Automatic Accent Classification Using Automatically Detected Pronunciation Variants

Aki Kunikoshi[1], *Herbert Teun Kruitbosch*[1], *David van Leeuwen*[2,3], *and Martijn Wieling*[1,4]

[1]University of Groningen, The Netherlands
[2]ICIS, Radboud University Nijmegen, The Netherlands
[3]NovoLanguage, Nijmegen, The Netherlands
[4]Haskins Laboratories, USA

{a.kunikoshi, h.t.kruitbosch, m.b.wieling}@rug.nl, d.vanleeuwen@science.ru.nl

Abstract

In this study we aim to identify the region of origin of speakers of Dutch speakers on the basis of their accent. We use a large crowd-sourced data set which contains recorded acoustic Dutch pronunciations from a large number of speakers, but also perceptual human judgements on where each speaker is from. Our approach consists of combining multiple sentence-level accent classifiers which were trained on the basis of alternative pronunciation variants (obtained using a forced alignment system incorporating pronunciation variation) for all words in the sentence. Our results indicated that our system is able to classify two Dutch regional accents (Groningen + Drenthe vs. Limburg) with an accuracy of about 78.6%. When distinguishing three Dutch regional accents (Groningen + Drenthe, Twente + Achterhoek, and Limburg), the overall accuracy was about 54.8%. While relatively low, the accuracy of our system was higher than human classification performance.

Index Terms: Dutch, accent recognition, shibboleths, pronunciation variation, forced-alignment

1. Introduction

Accents are fascinating, not only because they highlight how variable speech can be, but also because they provide us with a window into a speaker's background. Accent recognition has been therefore attracting researchers in the variety of fields, such as speech recognition [1, 2], or language recognition [3].

Accent recognition could be seen as a more subtle variant of language recognition, see, e.g., the discussion in the context of automatic accent location [4]. It is therefore also a tough problem, with varying perfomance for, e.g., seven American accents from TIMIT [5] and fourteen UK accents from the *Accents of the British Isles* corpus [6].

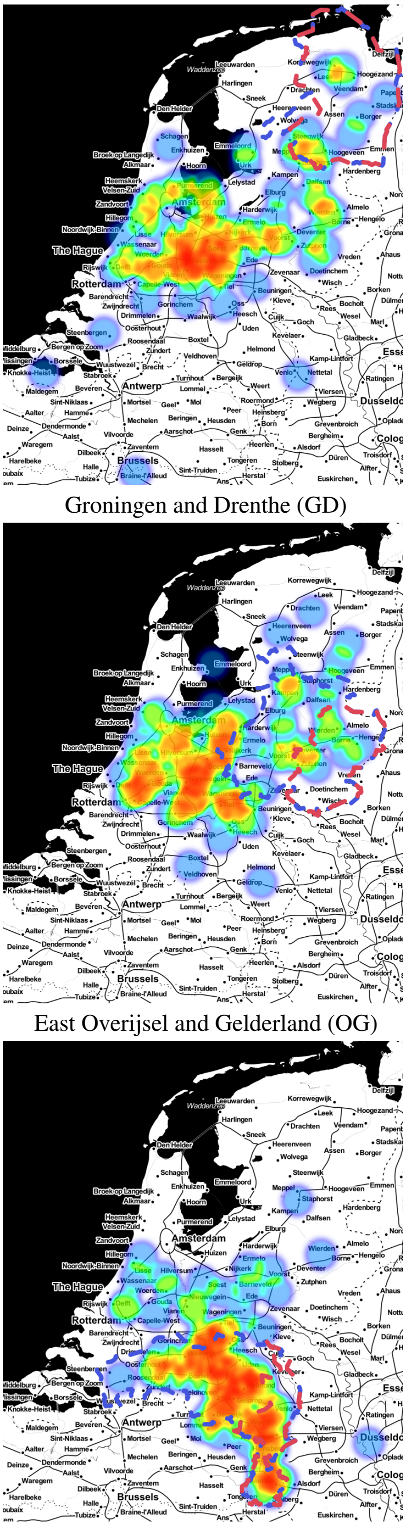
In this study, we will focus on Dutch accents. Various studies have investigated accented speech in the Dutch language. For example, Adank and colleagues [7] carried out an acoustic analysis of regional variability in the pronunciation of vowels in the Netherlands and found that accents could be classified on the basis of formant measurements with an accuracy of about 72 percent when investigating the speech of speakers. Van Leeuwen et al. [8] attempted to automatically distinguish Dutch spoken in The Netherlands versus that spoken in Belgium on the basis of short speech samples (ranging from 3 seconds to 30 seconds). Their error rate ranged from 3 percent (for 30-second-samples) to 16 percent (for 3-second-samples).

While the aforementioned studies show that Dutch accents can be distinguished rather well, these studies did not compare their system's performance to that of human listeners. Consequently, we also attempt to develop a Dutch accent classifier (using a different approach), but importantly we will compare its performance to that of human listeners. Our automatic approach proceeds by identifying for each word the specific (regional) variant the speaker uses (using an ASR system to determine the pronunciation variant), and subsequently feed this information into an a Dutch accent classification system.

2. Data set

In this study, we will use acoustic and perceptual data from the "Sprekend Nederland" (SN) project [9]. This project was initiated by the Dutch public broadcasting company NTR at the end of 2015. The intention was to record the regional varieties of spoken Dutch speech and measure the attitude of listeners towards these variants. By using a smartphone app, over 4000 participants have supplied their speech (over 200 000 recordings, 500 hours of speech). Besides providing their own speech (both spontaneous and read speech) and meta-data about their own background, participants judged the speech of other people, thereby assessing the region of origin of the speaker (indicated on a map with a variable zoom level), but also personal characteristics of the speaker. Due to the characteristics of the SN app, people recorded their utterances (and provided their meta data) in a fixed sequence. Consequently, most recordings are associated with items which were presented in the beginning of the sequence [9]. For this study, we have used the 10 most-frequently recorded sentences, or sequences of words (see Table 1).

Rather than focusing on all regions in the Netherlands, we will focus on three specific regions: the north of the Netherlands (the province of Groningen and Drenthe, henceforth "GD"), the east of the Netherlands (the eastern part of the provinces of Overijssel and Gelderland, henceforth "OG") and the southern part of the Netherlands (the province of Limburg, henceforth "LB"). All three areas are indicated in Figure 1 by red outlines. The reason for focusing on only a few regions is that the majority of the speakers who recorded their voice are relatively young [9], and highly educated and therefore tend to have relatively weak regional accents. Informal listening indeed suggests that participants did not show a large deviation in accent from what might be considered the Dutch norm. A hypothesis for the reason is that speakers were aware that they would be judged by others on their accent, but this is not further tested in this study. Figure 1 visualizes where Dutch listeners thought the speakers were from, together with where they were actually from (marked with a red outline; the meaning of the blue dashed line is explained further below).



Groningen and Drenthe (GD)

East Overijsel and Gelderland (OG)

Limburg (LB)

Figure 1: *Listener's judgments (heat map) of where the speakers are actually from (red outline). Listener's judgments outside of the area demarcated with a blue dashed line were assumed to be random guesses and equally distributed over all areas. See text for details.*

While listeners seem to be able to identify the region of origin from the LB speakers, this is not the case for the other two regions. Note that there were several questions in the SN app to assess region, such as "where are you from?", or "where have you lived for the longest period?", etc. We assigned a speaker to one of the three regions if the location (i.e., longitude and latitude) associated with the two aforementioned question was less than 30 kilometers apart and the middle point between the two locations was positioned in one of the three regions.

There were a few hundred speakers who satisfied the above criteria, but in order to create a balanced data set with the same number of speakers per region, we had to restrict the number of speakers to 207 per region. Out of this set 42 speakers per region were randomly assigned to the test set, whereas the remaining 165 speakers were assigned to the training set. Table 1 shows for each region how many of the 165 speakers per region provided a recording for each sentence. It is clear that this number diminishes quickly, which is due to the fixed-sequence-setup of the app through which the data was collected, and the participant's option to quit at any moment.

3. Accent classification system

Our text-dependent accent classification system consists of two parts. In the first part we identify the specific pronunciation variant for each word in a sentence. In the second part, we use the identified word pronunciations as features in a sentence-level classifier of the regional accent. Since we have a total of 10 sentences (or word sequences) available, we train 10 sentence-level classifiers. The final judgment is based on a majority vote of the 10 classifiers.

3.1. Identifying pronunciation variants

Our approach to obtaining the phonetic transcriptions is to use a forced alignment procedure, with many pronunciation variants per word which were generated using phonological and reduction rules [10]. In this way, the most likely pronunciation will be associated with the acoustic recording by the forced aligner. For words not occurring in the pronunciation dictionary, we use a Dutch grapheme-to-phoneme algorithm [11] augmented with the pronunciation variants obtained using the aforementioned approach [10].

We used HTK [12] to train a Dutch acoustic model on the basis of the Corpus Gesproken Nederlands (CGN) [13]. Specifically, we used the components of CGN containing read speech, as we are also classifying read speech here. Sentences which included non-linguistic sounds or foreign words were excluded. We used a phoneset of 38 phones (cf. Table 2), training an HMM acoustic model as tristate monophones with GMM output probabilities with 16 mixtures per state. We used 13 MFCCs and their first and second order derivatives as features using a frame shift of 5 ms and a window length of 25 ms.

The task of the forced aligner is to determine the most likely pronunciation given the alternatives in the pronunciation dictionary.

3.2. Classification system

For each of the 10 sentences a separate accent classifier was trained. The features for each sentence consisted of a one-hot encoding vector for each word in the sentence. The length of each one-hot encoding vector was equal to the number of distinct pronunciation variants per word detected by the forced alignment approach in our data set. For example, suppose our

Table 1: *Sentences and word sequences used for classification. The numbers indicate the number of recordings in the training set for each region.*

Sentence	GD	OG	LB
De sprei die jij bracht heeft Suzan altijd al gewild, omdat die zo leuk staat bij het gele laken.	165	164	165
Na de tocht in de regen vaart de kapitein blij weg.	164	165	165
Het echtpaar adopteert voor veel geld een tweejarig weesje.	153	158	153
Jeannette lacht nadat ze door een onbekende persoon is gekust.	116	109	106
takt, ik reken, verwart, erudiet, een bij.	74	71	82
Zacht knort het varken in de wei een karakteristiek geluid.	72	73	82
Na zijn reis maakt hij een kaart van de onverharde wegen en die zet hij direct op een DVD.	71	71	80
Kort na het wielerfestijn verbruikte Dries de rest van de deodorant.	49	54	67
pistool, geraapt, toveren, ijs, verkeerd.	50	56	61
Natuurlijk ga je van hard werken heus niet dood, maar in andere opzichten is de prijs soms hoger dan je dacht.	47	53	61

Table 2: *The 38 Dutch phonemes recognized*

Phoneme	IPA	Phoneme	IPA
@	ə	m	m
a	a: / ɑ̃:	n	n
ac	ɑ	nc	ŋ
au	ʌu	o	o:
b	b	oc	ɔ / ɔ: / ɔ̃:
d	d	p	p
e	e:	r	r
ec	ɛ / ɛ: / ɛ̃:	s	s
ei	ɛi	sc	ʃ
eu	ø:	(ssil)	(silence)
f	f	t	t
g	g	u	u
gc	ɣ	ui	œy
h	ɦ	v	v
i	i:	w	ʋ
ic	ɪ	x	X
j	j	y	Y
k	k	yc	ʏ / œ: / œ̃
l	l	z	Z

data set contains the following four different pronunciation variants for the word *heeft* “have”: /ef/, /eft/, /hef/ and /heft/, the pronunciation ‘eft’ is coded as $[0, 1, 0, 0]$. The final feature vector for the sentence is a concatenation of the one-hot vectors of the individual words.

Using the one-hot encoded features, an Adaptive Boosted (AdaBoost) Decision Tree classifier was trained per sentence separately. The final classification of the speaker’s accent is subsequently based on combining the sentence-level classifiers by majority vote. Note that the number of classifiers used depends on the amount of test sentences available for each speaker. In case there is a tie, the vote of each sentence level classifier is weighted by the precision with respect to the assigned class (obtained on the basis of 10-fold-cross-validation using the training set).

In the following section, the performance of our system is evaluated on the held-out test set.

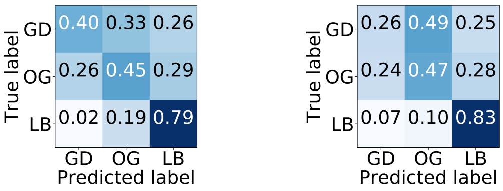
4. Human accent classification

To evaluate our automatic system, we are interested in assessing how well humans are able to determine where someone is from on the basis of their accented speech.

In the app, participants were not asked to choose from the three regions (i.e. the areas demarcated with a red line in Figure 1, henceforth the red areas), but rather to indicate freely on a map where they thought the speaker was from. To make this data more comparable to the three-class classifier, we converted the rater-provided map coordinates to the appropriate region. Accent judgments that were located inside the the areas in Figure 1 which are demarcated by a dashed blue line (i.e. blue areas, regions with a similar accent) were registered as a recognition of the accent spoken in the (neighboring) red area. Accent judgments located outside of the blue (and thus also red) areas were considered to be wrong (as different accents are spoken in those regions) or just a random guess, and therefore distributed equally over all three regions (i.e. in line with a forced-choice experiment). The size of these blue areas clearly differs, but this is linguistically motivated, as explained below.

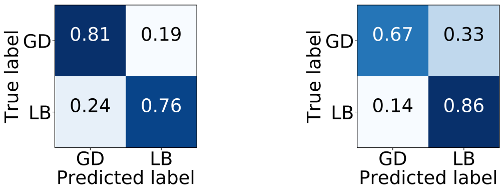
The areas GD and OG are located in the Low Saxon dialect area, consisting of the provinces of Groningen, Drenthe, Overijssel, a major part of Gelderland (i.e. excluding the “Betuwe”), and a small part of the northeastern province of Friesland (i.e. the “Stellingwerven”; in the remaining part of Friesland, the Frisian language is spoken). Any judgments inside the Dutch Low Saxon dialect area, are assumed to be choices for either the GD area or the OG area. The OG area (middle map of Figure 1) consists of the eastern part of the provinces of Overijssel and Gelderland, and accent judgments positioned in the rest of Overijssel or the rest of the Low-Saxon part of Gelderland were assumed to be votes for the OG area. The remaining part of the Dutch Low Saxon area consists of the province of Groningen, Drenthe, and the “Stellingwerven” and judgments inside this area were assumed to be votes for the GD area (top map of Figure 1). For the bottom map, accent judgments positioned in the province of Noord-Brabant in the south were also classified as a choice for LB. The reason for this is that accents in these two regions are relatively similar, and very different from the OG and GD accents.

We only could include speakers from the original set (i.e. the 207 speakers per region) who had associated accent location judgments (inside the areas demarcated by the dashed blue line). This resulted in a (balanced) set of 113 speakers per region. For these 339 speakers, a total of 905 estimated locations (i.e. answers to the question: “where do you think this person is from?”) were provided by 817 raters in the Sprekend Nederland data set. Some speakers were rated by multiple participants and in that case we used majority vote (first per sentence, if there were multiple raters for a single sentence, then



(a) (b)

Figure 2: *Accent classification performance in three regions by (a) our system and (b) humans.*



(a) (b)

Figure 3: *Accent classification performance in two regions by (a) our system and (b) humans.*

across sentences if multiple sentences for a single speaker were rated). If there was a tie assigning a region to one of the 339 speakers, $\frac{1}{2}$ or $\frac{1}{3}$ was added to the corresponding fields of the confusion matrix. As mentioned before, judgments outside of the blue areas were equally distributed over the three regions (operationalized by adding a frequency of $\frac{1}{3}$ to all three regions per judgment).

5. Results

The result of applying our system to the held-out training data is visualized in the normalized confusion matrix in Figure 2 (left). For comparison, the confusion matrix for the human performance is shown in the same figure (right). The overall accuracy of our system was 54.8%, whereas it was 47.5% for the human judgments.

Given that the region GD and OG both are located in the Dutch Low-Saxon dialect area and are more similar to each other than the LB accents, we also trained a classifier distinguishing only GD from LB. The performance of our system and the corresponding human performance is shown in Figure 3. The overall accuracy of our system was 78.6%, whereas it was 69.9% for the human judgments. In this case all judgments inside the Dutch Low Saxon dialect area, i.e. the areas demarcated by a blue dashed line in the top and middle map of Figure 1, were counted as a choice for the GD area.

6. Discussion

Our experiments revealed that while the performance of our accent recognition system is by no means high (compared to e.g., [7] and [8]), it is better than human recognition performance. Given the nature of our data set (relatively young, generally highly educated speakers, who do not exhibit strong regional accents), it is not unexpected that regional provenance is less clearly detected than in the studies of Adank [7] and Van Leeuwen [8].

The confusion matrices show that both our system and also humans are best able to identify the Limburg accent. This may be partly due to the larger number of acoustic recordings for the speakers from the LB area (see Table 1). However, it is likely that this is also due to a highly salient feature of the Limburg accent (and other southern Dutch accents): the soft "g". This sound occurs in many of the sentences included in our study (see Table 1). Our automatic system differs in two aspects from human classification performance. Humans are better able to identify the Limburg accents, whereas our automatic system is better able to distinguish the GD from the OG accents.

We should keep in mind, however, that humans were not asked to choose from a set of three regions to assign a speaker to, but rather asked to choose freely on a map. We have tried to convert these locations to one out of three (or two) regions by considering choices close to one of the three (or two) regions as a choice for the neighboring region with a similar accent. If an accent judgment was located in an area with a different accent, this was considered to be an equal vote for all three locations (i.e. similar to a random guess). Nevertheless, the actual human performance might have been somewhat different in a real forced-choice experiment.

In future work, we would like to use the actual phonemes supplied by the transcription of the forced alignment system as features and experiment with extracting additional acoustic features as well.

7. Conclusion

In this study we attempted to identified the region of origin of speakers of Dutch on the basis of their accent. We first implemented a forced alignment system to automatically identify the specific regional variant used by the speaker. We subsequently used these variants as features in an accent classification system, consisting of multiple sentence-level classifiers. Our automatic system was able to classify three Dutch regional accents with an accuracy of about 55%, whereas this increased to about 79% when distinguishing only two regions.

8. Acknowledgements

We would like to thank Barbara Schuppler for providing the code for obtaining the alternative phonological and reduced pronunciations. We also thank Gosse Bouma for providing us with his Dutch grapheme-to-phoneme application. Finally, we thank John Nerbonne and Marc van Oostendorp for their valuable feedback. This work is supported by a Digital Humanities Exploratory Project grant from the University of Groningen.

9. References

[1] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*. Columbia University, 2011.

[2] A. Huggins and Y. Patel, "The use of shibboleth words for automatically classifying speakers by dialect," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4. IEEE, 1996, pp. 2017–2020.

[3] L. M. Arslan and J. H. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[4] D. A. van Leeuwen and R. Orr, "The "Sprekend Nederland" project and its application to accent location," in *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*. Bilbao: ISCA, 2016, pp. 101–108.

[5] J. H. L. Hanson, U. H. Yapanel, R. Huang, and A. Ikeno, "Dialect analysis and modeling for automatic classification." in *INTERSPEECH*, 2004.

[6] A. Hanani, M. J. Russell, and M. J. Carey, "Human and computer recognition of regional accents and ethnic groups from british english speech," *Computer Speech & Language*, vol. 27, no. 1, pp. 59–74, 2013.

[7] P. Adank, R. Van Hout, and H. v. d. Velde, "An acoustic description of the vowels of northern and southern standard dutch ii: Regional varieties," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 1130–1141, 2007.

[8] D. A. van Leeuwen, D. Henselmans, and T. Niesler, "Spoken language variety recognition in dutch," 2014.

[9] D. A. van Leeuwen, F. Hinskens, B. Martinovic, A. van Hessen, S. Grondelaers, and R. Orr, "Sprekend Nederland: a heterogeneous speech data collection," *Computational Linguistics in the Netherlands Journal*, vol. 6, pp. 21–38, December 2016.

[10] B. Schuppler, M. Ernestus, O. Scharenborg, and L. Boves, "Acoustic reduction in conversational dutch: A quantitative analysis based on automatically generated segmental transcriptions," *Journal of Phonetics*, vol. 39, no. 1, pp. 96–109, 2011.

[11] G. Bouma, "A finite state and data-oriented method for grapheme to phoneme conversion," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 2000, pp. 303–310.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[13] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation." in *LREC*, 2000.