Natives' and non-natives' pronunciation of Dutch, and it's influence on automated speech recognition

Joost Baas (joost.baas@phil.uu.nl)

October 29, 2010

Bachelor thesis Cognitive Artificial Intelligence, Utrecht University

Supervisor: Dr. Ir. Gerrit Bloothooft, Utrecht University

Abstract

The Autonomata and Autonomata TOO research programs have produced large annotated speech corpora mainly targeted for Personal Navigation Devices. In this paper these corpora are analyzed. Systematic relationships between subjects' mother language and the deviation of the realized pronunciations to the standard pronunciation are identified, and the effects this has on the accuracy of the Automated Speech Recognition system that was part of the Autonomata projects.

Contents

1 Introduction

Automated speech recognition (ASR) is the automated process of transforming speech into (a digital representation of) text, sometimes described as speech-to-text (STT). ASR is a difficult task. Speech is the most natural form of language, and intimately connected with it. To reach human-like proficiency levels in language requires a lot of other cognitive skills to be on par with humans too, because language is intimately tied with cognition in general. For this reason, language is sometimes called an AI-complete problem [13], meaning all Artificial Intelligence related problems have to be solved to solve this problem. Arguably, because of it's strong ties with language, the same can be said about speech recognition. Developing recognition systems with similar levels of accuracy and flexibility that humans possess is therefore out of the question for the foreseeable future.

However, by sacrificing flexibility, it is possible to achieve results suitable for practical applications. Those systems usually focus on a particular purpose, limited in scope, and therefor have a relatively small number of lexical items and grammar rules, are sometimes speaker-dependent, and often only work for only one language. Background noise can often limit the practical use of such applications. In this paper, one particular speech recognition system is analyzed, VoCon system by Nuance.

1.1

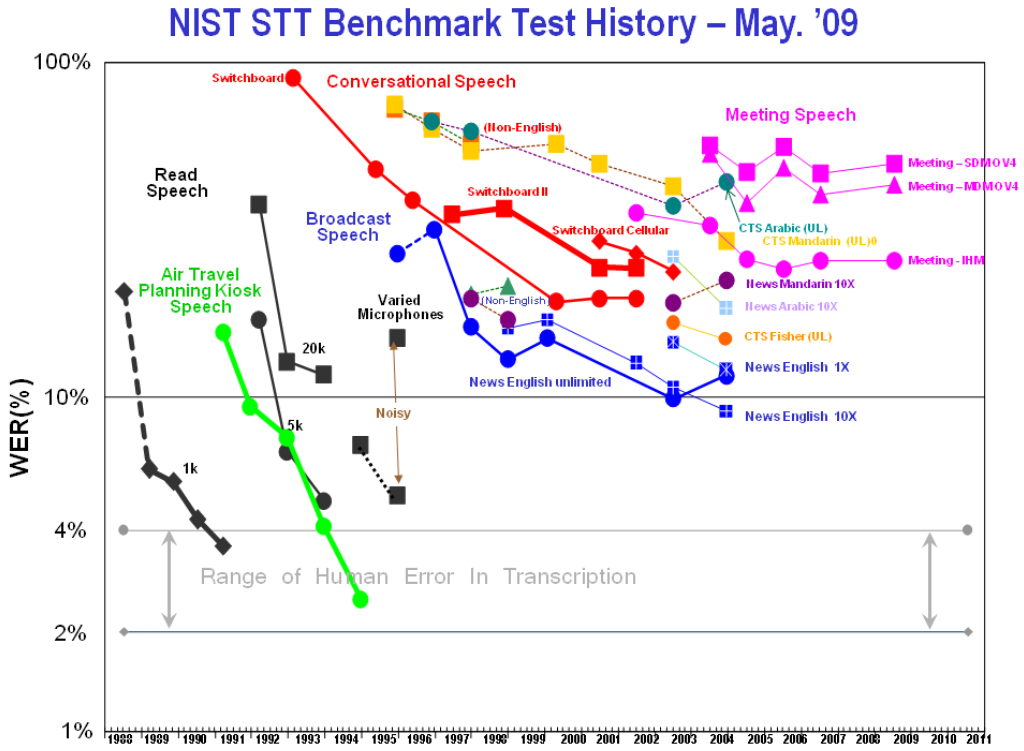


Figure 1: This graph gives an overview of word error rates (WERs) for different types of speech. Taken from Fiscus et al. (2007) [3]. Only specialized systems with limited lexicons achieve scores in the range of human error rates.

Though research had already been done on speech modeling in the preceding decades [2], the first real ASR system was developed in the early 50s [1]. A system designed by Bell Telephone Laboratories was able to recognize spoken digits. This system could be used with speech encoded in telephone quality sound and could achieve more than 97% accuracy rates, but because it had to be adjusted to the speaker it was not feasible to use in production systems.

Other early ASR systems experimented with statistical methods for syntax [4], segmenting the signal before attempting to recognize it, and trying to align speech patterns [6].

In the early 1970s, the first commercial ARS system became available, the VIP 100 system by Threshold Technology. This system was mainly used in industry for specific applications, and did not see widespread use. It did provide a taste of the possibilities of ASR, and the Advanced Research Projects Agency (ARPA) of the US Department of Defense started funding ASR projects. In the following decades, ASR systems gradually started using more refined methods to segment the signal and search for candidates in the vocabulary, with a shift from template matching to statistical frameworks, with Hidden Markov Models becoming the prevalent method in the late 1980s.

In the 1990s and 2000s increasing processing power and more refined methods for machine learning enabled wide-spread applications of ASR systems in both consumer and industrial markets [Figure 1].

1.2 Current applications of speech recognition systems

Commercial ASR systems are available that can be used to dictate larger texts. ASR can be a useful tool for people with disabilities like RSI or dyslexia, or in work settings where people don't have their hands available, like in medical settings [5]. Drawbacks are that these systems usually have to be used with a high quality microphone that is suitably positioned (like a headset), and that the software has to be trained to one particular speaker's voice, which takes time and makes it less flexible.

Some organizations are using ASR to transcribe large collections of audio or video recordings, usually to make an index of keywords or phrases that can be searched.

Many organizations provide information that can be accessed over the telephone. Especially in cases where the lexicon is highly structured and/or limited in size, ASR systems have replaced human operators. In spite of low audio quality over telephone lines, these applications usually have high accuracy rates because of their limited lexicon. It remains to be seen whether such applications will still be regularly used when smartphones with always-available internet connections become more common.

1.3 Autonomata projects

The Autonomata project is a cooperation between the universities of Gent (Belgium), Nijmegen (the Netherlands) and Utrecht (the Netherlands), and the companies TeleAtlas and Nuance [14]. In this project a corpus was created of spoken

name and address utterances, with phonemic transcriptions. Subjects had different native languages, and were living in Dutch-speaking areas for different amounts of time. Goal of this project was to provide a corpus that can be used to improve ASR systems for Dutch input.

In the Autonomata TOO project the same partners extended their goals to include native non-standard and multilingual terms [12]. In this project a spoken corpus was also created (be it smaller in size), only this time it included points of interest instead of person and geographical names. Unlike the first Autonomata project, the Nuance ASR system was used real-time, giving the subject feedback on what was recognized, and the opportunity to try again. Subjects had different mother languages as well.

The speech recognizer used in the projects is the Nuance VoCon recognizer. Because this is a proprietary, closed source software program, it is difficult to investigate how the recognizer gets its results. We know that VoCon uses language specific acoustic models, structured in Hidden Markov Models [15], and that it doesn't use the phrases in it's grammar/lexicon file directly, instead it converts the graphemes to phonemes (abbreviated g2p).

1.4 Objectives

This paper attempt to answer two questions using the data acquired by the Autonomata projects:

1. Which pronunciation errors are most likely to be made by subjects, both native and non-native Dutch speakers?
2. How do these errors influence automated speech recognition accuracy?

1.4.1 Speaker performance analysis

A well-performing ASR system can accurately determine what people tried to say. However, when people make mistakes when pronouncing words, the system has to be able to deal with deviations from the pronunciation the system expects. It can be expected that some of the mistakes are systematic, so if the ASR system can take into account what mistakes can be expected, it can better guess what the speaker *meant* to say.

Special consideration will be given to the differences between subjects with different mother languages.

For this question, it is hypothesized that there will be two main classes of errors:

- Realizing the phoneme usually associated with a given grapheme in the speakers' mother language, instead of the appropriate Dutch phoneme.
- Errors pertaining to the realization of phonemes uncommon to the speakers' mother language.

Dutch speakers will likely make errors when pronouncing phrases too, in the corpora there are many phrases containing non-Dutch words or names, in those cases Dutch speakers could make errors of the first kind.

1.4.2 Recognizer performance analysis

If we know the effect the pronunciation errors have on the performance of an ASR system, this provides valuable information on how to improve the system.

It is expected that more pronunciation errors will make it more difficult for the system to recognize the word correctly. Non-Dutch native speakers will likely realize phonemes in a slightly different way, and since ASR systems are trained on speech from native Dutch speakers, this might make things harder for the recognizer.

2 Methods

2.1 Data sets

Two similar data sets have been used, one generated by the Autonomata project (in this paper also referred to as Autonomata I), the other by the Autonomata TOO project (also referred to as Autonomata II). Both data sets contain short phrases, so there is no structural information between the words in one phrase, like sentence structure. The structure of the corpus itself differs in a subtle way, which provides some challenges to consolidate the data from the two projects, and interpret the results.

The alphabets used for phonemic transcription differ slightly between the projects. In Autonomata I the SAMPA set is used, like in the Corpus Gesproken Nederlands [8]. In Autonomata TOO a similar alphabet is used originally, but since there is a one-to-one mapping between equivalent sounds, for the purposes of this paper the transcriptions of Autonomata TOO have been rewritten to SAMPA form.

In the transcriptions (both realized and standard) the syllable or syllables with stress were marked with an apostrophe ('), but these have not been taken into account when calculating the distance between the transcriptions. The focus in this paper is on vowel substitutions, and the spectral differences caused by prosody are small, so ASR systems generally make little use of it compared to other characteristics.

2.1.1 Autonomata I

The Autonomata I project produced a corpus of names, pronounced by 240 participants. All participants pronounced a list of names and a list of words (mostly words used for driving a PND, or numbers). Because only the pronunciation of the words have been manually transcribed, the data from the words have not been used in this analysis. There were 20 different lists of names, 10 for each region (holland and belgium). Each participant pronounced one list of names. On each list there were between 169 and 179 names.

Project	Background	M	F	Total
I	Dutch	59	58	117
I	English	20	20	40
I	Moroccan	20	20	40
I	Turkish	10	10	20
I	French	10	8	18
II	Dutch	21	20	41
II	English	5	6	11
II	French	4	4	8
II	Turkish	2	4	6
II	Moroccan	0	3	3

Table 1: The subjects in the Autonomata I and II projects that have been used for this paper. Due to data corruption on the DVD that was used, data from 5 subjects on Autonomata I was not used. The data from Autonomata II was from a pre-final version, where not all pronunciations were transcribed yet.

Due to data corruption some data could not be used in this experiment, one participant pronounced 174 names, but only 136 were used, and data from 5 participants could not be used altogether. All data corruption occurred with flemish speakers.

2.1.2 Autonomata TOO

In this project lists of points of interest (POIs) were pronounced. The native Dutch speakers (40) pronounced one of four lists, composed of Dutch, French/English, or mixed POIs. Mixed language POIs are made up of Dutch words combined with French and/or English words. All subjects with foreign language backgrounds were given the same list, composed of Dutch and mixed language POIs. In this paper people with a Berber language are categorized under Moroccan. A pre-final version of the data was used, where not all subjects' pronunciations were transcribed yet, those subjects have not been used for this paper. One subject was originally categorized under the foreign speakers, but since he was born in the Netherlands, and his mother language was listed as Dutch, for this paper this subject was categorized with the Dutch native speakers.

2.2 VoCon accuracy tests

The VoCon recognizer was run on all samples. For each sample a maximum of 15 hypotheses were returned, with a confidence (score) for each one. The sample was recognized correctly if the hypothesis with the highest score was the phrase that the subject was trying to pronounce in this sample. The accuracy of the recognizer is the number of correctly recognized samples divided by the total number of samples. If the correct hypothesis is not the hypothesis with the highest score, but the score

of the correct hypothesis is close to the score of the first hypothesis, this means that although the sample wasn't correctly recognized, the recognizer was close. Therefore the ratio of the scores provides more information on the performance of the recognizer. The score ratio was calculated by dividing the score of the correct hypothesis by the score of the highest incorrect hypothesis. Sometimes the machine did not find any alternative hypotheses beside the correct hypothesis, in that case the ratio could not be calculated (dividing by 0 is not allowed), so it was set on 2.0. If the score ratio is below 1.0 this means the sample was not recognized correctly, if it is above 1.0 the sample was recognized correctly.

2.3 Vowel substitutions

A vowel substitution is counted as such if the preceding character (if any) and following character (if any) are the same between the standard and realized transcription, but the vowel has been substituted by another vowel. For convenience, only the most prevalent vowels have been considered. Diphtong vowels have also not been taken into account. Some examples of what counts and what does not count as a vowel substitution can be found in table 2.

Orthography	**Standard**	**Realized**	**Substitutions**
Mieke Moens	'mi-k@ 'muns	'mi-k@ 'mOns	u ⇒ O
Florian Verbeeck	'flo-ri-An v@r-'bek	'flO-ri-an 'v@r-bek	o ⇒ O, A ⇒ a
Tijn Danneel	'tE+n dA-'nel	'tAjn dA-'nel	

Table 2: Examples to clarify what counts as a vowel substitution. In the third example no vowel substitutions are counted as such, since the character immediately following the vowel is different in both strings.

From these vowel substitutions a matrix is created, which can be considered a similarity matrix, vowels which are similar are more likely to be substituted for each other. When replacing the numbers in the matrix by it's multiplicative inverse a dissimilarity matrix is formed (also known as a distance matrix). After making the matrix diagonally symmetrical, the Sammon multidimensional scaling algorithm[11] has been used to create a two dimensional representation of vowel distances.

2.4 Levenshtein distance

The Levenshtein distance is a metric for the distance between two strings of characters[7]. It represents the number of operations that are necessary to transform one string into another, where the available operations are to insert, delete or substitute a character. The distance between "live" and "loves" for example is 2, since if you substitute the "i" with an "o" in the first string, and insert an "s", the result is the second string.

The distance between the phonetic transcription of the realized pronunciation and the standard pronunciation have been calculated. In most cases the distance has been divided by the length of the standard transcription, to correct for the

length of words, obviously one substitution in a 40 character words should have less effect then one substitution in a 3 character word.

The levenshtein distance does not take into account the acoustic similarity of the characters, for instance the levenshtein distance between the pairs

⟨ 'lO-p@m , 'lO-pEm ⟩ (Loppem)

and

⟨ 'stA+-v@m-bErx, 'stY+-v@m-bErx ⟩ (Stuivemberg)

is 1 in both cases, but the acoustic difference is bigger in the second pair [9]. There are distance algorithms specifically designed to give an indication of the acoustic distance between two strings, but these have not been used because they are not meant for transcriptions, and don't handle input in multiple languages well [10].

To correct for word length, the levenshtein distance is divided by the length of the goal string, hereafter called the relative (levenshtein) distance. Presumably the recognizer will have less trouble with a single substitution, insertion or deletion if it is dealing with a long phrase than if it's concerning a short phrase. The number we get basically represents the proportion of phonemes that are pronounced correctly. Because the relative distance is calculated by dividing it by the length of the goal string, it is possible to get a number higher than 1, if there are a lot of substitutions, insertions or deletions, and the realized string is longer than the goal string.

3 Results

3.1 VoCon accuracy tests

In figure 2 we can see that the accuracy per mother language differs from about 0.72 for Turkish speakers to 0.87 for native Dutch speakers.

3.2 Vowel substitutions

The vowel substitutions [Figure 3] show one main pattern: a, e, i and o are often substituted for A, E, I and O respectively. These substitutions occur often in both directions, though slightly more often towards the short vowels than the other way around, except for the schwa (@), it seems that the subjects used less schwas than the standard transcriptions, mostly pronouncing e or E instead.

When we look at the heat maps from the languages individually [Figure 4], we see that the Dutch substitute the least number of vowels, as expected. Turkish and Moroccan subjects substitute vowels considerably more often. They don't use the schwa as much, more often than the other subjects they substitute it for E, I, O and e. English speakers substitute e for i and vice versa, or o for u, which could mean they have trouble applying Dutch g2p rules in stead of their native g2p rules. Most remarkable about the French is their substituting of I with i more often than the other way around, which they do less often than average.

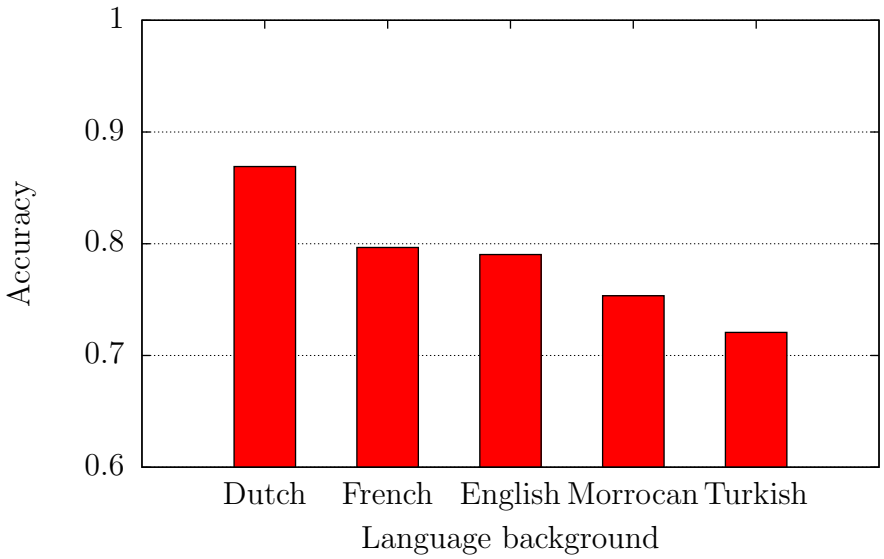


Figure 2: The accuracy of the VoCon recognizer per language background.

In figure 5 the vowel substitutions are shown after the Sammon's projection algorithm has determined the best 2D placement of the vowels. Sammon's algorithm gives a result that can be rotated or mirrored at will, since the dimensions don't represent anything. The scales however do matter if we want to compare the images to each other.

We can see clustering of the vowels into the clusters {u, o, O}, {A, a}, {@, E, e} and {i, I}. For all languages we see the cluster {u, o, O} far away from the other vowels, meaning there are few substitutions between that cluster and other vowels. The Moroccans show less clustering than the other languages, meaning the substitutions they made were less uniform.

We see relatively high, and consistent stress levels for all languages, except for the French speakers. This means we can compare the mother languages to each other.

3.3 Levenshtein distance

There were some large differences between participants regarding the average levenshtein distance. Figure 6 shows the distance for the different mother languages and per project. The distance for Autonomata I is higher than for Autonomata TOO, except for the Dutch. We can attribute this difference to the easier lists for non-native Dutch speakers, and harder lists for Dutch speakers, since the non-native speakers were only given Dutch and mixed (Dutch combined with French and/or English) phrases, and the native Dutch speakers were given only non-Dutch phrases to read. It could also be the case that the non-natives in the Autonomata TOO project were more proficient than the non-natives in Autonomata, for instance because they have been living in holland for a longer time, but this wasn't tested.

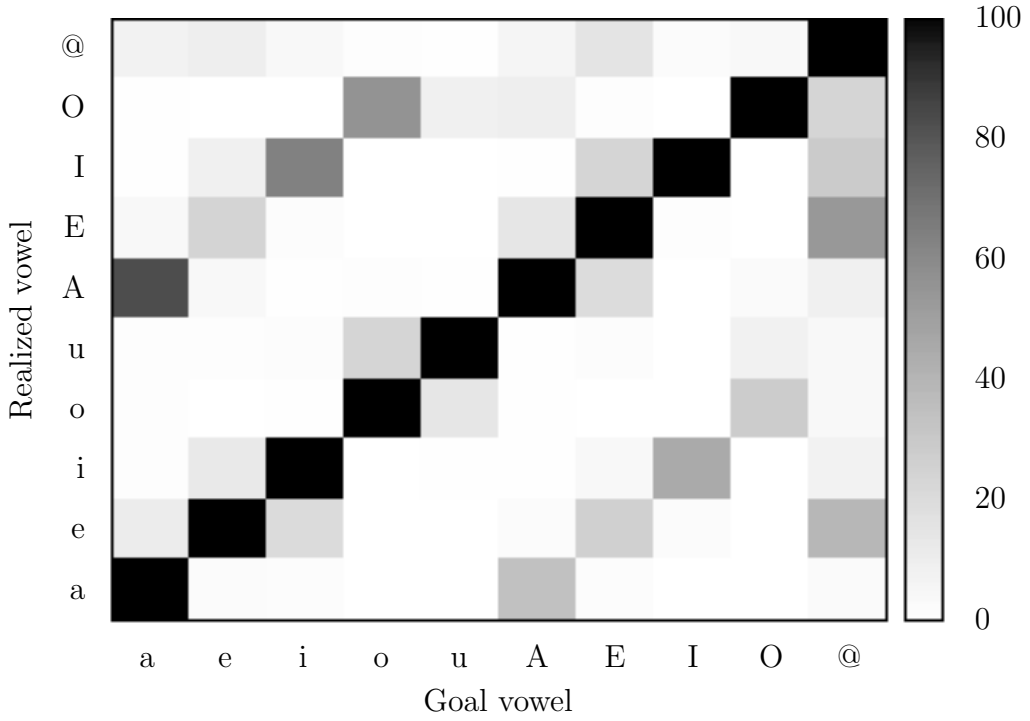
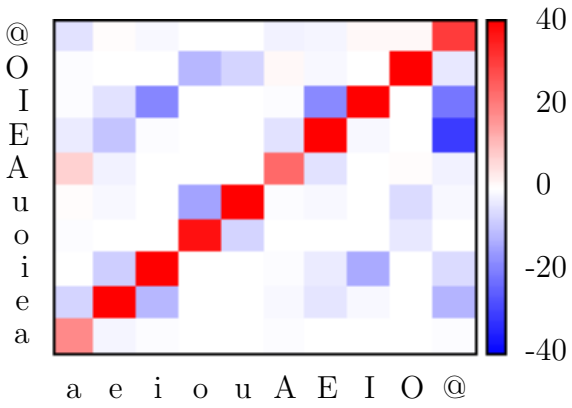
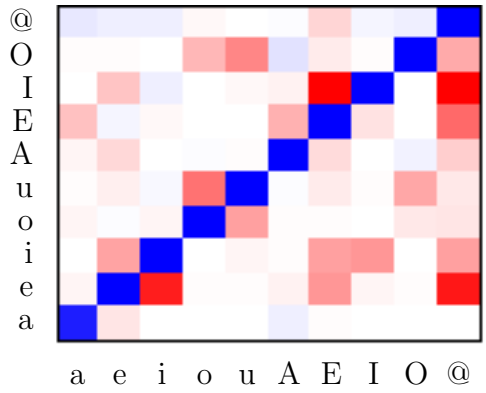


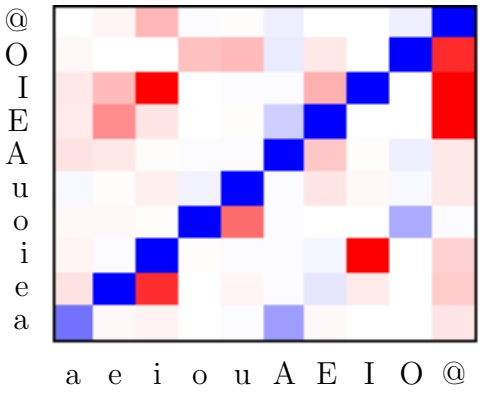
Figure 3: A heat map showing the vowel substitutions of all participants. On the vertical axis the goal vowel is shown, on the horizontal axis the produced vowel. The numbers in the right column represent per mil numbers, calculated as the number of vowel substitutions from goal to realized divided by the total number of goal vowels substituted. The darker the color of the cell, the higher the percentage of the goal vowel on the vertical axis that are pronounced as the vowel on the horizontal axis.



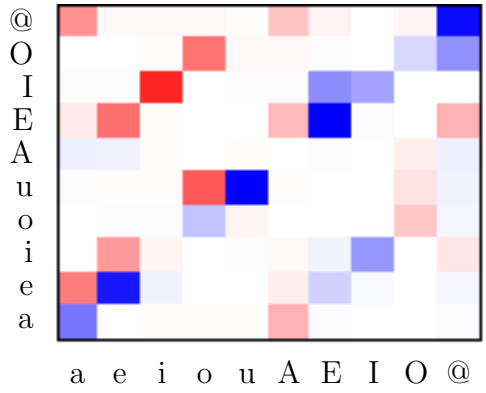
(a) Dutch



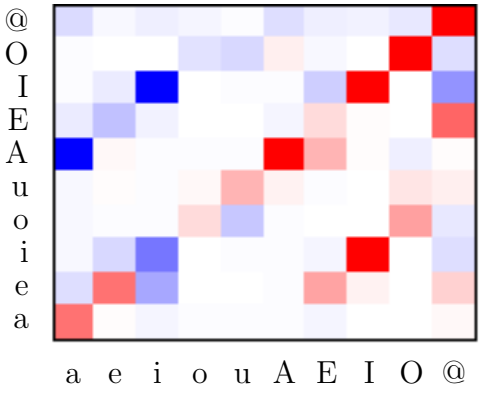
(b) Moroccan



(c) Turkish

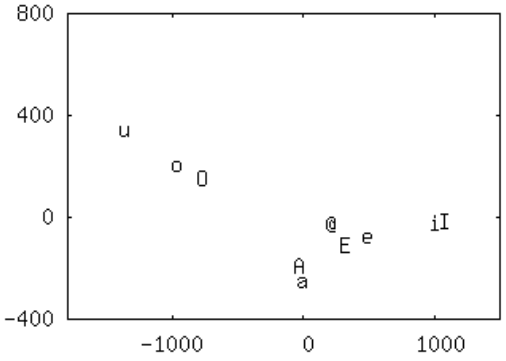


(d) English

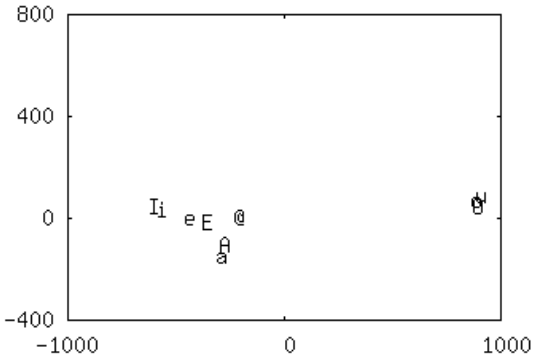


(e) French

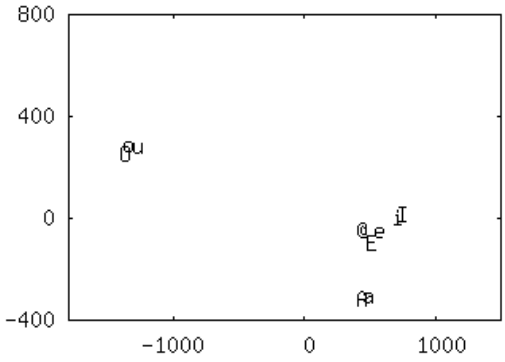
Figure 4: Vowel substitution heat maps for all mother languages individually, compared to the average. The horizontal axis shows the goal vowel, the vertical axis the realized vowel. Blue means there were less substitutions of that kind, red means more. The scale is the same for all heat maps, but only shown for the Dutch one. On the diagonal, more substitutions means the vowel was pronounced correctly more often. For more information about the format see the caption of figure 3.



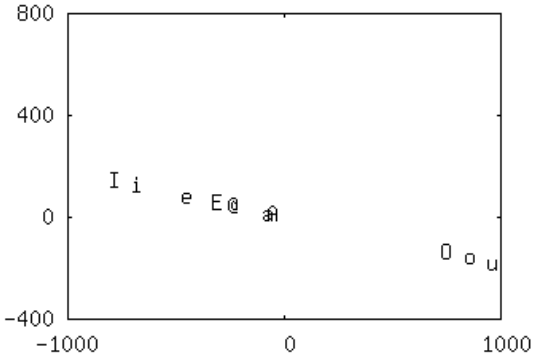
(a) Dutch
stress = 0.72



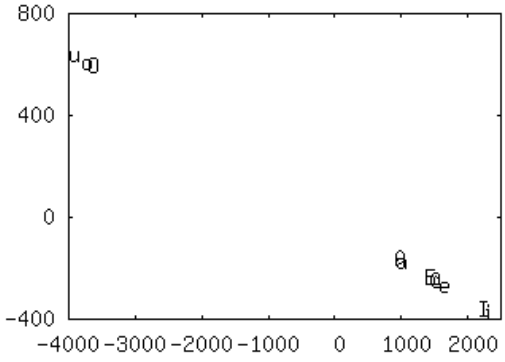
(b) Moroccan
stress = 0.61



(c) Turkish
stress = 0.65



(d) English
stress = 0.73



(e) French
stress = 0.45

Figure 5: Two dimensional representation of vowel distances for all mother languages, as determined by the Sammon multi-dimensional scaling algorithm. Two vowels that are often substituted for each other are placed close together. Note the differences in scale.

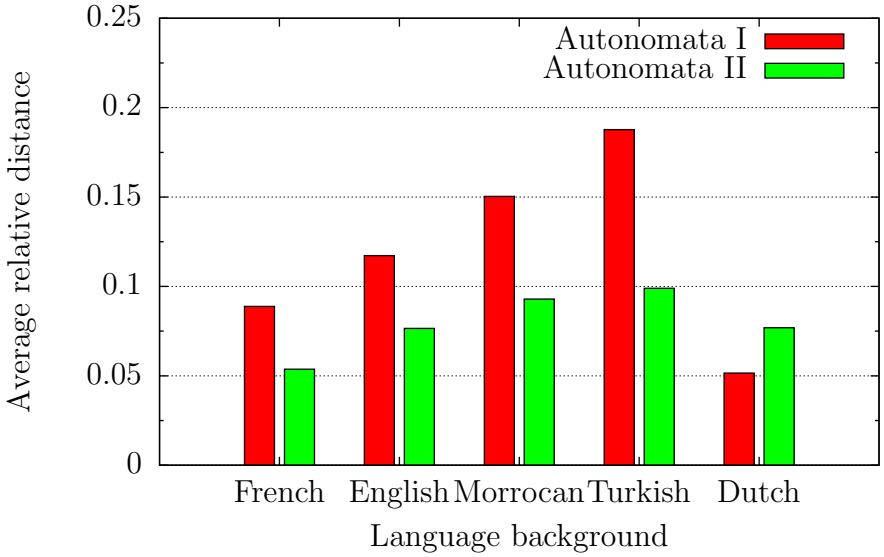


Figure 6: In this histogram the average levenshtein distance is shown per language background and per project.

The plot of all data points [Figure 7] does not show a recognizable pattern. This is because there are many points on the score ratio = 0 and score ratio = 2 lines, causing the points to be drawn on top of each other, making it impossible to determine how many there are. Figure 8 shows average values, and the plot is approximating the function of each line. To test whether the values for relative distance = 0 are equal, the Wilcoxon's Signed-Rank test was performed comparing Dutch and English (p-value=2.2e-16), showing that the values are significantly different.

4 Discussion

The differences in overall accuracy between the groups are quite high [figure 2]. This could be because some groups make less mistakes (what they say), or that they articulate more clear (how they say it), or a combination of the two. From figure 6 we see that distances differ between the groups, but from figure 8 we can also infer that for the same distance the score ratio (and thus accuracy) also differs.

Some of the vowel substitutions can be explained by an inappropriate g2p rule set. Because of the way the corpus was constructed, it is possible the standard transcription was not realized because subjects tried harder than usual to articulate clearly, producing somewhat exaggerated results. This could explain the fact that the sjwa (@) was more often pronounced as an E than in the standard transcription. Though the same subjects might pronounce slightly differently in everyday speech, it is likely they would pronounce the same way when using a voice interface.

The high stress levels for the multidimensional scaling plots means the data does

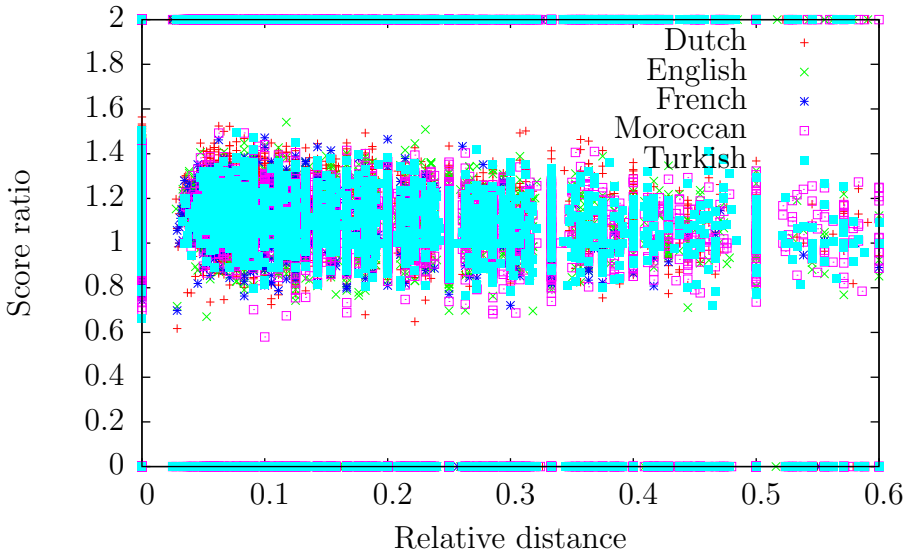


Figure 7: A plot of all data points, representing the relative distance vs the score ratio. Since many points are drawn on top of each other, it is difficult to see a pattern. A score ratio of 0 means the correct phrase was not in the top 15 hypotheses, a score ratio of 2 means there were no alternative hypotheses other than the correct one.

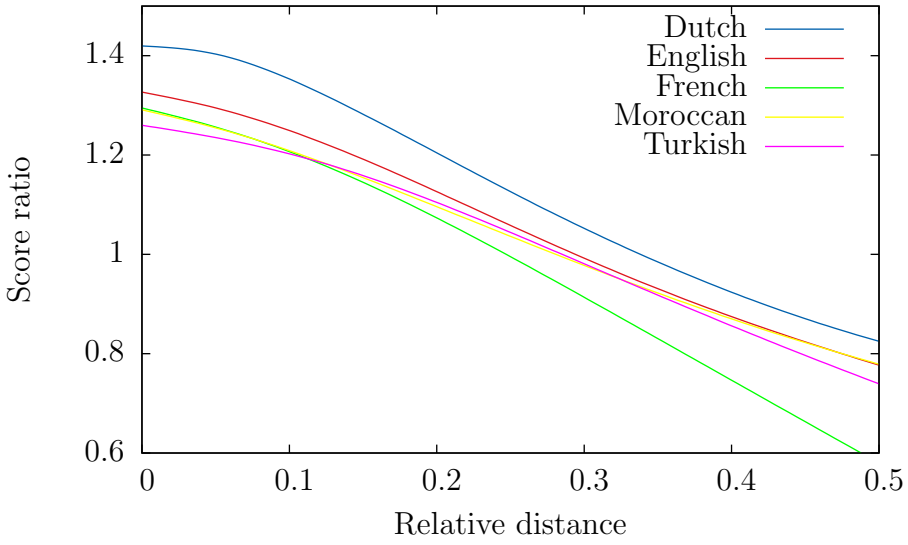


Figure 8: A smoothed plot of relative distance vs. score ratio. Recognizer performance varies significantly for people from different language backgrounds, even when corrected for distance.

not show a euclidian pattern. This could mean the pattern substitutions were either very erratic or uniform. A uniform pattern would mean vowels from one cluster don't get substituted with vowels in another cluster very often. By looking at figure 4 we can decide that the stress is mostly caused by the uniformity of the substitutions, and this explains why the stress for Dutch speakers is one of the highest.

A surprising result is that for all languages the variances of the vowel positions on the multidimensional scaling plots are mostly confined to one dimension. The vertical axes scales are about half that of the horizontal ones, and for the French it is only $\frac{1}{6}$th. One would expect the vowels to be placed more along two dimensions, one representing the first formant, and the other the second formant [9]. The first formant does not seem to matter much for the vowel's positions at all, the second formant is highly dominant here. Only the native Dutch and the Turkish speakers show some variances in the first formant. For the others, the positions only reflect the placement of the tongue during articulation. Further research would be necessary to explain this.

A remarkable result is that the (relative) distance is not the only thing that matters: even when the distance is the same, the score ratio is highest for the Dutch natives [Figure 8]. This also goes for the cases where there were no mistakes, which means the difference cannot be explained by assuming the Dutch make different *types* of mistakes, leading to the same distance, but posing less problems to the recognizer. This can only mean that the Dutch' articulation conforms better to the acoustic models of the recognizer compared to the non-native Dutch speakers.

5 Conclusions

The results show that people with different mother languages make more errors when pronouncing Dutch phrases. Since we see clearly distinct patterns in vowel substitutions, people with different language backgrounds make different types of mistakes.

These errors have significant influence on the ASR's accuracy. However, it seems at least as important to conform as best as possible to the acoustic model of the machine, since accuracy differs significantly even when the realized transcription is equal to the standard transcription.

6 Acknowledgements

I would like to thank the people who created the Autonomata projects, Marijn Schraagen in particular for providing all necessary data and tools. Special thanks to Gerrit Bloothooft for constant patience and encouragement, and Jenny Lazebnik for help with the statistic analyses and proofreading.

References

[1] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[2] H. Dudley. The vocoder. *Bell Labs Record*, 17:122–126, 1939.

[3] J.G. Fiscus, J. Ajot, and J.S. Garofolo. The rich transcription 2007 meeting recognition evaluation. In *MTPH07*, pages xx–yy, 2007.

[4] D. B. Fry and P. Denes. The design and operation of the mechanical speech recognizer at university college london. *J. British Inst. Radio Engr.*, 19(4):211–229, 1959.

[5] Lee Honeycutt. Researching the use of voice recognition writing software. *Computers and Composition*, 20(1):77 – 95, 2003.

[6] B. H. Juang and Lawrence R. Rabiner. Automatic speech recognition – a brief history of the technology development abstract. In *Encyclopedia of Language and Linguistics, 2nd Edition*. Elsevier, 2005.

[7] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10, 1966.

[8] Nelleke Oostdijk. Building a corpus of spoken dutch, 1999.

[9] L. C. W. Pols, H. R. C. Tromp, and R. Plomp. Frequency analysis of Dutch vowels from 50 male speakers. *The journal of the Acoustical Society of America*, 53:1093, 1973.

[10] Hema Raghavan. Using soundex codes for indexing names in asr documents. In *Library and Information Science Research*, pages 22–27, 2004.

[11] Jr. Sammon, J.W. A nonlinear mapping for data structure analysis. *Computers, IEEE Transactions on*, C-18(5):401 – 409, may. 1969.

[12] Marijn Schraagen and Gerrit Bloothooft. Evaluating repetitions, or how to improve your multilingual asr system by doing nothing. In *Proceedings of the seventh international language resources and evaluation (LREC 2010)*, 2010.

[13] Stuart C. Shapiro. Artificial intelligence, 1982.

[14] Henk van den Heuvel, Jean-Pierre Martens, Bart D'hoore, Kristof D'hanens, and Nanneke Konings. The autonomata spoken names corpus. In *Proceedings of the sixth international language resources and evaluation (LREC 2008)*, 2008.

[15] Steven Wegmann and Larry Gillick. Why has (reasonably accurate) automatic speech recognition been so hard to achieve? *CoRR*, abs/1003.0206, 2010.