Using Non-Native Error Patterns to Improve Pronunciation Verification

Joost van Doremalen, Catia Cucchiarini, Helmer Strik

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands

{j.vandoremalen, c.cucchiarini, h.strik}@let.ru.nl

Abstract

In this paper we show how a pronunciation quality measure can be improved by making use of information on frequent pronunciation errors made by non-native speakers. We propose a new measure, called weighted Goodness of Pronunciation (wGOP), and compare it to the much used GOP measure. We applied this measure to the task of discriminating correctly from incorrectly realized Dutch vowels produced by non-native speakers and observed a substantial increase in performance when sufficient training material is available.

Index Terms: pronunciation error detection, computer-assisted language learning, confidence measures, weighted GOP

1. Introduction

Adult second language (L2) learners are known to experience difficulties in learning to pronounce the sounds of an L2 (see [1] for reviews). The majority of L2 learners never acquire native-like performance and many of them have problems even in attaining a level of comfortably intelligible speech. An important limiting factor in acquiring the pronunciation of an L2 is considered to be the phonology of the mother tongue (L1).

Theories that attempt to explain L1-L2 interference in speech perception and production are based on the tenet that the perceptual salience of phonetic detail becomes tied to the distinctions that are relevant in L1 [2] [3]. This form of L1 entrenchment leads to "deafness" to phonetic distinctions in the L2 and causes difficulties in learning to perceive and produce L2 speech sounds. However, the positive finding is that new distinctions in an L2 can be learned, but this requires intensive feedback [4] [5].

Since in general it is not possible to offer intensive feedback on pronunciation in L2 classrooms, there is growing interest for Computer Assisted Pronunciation Training systems that make use of automatic speech recognition to provide feedback on pronunciation. This is also the aim of our DISCO project [6]. An important requirement for such systems is that pronunciation errors are reliably detected. For this purpose various measures of pronunciation quality have been developed [7] [8]. Although in general acceptable levels of performance can be achieved with these measures, it is our impression that better performance could be achieved by using pronunciation quality measures that take more account of the specific pronunciation errors that are made in the L2. More specifically, the research reported on in this paper evaluates a newly developed pronunciation quality measure on a set of Dutch vowels spoken by L2 learners.

In this paper we first provide a brief overview of the most used pronunciation quality measures and try to explain how such measures could be made more sensitive to error patterns (Section 2). We then go on to describe the case of vowel pronunciation error detection in Dutch (Section 3). In the following sections we report on experiments in which the performance of our new measure is compared to the much used GOP measure introduced in [7].

2. Pronunciation Quality Measures

Most pronunciation quality measures are segmental confidence measures. These confidence measures try to estimate the posterior probability of a phone:

$$\mathrm{P}(p|O) = \frac{\mathrm{P}(O|p)\mathrm{P}(p)}{\mathrm{P}(O)} \quad (1)$$

where p is the target phoneme and O the observation matrix. If this confidence measure is below a certain predefined threshold the phone is flagged as incorrectly realized. One well known instantiation of this notion is the Goodness of Pronunciation (GOP) algorithm [7] in which conditional probabilities are calculated using Hidden Markov Models (HMM) trained on native speech material.

In the applications of this algorithm an equal prior distribution is often assumed and the denominator $\mathrm{P}(O)$ is approximated by calculating the likelihood of the most likely phone sequence in the specific segment. In addition, transforming to a log scale and normalizing by phone duration dur yields [7]:

$$GOP(p) = \frac{\log\{\mathrm{P}(O|p)\} - \max_i \log\{\mathrm{P}(O|p_i)\}}{dur} \quad (2)$$

The decision of accepting or rejecting the phone as a correct pronunciation of the target phoneme is made by simple thresholding, which is determined separately for each target phoneme. This threshold can be calibrated on real non-native speech material or native material in which artificial errors have been introduced [9].

In [8] the posterior probability is estimated by:

$$\mathrm{P}(p|O) = \frac{\mathrm{P}(O|p)\mathrm{P}(p)}{\sum_i^N \mathrm{P}(O|p_i)\mathrm{P}(p_i)} \quad (3)$$

where the summation in the denominator runs over all N phonemes. The priors $\mathrm{P}(p)$ and $\mathrm{P}(p_i)$ represent the prior probability of the specific phoneme estimated from native speech material. Other approaches to pronunciation verification involve discriminative training methods such as Support Vector Machines [10] in which the posterior probability is estimated directly.

The research presented in this paper is grounded in the generative modeling approaches taken in [7] and [8]. We have found that the GOP scoring algorithm has difficulties in detecting errors in target phonemes with multiple acoustically close "neighbouring" phonemes. This is specifically the case in the Dutch vowel system, as explained in more detail in the next section. These difficulties are mainly caused by the fact that the denominator in Eq. 2 only takes into account the maximum

likelihood phone sequence, which might be an underestimation if there is more than one competing phoneme. In Eq. 3 this problem does not arise, but we think that weighting the likelihoods of competing phonemes $\mathrm{P}(O|p_i)$ based on how important they are for predicting an error in the target phoneme might improve this measure.

Therefore, we propose to combine multiple likelihood ratios from the target phoneme with all competitor phonemes in a logistic regression model. This regression model is trained on manually annotated non-native speech material. The measure, which we call *weighted GOP* (wGOP) is explained in more detail in Section 5.2.

3. Dutch Vowel System

The Dutch vowel inventory is relatively complex: it contains thirteen monophthongs, three diphthongs, and some additional vowels found mainly in loan words [11] [12] (see Fig. 1 for a vowel chart, SAMPA [13] is used in the current paper). In addition, there are relatively many vowels in the mid-to-high, front-central area of the vowel space:

- /I/ (as in /bIt/,"bid"; "pray")
- /Y/ (as in /pYt/, "put"; "well")
- /y/ (as in /byr/, "buur"; "neighbour")
- /2:/ (as in /l2:k/, "leuk"; "nice")
- /e:/ (as in /be:t/, "beet"; "bite")

Research has shown that in the case of Dutch, vowels pose particular problems to L2 learners [14]. The difficulties experienced by Dutch L2 learners in perceiving Dutch vowels do indeed appear to be connected to the relationship between the Dutch vowel system and that of their mother tongue [5], in the sense that L2 learners find it difficult to distinguish vowels that differ along dimensions that are not relevant in their mother tongue. New distinctions can however be learned if intensive feedback is provided [5].

With respect to production there is a compounding problem, because acoustic similarity is not the only influencing factor, orthography also plays a role in the sense that the orthography of the mother tongue is going to interfere with the way Dutch vowels are pronounced [14]. Moreover, in Dutch orthography the same grapheme is sometimes used to indicate two different phonemes, which might cause extra confusions.

Automatic classification of Dutch vowels produced by non-natives turned out be less successful than classification of vowels produced by native speakers [15]. Because of its characteristics — relatively many vowels with concentrations in a specific area of the vowel space — the Dutch vowel system is particularly suited to test the effectiveness of our newly developed pronunciation quality measure.

4. Material

The non-native speech material for the present experiments was taken from the JASMIN speech corpus [16].This material was recorded from speakers of many different mother tongues with relatively low proficiency levels, namely A1, A2 and B1 of the Common European Framework (CEF). For the experiments reported on in this paper we used the read speech material.

The material is obtained from 45 speakers reading the same set of phonetically rich sentences. In total there are 3669 chunks with a duration ranging from 5 to 15 seconds. 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.

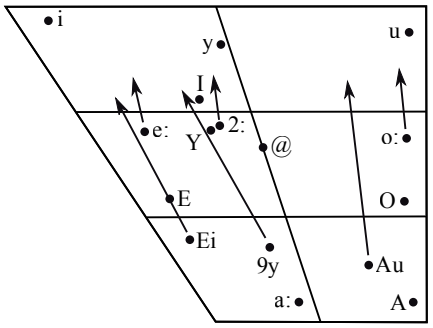

Figure 1: Dutch vowel chart

κ intra T_1	0.975
κ intra T_2	0.948
κ inter $T_1 - T_2$	0.913

Table 1: Transcription correction statistics

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 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 the inter and intra transcriber agreement, respectively. These agreement scores are shown in Table 1. 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.

5. Method

5.1. Phonetic Time Alignment

Firstly, an alignment between a canonical phonetic transcription using the CGN pronunciation lexicon [17] and the speech signal was created. This canonical transcription represents how the words should have been pronounced in Standard Dutch. Secondly, an alignment between the manually corrected phonetic transcription and the speech signal was created. The manually corrected transcription represents how the words have been realized.

The alignments were created by doing a Viterbi alignment with acoustic models trained using the SPRAAK package [18]. 47 3-state monophone Gaussian Mixture Models (GMM) were trained with native read speech material from the CGN speech database. For preprocessing purposes the input speech, sampled at 16kHz, is first divided into overlapping 32ms Hamming windows with a 10ms shift and pre-emphasis factor of 0.95. 12 Mel-frequency cepstral coefficients (MFCCs) plus C0, and their first and second order derivatives were calculated and cepstral mean subtraction (CMS) was applied.

The quality of these segmentations was checked semi-automatically. We observed that word-internal disfluencies

caused problems in the segmentation. These chunks could be detected relatively easily by spotting extremely long segments at the end of a chunk that were labelled as silence and that had low average acoustic likelihoods. We cleaned up the material by removing the 948 chunks that met these criteria.

In order to determine whether a certain vowel in the canonical transcription was correctly realized we checked whether more than 50% of the segment duration, as established in the canonical segmentation, contained the same vowel in the segmentation created from the manually corrected transcription. If this was not the case, then the vowel was flagged as incorrectly pronounced. Note that in this way, problems in the segmentation could lead to virtual pronunciation errors, which was the main reason to delete problematic chunks.

5.2. Likelihood Ratio Calculation

For the calculation of likelihood ratios for all vowel segments in the canonical transcription, we used the same monophone acoustic models with which we performed the Viterbi alignment. We calculated these likelihood ratios as:

$$\forall v \in \mathbf{V} : LLR_v^{v_t} = \frac{\log\{\mathrm{P}(O|v_t)\} - \log\{\mathrm{P}(O|v)\}}{dur} \quad (4)$$

where O is the observation matrix, v_t the target vowel sound and $\mathbf{V}$ the set of Dutch vowel phonemes. We will call these likelihood ratios $LLR_v^{v_t}$ *vowel scores*. The likelihoods of "competing" vowel sounds $\mathrm{P}(O|v)$ are simplified by following the same state level segmentation as the Viterbi path that was calculated for the target phone. That is, the competing vowels v switch states at the same times as the target vowel v_t.

Following Eq. 2, we also calculated the GOP measure, which we will denote with $LLR_{max}^{v_t}$. To calculate the likelihood of the optimal phone sequence in the segment we used an unconstrained free phone recognizer.

5.3. Model Training and Evaluation

Our baseline pronunciation verification system utilizes only the GOP score $LLR_{max}^{v_t}$. Our new measure, wGOP, combines the idividual vowel scores in a logistic regression model:

$$wGOP(v_t) = \frac{1}{1 + \exp\{-(\beta_0 + \sum_v \beta_v LLR_v^{v_t})\}} \quad (5)$$

These models are trained for each vowel phoneme separately. In these models, the dichotomous dependent variable, which represents whether the target phone was correctly or incorrectly pronounced, is predicted by the variables denoted as $LLR_v^{v_t}$, i.e. the vowel scores. To train a specific vowel model, we first extracted the segments for which this vowel appeared in the canonical transcription as a target phone. The number of segments per phoneme is shown in Table 2, together with the percentage of pronunciation errors. We also investigated whether adding the GOP score in the regression model as a predictor increased performance.

We trained and tested the models using leave-one-speaker-out cross-validation within the WEKA package [19]. That is, the β_v coefficients are first optimized using all segments of the first 44 speakers and afterwards tested on the segments of the remaining speaker. This is repeated until all segments are tested. The coefficients indicate to what extent the likelihood of a certain competing vowel is important in predicting whether the realized phone was correctly or incorrectly pronounced.

We evaluated the GOP score, the wGOP score and their combination using the equal error rate (EER), which is the point on the error curve where the false acceptance rate is equal to the false rejection rate.

phoneme	#inst	%errors	GOP	wGOP		Comb	
2:	235	44.68	32.34	24.69	+	25.55	+
9y	397	43.83	19.92	15.61	+	14.58	+
Ei	1204	40.78	26.42	22.60	+	22.18	+
Y	738	35.10	24.38	20.86	+	20.16	+
o:	1619	34.03	41.66	32.03	+	31.57	+
e:	1757	31.30	23.62	23.14	+	20.59	+
y	361	29.36	26.35	24.61	+	24.42	+
I	1715	29.16	29.13	24.42	+	21.40	+
A	2730	27.77	31.16	28.55	+	28.02	+
E	1695	17.05	25.57	24.45	+	22.91	+
i:	1637	16.56	23.70	24.29	-	23.26	+
a:	2131	10.00	29.65	22.35	+	22.58	+
Au	404	7.67	32.35	45.10	-	44.97	-
u	563	6.75	23.56	44.84	-	42.29	-
O	1426	4.98	33.13	44.53	-	34.10	-

Table 2: Overall results of the GOP measure and the weighted GOP score. Column descriptions: (1) Target phone, (2) Number of instances, (3) Percentage of incorrectly realized phones, (4) EER using GOP, (5) EER using wGOP, (6) Sign of EER difference between GOP and wGOP (7) EER using the combiation of GOP and wGOP (8) Sign of EER difference between GOP and the combination of GOP and wGOP.

6. Results

The EERs for each vowel are shown in Table 2. This list is ordered by the percentage of pronunciation errors per vowel. For the vowels for which the EER of wGOP measure is lower than the EER of the GOP measure, the improvement is 4.26% on average. This is not the case for /Au/, /u/, /O/ and /i:/, target vowels with very low percentages of pronunciation errors. Because the number of pronunciation errors for these phonemes is low, apparently no reliable regression models could be trained and the resulting EERs are therefore higher than those obtained using only the GOP measure. For vowels with many pronunciation errors and sufficient training material, our new method yields a substantial increase in performance. Combining the two methods is only beneficial for some vowels, most notably /e:/ and /I/.

To gain insight into these overall results, we investigated whether the phones that have been incorrectly realized were correctly rejected by the GOP and wGOP methods (Table 3). We did this for the three vowels with the highest percentage of pronunciation errors, /2:/, /9y/ and /Ei/, and for their three largest confusions. Here we see that the phones which are most often confused with the three targets — /y/ with /2:/, /Au/ with /9y/ and /a:/ with /Ei/ — benefit most from our new measure. Presumably this is caused by the weighting of the vowel scores based on frequent confusion patterns.

7. Discussion and Conclusions

From the results of the experiments we carried out we can conclude that tuning the wGOP measure with enough real non-native speech data considerably improves its discriminative ability compared to the GOP measure. An important concern in using this tuning data is the issue of generalizability to other speakers and tasks.

We trained the models speaker-independently, and the speakers in our material have widely varying L1s such as Turkish, Arabic, Spanish, Chinese, Persian, Hebrew, English etc.

target	realized	%of errors	GOP	wGOP
2:	y	32.38	18.10	23.81
2:	@	20.00	13.33	13.33
2:	Y	18.10	8.57	10.48
9y	Au	69.54	59.77	65.52
9y	a:	7.47	3.45	2.87
9y	A	5.75	4.02	4.59
Ei	a:	58.05	44.60	50.30
Ei	j	18.53	11.61	12.63
Ei	e:	8.35	5.91	3.87

Table 3: Distribution of realized phones in the correct rejects for the target phonemes /2:/, /9y/ and /Ei/. Column descriptions: (1) Target phoneme, (2) Realized phoneme, (3) Percentage of the total number of incorrectly pronounced phones for the target phoneme, (4) %correct rejects (%CR) at EER using GOP, (5) %CR at EER using wGOP.

Although these languages have a different phonology, apparently there is some systematicity in the error patterns of these speakers, at least enough for our measure to profit from it. This means that some phonemic confusions are quite stable across speakers. This was also observed in [14], where a number of phonemic confusions were identified that were common to L2 learners with varying L1s. On the other hand, we think that our measure could be further improved by using data from specific L1s or clusters of typologically similar L1s. With enough data available, our measure could be fine-tuned to the specific confusions that occur within a (type of) L1-L2 pair.

Another important aspect is the kind of task the speakers have to perform. We used read speech data, where the users had to read sentences from a computer screen. As stated in Section 2, there are some obvious phonemic confusions due to interference with the orthography in this task, which are not likely to occur when speakers are not reading but have to repeat spoken utterances. As this might lead to different error patterns, it follows that the tuning data has to be appropriate for the task it is employed for.

In our experiments we have treated all pronunciation errors on equal par, which might not be a valid assumption in all cases. Consider for example the pronunciation errors of /2:/ as /@/ and /Y/ (Table 3), which might be considered as less serious than the error of pronouncing /2:/ as /y/. In such cases one could argue about how "false" the false accepts of the system are and how this differs between the different pronunciation errors. Whether or not we have to treat these and other error patterns differently is in essence a matter of pedagogy, but ideally the technology should be able to deal with these requirements.

One way to approach the latter problem would be in the calibration of the threshold. In this paper we have used the EER as a measure of discriminative ability, but pedagogically the EER threshold might not be the optimal threshold. We could however optimize the threshold in such a way that it minimizes the total cost of erroneous decisions. This total cost can be calculated by weighting the different types of errors in a pedagogically sound way. It is not straightforward how these costs should be quantified and more research on this topic is needed to investigate this issue.

In the future we plan to investigate in which ways our method could be improved. For some vowel sounds discussed in this paper, this would involve handling their context dependence. Others aspects that could lead to improvement might be the initial segmentation, on which all local confidence scoring heavily depends, and speaker adaptation of the HMM models. Also we would like to investigate how our method generalizes to other sounds, such as consonants.

8. Acknowledgements

We would like to thank Laura Graaf, Floor Jansen, Eline van Buuren and Marieke Oenema for correcting the automatic phonetic transcriptions. The DISCO project is carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://taalunieversum.org/taal/technologie/stevin/).

9. References

[1] Strange, W., "Speech Perception and Linguistic Experience: Issues in Cross-Language Research," New York Press, Baltimore, MD, pp. 171-206, 1995.

[2] Best, C.T., "A direct realist view of speech cross language speech perception," In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. New York Press, Baltimore, MD, pp. 171-206, 1995.

[3] Flege, J.E., "Second language speech learning: theory, findings and problems," In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues. York Press, Timonium, MD, pp. 233-273, 1995.

[4] Logan, J. Lively, S. and Pisoni, D., "Training Japanese listeners to identify English /r/ and /l/: a first report," Journal of the Acoustical Society of America, vol. 89, pp. 874-886, 1991.

[5] Goudbeek, M., Cutler, A. and Smits, R., "Supervised and unsupervised learning of multidimensionally varying non-native speech categories," Speech Communication, vol. 50, pp. 109-125, 2008.

[6] http://lands.let.ru.nl/ strik/research/DISCO.

[7] Witt, S., "Use of speech recognition in computer assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[8] Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O., "Combination ofmachine scores for automatic grading of pronunciation quality," Speech Communication, vol. 30, pp. 121-130, 2000.

[9] Kanters, S., Cucchiarini, C. and Strik, H., "The Goodness of Pronunciation algorithm: a detailed performance study", In Proceedings of SLaTE 2009, Birmingham, 2009.

[10] Yoon, S-Y., Hasegawa-Johnson, M. and Sproat, R., "Automated Pronunciation Scoring using Confidence Scoring and Landmark-based SVM", In Proceedings of Interspeech, Brighton, United Kingdom, 2009.

[11] Booij, G., "The Phonology of Dutch". Oxford, Clarendon Press, 1995.

[12] Gussenhoven, C., "Dutch", in Handbook of the International Phonetic Association, Part II, Illustrations of the IPA, pp. 74-77. Cambridge, Cambridge University Press, 1999.

[13] http://www.phon.ucl.ac.uk/home/sampa/dutch.htm

[14] Neri, A., Cucchiarini, C. and Strik, H., "Selecting segmental errors inL2 Dutch for optimal pronunciation training," International Review of Applied Linguistics, vol. 44, pp. 357-404, 2006.

[15] Truong, K., Neri, A., Cucchiarini, C. and Strik, H., "Automatic Pronunciation Error Detection: An Acoustic-Phonetic Approach," In Proceedings of InSTIL, Venice, Italy, 2004.

[16] 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.

[17] Oostdijk, N., "The design of the spoken dutch corpus," in New Frontiers of Corpus Research, Peters, P., Collins, P. and Smith, A. Eds. Rodopi, pp. 105-112, 2002.

[18] Demuynck, K., Roelens, J., Van Compernolle, D. and Wambacq, P., "SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit," In Proceedings of ICSLP, p.495, 2008.

[19] Witten, I. and Frank, E., "Data Mining: Practical machine learning tools and techniques," Morgan Kaufmann, 2005.