Phonetic Features and Perception of Russian L2 English

Jessica Kuleshov - jjk2235

December 17, 2021

1 Introduction

For the author, the subject of Russian accented speech is most intriguing, as growing up around family members their accents have been a distinct marker of being an ”other” - in daily technological use such as voice commands, in social situations, and even when describing themselves. With an accent that seems to carry so much weight in the lives of its speakers, it is important to determine what qualitative and quantitative aspects there are in how this accent is represented physically versus how it is perceived. This paper explores the Russian-English accent from two distinct angles: A quantified approach in which the phonetic features of the accent are determined, and a qualitative approach in which the perception and sentiment of the Russian accent are assessed by both non-Russian speakers and the speakers themselves. In the first section, the author determines the quantitative phonetic content that the speaker's voice actually entails - a difference that then could be implemented in more smart devices and voice-activated technology, as well as be used as a baseline ground truth during a comparison of the perception of Russian-accented speech. In the second section, the author uses the information from the perception of the accent to posit that within Russian-speaking populations in English communities, there is a specific heightened awareness of one's own accent, a tendency for over-correction, and a strongly different perception of their own accent than those of the people around them.

2 Quantification of Russian accent phonetic features using ASR

2.1 Background

Automatic Speech Recognition (ASR) is the field of study that focuses on research in acoustic speech recognition using computational methods. The field is a strong combination of signals processing techniques and, as of recently, machine learning techniques. ASR has been a very strong field in the past several decades, strengthened most recently by a growing interest in incorporating spoken-word commands into a modern Internet-of-Things framework, where customers want to be able to ”talk” to their devices and give directions or ask questions without having to type. [Silverio-Fernández et al., 2018]

However, ASR has historically relied on very specific constraints in order to properly function in a given environment [O'Shaughnessy, 2008]. The task of speech recognition has been historically divided into two issues: One of Speaker-Independent Recognition, where the system can recognize multiple voices using the same general model, Speaker-Dependent Recognition, where the prosodic features of an individual's voice are taken into account. The advent of speaker recognition, however, tends to not take accents very well and is a known problem within the field. This problem is especially an issue in aforementioned ”smart” devices, where it is known to be so difficult that for a model to recognize a thick Scottish accent it is considered a gold-standard model. These supposedly gold-standard models typically work on datasets with a limited number of commands, where there can be enough training data and predictable variations in order to be able to effectively parse regardless of accent. However, when tasked with interpretation of a complicated command, such as asking Siri a complicated question about a technical subject, i.e. (”How do ergative-absolutive languages such as Basque function?”), these models often fail on non-clear or accented English speech. Considering that over 66% of English speakers are nonnative, this is a critical error that results in many frustrated device users.[Paul, 2017]

Multiple attempts have been made in order to create models that account for accented English speech. One recent paper from Yi et. al (2016) describes a multi-LSTM approach with CTC

(connectionist temporal classification) regularization for detecting multiple accents in Mandarin, showing significant improvements over models used in current devices. Other ASR models such as PLASER have been used in pronunciation assessment and improvement, where deviations in the data are treated as errors to be fixed and suggestions for pronunciation improvement are included.[Mak et al., 2003] Another approach from Chen et. al (2020) used a more by-hand approach in order to gather these deviations, transforming data into Praat TextGrids and then hand-selecting features, then calculating F1/F2 Euclidean distances in order to gather deviations, which they found to be very successful.[Chen et al., 2020] These approaches suggest that deviations from the data consist of the speaker's accent in the language, and if some kind of model can detect these changes, then it should be possible to derive a quantitative definition of what an "accent" is from these deviations.

It may seem like a fairly difficult problem to solve, as there are multiple indicators of an accent such as stress, pitch, tone choice, and segmental errors, fluency (or phonetic choice) has been found to be the most prominent errors, accounting for around 70% of the variance in data.[Kang, 2013] With this in mind, studying the phonetic content of a Russian accent using an ASR model such as those used previously or for pronunciation assessment is only a natural step forward.

2.2 Methodology

This work is based on the assumption that an accent is defined as a collection of somewhat organized phonetic and prosodic deviations from a "standard" lect of a language. If this is the assumption, then mathematically the difference between the phonetic features of accented- and non-accented speech would result in the set of deviations from the accent. In order to obtain these deviations, one must first extract the pitch features, compare accented- and non-accented speech to the same baseline to normalize the data, and then compare these two differences and phone-level confidences to determine the places where pronunciations differ.

Data Set	Usage	Contents for this project	Source
TED-LIUM	Training set	TED Talks by English speakers, native and nonnative, 118 hours	openslr.org
LibriSpeech	Forced alignment	Audiobook corpus of native English speech, 100 hours	openslr.org
Speech Accent Archive	Test set	48 recordings of Russian-diaspora speakers reading same passage	kaggle.com
Russian-American English		7 recordings of Russian-American immigrants reading same passage	Gathered by author

Figure 1: Information about datasets used for this project.

2.2.1 Datasets

All ASR processes were done using Kaldi, an open-source speech recognition toolkit that allows users to design their own models or use pretrained models to perform decoding [Povey et al., 2011].

Three datasets are used in this process for different steps. The TED-LIUM 3 dataset consists of 118 hours of TED Talks, sampled at 16 kHz[Hernandez et al., 2018]. This data is very clean and consists of speakers with clear English, determined to be very coherent and understood by the audience, since TED typically selects for speakers with clear presenting skills. This dataset was used for training the language model which the other two datasets are scored against. The model was downloaded pretrained with phonetic and pitch features, and all other required programs and scripts are included within the native Kaldi toolkit.

The LibriSpeech corpus is a corpus of audiobook recordings from native English speakers (both British and American English) and consists of two sections: one for 100 hours of clean training data, and one for 360 hours of semi-clean training data, though for this use case only the 100-hour set is employed [Panayotov et al., 2015]. This is generally also high-quality data and sampled at various rates, so for this use case all audio was downsampled to 16 kHz. The LibriSpeech dataset is treated as a baseline for English speech in this project and will be henceforth

referred to as the ”baseline set”.

The final corpus is comparatively very small, and is a synthesis of two smaller datasets. For ease of understanding, this dataset will be referred to as the ”test set”. This totals only 55 recordings of native Russian speakers speaking English. 48 of the recordings are sourced from the Speech Accent Archive, a larger corpus of speakers from many native languages all reading the same passage shown in (1), and the other 7 recordings are from native Russian speakers used in a previous paper that the author has written shown in (2). The speakers have widely-varying ages of onset and ages at time of recordings, and though biased towards cities, it also has diaspora Russian speakers from post-Soviet-bloc countries. The average age of speakers was 35.7, with an average age of onset at 13.29. 55.6% of speakers were female, and 61.1% of speakers were from Russia itself.

The passage for the first 48 speakers is as follows:

(1) 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.

(2) That morning, I sat down at the beach and set my towel on the sand. I filled my bucket with water for my sandcastle and put it next to my phone. I did bring four sandwich halves, because I didn’t know which flavors to pick. Then, while eating lunch, I looked west and saw the whale! It was very much stranded, and really quite heavy, so I didn’t know what to do at the time. Without your assistance, I could not have moved it off the shore, so thank you again for helping me!

The two paragraphs elicited were built for different purposes. In the case of (1), the Speech Accent Archive paragraph was constructed to test for all of the most common English phonetic constructions and was tested by native speakers of various languages, so it could not be designed to be tested on what the speakers necessarily would find ”difficult” or not native to their own language. In the case of (2), this paragraph was constructed specifically with native Rus-

sian speakers in mind, understanding that Russian speakers tend to struggle with certain sounds [Tumshevits, 2019]. These sounds are: u vs. ɯ, u vs. ʊ, ʌ vs. a, æ vs. a vs. ɛ, i vs. ɪ vs. ɨ, l vs. ʟ, r vs. ɹ, h vs. x, palatalization, and final devoicing.

By combining these two datasets, this should be able to give a more holistic view of what a Russian accent may truly appear as. It is also important to know that these recordings also suffer from a common issue in elicited speech - often, the recordings sound very purposeful, as if the speaker is attempting to enunciate each sound separately. This can somewhat skew the data, but ideally this over-enunciation brings out what is usually more subconsciously noticed by native English speakers as the accent.

2.2.2 Method

The initial step is to format the data such that the Kaldi framework can work with it later on. Here, every file in the test set is an mp3 and downsampled to 16 kHz so that a LibriSpeech script can prepare the data and outut a wav.scp formatting file that matches each file to its speaker id and bulk-converts them to a WAV formatting. During this step, other files, such as spk2utt and utt2spk are generated, where the utterances for each speaker are mapped bidirectionally (necessary for different internal scripts later in the process).

After initial data preparation, the next step is to extract phonetic and pitch features for both the baseline set and the test set. This was done using a script called ”make_mfcc_pitch.sh”, which is run on both sets. After these steps, the Cepstral Mean and Variance (CMVN) statistics are computed for each speaker to perform an initial normalization across recordings.

Once phonetic and pitch features are extracted for each recording, these features can be compared to the language model (LM) that was trained on the slight-to-no-accent speech from the TED-LIUM dataset. It is important to note here that Kaldi’s phonetic features are known to not be entirely accurate as the phonetic alignments are typically restricted to a certain set of preset phones, but the confidence values themselves can still be taken into account for giving these differences. The first step is to re-align the TED-LIUM dataset to the data from LibriSpeech, or

re-computing the weights for each given triphone given its environment. This establishes the native English baseline necessary to compare the data in our test set. From there, the newly-aligned and retrained model is used to decode the test set. The decoding was done using feature space maximum likelihood regression (fMLLR) for normalizing across speaker-adapted features, though it is unclear currently whether removing the normalization for speaker-adapted features would have an effect on the results. This process outputs structures native to Kaldi called lattices, which contain information to determine the best phonetic pronunciations given the environment, and each phone also has a corresponding ”confidence level” to show how likely that phone is.

At this stage, the resulting phones per-speaker can then be compared to the transcripts. For each given phone, its F1 and F2 distances can be calculated from the proper phone if the phones are not the same, and the confidence levels in phones ideally can also show how close the pronunciation is. Throughout this comparison process of phones to their ground truth, a dictionary is kept in order to keep track of the frequency of phones that are less certain across speakers.

2.3 Results

Though it was difficult to obtain the F1 and F2 distances through Kaldi directly, the phone confidence levels and errors in phones were obtainable. However, the confidence levels were not as varied as would have been expected - most phones were predicted with a confidence level of 1, with an average across speakers of .989. This is not enough variation to make the confidence level outputted by Kaldi to be a useful metric for understanding deviations in accents. However, seeing as the word error rate (WER) outputted by the program was still 54% on average, this means that the phones predicted - though confidently - were confidently incorrect and thus can still be used for analysis.

Since Kaldi does not use IPA symbols in their phonetic representations, they typically employ a combination of two phones to create the sound. With this representation, it is still possible to get a distinction between ”close-sounding” phones, as a comparison of the first letter may be enough to show the similarity between phones. This system has been difficult to implement, but

phone	err_rate	near_phones	error_phones
AA	0.3333	[]	['OW']
AW	0.5714	"['AE', 'AH']"	"['EH', 'OW']"
DH	0.4286	[]	"['K', 'SH', 'D', 'AY', 'W', 'HH', 'AH', 'AO']"
Y	0.6667	"['G', 'G', 'F', 'F']"	[]
HH	0.4167	[]	"['G', 'AH', 'B', 'K', 'F']"
CH	0.25	[]	"['D', 'EY', 'D']"
EH	0.1429	[]	"['IH', 'AO', 'UH']"
NG	0.3636	[]	"['N', 'AH', 'CH', 'OW']"
TH	0.5556	[]	"['Y', 'Y', 'R', 'R', 'F']"
IY	0.5152	['IH']	"['AH', 'UH', 'AO', 'EH', 'P', 'L', 'AA']"
B	0.25	"['W', 'D']"	['AY']
AE	0.6154	"['AA', 'AY', 'AH', 'AW']"	"['UH', 'UW', 'EH', 'IY', 'OW', 'IH', 'L']"
D	0.2368	"['T', 'T', 'T', 'L', 'N', 'L', 'T']"	"['DH', 'AY']"
G	1.3333	"['F', 'F']"	"['UH', 'IY']"
F	0.4762	"['W', 'B', 'M', 'M', 'B', 'K', 'G']"	"['IH', 'HH', 'HH']"
AH	0.3784	"['AE', 'AY', 'AE', 'AE', 'AE', 'AY']"	"['EH', 'IH', 'F', 'F', 'N', 'HH']"
K	0.35	"['P', 'P', 'R', 'L', 'T', 'N', 'F']"	[]
M	0.32	"['N', 'K', 'F', 'T', 'N']"	"['DH', 'HH', 'HH']"
L	0.36	"['Y', 'V', 'N', 'D', 'K', 'N', 'W']"	"['UW', 'UH']"
AO	0.5385	"['AH', 'AH', 'AH', 'AE', 'AH', 'AH']"	"['OW', 'EH', 'OW', 'ER', 'R', 'EY']"
N	0.3393	"['D', 'L', 'W', 'L', 'P', 'M', 'M', 'R']"	"['AH', 'IH', 'TH', 'OW', 'IY']"
IH	0.3922	"['IY']"	"['AO', 'AH', 'EH', 'OW', 'AE', 'UH', 'JH']"
S	0.375	"['F', 'F', 'F', 'K', 'N', 'D', 'W', 'F', 'R']"	"['TH', 'TH', 'HH', 'AE', 'UW', 'SH']"
R	0.4074	"['D', 'L', 'L', 'Y', 'D', 'F', 'G']"	"['ER', 'AY', 'IH', 'UW']"
EY	0.25	['ER']	['AH']
T	0.3425	"['D', 'F', 'V', 'N', 'K', 'R', 'P', 'W']"	"['TH', 'AH']"
W	0.25	"['D', 'F', 'T', 'M', 'Y']"	"['DH', 'IY', 'AH']"
V	0.4375	"['P', 'G', 'L', 'K', 'D', 'F', 'N']"	[]
AY	0.359	"['AH', 'AW', 'AH', 'AH']"	"['EY', 'EH', 'IH', 'UH', 'F', 'D', 'L', 'IH']"
Z	0.2222	['S']	['CH']
ER	0.5556	['EH']	"['M', 'L', 'M', 'L']"
P	0.4286	['K']	"['AE', 'DH']"
UW	0.7778	[]	"['AH', 'IH', 'EH', 'G', 'IH', 'AH', 'AE']"
SH	1.0	[]	"['S', 'S', 'DH']"
UH	0.3333	[]	"['N', 'EH', 'IH', 'W']"
OW	0.3333	[]	"['AE', 'AE', 'AY', 'IY', 'AO']"

Table 1: List of expected phones and errors, as well as error rates for the given phones.

preliminary results for one speaker can be shown in Table 1. This speaker, lbi-007, was reading the passage in (2). From this initial run, it can be shown from the phone error rate which phones

were more commonly misinterpreted. No phone has a 0% error rate; due to the method by which the information was extracted, i.e. computationally aligning the data and forgoing hand analysis, this is to be expected. Instead, more useful results can be gathered from comparing error rates against the average of the data set, as well as by the frequency of other phones being interpreted. Another important note would be that the error rates themselves, on such a small data set, cannot be fully trusted; for example, the error rate for ”SH” was 1.0 - this is solely because the three times it was interpreted, it was incorrectly matched or lined up.

There still is intriguing data to look at, especially in vowels, where most lined up fairly well with the transcription and carry more meaning. For ”AO”, for example, it can be seen that either it is interpreted as an ”a” sound or more rounded and rhoticized. Generally, the ”a” sounds outside of ”AO” are somewhat fronted and/or raised, with common mistakes being ”EY”, ”EH”, ”IH”, and ”IY”. This can be interpreted as a move of palatalization, as well as in line with the i/ɪ/ɨ indistinction that Russian speakers have been observed to have previously.

Back vowels are not interpreted very well, and velarization is also tricky to observe with this data. This is perhaps because the phones that are in the text file derived by the language directory preparation only contains English phones, but it is unclear whether that is the case. It would also be understandable if sounds like ”ɯ” or ”ʟ” would go undetected in a model trained on English, as they would instead align more with a higher-weighted phone close to it in frequency. Future attempts at modeling this can perhaps rely on output alignments from a model trained on Russian speech itself (not accented English), as that would potentially produce more phones that a Russian speaker’s speech would actually consist of.

Once the rest of the data for other speakers has been processed and cleaned, this section will have more complete and comprehensive results. For the moment, speaker lbi-007 is still a good representation of an overall phenomena.

2.4 Discussion

These results are exciting and seem to lend some credibility to ASR as an acceptable method for interpreting Russian-accented speech, as it follows changes that one would have expected from hand-analyzed observations. Remnants of palatalization, final devoicing, and vowel raising can all be observed in the data that currently has only been gathered from one speaker, and the hope is that with more data these patterns will continue.

If a successful model can predict accent deviations such as this, perhaps these deviations can be implemented in smart technology in a more effective manner - perhaps by recognizing from a specific set of deviations that this is a Russian accent and to interpret the command as such, knowing what sounds it should be expecting to hear in a native English accent instead. This method of quantification also lends itself well to ”scoring” speakers in a much less-biased manner. This method of quantification will be used in the next section where the perception of the accent will be tested against this score ”ground truth”.

Further work can be done here in order to more accurately get deviations from standard English speech. One potential extension would be to more accurately calculate F1 and F2 distances; even using a Euclidean distance would be beneficial in approximating the closeness of phones in a better method than taking the topmost-confident phone per sound byte. Because the confidence levels did not have high enough variance this made the analysis more difficult and the results still had to be more manually analyzed than what would have been necessary previously. Further work also will be done in order to fully interpret the data from the other speakers in order to get the most comprehensive and accurate results for phone error rates and common data.

References

[Chen et al., 2020] Chen, W.-H., Inceoglu, S., and Lim, H. (2020). Using asr to improve taiwanese efl learners' pronunciation: Learning outcomes and learners' perceptions. In Proceedings of the 11th Annual Pronunciation in Second Language Learning and Teaching Conference, pages 37–48. Northern Arizona University.

[Hernandez et al., 2018] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. Lecture Notes in Computer Science, page 198–208.

[Kang, 2013] Kang, O. (2013). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. undefined.

[Mak et al., 2003] Mak, B., Wong, J., Lo, J., Siu, M., Ng, M., Tam, Y.-C., Chan, Y.-C., Chan, K.-W., Leung, K.-Y., Ho, S., and Chong, F.-H. (2003). Plaser: pronunciation learning via automatic speech recognition. In Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing -, volume 2, page 23–29. Association for Computational Linguistics.

[O'Shaughnessy, 2008] O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition, 41(10):2965–2979.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

[Paul, 2017] Paul, S. (2017). Voice is the next big platform, unless you have an accent | backchannel. Wired.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely,

K. (2011). The kaldi speech recognition toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[Silverio-Fernández et al., 2018] Silverio-Fernández, M., Renukappa, S., and Suresh, S. (2018). What is a smart device? - a conceptualisation within the paradigm of the internet of things. Visualization in Engineering, 6(1):3.

[Tumshevits, 2019] Tumshevits, A. (2019). Perception of russian-accented speech by native and non-native speakers of english. Charles University Department of the English Language and Literature.