blogs/profhacker/ 5- easy-speech- to- text- solutions/23016 , March 2010.

[8] W. Meisel, \Comparative evaluation of voicemail-to-text services." http://techcrunch. com/2010/01/28/phonetag- voice- to- text- 86- percent- accurate- google- voice/, January 2010.

[9] C. Breen, \First look: Adobe Soundbooth CS4 Beta." http://www.pcworld.com/article/ 146346/first_look_adobe_soundbooth_cs4_beta.html, May 2008.

[10] Bureau of Labor Statistics, \Occupational outlook handbook, 2010-11 edition: Medical transcriptionists." http://www.bls.gov/oco/ocos271.htm, 2011.

[11] \Secure transcription services." http://www.securetranscription.com/ transcription- rates.html, March 2011.

[12] \Games With A Purpose." http://www.gwap.com/gwap/, March 2011. [13] G. Little, L. Chilton, M. Goldman, and R. Miller, \Exploring iterative and parallel human computation processes," in Proceedings of the ACM SIGKDD workshop on human computation , pp. 68{76, ACM, 2010.

[14] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, \On the use of information retrieval measures for speech recognition evaluation," Research Report , pp. 04{73, 2005.

[15] \Levenshtein." http://php.net/manual/en/function.levenshtein.php, March 2011. [16] M. Al-Aynati and K. Chorneyko, \Comparison of voice-automated transcription and human

transcription in generating pathology reports," Archives of pathology & laboratory medicine, vol. 127, no. 6, pp. 721{725, 2003.

[17] \Mechanical Turk." http://www.amazon.com/gp/help/customer/display.html?nodeId= 16465291, March 2011.

[18] \Crowdsourcing." http://en.wikipedia.org/wiki/Crowdsourcing, March 2011. [19] \List of crowdsourcing projects." http://en.wikipedia.org/wiki/List_of_

crowdsourcing_projects, March 2011. [20] W. Mason and D. Watts, \Financial incentives and the performance of crowds," ACM SIGKDD

Explorations Newsletter, vol. 11, no. 2, pp. 100{108, 2010. [21] J. Horton and L. Chilton, \The labor economics of paid crowdsourcing," in Proceedings of the

11th ACM conference on Electronic commerce, pp. 209{218, ACM, 2010. [22] J. Ferreira, \Mechanical Turk case study: Knewton." http://aws.amazon.com/solutions/

case- studies/knewton/, March 2011. [23] \Amazon Mechanical Turk: Requester web site FAQs." https://www.mturk.com/mturk/

help?helpPage=requester#what_happens_reject, March 2011. [24] C. Passey, \Turning audio into words on the screen: We hire services to transcribe (somewhat

complicated) speech; nailing 'turducken'," Wal l Street Journal, p. D4, October 2008. Published online on October 9, 2008 at http://online.wsj.com/article/SB122351860225518093. html#articleTabs\%3Darticle.

[25] \CastingWords." http://castingwords.com/, March 2011. [26] \CastingWords hit." https://www.mturk.com/mturk/preview?groupId=

2RGQ2GMMEGOOODN6XJ9AE2SGM3V995, March 2011. [27] L. Von Ahn and L. Dabbish, \Designing games with a purpose," Communications of the ACM,

vol. 51, no. 8, pp. 58{67, 2008. [28] L. Von Ahn, \Games with a purpose," Computer, vol. 39, no. 6, pp. 92{94, 2006.