CrossMark click for updates

Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch

Joost van Doremalen,[a] Catia Cucchiarini, and Helmer Strik
Centre for Language and Speech Technology, Radboud University Nijmegen, Erasmusplein 1, 6525HT, Nijmegen, Netherlands

(Received 11 January 2012; revised 21 February 2013; accepted 25 June 2013)

This research is aimed at analyzing and improving automatic pronunciation error detection in a second language. Dutch vowels spoken by adult non-native learners of Dutch are used as a test case. A first study on Dutch pronunciation by L2 learners with different L1s revealed that vowel pronunciation errors are relatively frequent and often concern subtle acoustic differences between the realization and the target sound. In a second study automatic pronunciation error detection experiments were conducted to compare existing measures to a metric that takes account of the error patterns observed to capture relevant acoustic differences. The results of the two studies do indeed show that error patterns bear information that can be usefully employed in weighted automatic measures of pronunciation quality. In addition, it appears that combining such a weighted metric with existing measures improves the equal error rate by 6.1 percentage points from 0.297, for the Goodness of Pronunciation (GOP) algorithm, to 0.236. © *2013 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4813304]

PACS number(s): 43.71.Gv, 43.72.Ne, 43.70.Kv [AA] Pages: 1336–1347

I. INTRODUCTION

Adult second language (L2) learners are known to experience difficulties in learning to perceive and produce the sounds of an L2 (see Flege, 1987, 1995, 1999; Best, 1995; Best *et al.*, 2001; MacKay *et al.*, 2001). The majority of adult L2 learners never acquire a native-like pronunciation and many of them retain a distinct foreign accent (Long, 1990). Incorrect pronunciation can hamper communication. Even speech that is intelligible, but characterized by a strong foreign accent, can elicit negative reactions in native speakers (Brennan and Brennan, 1981; Lippi-Green, 1997).

A major problem with pronunciation teaching is that it requires more practice time and teacher feedback than what is feasible in most language classrooms. For this reason, interest in Computer Assisted Pronunciation Training (CAPT) applications that make use of Automatic Speech Recognition (ASR) has been growing. Such systems would allow L2 learners to practice pronunciation in a private, stress-free environment in which they can access virtually unlimited input, practice at their own pace and, through the integration of ASR, receive individualized, instantaneous feedback anytime and anywhere.

Although ASR-based CAPT systems may appear particularly appealing, an important question that should be answered is to what extent such systems manage to identify pronunciation errors reliably and accurately. A large body of research has been devoted to the problem of automatic speech sound classification. An early example is Pols *et al.* (1973), in which the automatic classification of Dutch monophthongs was investigated. More recently, research specifically targeted toward automatic pronunciation quality measures that can be employed in ASR-based CAPT systems has focused on confidence scoring (Witt, 1999; Franco *et al.*, 2000; Yoon *et al.*, 2010; Wei *et al.*, 2009; van Doremalen *et al.*, 2009) using ASR-based techniques. This type of research has shown that pronunciation errors can be accurately detected to a certain extent (Witt, 1999; Franco *et al.*, 2000; Cucchiarini *et al.*, 2009; Wei *et al.*, 2009) and that difficulties may arise when it comes to identifying pronunciation errors that are based on subtle acoustic differences (Strik *et al.*, 2009).

In this paper we address the problem of automatic pronunciation error detection in L2 speech and investigate whether current automatic measures of pronunciation quality can be refined to capture subtle acoustic differences. Based on our previous research on automatic pronunciation error detection in Dutch L2, we developed the idea that the specific pronunciation error patterns produced by L2 learners might carry important information that could be exploited to improve error detection.

Pronunciation problems may be related to difficulties in perception, production, or both (Flege, 1987, 1999). An important limiting factor in acquiring the pronunciation of an L2 is considered to be interference from the mother tongue (L1). Theories that attempt to explain L1–L2 interference in speech perception are based on the tenet that the perceptual salience of phonetic detail becomes tied to the distinctions that are relevant in L1 (Kuhl *et al.*, 1992; Kuhl and Iverson, 1995; Best, 1995; Flege, 1995; Iverson *et al.*, 2003). This form of L1 entrenchment leads to "deafness" to phonetic distinctions in the L2 and may cause difficulties in learning to perceive and produce L2 speech sounds (Flege, 1995; Kuhl and Mellzoff, 1996). In the particular case of adult, literate learners there is another, less explored, but nonetheless influential factor that may affect the pronunciation of L2 sounds: The exposure to written language input and the influence of orthography that can derive from it (Young-Scholten, 2002; Erdener and Burnham, 2005; Bassetti, 2006). Adult learners

[a]Author to whom correspondence should be addressed. Electronic mail: j.vandoremalen@let.ru.nl

 0001-4966/2013/134(2)/1336/12/$30.00 © 2013 Acoustical Society of America

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

in instructional settings are often exposed to orthographic input and this may influence their pronunciation of specific L2 sounds.

Because the L1 phonology and orthography influence the pronunciation of L2 sounds, patterns of pronunciation errors in an L2 might differ depending on the speakers L1 but also the type of speech elicited. For instance, in read speech the influence of orthography is likely to be stronger than in spontaneous speech and possibly different error patterns may emerge. In this paper we argue that the error patterns that can derive from such interference are factors that can be utilized in the computation of automatic measures of pronunciation quality to improve their performance. So far various measures of pronunciation quality have been proposed (Witt, 1999; Franco *et al.*, 2000) that manage to identify relatively conspicuous errors. However, in our own research, we found that the widely used GOP scoring algorithm (Witt, 1999) has difficulties in detecting subtle errors in target phonemes with acoustically close "neighboring" phonemes (Strik *et al.*, 2009). This appeared very clearly in the case of the Dutch vowels, where relatively subtle acoustic differences are associated with different phonemic categories. Because of its characteristics—relatively many vowels, some of them distinguished by phonetic properties that are not employed in many languages, and concentrations in a specific area of the vowel space—the Dutch vowel system seems suited to investigate the performance of pronunciation quality measures.

The research reported on in this paper is aimed at analyzing the problem of automatic pronunciation error detection and at exploring possible improvements in detecting pronunciation errors that are caused by relatively subtle acoustic differences in speech sounds. A first stage in this research (Study 1, described in Sec. III) is to investigate which vowel errors are made by learners of Dutch as a second language (DL2) and their confusion patterns. This study is important to provide insight into the nature of the pronunciation errors that have to be detected. As will become clear, these errors concern subtle acoustic differences that are particularly challenging for automatic detection. In general, studies on mispronunciation detection do not provide such detailed information on the nature of the pronunciation errors and the speech data employed in the experiments. However, to clearly understand how the various measures perform it is necessary to know on which material they were tested. For instance, it should be made clear how detailed the annotations of the mispronunciations were, to what extent error gravity can be inferred from the annotations, and to what extent human labelers agreed with each other when labeling such mispronunciations.

In the second part of the paper, we go on to investigate how pronunciation errors can be detected by employing different pronunciation quality measures (Study 2, described in Sec. IV). We use two existing measures that have been previously applied by various authors to different languages. In addition, we use a pronunciation quality measure which should be able to capture subtle acoustic differences more appropriately. We test this in experiments in which we aim at detecting the vowel pronunciation errors made by DL2 learners observed in Study 1 by using all three measures. We evaluate and compare the performance of these measures and combinations thereof. We then discuss the differences observed and try to interpret the results obtained. The combination of Study 1 and Study 2 provides new insights into the ability of the different measures to detect subtle acoustic differences and into the relationship between informative predictors on the one hand and the observed error patterns on the other. Section V presents a general discussion of the results of the two studies while conclusions are drawn in Sec. VI.

II. A CASE IN POINT: THE DUTCH VOWEL SYSTEM

The Dutch vowel inventory is relatively complex: It contains 15 full vowels (12 monophthongs and 3 diphthongs), schwa and some additional vowels found mainly in loan words (Booij, 1995; Gussenhoven, 1999). In Fig. 1 a vowel chart is shown in which all full vowels of Dutch are represented by the average first and second formants (F1 and F2) measured in Adank *et al.* (2007). A feature chart of the Dutch monophthongs is shown in Table I.

The front vowels /ʏ/, /y/, and /øː/ are rounded. Furthermore, Dutch vowels can be divided into lax (/ɪ/, /ʏ/, /ɛ/, /ɑ/, and /ɔ/) and tense (/i/, /y/, /eː/, /aː/, /u/, /oː/, and /øː/) vowels. Phonologically, the tense vowels are long, but phonetically the high tense vowels /i/, /y/, and /u/ are long only before /r/ (Booij, 1995; Van der Harst, 2011). In this chart length is also indicated through IPA notation. Diphthongs are represented by arrows which indicate the glide from the initial to the final target position. Furthermore, the vowels /øː/, /eː/, and /oː/ are also slightly diphthongized.

Research with L2 learners has shown that, in the case of Dutch, vowels pose particular problems (Neri *et al.*, 2006). This is not surprising considering that the complexity of the L1 vowel system relative to that of the L2 may have consequences for L2 vowel acquisition (Iverson and Evans, 2007).

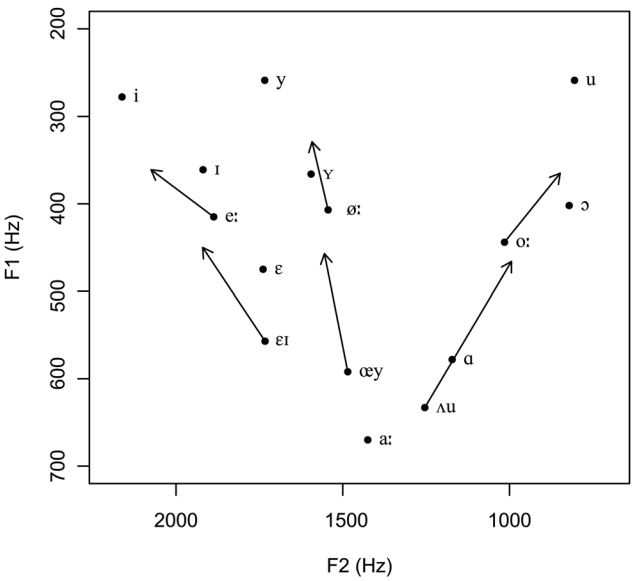

FIG. 1. Dutch vowel chart based on formant measurements described in Adank *et al.* (2007).

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

TABLE I. A feature chart containing the Dutch monophthongs adapted from Booij (1995). The features are consonant (cons), height (high and mid), backness (back) and roundedness (round).

	i	y	u	ɪ	eː	ʏ	øː	oː	ɔ	ɛ	ɑ	aː
cons	-	-	-	-	-	-	-	-	-	-	-	-
high	+	+	+	+	+	+	+	+	-	-	-	-
mid	-	-	-	+	+	+	+	+	+	+	-	-
back	-	-	+	-	-	-	-	+	+	-	+	+
round	-	+	+	-	-	+	+	+	+	-	-	-

The difficulties experienced by DL2 learners in perceiving Dutch vowels appear to be connected to the relation between the Dutch vowel system and that of their L1 and in particular to how L2 vowels map on to vowels in the native phonology (Escudero and Boersma, 2004; Goudbeek *et al.*, 2008). In general, distinctions based on dimensions that are not relevant in the L1 are likely to be more difficult than distinctions that hinge on cues that are exploited in the native phonology.

For example, in their study on Dutch vowels, Goudbeek *et al.* (2008) found that Spanish listeners had more difficulties in acquiring duration based distinctions, which are not exploited in their native phonology, than English listeners who are familiar with such distinctions. In addition, Goudbeek *et al.* (2008) found that learning a unidimensional distinction like the one between the Dutch vowels /ʏ/ and /øː/ as in the Dutch words "fut" and "feut," which differ essentially along the duration dimension, was easier for non-native listeners than acquiring a distinction based on two dimensions, like that between the Dutch vowels /y/ and /øː/ as in the Dutch words "fuut" (/fyt/) and "feut" (/føːt/), which differ with respect to F1 and duration, even if both dimensions are employed in the native phonology.

Vowel production data analyzed in Neri *et al.* (2006) are in line with the vowel perception data reported on in Goudbeek *et al.* (2008) in suggesting that distinctions based on two dimensions are problematic for DL2 learners. For example, in Neri *et al.* (2006), DL2 learners with different language background appeared to confuse /ɑ/ with /aː/, /ɪ/ with /i/ and /ɔ/ with /oː/. These pairs of tense and lax vowels are distinguished by both duration and spectral envelope (Adank *et al.*, 2004). In addition, if a learner's L1 possesses only one of the vowels in a pair, the two Dutch vowels are likely to be mapped to only one category. In such cases discrimination is difficult (Best, 1995), and this may affect production (Flege, 1995).

With respect to production there is a compounding problem besides acoustic similarity and assimilation to L1 categories. As mentioned above, orthography also plays a role, especially in read speech, in the sense that the orthography of a target language is likely to affect speech production in the target language (Young-Scholten, 2002; Bassetti, 2006; Erdener and Burnham, 2005). For example, problems in pronouncing /y/ and /ʏ/ correctly may be related to their being represented by the grapheme "u," which in other languages, e.g., Spanish and Italian, represents the phoneme /u/ instead of /y/ or /ʏ/. Moreover, in Dutch orthography the same grapheme is sometimes used to indicate two different phonemes, which might cause extra confusions. For instance, in the words "bomen" (trees) and "bom" (bomb) the grapheme "o" stands for the phoneme /oː/ in the first word and for /ɔ/ in the second. Similarly, in the words "buren" (neighbors) and "bussen" (buses) the grapheme "u" represents the phoneme /y/ in the first word and /ʏ/ in the second. Indeed, in Neri *et al.* (2006), errors made by DL2 learners in pronouncing the schwa sound appeared to be related to its being represented as "e" in Dutch orthography. Previous research we carried out on Dutch vowel production by L2 learners in read and spontaneous speech indicated that vowel errors in read speech may differ from those observed in spontaneous speech (van Doremalen *et al.*, 2010).

III. STUDY 1: VOWEL ERRORS BY DL2 LEARNERS

In this study we investigate the types of pronunciation errors made by DL2 learners in a database of read speech material collected from learners with different L1s. We first describe this speech database and the procedures applied to obtain accurate transcriptions for the present study. We then go on to present the results and relate them to those of previous research.

A. Material and method

1. Material

The L2 speech material for the present experiments was taken from the JASMIN speech corpus (Cucchiarini *et al.*, 2008). This material was recorded from L2 learners with many different mother tongues of which Arabic, Turkish, Chinese, and Hebrew are the most frequent. The learners have relatively low proficiency levels, namely, A1, A2, and B1 of the Common European Framework (Council of Europe 2001, 2001). For the experiments reported on in this paper we used the read speech material component of the database, which contains about 5 h of speech.

The material was elicited from 45 L2 learners, 18 males and 27 females, who read the same set of 40 phonetically rich sentences. The corpus comes with automatically generated phonemic transcriptions. These include disfluency phenomena such as filled pauses, restarts, and repetitions. More details on these transcriptions and the whole corpus can be found in Cucchiarini *et al.* (2008).

Because the automatically generated phonemic transcription can contain errors, for the present study we had two transcribers manually correct the phonemic transcriptions. The transcribers, who were students training as speech therapists, were instructed to correct the automatically generated phonemic transcription whenever they thought that a transcription was clearly wrong. For these corrections they were given the possibility of extending the set of phonetic symbols (SAMPA) (Wells, 1997), but eventually the transcribers used only the SAMPA symbols for Dutch. All phonemic transcriptions were corrected. The utterances were divided in chunks, stretches of around 5s of contiguous speech. The total number of chunks was 3669. Equal numbers of chunks were assigned to the two transcribers, who checked them in a random order. To be able to calculate intertranscriber agreement, we assigned 10% of the chunks to both

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

TABLE II. The most frequent phonemic substitutions, i.e., pronunciation errors, produced by L2 learners per target vowel.

Phoneme	*N*	%Correct	%Substitutions				
øː	276	53.68	y 14.47	ə 8.94	ʏ 8.09	oː 3.83	u 2.98
œy	423	55.19	ʌu 30.48	aː 3.27	ɑ 2.52	oː 1.51	ɔ 1.26
ɛi	1384	56.16	aːj 31.32	eː 3.41	œy 1.33		
ʏ	883	62.80	u 11.79	ə 6.64	ɔ 5.56	y 4.34	
oː	1749	64.24	ɔ 27.05	ə 2.72	u 1.42		
eː	2168	64.53	ɛ 14.34	ɪ 6.03	i 5.07	ə 1.94	ɛi 1.20
y	402	68.63	u 8.03	øː 5.26	ə 5.26	ʏ 3.88	i 1.66
ɪ	1907	69.13	i 22.33	ɛ 3.27			
ɑ	3253	72.24	aː 26.08				
ɛ	2092	82.03	ɪ 5.31	ə 3.24	eː 2.30	aː 2.01	ɛi 1.83
i	1883	87.08	ɪ 8.06	eː 4.00			
aː	2485	87.44	ɑ 10.73				
ʌu	419	92.33	oː 1.98	ɑ 1.73	aː 1.49	ə 1.49	
u	582	92.96					
ɔ	1617	94.17	oː 2.31				

transcribers. To check intratranscriber agreement, we had each transcriber correct 10% of the chunks twice. After removing 884 erroneously aligned chunks (for details see Sec. III A 4), the number of target vowel segments was 21 523.

The number of segments for each target phoneme is shown in Table II.

2. Phonetic time alignments

In order to detect vowel errors in this speech material, we automatically created a time alignment between the speech signal and a canonical phonemic transcription in a forced alignment process. First, this canonical phonemic transcription was generated utilizing the CGN pronunciation lexicon (Oostdijk, 2002) which contains pronunciation variants of the words as uttered by native speakers. This canonical transcription represents how the words are usually pronounced in Standard Dutch. If there are multiple acceptable pronunciation variants of a word the acoustically most likely variant is automatically selected. Second, an alignment between the speech signal and the manually corrected phonemic transcription was generated. The manually corrected transcriptions represent how the words have been realized by the L2 learners.

3. Acoustic models

Alignments were created through a Viterbi alignment using acoustic models trained with the SPRAAK package (Demuynck *et al.*, 2008). Forty-seven 3-state Gaussian mixture monophone Hidden Markov Models (HMMs) were trained with 42 h of native read speech material from the CGN speech database (Oostdijk, 2002). The total number of Gaussian components, which was shared among the monophone models, was 32 738. The average number of Gaussian components per state was 435.7. For preprocessing purposes, the input speech, sampled at 16 kHz, is first divided into overlapping 32 ms Hamming windows with a 10 ms shift and pre-emphasis factor of 0.95. Twelve Mel-frequency cepstral coefficients (MFCCs) plus C0, and their first and second order derivatives were calculated and cepstral mean subtraction (CMS) was applied.

4. Alignment verification

The quality of the alignments was checked semi-automatically. We observed that word-internal disfluencies caused problems in the alignment. Chunks containing such disfluencies could be detected relatively easily by spotting extremely long segments at the end of a chunk that were labeled as silence and that had low average acoustic likelihoods. We cleaned up the material by removing the 884 chunks that met these criteria, ending up with 2785 chunks in total. In order to determine whether a vowel was correctly realized, we checked whether more than 50% of the segment in the canonical segmentation overlapped with the same symbol in the segmentation created from the manually corrected phonemic transcription. If this was not the case, then the vowel was flagged as incorrectly pronounced.

B. Results

1. Transcriber agreement

Inter- and intrarater agreement over all sounds (including consonants) in terms of Cohen's κ are shown in Table III. Both transcribers changed less than 10% of the segments and there is quite some overlap in the segments they changed, which together explain the high agreement levels.

TABLE III. Transcription correction statistics of transcriber 1 (T_1) and 2 (T_2). $T_1 \cup T_2$ defines the set of segments which was corrected by either T_1 or T_2 (or both). Intra- and interrater agreements were calculated using Cohen's κ.

	Value
T_1 %segments changed	3.4%
T_2 %segments changed	8.2%
$T_1 \cup T_2$ %segments changed	8.7%
Cohen's κ intra T_1	0.975
Cohen's κ intra T_2	0.948
Cohen's κ inter $T_1 - T_2$	0.913

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

TABLE IV. Confusion matrix based on transcriptions made by the two transcribers for a subset containing 10% the material. The agreement coefficients per vowel were calculated by dividing the element on the diagonal by the sum of all the elements on that row and column, respectively.

	i	ɪ	eː	ɛ	aː	ɑ	oː	ɔ	u	y	ʏ	ɛi	ʌu	øː	œy	Agr.
i	382	20	18	-	-	-	-	-	-	-	-	-	-	-	-	0.845
ɪ	24	279	6	7	-	-	-	-	-	-	1	1	-	-	-	0.773
eː	6	13	267	11	-	-	-	-	-	-	-	3	-	1	-	0.788
ɛ	-	8	11	381	1	2	-	-	-	-	-	5	-	-	-	0.878
aː	-	-	-	1	440	57	-	1	-	-	1	7	-	-	-	0.789
ɑ	-	1	-	3	46	537	-	2	-	-	-	2	1	-	-	0.822
oː	-	-	-	-	-	-	254	29	5	-	-	-	2	2	1	0.767
ɔ	-	-	-	-	1	-	31	288	2	-	4	-	2	-	-	0.787
u	-	-	-	-	-	-	2	2	181	4	2	-	-	2	-	0.879
y	2	-	-	-	-	-	-	-	1	75	5	-	-	3	-	0.735
ʏ	-	1	-	-	1	-	-	-	2	5	76	-	-	-	-	0.776
ɛi	-	-	1	4	2	-	-	-	-	-	-	205	-	-	1	0.887
ʌu	-	-	-	-	1	1	5	4	-	-	-	-	109	-	6	0.773
øː	-	-	2	-	-	-	-	-	3	7	-	-	-	41	-	0.672
œy	-	-	-	-	-	1	-	-	-	-	-	-	10	-	40	0.678

We also calculated agreement measures for the individual vowels, shown in Table IV. This table also shows the confusion matrix of the vowel annotations used for interrater agreement calculation. For example, in 20 cases in which the first transcriber labeled a segment as /i/, the second transcriber labeled it as /ɪ/.

2. Pronunciation errors

Table II shows the proportions of correct pronunciations per target vowel in ascending order, as assessed by the native transcribers. The right hand part of this table indicates the substitutions with the highest relative frequency (>1%) made by the speakers for each vowel phoneme. The most frequent errors are found in the diphthongs /œy/ (as /ʌu/) and /ɛi/ (as /aːj/), although it has to be mentioned that the latter can also be considered a regional variant. Other frequent errors concern the confusion between tense and lax vowels such as /aː/-/ɑ/, /oː/-/ɔ/, /i/ - /ɪ/, /e/ - /ɛ/, and the vowel pairs /y/ and /øː/ and /y/ and /u/. Most of the other vowels have rather diffuse patterns of errors.

C. Discussion

1. Transcriber agreement

The level of agreement between the two transcribers (Table IV) varies for the different vowels. Relatively many disagreements concern contrasts like /aː/-/ɑ/, /oː/-/ɔ/, /y/-/ʏ/ and the cluster /i/, /ɪ/, /e/ and /ɛ/. This might have to do with the fact that L2 learners, like native speakers for that matter, realize vowels somewhere on a continuum between two phonemic classes. However, non-native speakers do this more often and differently from native speakers. For instance, they may realize a vowel with the quality characteristics of /ɔ/ and the duration of /oː/. It is not surprising that native transcribers find it difficult to categorize such sounds. Additionally, the transcribers might have different thresholds for deciding whether a phone is not realized canonically.

To gain insight into the relation between acoustic similarities in vowels as spoken by native speakers and the agreements of the transcribers in transcribing vowels spoken by non-native speakers, we have tried to visualize the differences using Principal Coordinates analysis (PCoA), also known as multidimensional scaling. For the acoustic similarities in native vowels, we calculated a distance matrix between the acoustic models and projected the vowels in a two-dimensional space using these distances (see Fig. 2). The Kullback–Leibler divergence between the second states of two HMM models containing Gaussian Mixture Models f and g was approximated using Monte Carlo simulation (Hershey and Olsen, 2007). We calculated

$$D_{\mathrm{MC}}(f \parallel g) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(x_i)}{g(x_i)} \tag{1}$$

using $n = 10\,000$ i.i.d. samples. Note that for diphthongs and diphthongized monophthongs, calculating only the distances

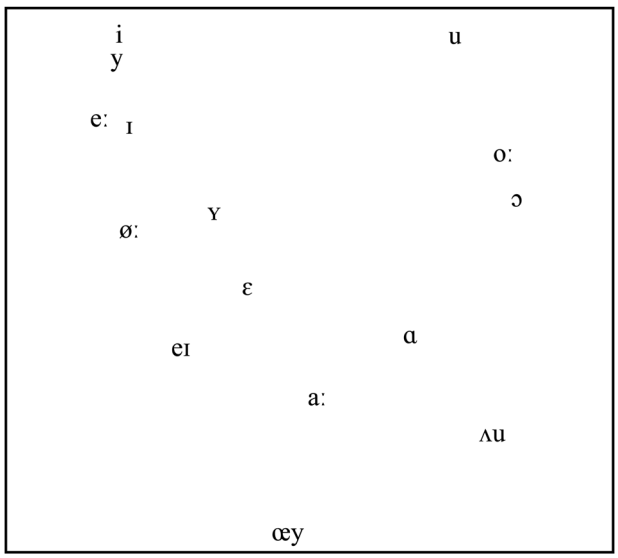

FIG. 2. Two-dimensional mapping based on a Principal Coordinates Analysis of Kullback–Leibler divergences between the acoustic vowel models. See Sec. III C 1 for details.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

between the second states of the models is an oversimplification, as these are less static sounds than monophthongs.

For the agreements regarding non-native vowels, we transformed the confusion matrix M shown in Table IV into a distance matrix D. First the elements were normalized for their frequency,

$$M'_{ij} = \frac{1}{2}\left\{\frac{M_{ij}}{\sum_k M_{kj}} + \frac{M_{ij}}{\sum_k M_{ik}}\right\}. \quad (2)$$

Then, these normalized agreement coefficients were transformed so that they could be interpreted as distances

$$D_{ij} = -\log(M'_{ij} + c), \quad (3)$$

where $c = 0.01$. All distances on the diagonal were set to 0. PCoA was carried out on the resulting distance matrix. The result is shown in Fig. 3. Although this particular projection is based on only few data points (as can be seen from Table IV) and the transformation from a confusion matrix to a distance matrix is not trivial, these representations seem to reveal some interesting patterns.

In Fig. 2 it can be observed that for example /oː/ and /ɔ/ are acoustically very similar, and these are also sounds that are often confused by the transcribers (see Fig. 3).

As stated before, non-native speakers tend to realize certain sounds in a continuum between two phonemic classes more often than native speakers. This specifically seems to be the case for /œy/ and /ʌu/ which, albeit acoustically distinct, are often confused by the native transcribers in our experiment. This seems to suggest that it would not be difficult for native listeners to discriminate /œy/ and /ʌu/ spoken by native speakers, but it is difficult in the case of non-native speech as some of these speakers tend to blur the distinction between /œy/ and /ʌu/.

The finding that for several contrasts (/aː/-/ɑ/, /oː/-/ɔ/, /y/-/ʏ/ and the cluster /i/, /ɪ/, /eː/ and /ɛ/) the agreement between two transcribers is low can be considered as a sort of benchmark for automatic error detection. In other words, it is not possible for a pronunciation error detection system to achieve 100% accuracy, when human transcribers do not agree perfectly.

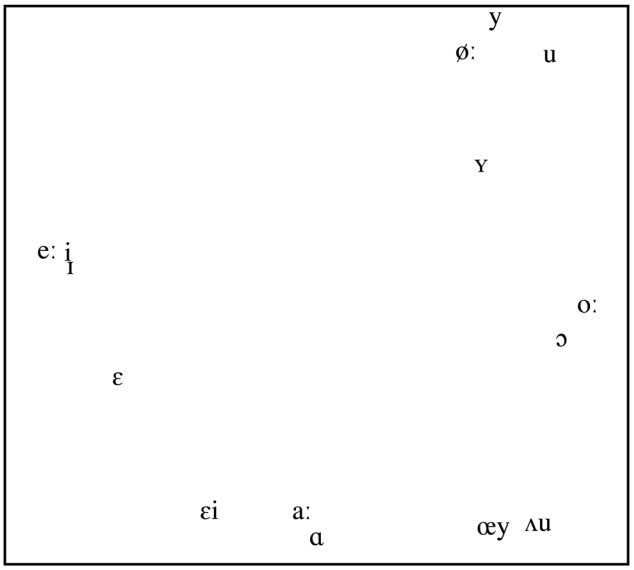

FIG. 3. Two-dimensional mapping based on a Principal Coordinates Analysis of interrater confusions. See Sec. III C 1 for details.

2. Pronunciation errors

The results concerning the most frequent vowel pronunciation errors (Table II) partly confirm those obtained in previous research with L2 learners by Neri *et al.* (2006), which showed that the most problematic vowels for L2 learners of Dutch with different L1s were: /ɑ/, /œy/, /aː/, /y/, /ɛi/, and /ʏ/. In line with results presented in Goudbeek *et al.* (2008), which showed that unidimensional contrasts were less problematic than multidimensional contrasts, we find that vowels in a pair that differ in two dimensions are difficult to keep apart, as is attested by the confusions between /oː/-/ɔ/, /ɑ/-/aː/, /ɪ/-/i/, and /y/-/øː/, which differ both in spectral envelope and duration.

Another finding that is partly in line with previous research is that the confusions between vowels tend to be asymmetric. For instance, the diphthong /œy/ was often realized as /ʌu/, while /ʌu/ was never realized as /œy/, /ɑ/ was more often realized as /aː/ than /aː/ as /ɑ/ and /ɪ/ was more often realized as /i/ than vice versa. An interesting asymmetry was also observed with respect to the vowels /y/ and /ʏ/ which were often realized as /u/ (8.03% and 11.79%, respectively) while /u/ was seldom realized as either /y/ or /ʏ/.

As anticipated in Sec. II, some of these errors may be ascribed to assimilation to L1 categories, for instance, because the learner's mother tongue has only one of the vowels in the pairs of tense and lax vowels. Assimilation to L1 categories could also be responsible for errors in which the diphthong /œy/ is realized as /ʌu/. An additional explanation for some of the common errors and error patterns may be interference from Dutch orthography. This could apply in the case of /y/ and /ʏ/ being realized as /u/. Similarly, it could hold for /øː/, which is represented by the grapheme "eu," being realized as /y/, /ʏ/, or /u/. In van Doremalen *et al.* (2010) we found that such confusions were indeed more frequent in read speech than in spontaneous speech, where orthography will be less of an obstacle. The asymmetry observed in the confusions between /y/, /ʏ/, and /u/ seems to support this hypothesis.

As was to be expected, many of these errors concern subtle acoustic differences. For the purpose of automatic pronunciation training, it is important to be able to identify such errors and this requires a pronunciation quality measure that is capable of capturing such subtle differences. In the next section, we investigate the performance of some of these measures.

IV. STUDY 2: PRONUNCIATION ERROR DETECTION IN VOWELS UTTERED BY DL2 LEARNERS

In this study, we investigate the performance of various pronunciation quality measures in detecting Dutch vowel pronunciation errors. First, we discuss pronunciation quality measures in general and give two examples of widely used

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

measures. Subsequently, we describe a measure designed to be sensitive to relevant subtle acoustic differences. We then proceed to study its performance in comparison to that of the other two measures.

A. Automatic pronunciation quality measures

Several methods have been proposed to automatically assess segmental pronunciation quality. One prevalent method is to calculate a segmental confidence measure that indicates the confidence we can have that the realized phone belongs to the same phonemic class as the one that should have been uttered. If this confidence is too low, the segment is considered as erroneously realized.

Most confidence measures estimate the posterior probability of a symbol, e.g., a word or a phone, given some set of acoustic observations. In the case of segmental pronunciation quality, this estimation is usually carried out for individual phones,

$$\mathrm{P}(p|O) = \frac{\mathrm{P}(p|O)\mathrm{P}(p)}{\mathrm{P}(O)} \tag{4}$$

$$\mathrm{P}(p|O) = \begin{cases} \geq \theta & : \text{correct} \\ <\theta & : \text{incorrect}, \end{cases} \tag{5}$$

where p is the target phoneme and O a set of acoustic observations. In practice, the prior probability $\mathrm{P}(p)$ is often discarded. If the resulting value is below a certain predefined threshold θ the phone is flagged as incorrectly realized; otherwise it is regarded as correctly realized. It is in general very difficult to determine the denominator $\mathrm{P}(O)$ in Eq. (4), so various procedures have been proposed to estimate it. Below, we discuss two approaches to factoring out $\mathrm{P}(O)$, i.e., the Goodness of Pronunciation (GOP) measure and the Average Posterior probability Estimator (APE). We also present an alternative measure, weighted Phone Confidence (wPC).

1. Goodness of pronunciation

One well known method to approximate the denominator in Eq. (4) is the one used in the Goodness of Pronunciation (GOP) algorithm (Witt, 1999; Witt and Young, 2000). In this method, Hidden Markov Models (HMMs) are used to model the likelihood of the acoustic observations, such as mel-frequency cepstral coefficients (MFCCs) or perceptual linear predictive coefficients (PLPs), given the phonemic class to which the phone belongs. These phone models are usually trained on native speech material.

In this algorithm the ratio of the likelihood of the target phoneme and the likelihood of the acoustically most likely phoneme is calculated for each frame. This normalization is intended to approximate the denominator in Eq. (4), $P(O)$. The resulting measure is normalized by the duration of the segment and transformed to a log scale, which yields

$$\mathrm{GOP}(p) = \frac{1}{t_e - t_b} \sum_{t=t_b}^{t_e} \log \left[\frac{\mathrm{P}(O_t|p)}{\max_i \mathrm{P}(O_t|p_i)} \right], \tag{6}$$

where p is the target phoneme, t_b and t_e the beginning and ending times of the target segment, respectively, and O the acoustic observations. The higher the value of the GOP measure, the higher the likelihood that the target phoneme was indeed uttered by the speaker. The decision of accepting or rejecting the phone as a correct pronunciation of the target phoneme is made by simple thresholding. These thresholds are determined separately for each target phoneme and can be calibrated on real non-native speech material or on native material in which artificial errors have been introduced (Witt, 1999; Kanters *et al.*, 2009).

2. Average posterior probability estimator

A related method to estimate the posterior probability, which we denote as the average posterior probability estimator (APE), is introduced in Franco *et al.* (2000),

$$\mathrm{APE}(p) = \frac{1}{t_e - t_b} \sum_{t=t_b}^{t_e} \log \left[\frac{\mathrm{P}(O_t|p)}{\sum_i^N \mathrm{P}(O_t|p_i)} \right], \tag{7}$$

where the summation in the denominator runs over all N phones. The main difference with the GOP measure is that the denominator is estimated by the summation over all phones instead of the maximum likelihood phone sequence.

3. Weighted phone confidence

To take due account of subtle relevant acoustic differences between realizations of target speech sounds, we use an alternative measure designed to be more sensitive to these differences. In this measure, we combine the ratios of the likelihood of the target phoneme and all other (relevant) "competing" phonemes in a logistic regression model. It is important to realize that what are competing phones may differ depending on the language background of the L2 learners, their degree of proficiency, and whether we are dealing with read speech or spontaneous speech. Wei *et al.* (2009) also adopt a combination of these types of scores, but they employ this measure to detect non-standard variants in native speech by using a Support Vector Machine (SVM).

The rationale behind the present approach is that the individual scores capture the discrepancy between the L2 target phoneme and other, "competing," phonemes. For instance, in the GOP measure a categorical choice is made between possible realizations of the target phone: The most likely phone is chosen and the rest of the information is lost. In the wPC measure, on the other hand, various options are kept open and the information on the distance between the target phoneme and its competitors remains available. The individual scores are weighted and summed so that the impact of each likelihood ratio on the dependent variable, the correctness of a phone, can be taken into account. The weights are obtained by training logistic regression models with a ridge estimator (le Cessie and van Houwelingen, 1992) on non-native speech data. Each phone is categorized as either correct (1) or incorrect (0). The specifics of the training and implementation are presented in Sec. IV B 1.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

We call the resulting metric the *weighted Phone Confidence* (wPC).

We denote the individual phone confidence (PC) scores for a target phoneme p_{targ} with a competitor phoneme p_i as $\mathrm{PC}_{p_i}^{p_{\mathrm{targ}}}$, which is defined as

$$\text{for every } p_i \text{ in } \mathbf{P}: \ \mathrm{PC}_{p_i}^{p_{targ}} = \frac{1}{t_e - t_b} \sum_{t=t_b}^{t_e} \log \left[\frac{\mathrm{P}(O_t|p_{targ})}{\mathrm{P}(O_t|p_{targ}) + \mathrm{P}(O_t|p_i)} \right], \tag{8}$$

where O is the observation matrix, p_{targ} the target phoneme and $\mathbf{P}$ the set of phonemes that is hypothesized to be in competition with the target phoneme. Note that the denominator for each PC score is a stable term. This is in contrast with the GOP score, where the denominator changes when the acoustically most likely phone changes. As mentioned above, the PC scores are combined in a logistic regression model

$$\mathrm{wPC}^{p_{trag}} = \frac{1}{1 + \exp\left\{ -\left(\beta_0 + \sum_i \beta_i \mathrm{PC}_{p_i}^{p_{trag}} \right) \right\}}. \tag{9}$$

These models are trained for each phoneme separately. In these models, the dichotomous dependent variable, which represents whether the phone was correctly or incorrectly realized, is predicted by the combination of likelihood ratios.

B. Material and method

For these experiments we used the same speech material as in Study 1 (see Sec. III A 1). This consists of the speech signals and the corresponding alignments of the canonical transcription with a detailed transcription that had been manually corrected by trained transcribers (see Sec. III A 4 for details). The baseline pronunciation quality measures we evaluated on this material are the GOP and APE measures. In addition, we evaluate the wPC measure and different combinations of these measures.

In Sec. IV B 1, we explain how we implemented and evaluated the GOP, APE, and wPC measures. In Sec. IV B 2, we discuss how we automatically selected the most informative predictors. This was done in order to obtain models that better generalize to unseen data and that are easier to interpret.

1. Pronunciation quality measure implementation and evaluation

For the calculation of the GOP, APE, and PC scores, we employed the acoustic monophone models discussed in Sec. III A 3. We calculated the GOP measure following Eq. (6). To obtain the denominator, the likelihood of the optimal phone sequence, we employed an unconstrained free phone recognizer which was used to decode whole audio files. The APE measure was calculated following Eq. (7).

The wPC measure was implemented following Eqs. (8) and (9). For the target vowel phonemes, we chose all the other 15 Dutch full vowel phonemes (see Fig. 1), schwa, and a silence model as potentially competing phonemes. These PC scores were calculated, and the likelihoods of these competing phonemes are simplified by following the same state level segmentation as the Viterbi path that was obtained for the target phoneme. That is, the competing phonemes begin, end, and switch states at the same times as the target phoneme.

Subsequently, the regression models are trained for each vowel phoneme separately. To train a specific regression model of a target vowel, we extracted the segments for which this vowel appeared in the canonical transcription as a target phoneme and calculated the 17 PC scores for these segments. Then, we trained and tested the models using leave-one-speaker out cross-validation within the WEKA package (Witten and Frank, 2005). That is, the coefficients are first determined using all tokens of 44 speakers and afterward tested on the tokens of the remaining speaker. This is repeated until all tokens are tested. The number of tokens per phoneme is shown in Table V, together with the percentage of pronunciation errors.

We evaluated the pronunciation quality measures on the basis of the equal error rate (EER), which is the point on the receiver operating characteristic (ROC) curve where the false positive rate is equal to the false negative rate.

2. Model selection

Although the GOP, APE, and wPC measures are all intercorrelated, they might still carry different information. For this reason, we also evaluated models in which both the GOP and APE measures and all PC scores are included in the logistic regression models discussed in the previous section. Some of the 17 PC scores regarding acoustically similar vowels are also highly correlated. Furthermore, some scores may not be informative at all. This can be a problem,

TABLE V. Overall results of the GOP measure, the APE measure and the wPC score. BI = best individual predictor, All=all predictors, BS=best subset of predictors.

			Equal Error Rate					
Phoneme	*N*	%Errors	GOP	APE	wPC	BI	All	BS
øː	276	46.32	0.331	0.315	0.277	0.292	0.246	0.223
œy	423	44.81	0.215	0.220	0.161	0.174	0.170	0.148
ɛi	1384	43.84	0.271	0.247	0.229	0.247	0.216	0.212
ʏ	883	37.20	0.269	0.251	0.205	0.251	0.209	0.196
oː	1749	35.76	0.422	0.414	0.325	0.341	0.312	0.310
eː	2168	35.47	0.242	0.277	0.229	0.242	0.205	0.200
y	402	31.37	0.282	0.255	0.254	0.254	0.247	0.231
ɪ	1907	30.78	0.292	0.318	0.240	0.254	0.216	0.202
ɑ	3253	27.76	0.301	0.305	0.281	0.295	0.280	0.275
ɛ	2092	17.97	0.262	0.262	0.243	0.262	0.228	0.220
i	1883	12.92	0.233	0.255	0.233	0.232	0.233	0.221
aː	2485	12.56	0.336	0.286	0.231	0.230	0.221	0.210
Avg.			0.288	0.284	0.242	0.256	0.232	0.221
ʌu	419	7.67	0.373	0.354	0.424	0.354	0.410	0.323
u	582	7.04	0.299	0.263	0.451	0.263	0.356	0.263
ɔ	1617	5.83	0.319	0.348	0.423	0.319	0.352	0.309
Avg.			0.297	0.291	0.280	0.267	0.260	0.236

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

because the number of instances on which the models are trained is quite low. As this can lead to overfitting, decorrelating, or removing predictors can actually increase the generalizability. Moreover, to interpret the models, it would be interesting to observe how the selected predictors relate to the error classes observed in Study 1. Therefore, we investigated the effects of automatically selecting the most informative set of predictors.

A method for efficiently evaluating an important subset of the total number of alternative models is stepwise regression. In this framework, predictors are either iteratively added to an empty set or dropped from the full set of predictors (or a combination) based on their contribution to the prediction of the independent variable. Stepwise insertion and stepwise removal can yield different results, because a specific set of predictors affects the predictor that will subsequently be selected when the predictors are intercorrelated. However, for our data set, all methods seemed to yield the same results. Stepwise regression was carried out using the R software package (R Development Core Team, 2010).

We evaluated the full set of predictors, the individual predictor with the highest goodness-of-fit and the selected subset of predictors using stepwise regression and compared them to the measures described in the previous section.

C. Results

ROC curves are shown in Fig. 4. The EERs for the different measures and the different vowels are shown in Table V. These EERs are calculated over the full dataset, because each token is evaluated through leave-one-speaker-out cross-validation. The average over the vowels, which is not weighted by the number of tokens per vowel, is also shown. The list is ordered by the percentage of pronunciation errors per vowel. Although the EER of the APE measure (0.297) is somewhat lower than that of the GOP measure (0.291), the difference is not statistically significant beyond the 0.95 confidence interval. The wPC measure, however, performs significantly better than the two other measures, with an overall EER of 0.280, which is a relative improvement of 3.9%.

In addition, for the three vowels /ʌu/, /u/, and /ɔ/ the number of pronunciation errors was quite low with a relative frequency lower than 10%. For these vowels, the EER of the wPC measure is much higher than the EER of GOP and APE. Apparently, no reliable regression models can be trained when such a small portion of the segments have been incorrectly realized. In other words, the wPC metric performs better for the errors that are more frequent and since frequency is considered to be an important criterion for error selection in L2 pronunciation training (Neri *et al.*, 2006), we chose to calculate the average EER without the three vowels with the lowest relative error frequency. If we do not take these vowels into account, the relative improvement in wPC with respect to APE is 15.0%, or 5.5 percentage points. The improvement in wPC relative to the best individual measure (GOP or APE) for a given vowel is highest for /œy/: [25.1% (21.5–16.1)/21.5], /øː/: (23.6%), /oː/: (21.5%), /aː/: (19.2%), /ʏ/: (18.3%) and /ɪ/: (17.8%).[1]

To gain insight into which PC scores are important in the wPC models, and to study the effect of combining wPC with GOP and APE, we carried out model selection. We evaluated three additional measures: The best individual (BI) predictor out of the total set of 19 predictors, a combination of all predictors (*All*) and a combination of the best subset (BS) of predictors selected per vowel through stepwise regression. The selected predictors BI and BS are shown in Table VI. The performance of these three measures is shown in Table V. Overall, the BI predictor performs better than the APE and GOP measures. The reason for this is that for some vowels one of the confidence scores performs better than both GOP and APE. This is the case for /aː/, /ɑ/, /i/, /ɪ/, /oː/, /œy/, and /øː/. In Sec. IV D 2 we discuss this in more detail.

Including all predictors in the regression model decreases the EER over wPC from 0.242 to 0.232 for the vowels with more than 10% pronunciation errors. As said, not all of these predictors carry useful information. The

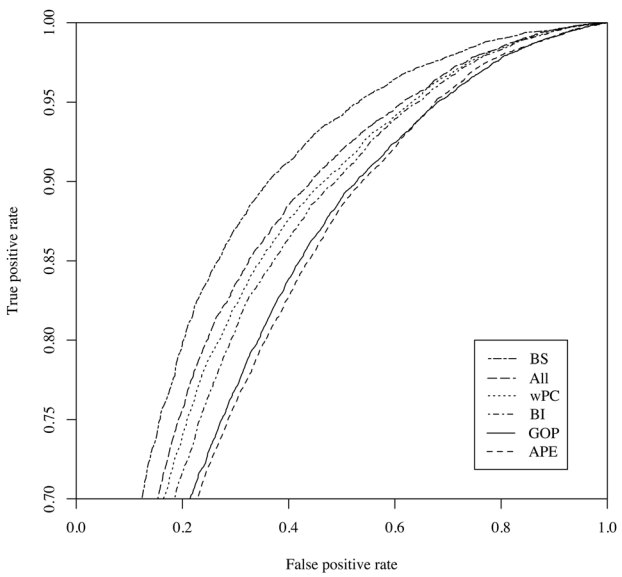

FIG. 4. ROC curves of the six different measures.

TABLE VI. The best individual (BI) predictor and the best subset (BS) of predictors after carrying out stepwise regression per vowel.

Phoneme	BI	BS
ø	PC_y	PC_y APE $PC_ʏ$
œy	$PC_{ʌu}$	$PC_{ʌu}$ APE $PC_{aː}$
ɛi	APE	APE $PC_{aː}$ $PC_ɛ$ $PC_{œy}$
ʏ	APE	APE PC_u PC_y $PC_@$ $PC_ɔ$
oː	$PC_ɔ$	$PC_ɔ$ APE
eː	GOP	APE PC_i $PC_ɪ$ $PC_ɛ$
y	APE	APE PC_u $PC_{øː}$ $PC_ə$
ɪ	PC_i	PC_i APE $PC_ɛ$ $PC_{eː}$
ɑ	$PC_{aː}$	APE $PC_{aː}$ $PC_{oː}$ $PC_{ɛi}$
ɛ	APE	APE $PC_ɪ$ $PC_{aː}$ $PC_ə$
i	$PC_ɪ$	$PC_ɪ$ $PC_ə$
aː	$PC_ɑ$	$PC_ɑ$ $PC_{ɛi}$
ʌu	APE	$PC_{oː}$ $PC_{œy}$
u	APE	APE
ɔ	GOP	APE $PC_{oː}$

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

number of predictors obtained through stepwise regression is only 3 (on average), and the overall performance of these subsets *BS* is 0.221, which is significantly better than *All*, as can be derived from their confidence intervals. Probably *BS* performs better than *All* because of overfitting in the case of *All*. For *BS* we also calculated the average recall of pronunciation errors at precisions of 0.600, 0.700, and 0.800. The average recall values weighted by the frequency of the vowel classes are 0.645, 0.612, and 0.526, respectively. The unweighted average recall values are 0.707, 0.620, and 0.529.

It is interesting to note that the selected PC scores are similar to the target vowel substitutions (see Table II). For example, the selected PC scores for /eː/ are those relative to /i/, /ɪ/ and /ɛ/ and these are also the vowels with which /eː/ is often confused. We elaborate on this finding in the following section.

D. Discussion

1. Selected PC scores and error patterns

The confidence scores of the vowels with which the targets are most frequently substituted (shown in Table II) are always present in the automatically selected subset (shown in Table VI). For example, /øː/ is most often substituted with /y/ (14.5%, see Table IV) and its PC score is among the selected subset of predictors, as well as /ʌu/ for /œy/ (30.5%), /ɔ/ for /oː/ (27.1%), /aː/ for /ɑ/ (26.1%) and /i/ for /ɪ/ (22.3%). The PC scores of other frequent confusions are also often present in the selected subset. It appears that specific pronunciation errors found by the transcribers coincide with the PC scores obtained through stepwise regression. This indicates how specific error patterns may be relevant for error detection, in the sense that these patterns indicate important features relevant for pronunciation error detection. Besides the PC scores, APE is also often included in the best subset of predictors, except for /i/, /aː/, and /ʌu/.

2. GOP and the most informative PC score

As can be seen from Table VI, for some vowels the best individual predictor is a score relative to only one other vowel (one PC score). In contrast, in the GOP algorithm, the score in the denominator is always the most likely phone. So, the denominator in the GOP score can change at points where the most likely phone switches from one to another. When analyzing subtle errors, for example, when a target sound a is realized on a continuum between a and another sound b, this might not be a desirable property. This can best be illustrated by a simplified example.

Suppose we have a one-dimensional acoustic observation vector O and two hypothetical phonemic classes a and b, modeled by Gaussian distributions [Fig. 5(A)]. In Fig. 5(B) the GOP measure for the target phoneme a is shown. We can see that it is zero everywhere where $P(O|x=a) \geq P(O|x=b)$. If $P(O|x=a) < P(O|x=b)$ the GOP measure drops abruptly, whereas this effect does not seem to reflect the gradual acoustic change. This happens because as the most likely phone switches, the denominator in Eq. (6) also suddenly changes.

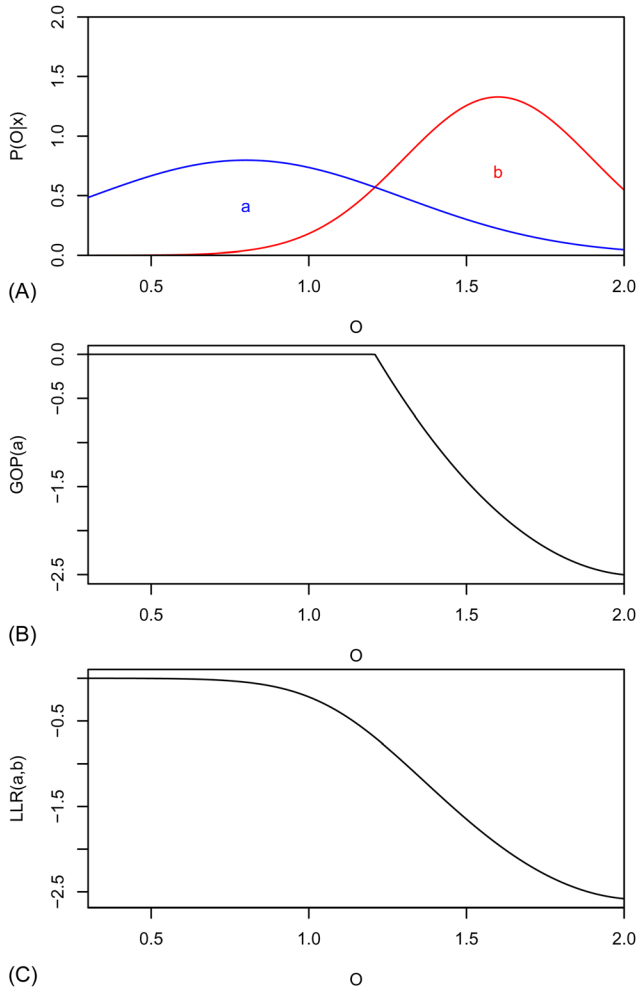

FIG. 5. (Color online) Hypothetical example of how different measures represent acoustically close observations. (A) Two Gaussians representing phones a and b are shown. (B) The GOP score for phone a and (C) the PC score of phone a relative to phone b are shown. See Sec. IV D 2 for more details.

However, if we calculate the likelihood ratio between $P(O|x=a)$ and the stable normalization factor $P(O|x=a) + P(O|x=b)$ as in the PC scores, we represent this situation in a more gradual manner, as shown in Fig. 5(C). We would expect the best individual PC score only to work better than GOP for target phonemes with errors concerning only one phoneme or a cluster of similar phonemes. This is corroborated by the finding that in our experiment most of the target vowels for which one PC score is better than GOP, /aː/, /ɑ/, /i/, /ɪ/, /oː/, /œy/ and /øː/, are frequently confused with only one other vowel.

V. GENERAL DISCUSSION

The research reported on in this paper was inspired by the idea that information on error patterns in L2 pronunciation might be useful for developing improved automatic measures of pronunciation quality.

To investigate whether automatic pronunciation error detection can be improved by employing quality measures that take account of the specific error patterns observed in an L2, we conducted two studies on Dutch vowel pronunciation.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

The rationale behind the choice for Dutch vowels was that Dutch vowels constitute an interesting and illustrative example for this kind of research because of the complex error patterns they induce.

The results of Study 1 do indeed reveal complex error patterns, which in part can be ascribed to the mismatch between the Dutch vowel phonology and those of the L1s and to interference from L2 orthography.

In Study 2 we compared the performance of three different measures of pronunciation quality: GOP, APE, and wPC. The GOP and APE measures are not targeted toward modeling specific error patterns, whereas the wPC measure is, as it is trained on a corpus of manually annotated speech. The relative improvement of wPC over APE is 15.0% and the combination of automatically selected informative predictors among PC scores, GOP and APE yields a relative improvement of 22.2% over APE. The average EER of this last measure is 0.221. This means that when a threshold is set at this point on the operating curve, the false negative rate is 22.1%, and the true negative rate is 77.9%. As false negatives in CAPT systems are usually regarded as more detrimental than false positives, this threshold should be changed to reduce the number of false negatives at the expense of also reducing the number of true negatives.

Another important concern in using the wPC measure in applications is the issue of generalizability to other speakers and tasks. We trained the acoustic models speaker-independently, and the L2 learners in our material have widely varying L1s. Although these languages have different phonologies, apparently there is some systematicity in the error patterns of these L2 learners, at least enough for our measure to profit from it. This means that some phonemic confusions are quite stable across L2 learners. This was also observed in Neri *et al*. (2006), where a number of phonemic confusions were identified that were common to L2 learners with varying L1s. On the other hand, it is reasonable to assume that our measure could be further improved by using data from specific L1s or clusters of typologically similar L1s, as this might lead to more specific confusions and therefore more accurate regression models.

In this connection, another important element is the kind of task the L2 learners have to perform. We used read speech data, where the speakers had to read sentences from a computer screen. As stated in Sec. II, there are some obvious phonemic confusions due to interference with the orthography in this task, which are less likely to occur when L2 learners are not reading but, for example, have to repeat spoken utterances. This might lead to different error patterns. Since we have seen that error patterns bear information that is useful in computing pronunciation quality measures, the speech data used for training the error detection algorithm should be of the same type—with similar error patterns—as those in which pronunciation errors will have to be detected.

There are several ways in which the wPC measure could be improved. For example, in our specific use case of Dutch vowels, one important characteristic which we did not model is duration (Booij, 1995), which should be taken into account explicitly when assessing the pronunciation quality of a phone. However, it is generally difficult to model phone duration because of a normalization problem. This normalization can be performed on different levels, and it is not directly clear which option is optimal. This is a problem that should be explored in further research.

Another important property of phonemes in general is their context dependence. In this research we did not employ context dependent models, but for some phonemes this might be crucial to assess their quality. For example, the phonemes /o:/, /e:/, and /ø:/ are diphthongized when they are pronounced before certain consonants (/r/, /l/, /j/, and /w/). Initial experiments in which this contextual knowledge was included into the classifiers yielded very promising results.

A final aspect that could lead to improvement is the segmentation of the speech signal into phones. Since all local confidence scoring heavily depends on it, it follows that improving the segmentation is likely to result in better detection performance.

VI. CONCLUSIONS

In this paper we have studied the nature and frequency of vowel pronunciation errors produced by learners of Dutch as a second language. This study has revealed that many of these errors concern relatively subtle acoustic differences. We then investigated how to automatically detect these pronunciation errors. We compared well-established pronunciation quality measures (GOP and APE) with an alternative measure (wPC) that takes account of error patterns to capture relevant acoustic differences. We found that the proposed measure performed significantly better than the two other measures. From additional experiments involving model selection techniques, we observed that the predictors in the selected models do indeed coincide with frequently observed pronunciation errors.

ACKNOWLEDGMENTS

The DISCO project was carried out within the STEVIN program which was funded by the Dutch and Flemish Governments.

[1]We also trained and tested Support Vector Machine (SVM) models with Radial Basis Function (RBF) kernels and these results were comparable to the results obtained with the logistic regression models.

Adank, P., Smits, R., and van Hout, R. (**2004**). "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**, 3099–3107.

Adank, P., van Hout, R., and van de Velde, H. (**2007**). "An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties," J. Acoust. Soc. Am. **121**, 1130–1141.

Bassetti, B. (**2006**). "Orthographic input and phonological representations in learners of Chinese as a foreign language," Written Language Literacy **9**, 95–114.

Best, C. (**1995**). "A direct realist view of speech cross language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (New York Press, Baltimore, MD), pp. 171–206.

Best, C., McRoberts, G., and Goodell, E. (**2001**). "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listeners native phonological system," J. Acoust. Soc. Am. **109**, 775–794.

Booij, G. (**1995**). *The Phonology of Dutch* (Clarendon Press, Oxford, UK), pp. 1–205.

Brennan, E., and Brennan, J. (**1981**). "Accent scaling and language attitudes: Reactions to Mexican American speech," Language Speech **24**, 207–221.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57

Council of Europe 2001 (**2001**). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment* (Cambridge University Press, New York), pp. 1–258.
Cucchiarini, C., Driesen, J., Van hamme, H., and Sanders, E. (**2008**). "Recording speech of children, non-natives and elderly people for hlt applications: The jasmin-cgn corpus," in *Proceedings of LREC 2008*.
Cucchiarini, C., Neri, A., and Strik, H. (**2009**). "Oral proficiency training in Dutch l2: The contribution of asr-based corrective feedback," Speech Commun. **51**, 853–863.
Demuynck, K., Roelens, J., Compernolle, D. V., and Wambacq, P. (**2008**). "Spraak: An open source speech recognition and automatic annotation kit," in *Proceedings of ICSLP*, pp. 495–498.
Erdener, D., and Burnham, D. (**2005**). "The role of audiovisual speech and orthographic information in nonnative speech production," Lang. Learn. **55**, 191–228.
Escudero, P., and Boersma, P. (**2004**). "Bridging the gap between l2 speech perception research and phonological theory," Stud. Second Lang. Acquis. **26**, 551–585.
Flege, J. (**1987**). "A critical period for learning to pronounce foreign languages?," Appl. Ling. **8**, 162–177.
Flege, J. (**1995**). "Second language speech learning: Theory, findings and problems," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (Timonium: York Press, Baltimore, MD), pp. 233–273.
Flege, J. (**1999**). "Age of learning and-second-language speech," in *Second Language Acquisition and the Critical Period Hypothesis*, edited by D. Birdsong (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 101–132.
Franco, H., Neumeyer, L., Digalakis, V., and Ronen, O. (**2000**). "Combination of machine scores for automatic grading of pronunciation quality," Speech Commun. **30**, 121–130.
Goudbeek, M., Cutler, A., and Smits, R. (**2008**). "Supervised and unsupervised learning of multidimensionally varying non-native speech categories," Speech Commun. **50**, 109–125.
Gussenhoven, C. (**1999**). "Dutch," in *Handbook of the International Phonetic Association, Part II, Illustrations of the IPA* (Cambridge University Press, Cambridge, UK), pp. 74–77.
Hershey, J. R., and Olsen, P. A., (**2007**). "Approximating the Kullback–Leibler divergence between Gaussian mixture models," in *Proceedings of ICASSP*, pp. 317–320.
Iverson, P., and Evans, G. (**2007**). "Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration," J. Acoust. Soc. Am. **122**, 2842–2854.
Iverson, P., Kuhl, P., Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (**2003**). "A perceptual interference account of acquisition difficulties for non-native phonemes," Cognition **87**, 47–57.
Kanters, S., Cucchiarini, C., and Strik, H. (**2009**). "The goodness of pronunciation algorithm: A detailed performance study," in *Proceedings of SLaTE 2009*, Birmingham, Alabama, pp. 49–52.
Kuhl, P., and Iverson, P. (**1995**). "Linguistic experience and the perceptual magnet effect," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Timonium, MD), pp. 121–154.
Kuhl, P., and Mellzoff, A. (**1996**). "Infant vocalizations in response to speech: Vocal imitation and developmental change," J. Acoust. Soc. Am. **100**, 2425–2438.
Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B. (**1992**). "Linguistic experience alters phonetic perception in infants by 6 months of age," Science **255**, 606–608.
le Cessie, S., and van Houwelingen, J. C. (**1992**). "Ridge estimators in logistic regression," Appl. Stat. **41**, 191–201.
Lippi-Green, R. (**1997**). *English with an Accent: Language, Ideology, and Discrimination in the United States* (Routledge, New York), pp. 1–304.
Long, M. (**1990**). "Maturational constraints on language development," Stud. Second Lang. Acquis. **12**, 251–285.
MacKay, I., Flege, J., Piske, J., and Schirru, C. (**2001**). "Category restructuring during second-language (l2) speech acquisition," J. Acoust. Soc. Am. **110**, 516–528.
Neri, A., Cucchiarini, C., and Strik, H. (**2006**). "Selecting segmental errors in l2 Dutch for optimal pronunciation training," Int. Rev. Appl. Ling. **44**, 357–404.
Oostdijk, N. (**2002**). "The design of the spoken Dutch corpus," in *New Frontiers of Corpus Research*, edited by P. Peters, P. Collins, and A. Smith (Rodopi, Amsterdam, Netherlands), pp. 105–112.
Pols, L., Tromp, H., and Plomp, R. (**1973**). "Frequency analysis of Dutch vowels from 50 male speakers," J. Acoust. Soc. Am. **53**, 1093–1101.
R Development Core Team (**2010**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), pp. 1–3409.
Strik, H., Truong, K., de Wet, F., and Cucchiarini, C. (**2009**). "Comparing different approaches for automatic pronunciation error detection," Speech Commun. **51**, 845–852.
Van der Harst, S. (**2011**). *The Vowel Space Paradox* (Netherlands Graduate School of Linguistics/Landelijke, Utrecht, Netherlands), pp. 1–380.
van Doremalen, J., Cucchiarini, C., and Strik, H. (**2009**). "Automatic detection of vowel pronunciation errors using multiple information sources," in *Proceedings of ASRU*, Merano, Italy, pp. 580–585.
van Doremalen, J., Cucchiarini, C., and Strik, H. (**2010**). "Phoneme errors in read and spontaneous speech: Relevance for capt system development," in *Proceedings of SLaTE 2010*, Tokyo, Japan.
Wei, S., Hu, G., Hu, Y., and Wang, R.-H. (**2009**). "A new method for mispronunciation detection using support vector machine based on pronunciation space models," Speech Commun. **10**, 896–905.
Wells, J. (**1997**). "Sampa computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, edited by D. Gibbon, R. Moore, and R. Winski (Mouton de Gruyter, New York), pp. 684–732.
Witt, S. (**1999**). "Use of speech recognition in computer assisted language learning," Ph.D. thesis, University of Cambridge, Cambridge, UK.
Witt, S., and Young, S. (**2000**). "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Commun. **30**, 95–108.
Witten, I., and Frank, E. (**2005**). *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, Burlington, MA), pp. 1–629.
Yoon, S.-Y., Hasegawa-Johnson, M., and Sproat, R. (**2010**). "Landmark-based automated pronunciation error detection," in *Proceedings of Interspeech 2010*, pp. 614–617.
Young-Scholten, M. (**2002**). "Orthographic input in l2 phonological development," in *An Integrated View of Language Development—Papers in Honour of Henning Wode*, edited by P. Burmeister, T. Piske, and A. Rohde (Wissenschaftlicher Verlag, Trier, Germany), pp. 263–279.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 131.174.248.125 On: Tue, 07 Jan 2014 10:06:57