Foreign accented speech transcription and accent recognition using a game-based approach

Rio Akasaka '09

rakasak1@swarthmore.edu

Contents

1 Previous Research **4**

Games with a purpose 6

2 The Corpora **6**

3 The Game **7**

Task 1: Accent Recognition 8

Task 2: Transcription 9

Facebook as a research environment 10

4 Methodology **13**

5 Results **13**

Task 1: Accent Recognition 14

Task 2: Transcription 18

Sample transcriptions 19

6 Discussion **20**

7 Conclusions **21**

8 Future Work **22**

9 Acknowledgements **23**

Abstract

While significant improvements have been made in reducing sentence error rate (SER) and word error rate (WER) of automatic speech recognition (ASR) technology, existing systems still face considerable difficulty parsing non-native speech. Two methods are common in adapting ASR systems to accommodate foreign accented speech. In the first, accent detection and identification is followed by an accent-specific acoustic model (Faria 2006, Chen et al. 2001) or dictionary (Fung and Kat 1999). Accents have also been classified by severity (Zheng et al. 2005, Bartkova and Jouvet 2007). The alternative is to use acoustic or phonetic models from both native and non-native speech (Bouselmi et al. 2006, Matsunaga et al. 2003). It has been shown that the use of accent-specific data improves recognition rate (Arslan and Hansen 1996, Humphries et al. 1996) but success rates vary among languages. In either case, specific information needs to be obtained regarding particular accents, and the process of adapting existing corpora to train language models is both time-consuming and tedious, limiting advances in the field.

We introduce the Foreign Accented Speech Transcription Game (FASTGame) as a way to transform the transcription process into a more enjoyable format. The FASTGame is a 'game with a purpose' designed to obtain normalized orthographic transcriptions of foreign accented speech from naïve listeners. The FASTGame is accessible online through the social networking website Facebook and contains two tasks. The first asks the player to determine the native language of a foreign accented speaker of English from four available options as rapidly as possible. Players are incentivized by scores that reflect how well they perform. For this task they are based on accuracy and speed. In addition to examining the specific cues that trigger accent recognition, analysis can be made on the data about user responses to novel accents.

The second task asks the player to transcribe a phrase spoken by a foreign accented speaker of English. Their scores are calculated based on agreement with other users. In the event that transcriptions have not already been written, scores are assigned randomly. All transcriptions for a particular recording are then aggregated and the correct transcription will then be generated based on multiple agreement.

Existing continuous speech recognition software fail to accurately produce transcriptions for such recordings, which are also of varying audio quality and accent severity. By performing time-alignment on the transcriptions provided with this game, valuable training data can be used to improve language models for accented speech. In both tasks of the game, steps are taken in order to avoid repeated plays and undesirable data conditioning.

The FASTGame was created as an alternative to existing methods for obtaining tran-

scriptions, and its primary merit is in supplementing large speech corpora with additional data in a relatively inexpensive and effortless manner.

1 Previous Research

Insufficient research has been done on the role of naïve users in orthographic transcription of corpora. Regardless, transcription of spoken corpora is a time-consuming task: Chafe et al. (1991) suggested that it would take six "person-hours" to transcribe one minute of recorded speech to use in the Santa Barbara Corpus of Spoken American English. Even considering the need for annotating additional information beyond the orthographic transcription, this is still a large commitment of resources.

In addition, when a selected number of transcribers are used, individual errors can occur - for example, the transcriptions of the British National Corpus (BNC) contain numerous spelling and tagging errors (Mitton et al. 2007). A possible solution is through the use of multiple agreement, where several users contribute to a solution and a 'ground truth' is established when identical responses are given by different people. The best example of this is the CAPTCHA tool, with which human authentication and verification can be done by asking the user to input text that has been distorted but remains nonetheless readable. In Carnegie Mellon's reCAPTCHA, the CAPTCHA is reconfigured to have humans read distorted words that are scanned in from public domain books. Where optical character recognition software (OCR) fails, humans can read those words as part of the challenge response task. Multiple agreement can establish a ground truth for the actual orthographic content.

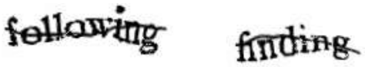


Figure 1. *A sample challenge task from reCAPTCHA.*

While Schlaikjer (2007) argued that speech transcription conducted in a similar manner would result in a variety of spelling and punctuation differences among listeners, making it difficult for validation tasks, it is important to highlight here that the goal of the game is not to validate users but to collect information that would otherwise be tedious or time-consuming to obtain. By requiring ground truth to be based upon a high agreement metric, individual spelling errors and inconsistencies can be ignored.

With regards to the accent recognition step, Arslan and Hansen (1996) demonstrated that when using isolated words, in general, longer words led to more accurate recognition of accent. Their experiment showed that the average classification rate for human listeners of isolated Turkish, German or Chinese accented English words was 52.3%. Vieru-Dimulescu and de Mareuil (2006) also demonstrated similar results, with a 52% identification rate for foreign accented French among six possible choices (English, Arabic, German, Spanish, Portuguese and Italian). Flege (1984) showed that there was no apparent difference in listener detection of foreign accent between read and spontaneous speech.

While a variety of experiments have concluded that differences in rating the *degree* of accent may (Thompson 1991) or may not (Flege and Fletcher 1992) exist between linguistically trained listeners and naïve listeners, no experiment has explored the relationship between linguistic training and accent recognition or between rating the degree of an accent and recognizing the same.

McDermott (1986) found that a variety of phonological factors influenced listener's judgments of accent, as well as listener background and exposure to foreign languages. Subsequent studies have considered the role of pronunciation (in Japanese liquids, such as the substitution of /l/ and /r/ (Riney et al. 2000), vowel quality (Munro et al. 1999), prosody (in Brazilian Portuguese, Major (1986)) and even comprehensibility (Ikeno and Hansen 2007), grammatical accuracy (Varonis and Gass 1982), and fluency (Anderson-Hsieh and Koehler 1988) in affecting listener perception of foreign accent.

While Magen (1998) analyzed the sensitivity of monolingual American English listeners to Spanish-accented English, there has been little research on multilingual listeners with varying degrees of exposure to multitudes of languages. This is due in large part to the difficulty in obtaining sufficient data so that individual variations of language experience and exposure can have less of an effect on the entire conclusions obtained. The FASTgame attempts to address this issue by using the social network nature of Facebook to encourage many players to play.

While Flege and Munro (1994) demonstrated that listeners who were unfamiliar with French nonetheless detected accent in English spoken by native speakers of French in a binary forced choice test (i.e. accented or not), it is not assumed that the same can be said of recognizing different variants of accent, i.e. English with an Italian accent as opposed to English with a German accent.

Previous studies in foreign accented speech perception have presented stimuli ranging from milliseconds (Flege 1984) to minutes (Elliott 1995), including word (Flege and Munro 1994), phrase (Magen 1998) and sentence (Thompson 1991) segments. The approach presented here is unique in that listeners will only listen to the recording for as long as they need to identify the native language of the speaker. This is encouraged by informing the player that the score is based on correctness as well as speed.

There are very few parallel studies that have been performed where participants are asked to recognize foreign accents as well as attempt to transcribe them. Ingram and Nguyen's study of Vietnamese accented English comes closest, where 169 native and non-native participants listened to recordings from 21 speakers and rated their intelligibility as well as accentedness. They were also asked to transcribe the recordings which they could listen to up to four times. The transcriptions were used to assess comprehensibility rather than for actual data retrieval, however.

Finally, Arslan and Hansen (1996) demonstrated that the knowledge obtained from accent classification is useful for improving speaker independent speech recognition systems. While the game only serves to provide transcriptions for foreign accented speech, the goal is to be able to use the information obtained to improve upon language models in ways that would not be possible with limited amounts of data.

Games with a purpose

The FASTGame is unique in adopting a game-based approach specifically for linguistic research, but the concept of games that provide useful information is not new. Recent research has delved into improving existing computer algorithms with what are known as 'games with a purpose' where players perform classification and description tasks online and obtain points when pairs of players agree with each other. In particular, these have been shown to be effective in applying descriptive labels for individual images, (*ESP Game* (von Ahn and Dabbish 2004) and *Phetch* (von Ahn et al. 2006), for example) as well as for popular songs (*MoodSwings* (Kim et al. 2008), *Listen Game* (Turnbull et al. 2007), *MajorMiner* (Mandel and Ellis 2007), among others).

2 The Corpora

The recordings used in this game were obtained from two different corpora:

George Mason University's Speech Accent Archive contains a continually expanding database of recordings from native and non-native speakers of English from 250 countries. Each speaker is asked to read an elicitation paragraph, the Stella Passage[1], which is designed to contain most of the consonants, vowels and clusters of American English while using relatively simple words (Weinberger 2005). The recordings are compressed at a sampling rate of

[1] The elicitation paragraph is as follows: *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."*

22kHz and a phonetic transcription is also provided. The Speech Accent Archive is used in Task 1 because maintaining the same elicitation paragraph will ensure that in general, the same amount of linguistic information will be conveyed in a given amount of time.

The CSLU Foreign Accented English (FAE) corpus contains 4925 continuous speech utterances from native speakers of 22 countries. Each individual is asked to talk about themselves in English for 20 seconds. In addition to the recordings, the corpus also contains information about the speaker's linguistic background as well as an assessment of the degree of accent by four native American English speakers. It is recorded with a sampling rate of 8kHz (telephone quality).

There are multiple challenges present with the FAE corpus. Firstly, it has not been transcribed. Additionally, while disfluencies are common in continuous speech corpora (Yu 2005, Shriberg 1996), the corpus contains native and non-native speakers of English with varying commands of the language. Each speaker is asked to talk about themselves, often referring to their own name, and interpersonal agreement about transcribing proper names without any standardization is likely to be low. Lastly, because each recording is exactly 20 seconds long, the recordings often cut the speaker off mid-phrase. While automated speech recognition software remain limited in their scope of knowledge about context and language, human transcribers are far better suited for the task of recognizing disfluent or interrupted speech. Consider the following example:

```
I live in the Northeast coast of the United States, New Yo-
```

An automated speech recognizer would be limited in its ability to recognize the complete word ("New York"), but a human transcriber can easily associate 'Northeast coast' and 'United States' with New York.

3 The Game

The game was created using ActionScript 3, a scripting language for Adobe Flash. Flash proved to be more suitable than comparable online game media (AJAX, Java) because it is supported by most computer systems and browsers[2] and it can handle small increments of time efficiently. Accurate time calculations are essential to Task 1, where recognition speed is measured. Furthermore, to prevent problems arising when multiple users are accessing

[2] "Adobe - Flash Player Statistics" http://www.adobe.com/products/player_census/flashplayer/ Retrieved November 29, 2008

the same files, ActionScript uses *event listeners* to ensure that tasks only begin after the server data has been completely retrieved, rather than at the precise moment that task was requested.

In order for the game to communicate efficiently with the server to store the information related to each session, the application is configured to use Extensible Markup Language (XML). XML is a specification that is practical in encoding data as well as information about the data (the markup) in a structured manner.

```
<?xml version=’1.0’ encoding=’UTF-8’?>
<userData>
    <languages>
        <data>russian</data>
        <limit>20</limit>
    </languages>
</userData>
```

Figure 2. *A portion of the flash.xml file containing markup about the Russian language set available for Task 1 of the game as well as the number of files that are available for it.*

The game contains two distinct tasks which are distributed throughout each round of the game as well as a set of instructions.

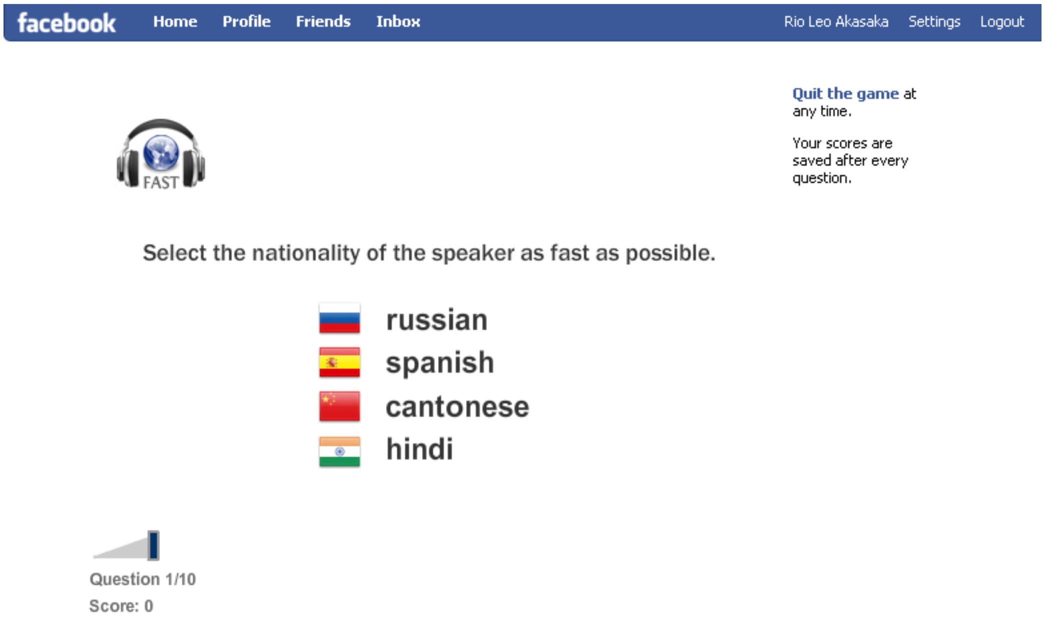
Task 1: Accent Recognition

In Task 1, players are asked to identify the native language of a foreign accented speaker of English as quickly as possible out of four randomly generated choices, one of which is the correct answer. The recordings are obtained from the GMU corpus. While they may take as much time as they want, players do not have the option to pause or replay a recording, and their score is lower the more time they spend responding. The scoring, however, is designed to be unobtrusive so as to prevent random or rushed decision-making. The score for an accurate response in this task is as follows:

$$round\left(\frac{1}{time} * 100\right) \tag{1}$$

where *time* is the time taken to answer the question and *round* rounds the value to the nearest integer. This equation was motivated by initial tests where on average it was found that the average player could respond within 5 seconds - 20 points is awarded, then, for average performance while quicker responses obtain more points. Inaccurate responses automatically deduct 20 points.

A pre-game questionnaire is also given to ask the player about their native language as well as the languages they speak and the languages they have had exposure to, but no linguistic training or background is required to participate. Furthermore, the term 'exposure' is deliberately loosely defined in the question (*"heard in passing, friends who speak, etc"*) so as to minimize over-interpretation of the term.



Quit the game at any time.

Your scores are saved after every question.

Select the nationality of the speaker as fast as possible.

- russian
- spanish
- cantonese
- hindi

Question 1/10

Score: 0

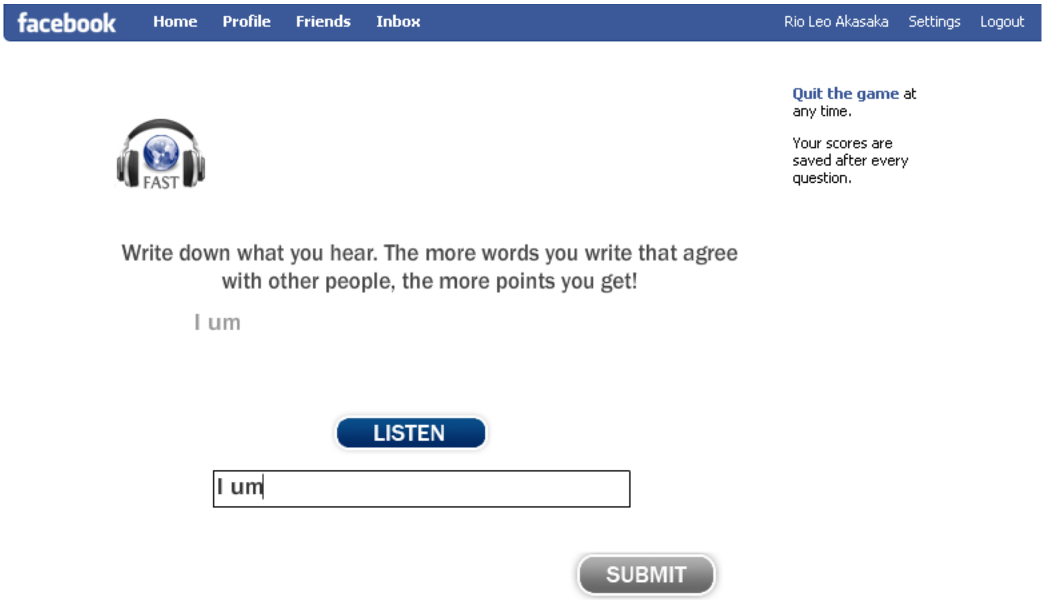
Figure 3. *A screenshot of the accent recognition task of the FASTGame as seen from a player's perspective on Facebook.*

Task 2: Transcription

In Task 2, players are asked to transcribe short recordings that are randomly selected from the ones available through the CSLU-FAE corpus. They are allowed to listen to the recordings as many times as they want or need to. The transcription is sent to the server, where they are compared with the transcriptions already provided by other users. An XML file is generated containing all the words in the transcription that agree with at least one other player. If the transcription is the first one for a particular recording, the score is determined randomly with points between 20 and 50. Otherwise, the score is determined as follows:

$$round\left(correct + \frac{correct}{length} * 10\right) \tag{2}$$

where *correct* is the number of words that agree with other users and *length* is the word count of the transcription - hence if the user transcribes a longer utterance and gets many of those words correct, their score will be considerably better than if they had transcribed a shorter section of that same transcription just as accurately. Since some of the recordings inadvertently do not contain any decipherable utterances (replaced instead by background noise or disfluencies), an option is given to the user to opt out of the transcription if they do not hear or cannot understand the recording. 10 consolation points are then awarded.



Quit the game at any time.

Your scores are saved after every question.

Write down what you hear. The more words you write that agree with other people, the more points you get!

I um

I um

Figure 4. *A screenshot of the transcription task of the FASTGame as seen from a player's perspective.*

Facebook as a research environment

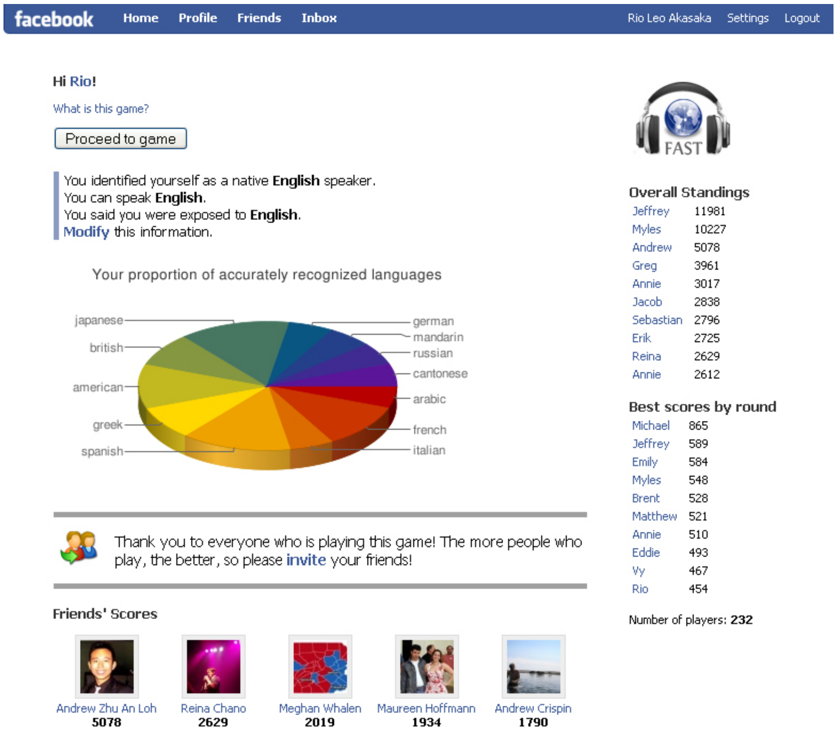
Facebook is increasingly being used as a platform for research[3], allowing for data retrieval and polling on a massive scale, with over 30 million users in the United States aged 18 and over and more than 120 million worldwide[4]. While personal details are self-reported, social networks can function as a powerful utility for information retrieval because of the ease with which individuals can participate and share their experiences. Turnbull et al. (2007) demonstrated the efficacy of a collaborative game with a Facebook interface where

[3] *"On Facebook, Scholars Link Up With Data"* `http://www.nytimes.com/2007/12/17/style/17facebook.html`, retrieved November 25, 2008

[4] *"Facebook Statistics"*, `http://www.facebook.com/press/info.php?statistics`, retrieved November 29, 2008

players provide descriptive tags and annotations for individual songs. Microsoft Research has released the Collabio application[5] where individuals write tags that describe each other, obtaining 'points' when multiple users agree.

The FASTGame uses an application programming interface (API), allowing it to access the resources and functionality available throughout the Facebook website. For example, players are able to see their friends' game scores, and the game also posts a small note on the user's profile with their own score so that other users of Facebook can find out about the game. To make the game interface more visually stimulating, a pie chart with the user's accurate language distribution is also added, which is drawn using the Google Charts API. In order to prevent players from trying to artificially improve their score, the pie chart is not published with specific percentages.



Hi Rio!

What is this game?

Proceed to game

You identified yourself as a native **English** speaker.
You can speak **English**.
You said you were exposed to **English**.
Modify this information.

Your proportion of accurately recognized languages

Thank you to everyone who is playing this game! The more people who play, the better, so please invite your friends!

Friends' Scores

Andrew Zhu An Loh	Reina Chano	Meghan Whalen	Maureen Hoffmann	Andrew Crispin
5078	**2629**	**2019**	**1934**	**1790**

Overall Standings

Jeffrey	11981
Myles	10227
Andrew	5078
Greg	3961
Annie	3017
Jacob	2838
Sebastian	2796
Erik	2725
Reina	2629
Annie	2612

Best scores by round

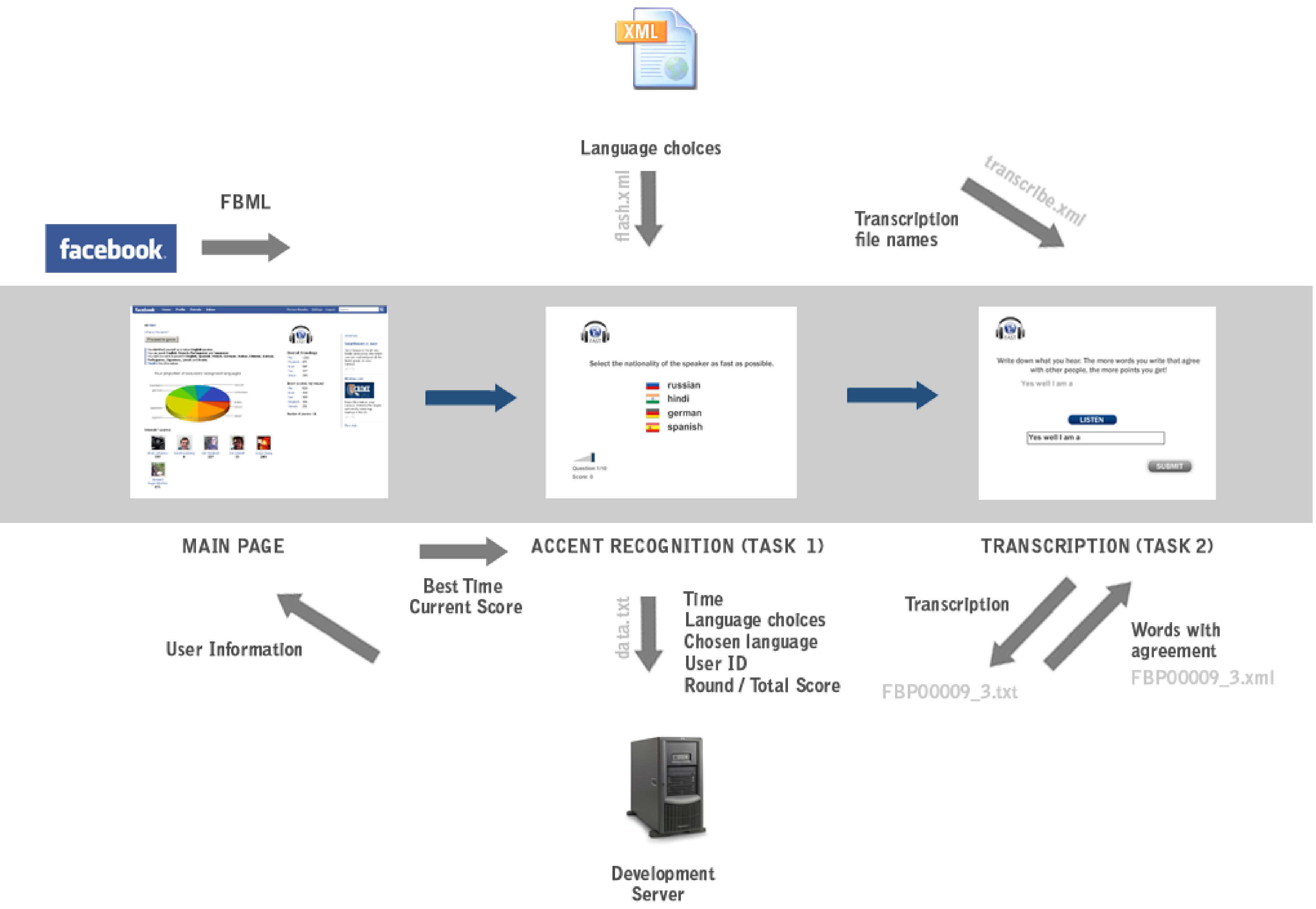
Michael	865
Jeffrey	589
Emily	584
Myles	548
Brent	528
Matthew	521
Annie	510
Eddie	493
Vy	467
Rio	454

Number of players: **232**

Figure 5. *A screenshot of the main game dashboard of the FASTGame.*

Figure 6. *The following page displays the overall layout of the game as presented to the user (with the gray background) and as handled by the server (above and below)*

[5] *"Collabio: Collecting Tags on Users"* `http://research.microsoft.com/cue/collabio/`, retrieved November 29, 2008


XML
Language choices
flash.xml
facebook
FBML
Transcription file names
transcribe.xml
Select the nationality of the speaker as fast as possible.
russian
hindi
german
spanish
Question 1/10
Score: 0
Write down what you hear. The more words you write that agree with other people, the more points you get!
Yes well I am a
LISTEN
Yes well I am a
SUBMIT
MAIN PAGE
Best Time
Current Score
ACCENT RECOGNITION (TASK 1)
TRANSCRIPTION (TASK 2)
User Information
data.txt
Time
Language choices
Chosen language
User ID
Round / Total Score
Transcription
FBP00009_3.txt
Words with agreement
FBP00009_3.xml
Development Server

4 Methodology

For the pilot study, 4925 speech recordings from the CSLU-FAE were used. Each file was segmented where pauses of a minimum duration of 0.4 seconds and maximum intensity of 50dB were detected in order to make each clip more manageable to transcribe. A script running on the phonetics and acoustics software Praat was used for this task. Of the segmented files generated, 1257 files were selected based on the amount of information each recording contained (determined by file size) - the transcription task requires fairly short recordings, and the parameters for determining what constitutes a pause do not apply to all recordings, which often result in a recording not being segmented at all. All files could be segmented if the duration and intensities are modified, but there is also the need to restrict the number of files to transcribe in order to ensure that different players will transcribe a particular recording.

The accent recognition task uses recordings of native speakers from 13 language sets: American, British, Hindi, Russian, Cantonese, German, Mandarin, Arabic, Spanish, Greek, Japanese, French and Italian. The number of recordings available for each language set varies, from an upper bound of 25 to a lower limit of 6. The game uses XML files in conjunction with a random number generator to select the four language choices, and then furthermore selects the actual audio recording with another random number in the range of number of files available for that language. Doing so ensures a random distribution of the 13 languages shown to all users despite the difference in number of available files. The effects of repeated playbacks of the same recording (for language sets with fewer individual files) are limited by confining analysis to the first 20 instances of gameplay.

Paired sample t-tests are performed to compare accent recognition speed between questions where the language in the recording they hear is familiar to them and not. Familiarity is defined by whether or not the language is within (1) the set of languages they speak and (2) the set of languages they have had exposure to, both of which are questions asked in the pre-game questionnaire.

5 Results

During the course of the three-week study, 368 players participated, of which 353 completed the questionnaire and played at least one question. 333 played at least a complete round of the game. There are 10 recognition questions in a round. Discarding the first question of a game for possibilities of variability (due to adjustment of volume and accustomization to the game), the mean play count was 61.06 with a standard deviation of 106.78 and a maximum of 606.

Task 1: Accent Recognition

Task 1 has been played a total of 11814 times, with 55.26% of the accents accurately identified (N=6528).

Player profiles

81.8% (N=301) of the 368 players declared they were native English speakers. The remaining reported nativeness in Chinese (9), Italian (9), Spanish (8) and Korean (8), among others (classified as *"Other"* below). 84.2% (N=310) reported being able to speak at least one other language, including various combinations of Spanish (173), French (130), Chinese (53), German (50), and Japanese (27). The average number of languages each player claims to speak is 1.61 (σ=1.22) and the average number of languages they claim to have had exposure to is 6.91 (σ=3.97). Given that the primary language interface of Facebook and of the game is English, fluency in English is not counted as one of the languages each player claims to speak or has had exposure to.

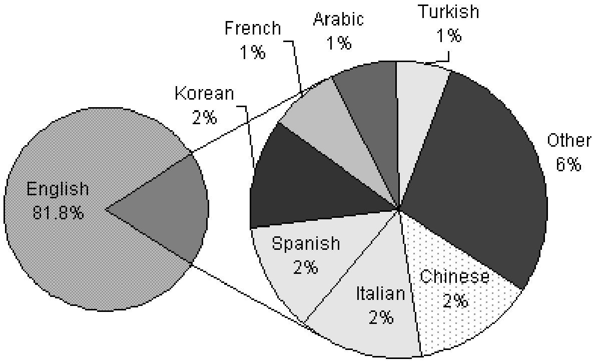


Figure 7. *A distribution of self-reported native language.*

Distribution of Languages

The distribution of languages that were presented to all the players in the game is shown below, along with the number of those accurately determined.

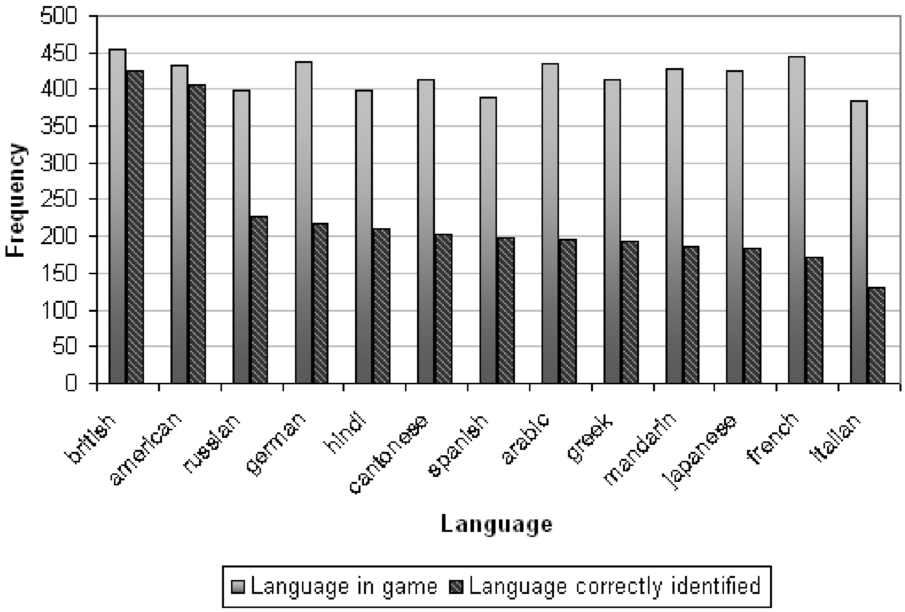


Figure 8. *Distribution of languages given in the accent recognition task of the game and number of those accurately determined*

Conditioning with multiple plays

A linear fit of the average speed of accurate responses against the number of plays provided a slope of -0.0046. While the slope and R^2 (0.017) are small, it could suggest the possibility that increased gameplay (on the order of 10 or 20) reduces speed - accent judgments have been demonstrated to be reasonably accurate even with speech recordings as brief as 30 ms (Flege 1984). Where appropriate, calculations have been made taking only the first 20 instances of game play into consideration.

Accuracy did not improve with increased playcount (slope = 0.0002). Limiting the recognition to the first 20 instances, accuracy drops down to 54.04% (2945 of 5450) from 55.26% (6528 of 11814).[6] It is evident that this result differs from mere chance.

[6] This value drops even further when we disregard American and British English responses: 47.89% with the 20 instance limit, 49.60% considering all.

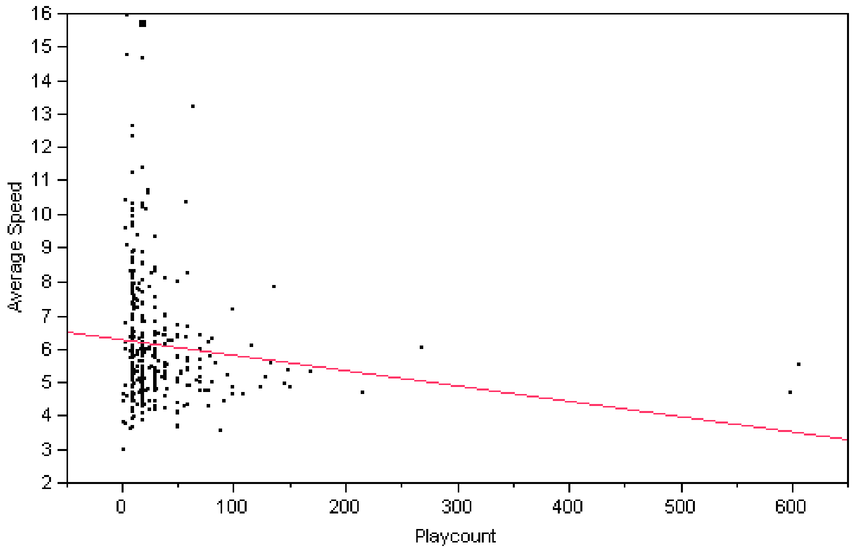


Figure 9. *Average speed per user against number of plays for all plays. Slope = -0.0046,* R^2 *= 0.017*

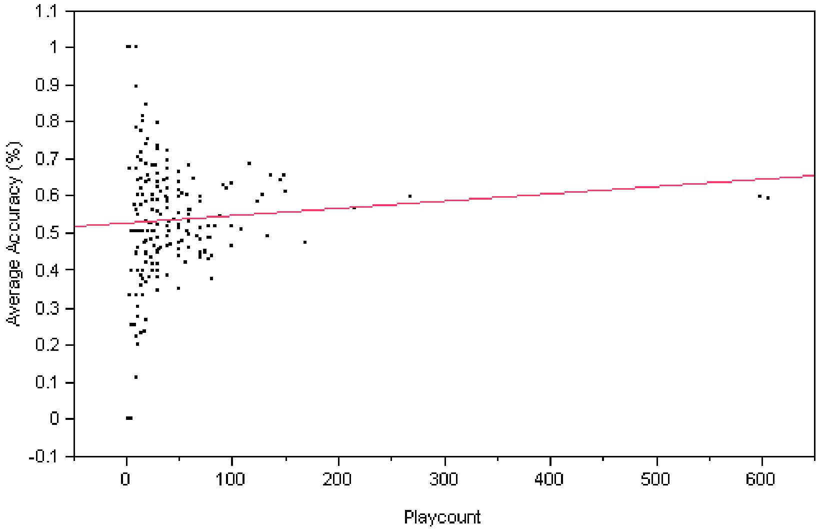


Figure 10. *Average accuracy per user against number of plays for all plays. Slope = 0.0002,* R^2 *= 0.005*

Accuracy and response times

The overall distribution of the response time to queries in Task 1 was μ=5.11 sec, σ=3.20 (N=5450). Correct responses averaged 4.68 seconds (σ=1.92), incorrect responses averaged 5.69 (σ=2.36, N=352) per user, and a paired-sample t-test confirmed the difference as statistically significant (t(351)=10.73, μ_{diff}=1.01, p<0.001).

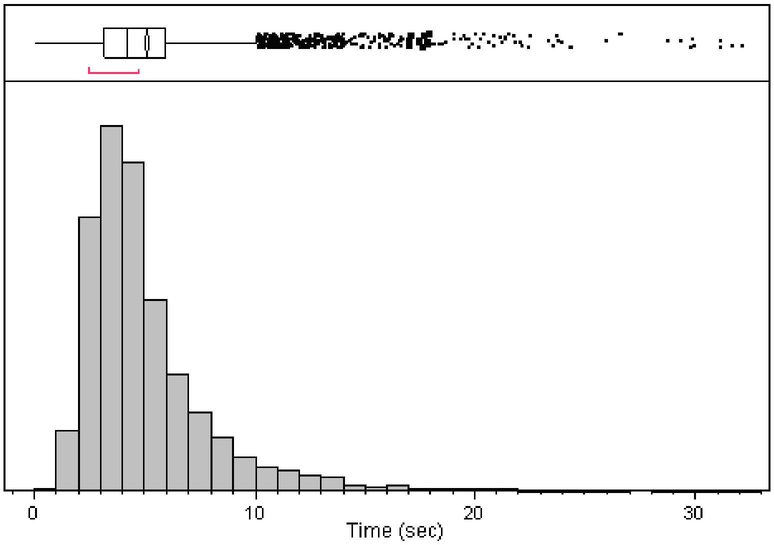


Figure 11. *Distribution of speed of responses. μ=5.11 sec, σ=3.20, N=5450. Because of the skew to the right, all computations involved a* log*(speed) calculation.*

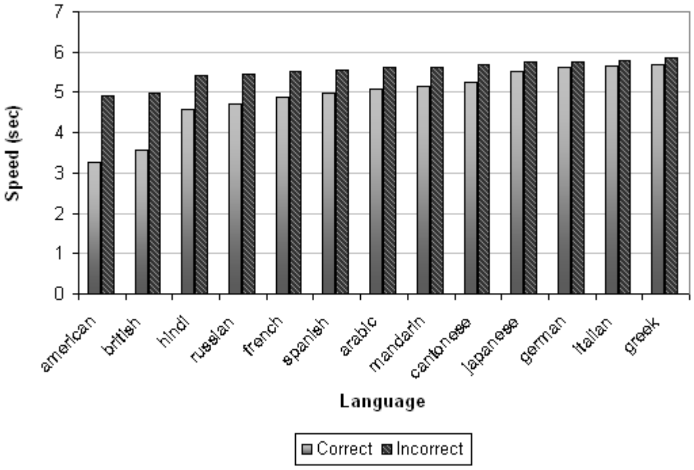


Figure 12. *Average speed of correct and incorrect response times for each language.*

One-way ANOVA using Tukey's HSD test identified accurate response times to British and American English recordings as significantly lower compared to the other languages (q=3.786, α=0.01).[7]

Accuracy as a function of speed

The following plot shows the average user accuracy against average response time per individual. The slope is -0.011, R^2=0.019. This suggests that it is highly likely that accuracy

[7] For further details as well as complete statistical output using JMP and SPSS, visit `wiki.rioleo.org`

decreases the longer a user takes to respond, a result to be expected since players who are confident about their choices will likely not wait to respond.

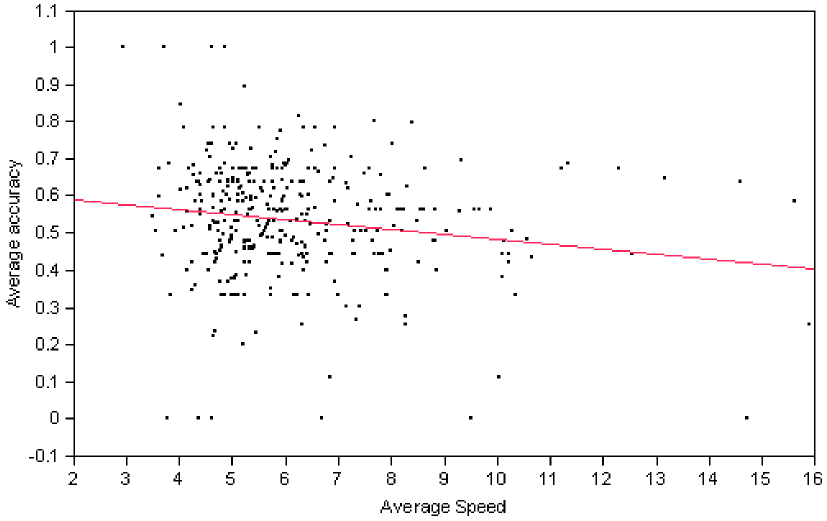


Figure 13. *Individual accurate percentage plotted against average response time for the game showed a slope of -0.0128.*

Accuracy and speed as a function of languages exposed to or spoken

A paired-sample t-test on the average speed and accuracy demonstrated for accurate responses, *exposure* to a language resulted in a quicker response (μ_{diff}=1.06, t(286)=7.89, p<0.001) as well as higher accuracy (μ_{diff}=0.12, t(286)=8.88, p<0.001).

This is not consistent, however, with *spoken* language: while accurate responses to recordings that were within a user's spoken language set were faster (μ_{diff}=1.52, t(317)=14.07, p<0.001), they were *less accurate* than if they were not (μ_{diff}=0.11, t(317)=10.62, p<0.001). This conclusion was also confirmed using smaller subsets of the data (e.g., native English speakers only) as well as larger datasets (including newer data, since the game continues to be played).

Task 2: Transcription

In the three weeks that followed the initial release of the game, 1093 recordings out of the 1257 available were transcribed with an average of 2.86 transcriptions per file (standard deviation of 1.58). The multiple agreement analyzer scores each transcription based on the number of other words that are located at the same position in the sentence as well as one position before and after, to accommodate for transcriptions that differ by one word towards the beginning of a sentence. It also takes into consideration the overall length of the sentence. In the example below, the second transcription is selected.

FSD00135_2		
Transcription	**Score**	**Length**
`So like i was saying in swedish`	7	38
`well like i was saying in swedish i moved here`	10	56
`oh like I was saying in swedish I moved here`	10	54

Table 1. *Three transcriptions and their respective scores for multiple agreement and length.*

The following is the transcription obtained via the game for FAR00524, a speaker of Arab descent. Each line designates a separate file segmented from the original based on pauses in the recorded utterance.

```
I'm twenty eight years old, I'm from Damascus, Syria
I've got a bachelor's in uh, electrical engineering with a masters in
I am a business admistration majoring in management information systems
```

The following is the same file transcribed through the transcription tool in the commercially available speech transcription software Dragon NaturallySpeaking^{TM} 9'.

```
28 years old and from there if you are on the bachelor degree in an
engineering with the monsters in for the nation majoring in management
information systems
```

Since the speaker from each recording cannot train the recognizer (training tasks require generating continuous speech based on text that is provided by the software), different scenarios involving very minimal training as well as complete and thorough training were tested for overall accuracy. The above was the most accurate and was generated using a training set provided by the author.

Sample transcriptions

The following highlights some of the transcriptions obtained as part of Task 2. Since baseline or ground truth data is not available, a qualitative analysis is presented.
The transcriptions for sentences with acronyms show high consistency. (*FSD00101_7*)

```
and I’m involved with a lot of AARP work
and i’m involved with a lot of aarp work
And I’m involved with a lot of AARP work
```

Proper nouns are shown here to have a variety of transcriptions, though it is difficult to ascertain which is the correct one. (*FRU00037_2*)

```
my name is elphina
my name is athena
my name is agina
```

The following transcriptions show the variability in transcribing fillers in speech. (*FBP00706_2*)

```
Yes, uh, I am
Yes uh I am
yes uh i am
Yes eh I’m
yes I am
```

Another point of concern is transcribing numbers: without specific training or instructions, players choose to transcribe as they see fit. This is also the case with the word ‘okay’. (*FSD00147_2*)

```
okay i’m thirty four years old
ok i’m 34 years old
ok i’m thirty-four years old
```

There are also instances where some words are completely replaced. (*FSP00367_1*)

```
well it’s incredible how much of my experience I have forgotten
well it’s incredible how much of my space i have forgotten
```

6 Discussion

The unexpected results of the relationship between accuracy and language exposure may be due to the lack of consistency between each player’s interpretation about what amounts to exposure to a particular language. In particular, it is important to note that while there were 27 different languages among the languages spoken by the 368 players, there were 44 different languages that they claimed having had exposure to. There is also the need to highlight the fact that while second language (L2) speakers may adopt the most salient (prosodic) features from their native language (L1) (Tomokiyo and Waibel 2001, Henry et al. 2006), a clear relationship between *speaking* a language and *recognizing* prosodic features of that language hasn't been shown.

The lack of correlation between accuracy and playcount can be explained by examining the role of speed (Figures 9 and 13) and realizing that because longer response times are correlated with reduced accuracy, and more plays demonstrate faster (but not necessarily more accurate) responses, players assume that in playing more they should improve, and thus respond faster. This assumption may be related to the question of language exposure, where players assume that if they speak a language (or identify among the answer choices a language they speak) they should be able to make judgments more quickly, but are less accurate in doing so.

It is clear that the results from Task 2 are less than perfect, given the wide variety of transcriptions, but it is also important to note that the number of transcriptions is limited by the number of players, and the number of files to transcribe far outweighed the number of players necessary for adequate multiple agreement. While word omissions and spelling mistakes do exist in individual transcriptions, when combining multiple users the result is remarkably robust, particularly when considering disfluencies and hesitations (*um, uh*):

```
I've got a bachelor's in uh, electrical engineering with a masters in
I've got a bachelors degree in a with a master in
i've got a bachelor's degree in engineering with a bachelor's in
```

Figure 14. *No single player obtains the correct transcription, but the combined information is nonetheless accurate*

While segmentation along pauses was necessary to divide the recordings for Task 2 into smaller files, there were instances where speech is cut off in between utterances. When combined, the transcriptions do not necessarily reflect the content of the original phrase, as is the case with FAR00524. The correct transcription is as follows:

```
I'm 28 years old, I'm from Damascus, Syria, I've got a bachelor's degree in uh
electrical engineering with a masters in uh business administration majoring in
management information systems [...]
```

7 Conclusions

During the three week study with FASTGame on Facebook, 368 users played Task 1 a total of 11814 times, correctly identifying the native language of an accented speaker 55.26% of the time. This study has also shown that recognition tasks involving a language that a user has been exposed to results in higher accuracy and faster speed, but that when considering languages that are spoken by the user, accuracy is lower despite faster recognition.

The players transcribed 3129 individual files from 1093 recordings in Task 2, providing data with an average of 2.86 transcriptions per file. Multiple agreement has been shown to perform remarkably well despite the limited number of transcriptions available, remaining robust to minor interplayer differences.

The FASTGame was demonstrated to be a novel method with which the time-consuming and tedious process of transcription can be instead conducted using 'crowdsourcing', where multiple individuals perform small portions of a larger task. Unlike similar experiments in accent perception and recognition, the FASTGame is not designed to isolate a particular feature or language in an attempt to support or refute a hypothesis. Rather, it serves as a powerful tool enabling research using an environment that is both flexible and realistic, providing a variety of useful data.

8 Future Work

The work completed here, while successful, is just one part of a multistep process of attempting to improve foreign accented speech transcription. Much remains to be seen with regards to how the transcriptions can now be time-aligned to the recordings and how conventional methods using n-gram Hidden Markov Models can be used to improve recognition. There is also much more that can be done using the existing data- for example, future analysis could examine the individual recordings that resulted in the most accurate responses.

During the course of the study, many alternative approaches to the experiment were considered but not implemented in order to maintain consistency with the data already obtained. For example, another study that could be performed using the same framework would be to tailor Task 1 to display languages and transcriptions that reflect the user's language background. A more sophisticated study could then be performed by including accents that the user may not have had exposure to but are linguistically or geographically related to the ones they know.

It would be ideal to incorporate additional fun tasks to make the game more entertaining and useful- for example, an addition to Task 2 would be to introduce a task where users would look at the possible transcriptions already made for a particular recording and select the one that best reflects the transcription (or make amendments as necessary). The scoring for the game should also be improved in order to retroactively add scores for individuals who transcribe a new file whenever other users later agree.

Lastly, many of the ‘games with a purpose’ extend the collaborative nature of the game even further by having real-time (or simulated) competitions between two players who must agree in their annotation or labeling tasks. Facebook can be used to bring together friends in playing the game, though considerations must be taken to prevent two players collaborating with the intent to artificially improve their scores.

9 Acknowledgements

Grateful acknowledgement is made to Dr. Jiahong Yuan at the University of Pennsylvania for his support, Drexel University ECE and MET lab advisor Dr. Youngmoo Kim for his encouragement, Dr. Etsuko Hoshino-Browne for her tireless assistance, Dr. Steven C. Wang for his ideas, and my thesis advisor Dr. K. David Harrison for helping me bring it all together.

References

J. Anderson-Hsieh and K. Koehler. The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, 38(4):561–613, 1988.

L. M. Arslan and J. H. L. Hansen. Language accent classification in American English. *Speech Communication*, 18(4):353–367, 1996.

L. M. Arslan and J. H. L. Hansen. Frequency characteristics of foreign accented speech. In *Proc. ICASSP '97*, volume 2, page 1123. IEEE Computer Society, 1997. ISBN 0-8186-7919-0.

K. Bartkova and D. Jouvet. Automatic detection of foreign accent for automatic speech recognition. In *Proc. ICPHS '07*, 2007.

G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton. Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints. In *Proc. ICSLP '06*, 2006.

W. L. Chafe, J. W. D. Bois, and S. A. Thompson. Towards a new corpus of spoken American English. In *English Corpus Linguistics*. New York: Longman, 1991.

T. Chen, C. Huang, E. Chang, and J. Wang. Automatic accent identification using Gaussian Mixture Models. In *IEEE Workshop on ASRU*, 2001.

A. R. Elliott. Field independence/dependence, hemispheric specialization, and attitude in relation to pronunciation accuracy in Spanish as a foreign language. *The Modern Language Journal*, 79(3):356–371, 1995.

A. Faria. *Accent Classification for Speech Recognition*, volume 3869. Springer Berlin / Heidelberg, 2006.

J. E. Flege. The detection of french accent by American listeners. In *Journal of the Acoustical Society of America*, pages 692–707, 1984.

J. E. Flege and K. L. Fletcher. Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America*, 9(1):370–389, 1992.

J. E. Flege and M. J. Munro. The word unit in second language speech production and perception. *Studies in Second Language Acquisition*, 16(4):381–411, 1994.

P. Fung and L. W. Kat. Fast accent identification and accented speech recognition. In *Proc. ICASSP '99*, pages 221–224, 1999.

J. H. L. Hansen and L. M. Arslan. Foreign accent classification using source generator based prosodic features. In *Proc. ICASSP*, pages 836–839. IEEE, 1995.

G. Henry, A. Bonneau, and V. Colotte. Making learners aware of the prosody of a foreign language. In *Current Developments in Technology-Assisted Education*, 2006.

J. Humphries, P. Woodland, and D. Pearce. Using accent-specific pronunciation modelling for robust speech recognition. In *Proc. ICSLP '96*, volume 4, pages 2324–2327, 1996.

A. Ikeno and J. H. L. Hansen. The effect of listener accent background on accent perception and comprehension. In *URASIP Journal on Audio, Speech, and Music Processing*, page 76030, 2007.

K. Jesney. The use of global foreign accent rating in studies of l2 acquisition. Technical report, Language Research Centre, University of Calgary, 2004.

Y. E. Kim, E. Schmidt, and L. Emelle. Moodswings: a collaborative game for music mood label collection. In *Proc. ISMIR 2008*, 2008.

H. S. Magen. The perception of foreign-accented speech. *Journal of Phonetics*, 26:381–400, 1998.

R. C. Major. Paragoge and degree of foreign accent in Brazilian English. *Second Language Research*, 2(1):53–71, 1986.

M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. In *Proc. ISMIR 2007*, 2007.

S. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura. Non-native English speech recognition using bilingual English lexicon and acoustic models. In *Proc. ICME '03*. NTT Cyber Space Labs, 2003.

W. C. McDermott. *The Scalability of Degrees of Foreign Accent*. PhD thesis, Cornell University, 1986.

R. Mitton, D. Hardcastle, and J. Pedler. BNC! Handle with care! Spelling and tagging serrors in the BNC. In *Corpus Linguistics Conference*, 2007.

M. J. Munro, T. M. Derwing, and J. E. Flege. Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Studies in Second Language Acquisition*, 27:385–403, 1999.

T. J. Riney, M. Takada, and M. Ota. Segmentals and global foreign accent: The Japanese flap in EFL. *TESOL Quarterly*, 34(4):711–37, 2000.

A. H. Schlaikjer. A dual-use speech CAPTCHA: Aiding visually impaired web users while providing transcriptions of audio streams. Technical report, Carnegie Mellon University, 2007.

E. Shriberg. Disfluencies in switchboard. In *Proc. ICSLP*, pages 11–14, 1996.

E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak. Detecting nonnative speech using speaker recognition approaches. In *Proc. Odyssey Speaker and Language Recognition Workshop*, 2008.

I. Thompson. Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, 41(2):177–204, 1991.

L. M. Tomokiyo and A. Waibel. Adaptation methods for non-native speech. In *Proceedings of Multilinguality in Spoken Language Processing*, 2001.

D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *8th International Conference on Music Information Retrieval (ISMIR*, pages 535–538. Österreichische Computer Gesellschaft, September 2007.

E. M. Varonis and S. Gass. The comprehensibility of non-native speech. *Studies in Second Language Acquisition*, 4(2):114–36, 1982.

B. Vieru-Dimulescu and P. B. de Mareuil. Perceptual identification and phonetic analysis of 6 foreign accents in French. In *Proc. ICSLP '06*, pages 441–444, 2006.

L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004.

L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 79–82, New York, NY, USA, 2006.

S. H. Weinberger. Web accents. In *Proc. Phonetics Teaching & Learning Conference 2005*, 2005.

H. Yu. *Recognizing Sloppy Speech.* PhD thesis, Carnegie Mellon University, 2005.

Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon. Accent detection and speech recognition for Shanghai-accented Mandarin. In *Proc. Interspeech '05*, 2005.