Empirical Analysis of Bias in Voice-based Personal Assistants

Lanna Lima*
lima.lanna@edu.unifor.br
lannalima.br@gmail.com
Universidade de Fortaleza
Fortaleza, Ceará

Vasco Furtado
Universidade de Fortaleza
Fortaleza, Brazil
vasco@unifor.br

Elizabeth Furtado
Universidade de Fortaleza
elizabeth@unifor.br

Virgilio Almeida
Computer Science Department, UFMG
virgilio@dcc.ufmg.br

ABSTRACT

Voice-based assistants are becoming increasingly widespread all over the world. However, the performance of these assistants in the interaction with users of languages and accents of developing countries is not clear yet.

Eventual bias against specific language or accent of different groups of people in developing countries is maybe a factor to increase the digital gap in these countries.

Our research aims at analysing the presence of bias in the interaction via audio. We carried out experiments to verify the quality of the recognition of phrases spoken by different groups of people. We evaluated the behaviour of Google Assistant and Siri for groups of people formed according to gender and regions that have different accents. Preliminary results indicate that accent and mispronunciation due to regional differences are not being properly considered by the assistants we have analyzed.

CCS CONCEPTS

• **Human-centered computing → User studies**; **Laboratory experiments**; **Empirical studies in HCI**.

KEYWORDS

voice-based personal assistants, bias, machine learning.

ACM Reference Format:
Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical Analysis of Bias in Voice-based Personal Assistants. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317597

1 INTRODUCTION

Audio interaction will be the fastest and most convenient way for many purposes compared to using tactile monitors or remote controls [6]. Voice-basec assistants are rapidly becoming widespread and is reaching mainstream audiences in US and countries in Europe and Asia [8]. Voice-based wearables, audio, and video entertainment systems are examples of that. Also, voice control systems will soon be the norm in vehicles. Google Home and Amazon's Alexa are making popular the idea of a "`smart home`" across millions of households in the US. ComScore[8] predicts that by 2020, half of all our searches will be performed by voice. This scenario will be impactating in countries like Brazil, where a large part of the population is functionally illiterate [4]. The high adherence of Brazilians to Whatsapp (120 million active users/ month) using communication via audio is indicative of this [3].

Despite this enlightening perspective, voice-recognition technology has imperfections. The Wired Magazine recently exemplified how accent and gender differences might difficult human interaction [6]. A typical database of American voices, for example, would lack poor, uneducated, rural, non-white, non-native English voices. These kind of problems occurs in US with its large immigrant population, but is particularly critical in the developing world.

In Brazil and India, for instance, countries with large territories, geolinguistic variations are strongly related to the socioeconomic classes of the population. In Brazil, the southern region is known as the industrial center of the country, while the northern region possesses a less wealthy population. The economic issues influence the access to technology as well as the importance of the linguistic variations, that can provoke bias in the audio interaction.

Our research aims to analyze the presence of bias in the interaction via audio. The first step towards this was through an experiment to verify the quality of the recognition of phrases spoken in Portuguese by people of different genres and originating from different regions with different accents. Google Assistant and Apple's Siri were the assistants chosen.

Two questions have driven the research:

1: Does user's gender affect the understanding of the interaction system via audio?

2: Does the user's accent affect the understanding of the interaction system via audio?

An experiment was conducted in a Lab of a brazilian University with 20 volunteers, with the objective of elucidating these research questions. Preliminary results indicate that regional accent issues are not being properly considered by Google and especially by Siri.

2 RELATED WORK

Bias from audio interaction is a theme relatively few explored in the literature. It cannot be strictly classified as a bias of interaction

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-9999-9/18/06.
https://doi.org/10.1145/3308560.3317597

cf. [1], because the latter typically concentrate on the user interface and the user's own self-selected, biased interaction. Bias in the interaction is typically studied by analysing the user actions (like mouse movement and number of clicks). As for the audio interaction, this kind of technique is useless. The bias in the audio interaction comes from the process of training the assistant. It is a kind of data bias [1] originated from the creation of unbalanced databases that exclude parts of a population. Also, the interaction via audio depends rather on the user context and features than in their action and choices. Different speaking styles of human beings (i.e. the language accent, ethnic origin, emotion, gender, age) and the user's physical location are example of this.

Yanli Zheng and colleagues [12] worked on speech recognition for accents on Chinese. They evaluated the accuracy of Automatic Speech Recognition (ASR) of Interactive Voice Response systems (IVR), and proposed strategies to improve the accuracy of these systems. Similarly, reference [11] studied the difference in regional accent of English in England, America, Australia and India. A systematic view of ASR systems was done in [9] but the features used in the comparison of the systems are focused on their performance such as speed in real-time and response time. The study is a useful source for developers of devices like thermostats, doorbells, light bulbs and car accessories who want to add IVR (like virtual assistants) to the devices.

On the other hand, reference [10] evaluated the accuracy of speech recognition of a talker by a human (the listener). They show that listeners identify voices more accurately in their native language than in an unknown, foreign language. The talker identification [7]is studied, varying aspects related to the listener competence in the language to be spoken (such as her/his language familiarity, age, early language experience, cognitive skill, etc.).

Bellegarda [2] investigated the model behind Siri and its ability to infer knowledge. He concluded that ontology-based systems, such as Siri, are better suited for initial deployment in well-defined domains across multiple languages, but must be carefully tuned for optimal performance. Data-driven systems have the potential to be more robust, as long as they are trained on enough quality data.

Despite this related work, none in the literature has focus on bias due to regional accents within the country and the relation that this can imply in economic differences. Our study in this article with an experiment with participants speaking in Brazilian Portuguese for popular assistants is a step towards this.

3 EMPIRICAL EVALUATION

3.1 Methodology

An experiment was designed to verify the understanding of voice assistants in smartphones. The understanding was measured from the transcription of a spoken sentence by participants and the counting of the number of attempts required by the assistant to make the accurate transcription. In relation to the number of attempts, we assumed that a user would give up from being understood after a certain number of attempts. For the whole experiment, we considered that if after three attempts the transcription of the sentence was not accurate, the sentence would be classified as inaccurate. We selected for our experiments two smartphone assistants: Apple's Siri and Google Assistant. Twenty volunteers participated of the experiments: 10 of them interacted with Siri and 10 with Google. The profile of the participants varied in gender, nationality, accent, age, income range and education. Illiterate people were not included, but non-natives who could read portuguese and speak brazilian portuguese were.

Each participant was asked to read four sentences. The sentences contain words, chosen by HCI experts, that are often pronounced in the wrong way due to variation in pronunciation depending on the accent of the participants. For example, the word beneficent (in Portuguese "beneficente"), is commonly pronounced in the wrong way by people from certain regions of Brazil. Other words in which this occurs were used such as: "bicarbonato", "iogurte", "cabeleireiro", "problema", "padrasto", "cérebro", "entretido", "brócolis" and "crocante". In order to include a random factor in the process, seven sentences were elaborated and the participants made the draw of four of these seven. The sentences were elaborated as questions to more closely simulate the interaction with assistants. The seven sentences were the following:

"How do I get the privilege of going to a beneficent ("beneficente") event?"

"Is it bad to put baking soda ("bicarbonato") in yogurt ("Ãŕogurte")?"

"Where's the nearest hairdresser ("cabeleireiro")?"

"Where can I treat a thoracic problem ("problema")?"

"Is it coincidence meeting someone who has the same stepfather ("padrasto")?"

"What to do to make the brain ("cÃĽrebro") entertained ("entretido")?"

"How to cook the broccoli ("brócolis") to make it crispy ("crocante")?"

For each sentence read two metrics were collected, the accuracy of the sentence transcription and the number of attempts, the latter being limited to three attempts. All sessions were individual and conducted by two experts in Human Computer Interaction (HCI), during December 4-8th , 2018.

3.2 Characterization of the Participants

The twenty participants were divided into two groups of ten. In order to reduce learning effects it was conducted a Between-Subjects study [5], thereby each group interacted with only one of the assistants.

The participants profile containing their characteristics of gender and region were collected via a socioeconomic cultural questionnaire and anonymized.

The distribution of participants by gender and accent was not uniform. Out of the twenty participants seven were female and 13 male. Information about the accent was stated by the participant, usually this information was filled out with their place of birth, often the participants declared the city, state or country of origin, in the case of foreigners fluent in Brazilian Portuguese. As for the accent, native participants were aggregated by brazilian regions.

Eight participants declared an accent from the Northeast region, six from the Southeast region, two Foreigners who do not speak Portuguese natively and four of participants were aggregated as Others because of their lack of representativeness if they were considered by their proper regions, one was from the South, one from the Midwest and two were from the North .

Table 1: Distribution of Participants by Accent

Region	Participants	Proportions
Northeast	8	40%
Southeast	6	30%
Others	4	20%
Foreign	2	10%
Total	**20**	**100%**

3.3 Metrics of Evaluation

We defined two metrics to assess the quality of the recognition of the sentences spoken to the assistants. The first metric was the transcription, by the assistant, of the sentence spoken by the participant. Each transcript metric could assume two values: accurate and inaccurate. The label accurate was assigned in cases where the transcription of all words pronounced were done in correct Portuguese, even in cases where a word was pronounced incorrectly, for example, if the participant pronounced "beneficiente" and the assistant transcribed "beneficente" as should be in the formal Portuguese.

The label inaccurate was attributed to the transcripts wrongly done even in cases in which the assistant used a word similar to the one that was pronounced. For example, the assistant transcribed "yakult" (a brand of probiotic fermented milk) when the participant pronounced the word yogurt in the sentence.

The second metric is the number of attempts. It refers to the number of times the participant repeated the sentence if it was not transcribed exactly. There was a limitation of three attempts for each sentence, considering that in real-world interactions users have a tolerance of system errors until withdrawal. Thus, for example when the participant pronounced yogurt and the assistant transcribed "yakult" the participant could pronounce the sentence up to three times.

4 RESULTS

4.1 Number of Attempts and Quality of Transcriptions

Each of the twenty participants read four sentences, that should generate a base of 80 readings if all the readings had been accurate. As 24 sentences have been read a second time and 11 have been read a third time, the total base contains 115 readings. The proportion of accurate and inaccurate by tries of reading is presented in Figure 1. In the first round of readings (T1), 70% of the sentences have been transcribed accurately and 30%(24 sentences) in an inaccurate form. These 24 sentences were read again (T2) and 54%had an accurate transcription and 46%inaccurate. Finally, a third round of reading (T3) was performed with the eleven sentences that have been inaccurately transcripted. Since the maximum number of tries were limited to 3, six out them have not been transcripted correctly.

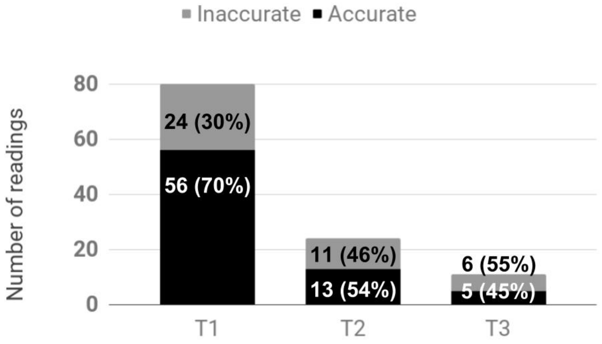

Figure 1: Proportion of accurate and inaccurate readings by attempts.

4.1.1 *Siri vs Google.*

Figure 2 shows the proportion of accurate and inaccurate readings for Siri (left side) and Google Assistant (right side). The number of times that Google assistant transcribed , in the first attempt (T1), a sentence read by the participants were greater than Siri's one (88% vs. 52%). It is worth noting that even though Google assistant has a better performance in transcribing the sentences correctly in the first attempt, both reach similar levels of sentences without correct transcription (3 transcriptions of sentences were inexact for each assistant). A qualitative analysis showed that the sentence " Como faço para ter o privilégio de ir a um evento beneficente?" (sentence a in Section 3.1) has been wrongly transcribed by both assistants. The three sentences that Siri transcribed non-exactly were pronounced with an accent from Northeast while Google assistant had two sentences pronounced by non-native speakers and one with accent from the group Others.

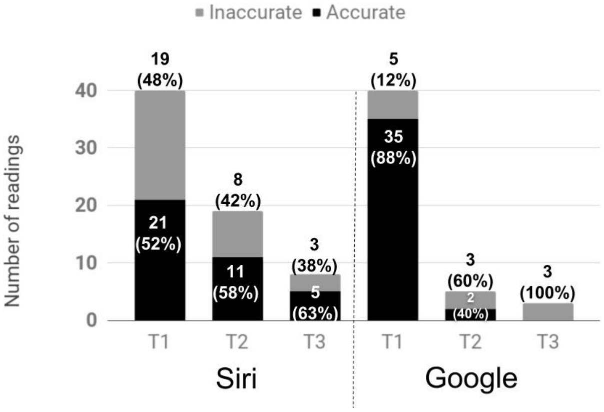

Figure 2: Proportion of accurate and inaccurate readings for Siri and Google.

4.1.2 *Female vs Male.*

Female participants were better understood by assistants than male participants. Figure 3 shows the proportion of the distribution of accurate and inaccurate by tries of reading by female participants (left side) and male participants (right side).

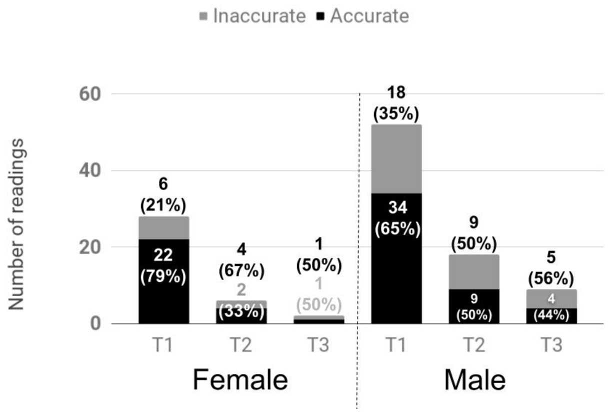

Figure 3: Proportion of accurate and inaccurate readings for Siri and Google.

4.1.3 *By Accent.* Participants with an accent from the Southeast were better understood by the assistants than participants from other regions. Figure 4 shows the proportion of accurate and inaccurate for the tries of reading by participants, grouped by the region of accent: Northeast, Southeast, Others and Foreign.

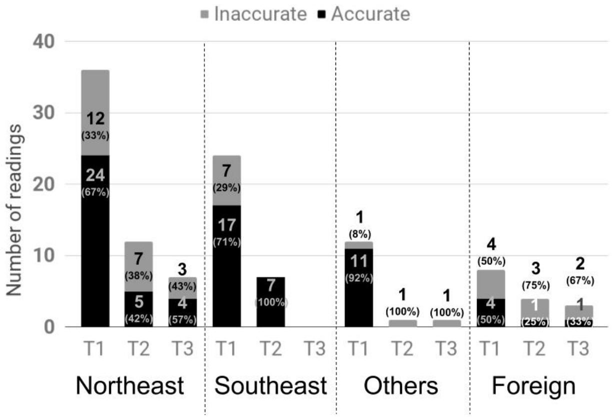

Figure 4: Proportion of accurate and inaccurate readings by region of the accent of the participants.

4.1.4 *Siri by Gender.* Siri had a better understanding of female participants than of male participants. Figure 5 shows the proportion of accurate and inaccurate for the tries of reading for Siri for female and male participants.

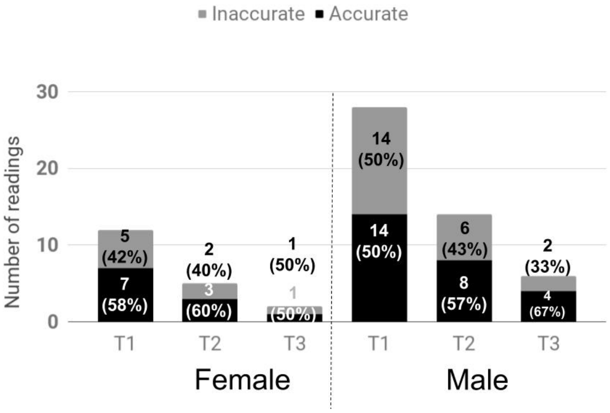

Figure 5: Proportion of accurate and inaccurate readings by female and male participants using Siri.

4.1.5 *Google Assistant by Gender.* Google also had a better understanding of female participants than of male participants. It was better than Siri for both female and male. Figure 6 shows the proportion of accurate and inaccurate by tries of reading for Google for female and for male participants.

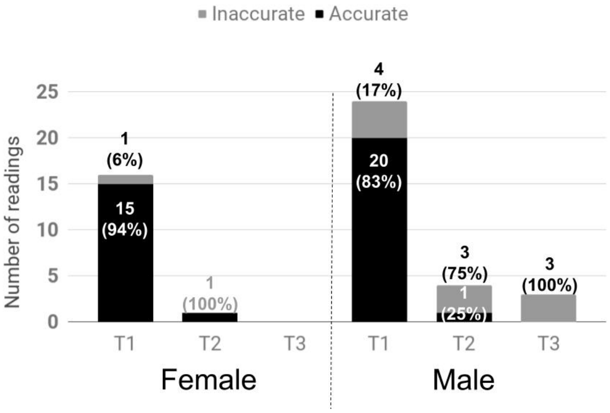

Figure 6: Proportion of accurate and inaccurate readings made by female and male participants using Google Assistant.

4.1.6 *Siri by Accent.* As aforementioned, Siri had a better understanding of participants of the Southeast region. Surprisingly it worked better even for Foreigners than for participants from the Northeast region. Figure 7 shows the proportion of accurate and inaccurate by tries of reading by participants from the different regions of accent. There were no participants from Others using Siri.

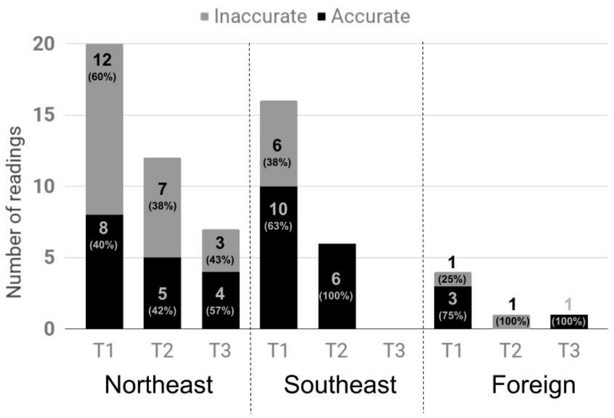

Figure 7: Proportion of accurate and inaccurate readings by region of the accent using Siri.

4.1.7 Google Assistant by Accent. Google had better understanding of participants of the Northeast. Figure 8 shows the proportion of accurate and inaccurate by tries of reading by participants from different regions.

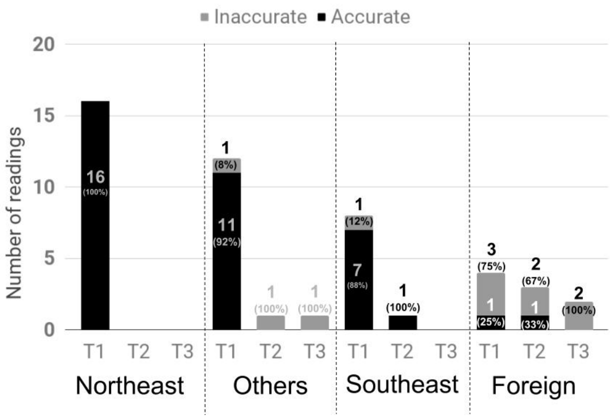

Figure 8: Proportion of accurate and inaccurate readings by accent using Google Assistant.

4.2 Summary

Table 2 presents the number of sentences read to each assistant, segmented by gender and accent. For instance, it shows that 12 sentences were read by female participants using Siri and 16 by female participants using Google. The arithmetic average of the number of attempts that the participants made to be understood (i.e., the number of tries to reach the exact transcription) is 1.50 for Siri and 1.06 for Google.

It evidences that, in average, the number of tries using Siri for participants with typical accent from the Northeast (2.00) is almost double the participants with accent from the Southeast (1.31). Considering gender segmentation, both assistants have better results for female participants, Google presented the best performance when transcribing female readers, with an average of 1.06 tries. We have applied the chi-square test to check whether transcriptions had a detectable difference regarding gender and accent depending on which assistant was used. The result of gender to Siri was χ^2(NA, N=66) = 1.02, p>0,05 and to Google χ^2(1, N=48) = 4.32, p<0.05 and of accent to Siri was χ^2(3, N=66) = 13.66, p<0.05 and to Google was χ^2(3, N=48) = 19.40, p<0.001. When comparing transcription between genders, there was a significant difference when the reading was made for Google Assistant. However, for Siri the difference in transcription regarding gender was not significant. Regarding the comparison between transcriptions by accent both Google Assistant and Siri presented a significant difference.

Table 2: Number of sentences read and average of tries by assistant, Gender (Gn) and Accent (Ac). (S=Siri, G=Google), Accent: Northeast (NE), Southeast (SE), Others (O) and Foreign (F)

		Number of Sentences		Proportion of Accurate Transcriptions		Arithmetic Average of Tries	
		S	G	S	G	S	G
Gn	F	12	16	58%	94%	1.50	1.06
	M	28	24	54%	68%	1.71	1.21
Ac	NE	16	16	44%	100%	2.00	1.00
	SE	16	8	73%	89%	1.31	1.13
	O	4	12	-	79%	1.75	1.00
	F	4	4	67%	22%	1.50	2.25

5 CONCLUSION

This study presents a preliminary analysis indicating that the training process of smartphone assistants, for the Brazilian Portuguese, can be biased towards voices of individuals from the most developed part of the country. It is particularly true for the case of Apple's assistant Siri when recognizing phrases with accent for the Southeast region. Variations in the quality of recognition on the basis of gender are also an indicator that further investigations need to be carried out in order to specify the reasons for it.

As future work, we plan to investigate if the results obtained here are also true regardless of whether the speech is composed of isolated words, short sentences, or long samples. We also plan to investigate variations on the transcriptions caused by the order of the sentences presented to the assistant, this can indicate whether the assistant has the potential to learn after unsuccessful tries. We consider that transcriptions are not the only way to evaluate the quality of the interaction via audio. The quality of a dialogue, for instance, depends on user's actions and choices. When an assistant fails to understand a word, an user tends to change (slightly or not) his/her intonation and/or timing, in the attempt to make the system understand the intended utterance. How much these changes bring prejudice to the dialog (e.g. making the dialog tedious) is a theme for further investigation.

Some limitations on the methodology of this preliminary study will be removed in the future, such as the unbalanced number of participants by gender and accent. We should construct the size of the samples proportionally to demographic indicators of the country. Semantics was not a variable analyzed in the scope of this study, which would encompass regionalisms and words employed by specific social classes.

This work provides useful information for the definition of rules and policies regarding the evolution of voice-based assistants and devices in the developing world. Such policies must protect vulnerable social groups, that are excluded due to social, economic or regional issues.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[2] Jerome R Bellegarda. 2014. Spoken language understanding for natural interaction: The siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*. Springer, 3–14.

[3] Estado de São Paulo Newspaper. 2017. WhatsApp chega a 120 milhões de usuarios no Brasil.

[4] Adriana de Souza e Silva, Daniel M Sutko, Fernando A Salis, and Claudio de Souza e Silva. 2011. Mobile phone appropriation in the favelas of Rio de Janeiro, Brazil. *New Media & Society* 13, 3 (2011), 411–426.

[5] Anthony G Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314.

[6] Reuters Institute. 2018. The Future of Voice and the Implications for News.

[7] The Intercept. 2018. Amazon's accent recognition technology could tell the government where youâĂŹre from.

[8] Wired Magazine. 2017. Voice Is the Next Big Platform, Unless You Have an Accent.

[9] Rami Matarneh, Svitlana Maksymova, Vyacheslav Lyashenko, and N Belova. 2017. Speech recognition systems: A comparative review. (2017).

[10] Tyler K Perrachione. 2017. Speaker recognition across languages. Oxford University Press.

[11] Moirangthem Tiken Singh, Abdur Razzaq Fayjie, and Biswajeet Kachari. 2015. Speech Recognition System for North-East Indian Accent. In *International Journal of Applied Information Systems (IJAIS)*. Vol. 9. Citeseer.

[12] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Daniel Jurafsky, Rebecca Starr, and Su-Youn Yoon. 2005. Accent detection and speech recognition for shanghai-accented mandarin. In *Ninth European Conference on Speech Communication and Technology*.