Phoneme Errors in Read and Spontaneous Non-Native Speech: Relevance for CAPT System Development

Joost van Doremalen, Catia Cucchiarini, Helmer Strik

Department of Linguistics, Radboud University Nijmegen, The Netherlands
{j.vandoremalen, c.cucchiarini, h.strik}let.ru.nl

Abstract

For the purpose of pronunciation assessment and training in a second language both read and spontaneous speech are employed. In this paper we present the results of a study on the nature of phoneme errors in Dutch read and spontaneous non-native speech and discuss the possible consequences and relevance of these findings for the purpose of developing Computer Assisted Pronunciation Training systems.

1. Introduction

Pronunciation is considered a difficult skill to learn in a second language (L2), with the majority of L2 learners never acquiring native-like performance and many of them having problems even in attaining a level of comfortably intelligible speech. Research has shown that serious pronunciation problems can hamper communication [1][2] or even put the learner at a social and professional disadvantage (see for reviews [3]). The emphasis on communicative effectiveness in language teaching has brought about a renewed interest in pronunciation training. In addition, various studies have shown that tailor-made training is effective in improving perception and production of L2 speech sounds [4].

However, the kind of specific and intensive training that is required for improving pronunciation cannot be generally applied in L2 classrooms because it is too time-consuming. In the classroom, as well as in natural interactions, feedback on pronunciation errors is provided incidentally and is not always interpreted correctly [5].

Computer Assisted Pronunciation Training systems that make use of automatic speech recognition would seem to constitute an interesting alternative, as such systems can offer virtually unlimited input, can provide individualized, instantaneous feedback and the possibility of practicing as much as possible.

In spite of these undoubtedly attractive features, there are also important technological limitations that should be reckoned with, since automatic speech recognition (ASR) of non-native speech is still problematic [6].

To partly alleviate the ASR problems caused by non-native speech, various strategies have been proposed to constrain the output of the learner so that the speech becomes more predictable and thus more tractable form the ASR point of view. A common strategy consists in eliciting constrained output from learners by letting them read aloud texts displayed on the screen. Even in this case, however, it is possible that learners read something different from what is presented to them, but the chances are high that they will do what they are asked to do.

Although this might be a safe option from the ASR point of view, one might argue whether this is pedagogically sound from the perspective of pronunciation learning. The type of speech that is elicited in this way is indeed read speech, which of course is different from the spontaneous speech that learners will have to use in communicating in the L2. On the other hand, learning to read aloud is one of the skills that have to be learned in the L2.

For the purpose of CAPT development, such choices are very important as they determine which sounds are identified as being problematic, which ones are selected for pronunciation training and how algorithms for error detection are designed and trained, as will be explained below. Within the DISCO project [11] [12], which is aimed at realizing an ASR-based training system for oral proficiency in Dutch L2, we decided to study the nature and frequency of phoneme errors in read and spontaneous non-native speech with a view to developing a sound pronunciation training component.

In this paper we first discuss the pros and cons of using either read or spontaneous speech for the purposes of pronunciation assessment and training (section 2). We then go on to present a study on the nature of phoneme errors in Dutch read and spontaneous non-native speech (section 3). We will end with a discussion of the results and some concluding remarks.

2. Pronunciation assessment and training: read speech vs. spontaneous speech

In the previous section we referred to the necessity of using read speech in ASR-based CAPT systems to make it easier for ASR to handle non-native speech. However, there are other reasons for using read speech when it comes to pronunciation assessment and training.

To start with, one could argue that learning grapheme-phoneme correspondences and being able to read aloud is a skill that should be learned in the L2 just as it is learned in the L1. In addition, for pronunciation assessment read speech offers a number of advantages. First, by eliciting read speech it is possible to control what the speakers will say and to have them produce the same words and sounds. This homogeneity in content ensures that pronunciation scores are comparable. When human judges are involved this has the additional advantage that raters are not influenced by oral production factors lying outside the domain of pronunciation such as grammar or lexicon.

Having the possibility of controlling the content of the utterances also has the advantage that phonetically balanced material can be used. In turn this is attractive because the pronunciation of all phonemes of a language can be evaluated.

So, although there are several good reasons for employing read speech when evaluating L2 pronunciation, this also has some drawbacks. For instance, the ability to read aloud in an L2 partly depends on the familiarity with L2 orthography. Interference from L2 orthography might cause specific phoneme errors thus providing a biased picture of pronunciation difficulties [7]. If some of the

errors observed in L2 read speech are simply decoding errors caused by insufficient knowledge of L2 orthography or interference from it, it is legitimate to ask whether such errors are pronunciation errors at all. In addition, research has shown that the nature and frequency of phoneme errors in non-native speech production are related to the specific relation between L1 and L2 orthography [8].

By employing spontaneously produced speech such forms of orthography interference could be avoided. The phoneme errors thus observed are more likely to give an indication of real pronunciation problems. However, since the content of the utterance cannot be controlled, it is questionable whether the speech elicited provides a complete representation of potential pronunciation problems.

Given that these choices play an important part in CAPT development, we decided to study the nature and frequency of phoneme errors in read and spontaneous non-native speech to be able to make optimal choices for pronunciation training within the framework of our DISCO system for L2 oral proficiency training.

3. Phoneme errors in read and spontaneous non-native speech: the case of Dutch

3.1. Non-native speech material

The speech material for the present experiments was taken from the JASMIN speech corpus [9], which contains speech from speakers with different mother tongues and relatively low proficiency levels, namely A1, A2 and B1 of the Common European Framework (CEF). The speech was collected in two different modalities: read speech and human-machine dialogues. The read speech material we used for this study consists of utterances produced by 45 speakers while reading aloud short texts from the screen and sets of phonetically rich sentences. The spontaneous speech material was derived from the human-machine dialogues which were collected through a Wizard-of-Oz-based platform [9].

Orthographic transcriptions were manually created and include fluency phenomena such as filled pauses, restarts and repetitions. From these orthographic transcriptions, phonetic transcriptions were automatically generated using a pronunciation lexicon with native and non-native pronunciation variants. Phonetic transcriptions for words which contain disfluencies were manually created. Because the automatically generated phonetic transcription can contain errors, we had two transcribers manually correct the phonetic transcriptions on the word level. They were instructed to change the phonetic transcription whenever they thought that an error had been made. For this correction, only the SAMPA symbols for Dutch (SAMPA [10] is also used in the remainder of this paper) were used.

Chunks were presented in a random order. 10% of the material was corrected by both transcribers and another 10% was transcribed twice by the same transcriber in order to calculate inter-transcriber and intra transcriber agreement, respectively. The inter-transcriber agreement is 0.91 (Cohens kappa) and the mean intra-transcriber agreement is 0.96. Both transcribers changed less than 10% of the segments, and there is quite some overlap in the segments they changed, which explains the high agreement levels. We refer to this manually corrected transcription as the *reference transcription*.

3.2. Automatic analysis

The speech material was automatically analyzed to identify phoneme errors. We aligned the reference transcription with a canonical transcription obtained from the CGN lexicon, as in [11]. The alignment was done using an algorithm that takes two phoneme sequences as input and calculates the optimal alignment on the basis of phonetic distances between pairs of phonemes [11].

The CGN lexicon provides some common pronunciation variants. We integrated these variants in the procedure by aligning the reference transcription with the canonical transcription that has the smallest phonetic distance to the reference transcription. In this way, some of the pronunciation variation that natives exhibit is taken into account.

After the alignment we calculated confusion matrices of phonemic substitutions and deletions.

4. Results

In Tables 1 and 2 all target phonemes are listed together with their frequency, percentage correctly realized and their three most frequent confusions. The phonemes are divided into six phonemic groups: diphtongs, monophtongs, plosives, fricatives, nasals and approximants.

4.1. Vowels vs. consonants in read and spontaneous speech

What appears from these tables is that, in general, more errors are produced in read speech than in spontaneous speech and that vowels cause more errors than consonants. Tables 3 shows average percentage correct scores for vowels and consonants not weighted by the individual cell frequencies while in Table 4 cell frequency is taken into account. Both measures indicate that vowels are more problematic than consonants in read speech, which is in line with results of previous research [7], while the results in Table 4 are mixed. In spontaneous speech consonants appear to be more problematic if cell frequency is taken into account.

Among the consonants, we see that the most frequently incorrectly pronounced consonants — /G/, /v/ and /z/ —, are often realized as their devoiced counterparts /x/, /f/ and /s/, respectively. In many regions in the Netherlands this phenomenon also occurs among native speakers and it is therefore questionable whether these should be regarded as pronunciation errors at all.

The data in Tables 3 and 4 also show that the difference in the number of errors between vowels and consonants is much smaller in spontaneous speech than in read speech.

All target vowels appear to be less problematic in spontaneous speech than in read speech, except for /O/., while the number of consonant errors is comparable in read and spontaneous speech.

In read speech the most problematic vowels, based on their relative error percentages, are /9y/, /Y/, /y/, /2:/, /e:/ and /E/. These vowels are much less error prone in spontaneous speech. This can be ascribed to different factors.

target	freq	%cor	error#1	error#2	error#3
Au	940	96.0	o::1.5	a:: 0.9	u: 0.5
Ei	2750	85.6	a:: 10.4	e:: 1.1	E: 1.1
9y	1094	51.7	Au: 38.0	o:: 4.8	O: 2.2
i	3852	93.3	I: 3.3	e:: 1.9	@: 1.0
e:	4206	75.7	E: 9.9	i: 7.4	I: 2.6
a:	5134	94.8	A: 4.2	@: 0.5	-: 0.3
o:	3703	88.0	O: 8.2	u: 2.1	@: 0.5
u	1367	95.5	y: 1.1	Y: 1.0	O: 0.9
y	961	67.4	u: 24.7	Y: 2.1	2:: 2.0
I	3845	84.9	i: 12.6	E: 2.1	@: 0.3
E	5366	84.7	@: 8.5	I: 4.2	-: 0.8
A	6461	91.6	a:: 7.0	@: 0.6	O: 0.3
O	3292	96.8	o:: 2.0	a:: 0.5	u: 0.3
Y	1656	61.6	u: 25.4	y: 7.5	O: 2.5
2:	627	72.9	y: 8.8	u: 5.6	@: 2.7
@	20745	94.0	E: 2.6	I: 1.2	-: 1.0
p	2847	96.4	b: 2.8	-: 0.8	
t	13899	90.2	-: 7.0	d: 2.5	s: 0.1
k	4751	96.2	g: 2.4	-: 0.7	x: 0.3
b	3149	99.7	p: 0.2	w: 0.1	
d	8909	99.2	-: 0.5	t: 0.3	
f	1688	89.0	v: 7.0	-: 3.7	w: 0.1
s	7041	91.4	z: 5.9	-: 1.6	S: 0.8
S	145	87.6	s: 11.7	j: 0.7	
x	3674	91.5	G: 3.0	-: 2.7	k: 1.3
v	4563	62.0	f: 37.4	w: 0.5	
z	2598	74.3	s: 25.7		
Z	254	81.5	x: 8.7	G: 4.3	s: 2.4
G	1075	50.6	x: 35.8	h: 6.5	g: 5.1
h	2984	95.4	-: 2.4	G: 1.0	x: 0.8
m	4212	99.2	-: 0.7		
n	16380	94.8	-: 3.1	N: 1.3	m: 0.5
N	1192	93.8	n: 3.0	-: 1.8	g: 0.8
j	2827	88.4	S: 9.1	-: 2.4	Z: 0.2
l	6941	98.6	-: 1.2	w: 0.1	j: 0.1
r	12199	92.7	-: 6.0	l: 1.1	j: 0.1
w	2524	98.9	v: 1.0		

Table 1: Phonemic substitutions and deletions in read speech.

target	freq	%cor	error#1	error#2	error#3
Au	206	98.1	a: 1.0	u: 0.5	o:: 0.5
Ei	1279	89.3	a:: 6.5	A: 3.0	i: 0.7
9y	196	61.2	Au: 28.6	O: 4.6	o:: 3.1
i	1966	92.9	I: 4.9	e:: 1.0	@: 0.7
e:	2430	89.1	E: 5.1	i: 4.2	I: 0.4
a:	3152	97.9	A: 1.2	@: 0.4	-: 0.2
o:	1591	95.5	u: 1.9	O: 1.1	2:: 0.5
u	804	96.1	Y: 1.7	2:: 0.6	O: 0.5
y	311	71.1	u: 17.7	@: 4.2	i: 3.5
I	4260	94.2	i: 4.8	E: 0.7	@: 0.2
E	2642	94.2	I: 2.3	@: 1.5	e:: 0.6
A	2685	94.1	a:: 3.6	E: 0.7	e:: 0.5
O	1274	92.4	o:: 3.8	Y: 1.2	A: 1.1
Y	342	65.8	u: 22.2	@: 8.8	I: 1.2
2:	167	80.8	@: 4.8	u: 4.2	Y: : 3.6
@	9712	96.5	-: 1.5	I: 0.7	E: 0.7
p	807	91.6	b: 7.6	-: 0.6	g: 0.1
t	6108	90.7	-: 4.9	d: 4.1	j: 0.1
k	4428	94.8	g: 4.1	-: 0.9	x: 0.1
b	962	100.0			
d	3107	98.5	-: 1.3	t: 0.1	j: 0.1
f	802	92.6	v: 6.5	-: 0.9	
s	3091	90.7	z: 6.8	-: 1.2	S: 0.9
S	73	89.0	Z: 8.2	j: 1.4	s: 1.4
x	1805	91.0	G: 3.7	-: 3.3	g: 0.9
v	1641	60.9	f: 38.8	-: 0.1	b: 0.1
z	1128	66.8	s: 32.8	S: 0.2	
Z	21	100.0			
G	585	56.4	x: 30.8	g: 5.0	-: 3.6
h	1093	82.5	-: 15.7	x: 1.2	d: 0.2
m	3424	99.4	-: 0.3	n: 0.2	b: 0.1
n	6912	94.1	-: 3.4	N: 1.9	m: 0.3
N	459	97.6	n: 1.5	g: 0.4	x: 0.2
j	1468	93.9	S: 3.5	-: 2.5	h: 0.1
l	3629	98.4	-: 1.5	s: 0.1	g: 0.0
r	4198	97.6	l: 1.4	-: 0.8	n: 0.0
w	1657	99.9	v: 0.1		

Table 2: Phonemic substitutions and deletions in spontaneous speech.

	Read	Spontaneous	Total
Vowels	83.4	88.0	85.7
Consonants	89.1	89.9	89.5
Total	86.6	89.1	87.8

Table 3: Average of %correct of vowels and consonants, read and spontaneous speech and their totals. These percentages are not weighted by cell frequencies.

	Read	Spontaneous	Total
Vowels	88.7	93.8	90.4
Consonants	92.0	92.4	92.1
Total	90.7	93.0	91.4

Table 4: Average of %correct of vowels and consonants, read and spontaneous speech and their totals. These percentages are weighted by cell frequencies.

First, these sounds are represented graphemically by 'ui' (/9y/), 'u' (/Y/), 'uu' or 'u' (/y/), 'eu' (/2:/), 'e' or 'ee' (/e:/) and 'e' (/E/). The use of the same graphemes, 'e' and 'u', to represent these phonemes might be responsible for the higher percentages of confusions in read speech as opposed to spontaneous speech, where orthography will be less of an obstacle.

This is corroborated by the finding that /y/ is more often realized when /2:/ and /Y/ are the target in read speech than in spontaneous speech.

Second, some problematic vowels occur much less often in the spontaneous material than in the read material. This is for example the case for /9y/, /Y/ and /2:/. This difference is probably related to the different composition of the read and spontaneous material. A requirement of the phonetically rich sentences contained in the read speech material used in this study is that all phonemes appear at least once in a set of sentences. The frequency of occurrence of the various phonemes can therefore be different in spontaneous speech where there are no such requirements.

Among the consonants the biggest differences between read and spontaneous speech are found for the fricatives /Z/ and /h/.

As noted in [7] the fricative /Z/ is very infrequent in normal Dutch, in which it represents 0.05% of the consonants while in the phonetically rich sentences used for this study, it represents 1% of the consonants.

The glottal fricative /h/ is more often deleted in spontaneous speech than in read speech. In this case it seems that orthography has the function of reminding the speaker of the presence of this phoneme, which is otherwise neglected, probably due to its relatively low salience.

4.2. Error patterns in read and spontaneous speech

A final remark concerns the confusion patterns associated with the various phonemes. In general there are many similarities between read and spontaneous speech, although some differences are also present. For instance, the diphthong /Ei/ is confused with /a:/, / e:/ and /E/ in read speech and with /a:/, /A/, and /i/ in spontaneous speech. The relation between the confusions in read speech and the grapheme representation 'ei' seems rather obvious. Similarly, /y/ is confused with /u/, /Y/, and /2:/ in read speech and with /u/, /@/, and /i/ in spontaneous speech. This also seems to be related to the grapheme representation 'u' or 'uu'. Finally, the vowel /Y/ is confused with /u/, /y/, and /O/ in read speech and with / u/, /@/, and /I/ in spontaneous speech. Again there seems to be a relation between the confusion pattern in read speech and the grapheme 'u'.

Among the consonants we see that the fricative /Z/ is confused with /x/, /G/, and /s/ in read speech whereas no confusions are found in spontaneous speech. The relation between the confusion pattern in read speech and the grapheme 'g' is evident also in this case.

5. Discussion and Conclusions

In the previous section we have seen that there can be differences in the occurrence of phoneme errors in read and spontaneous non-native speech. Frequency of occurrence, either absolute or relative, may be a criterion in selecting problematic phonemes that should be the focus of pronunciation training [7][12]. It follows that when making such selections one should take into account which type of speech material was used for assessing pronunciation, because this partly determines the results. Related to this, it is also important that the type of training be based on the nature of the errors identified. If errors are caused by interference from the orthography, rather than by a difficulty articulating the sound in question, then some training in phonics might be more appropriate than specific pronunciation training.

The degree to which phoneme errors are affected by the orthography will be related to the degree of orthographic transparency or orthographic depth of the L2 [13] and to the relation between the L1 and the L2 [8]. To fully appreciate the results of this study and the possible consequences for CAPT development it is important to point out that Dutch is considered to be a relatively transparent language with relatively low orthographic complexity [13],

Another aspect that should be taken into account when developing systems for pronunciation training concerns the error patterns. In a recent paper we have shown that knowledge of the error patterns can be used to develop more sensitive and more accurate metrics for pronunciation error detection [11]. Since error patterns may vary depending on whether read or spontaneous speech is used, it follows that this should be taken into account when designing CAPT systems.

6. Acknowledgements

The DISCO project is carried out within the STEVIN programme funded by the Dutch and Flemish Governments (http://taalunieversum.org/taal/technologie/stevin/).

7. References

[1] Flege, J., "The relation between L2 production and perception," In Proceedings of the XIVth International Congress of Phonetics Sciences, Berkeley, pp. 1273-1276, 1999.

[2] Van Wijngaarden, S.J., "The intelligibility of non-native speech," Doctoral dissertation, Free University, Amsterdam, 2003.

[3] Eisenstein, M., "Native reactions to nonnative speech: A review of empirical research," Studies in Second Language Acquisition, vol. 13, pp. 23–41, 1983.

[4] Lively, S.E., Logan J.S. and Pisoni D.B., "Training Japanese listeners identify /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," Journal of the Acoustical Society of America, vol. 94, pp. 1242–1255, 1993.

[5] Lyster, R., and Ranta, L., "Corrective feedback and learner uptake: Negotiation of form in communicative classrooms," Studies in Second Language Acquisition, vol. 19, pp. 37-66, 1997.

[6] Bouselmi, G., Fohr, D., Illina, I. And Haton, J., "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in Proceedings of ICSLP, 2006.

[7] Neri, A., Cucchiarini, C. and Strik, H. "Selecting segmental errors in L2 Dutch for optimal pronunciation training," IRAL - International Review of Applied Linguistics, vol. 44, pp. 357–404, 2006.

[8] Erdener D.V. and Burnham D.K., "The role of audiovisual speech and orthographic information in nonnative speech production." Language Learning, vol. 55, pp. 191–228, 2005.

[9] Cucchiarini, C., Driesen, J., Van Hamme, H. and Sanders, E., "Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus," in Proceedings of LREC, 2008.

[10] Wells, J. S, "SAMPA - computer readable phonetic alphabet," http://www.phon.ucl.ac.uk/home/sampa/.

[11] Van Doremalen, J., Cucchiarini, C. and Strik H., "Using Non-Native Error Patterns to Improve Pronunciation Verification", Submitted to Interspeech 2010.

[12] Cucchiarini, C., Neri, A. and Strik, H., "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," Speech Communication, 2009.

[13] Van den Bosch, A., Content, A., Daelemans, W., & De Gelder, B. (1994). Measuring the complexity of writing systems. Journal of Quantitative Linguistics, 1(3), 177–188.