Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources

Joost van Doremalen, Catia Cucchiarini, Helmer Strik

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

Frequent pronunciation errors made by L2 learners of Dutch often concern vowel substitutions. To detect such pronunciation errors, ASR-based confidence measures (CMs) are generally used. In the current paper we compare and combine confidence measures with MFCCs and phonetic features. The results show that the best results are obtained by using MFCCs, then CMs, and finally phonetic features, and that substantial improvements can be obtained by combining different features.

I. Introduction

The application of Automatic Speech Recognition (ASR) technology to second language (L2) learning, and in particular to pronunciation training, has received growing attention in the last decade [1]. Our institute has been involved in applying ASR to pronunciation training in several projects, e.g. [30] [31][32][33]. The research presented here is carried out within the framework of the DISCO project, which is aimed at developing a prototype of an ASR-based CALL application that provides feedback on Dutch L2 pronunciation, morphology, and syntax [30].

The use of ASR technology is especially advantageous when it comes to identifying specific pronunciation errors and providing corrective feedback to the learners. L2 learners do indeed appear to have difficulties in identifying their own pronunciation errors [2]. This suggests that Computer Assisted Language Learning (CALL) programs that can provide automatic corrective feedback on pronunciation are preferred to systems that can only offer the opportunity of listening and repeating L2 speech without corrective feedback. In line with these requirements several studies have addressed pronunciation error detection through ASR [3][4][5][6]. The main challenge in these approaches is to develop algorithms that achieve sufficient accuracy in error detection so that the feedback provided to the learners is not misleading. In general, achieving sufficient detection accuracy is particularly challenging exactly for those sounds that are easily confused or mispronounced by L2 learners.

Pronunciation errors in a second language can derive from several sources. An important limiting factor in acquiring the pronunciation of an L2 is considered to be interference from the first language (L1) [7], which can affect L2 speech production both at the prosodic and at the segmental level. L2 learners may have difficulties with the different syllable structure of the language to be learned, its rhythm and temporal organization, its phonemic inventory and phonotactics. Here we will focus on segmental aspects. L2 learners might insert or delete speech sounds, realize L2 phonemes incorrectly or even use phonemes from their L1. In particular, L2 learners may find it difficult to realize certain phonetic contrasts, either because they do not exist in their L1, or because they do exist but are not phonologically distinctive. Consequently, when trying to pronounce L2 phonemes, L2 learners may end up producing L1 phonemes that are somewhat similar but not identical. In such cases relatively subtle acoustic distinctions may lead to phonemic substitutions. Identifying those errors is of course more difficult than identifying substitutions of sounds that are acoustically more different.

For these reasons, various studies in pronunciation error detection have focused on sets of L2 phonemes that are very similar in acoustic and articulatory terms, in attempts to find accurate methods of identifying the mispronounced sounds. In general, ASR-based confidence measures (CMs) like posterior probabilities or the Goodness of Pronunciation measure (GOP) are used for pronunciation error detection [8][3][4][5]. These CMs give an indication of how confident the recognizer is that a given target sound was pronounced: the lower the confidence, the higher the chance that another sound was pronounced. Such measures have the advantage that they can be obtained fairly easily with an ASR system and they can be calculated in similar ways for all speech sounds.

However, since segmental pronunciation errors tend to concern specific phonetic contrasts that pose special difficulties to L2 learners, a promising approach to pronunciation error detection might be one that uses phonetic information related to the problematic contrasts. Along these lines, [9] and [10] developed dedicated classifiers to identify pronunciation errors that appeared to be frequent in Dutch L2 and that concerned relatively subtle distinctions such as that between fricatives and plosives and that between long and short vowels.

In [9] it was shown that good classification results can be obtained by using phonetic features; more specifically, by using more general features for vowels (formants, pitch and duration), and very specific features for differentiating a plosive from a fricative. In [10] a comparison was made of different approaches for differentiating a plosive from a fricative. A method in which phonetic features were used together with LDA performed better than GOP. However, similar results were obtained for MFCCs in combination with an LDA.

So, on the one hand phonetic features seem to be promising for classification, but on the other simply using MFCCs also provides good results. Furthermore, the results for these two methods were better than those for GOP, for the specific cases that were studied. These interesting results led to a number of questions: how these different methods would perform on other sounds, and whether something could be gained by combining different measures. In the current study we tried to answer these questions.

The outline of the paper is as follows. In section 2 we explain the background of this research. In section 3 we describe the material used and the method adopted in our experiments. The results are presented in section 4 and discussed in section 5.

II. Research background

Considering that L2 pronunciation errors are often related to interference from the L1, it seems very advantageous to have CALL systems that are designed for specific combinations of L1 and L2, and that can address the errors you would expect for those specific combinations, for instance German, Italian, Chinese or Japanese students learning English [11][5][12], or Americans learning French [13]. In general, using such fixed combinations of languages also has considerable advantages from the point of view of ASR technology: speech recognition is facilitated and pronunciations errors are more easily predictable. However, the feasibility of such systems heavily depends on the number of students and the approach used in L2 classes.

In the Netherlands, it is common practice to have heterogeneous L1 groups of learners in Dutch L2 classes. For this reason, in our research on ASR-based pronunciation training for Dutch L2 [6][10][14] we have focused on pronunciation errors that can be made by any learner, regardless of his/her L1. Although it is known that pronunciation errors are likely to be affected by the L1, in our research we also found that, at least for Dutch, it is possible to identify a set of phonemes that are particularly problematic for many L2 learners with different mother tongues [14]. This research and observations by Dutch L2 teachers indicate that, in general, vowels are more problematic than consonants [14], which may partly be due to the relatively high number of vocalic phonemes in Dutch compared to other languages [15] [16]: Dutch has 13 monophthongs, 3 diphthongs and some additional vowels found mainly in loan words [17].

The vocalic pronunciation errors, which concern almost all vowels and very often the diphthongs, appear to be related to difficulties with actually pronouncing the sounds and to orthographic interference [14]. In particular, vocalic errors are concentrated on realising a number of contrasts that many L2 learners are not familiar within their L1s, such as /a/ versus /A/, /e/ versus /E/, /o/ versus /O/, /i/ versus /I/, /u/ versus /y/, /u/ versus /Y/ and /y/ versus /Y/ (SAMPA notation [34]). The problems in realising such contrasts are not only related to their absence in the learner's L1, but also to Dutch orthography, as sometimes the same grapheme is used to indicate two different phonemes. For instance in the words "bonen" (beans) and "bom" (bomb) the grapheme "o" stands for the phoneme /o/ in the first word and for /O/ in the second word, Similarly, in the words "buren" (neighbours) and "bussen" (buses) the grapheme "u" represents the phoneme /y/ in the first word and /Y/ in the second word

The vowels /a/, /e/, /o/, and /i/ are generally longer than their short counterparts /A/, /E/, /O/ and /I/, but the distinction between long and short vowels seems to be based more on phonological grounds than on phonetic ones [18]. /e/ and /o/ are longer than /E/, /O/ respectively, while the high vowels /i/, /y/ and /u/ are longer than /I/ and /Y/ only when they are followed by /r/ [18]. According to [19] the difference in length between the long and the short vowels only appears in prosodically strong positions, a strong syllable in a foot.

In addition, duration is not the only characteristic that distinguishes the long vowels from their short counterparts, as the spectral characteristics also vary [18][17]. The vowels /e/, /o/, /i/, /u/ and /y/ are higher than /E/, /O/, /I/, and /Y/, respectively. /y/ and /Y/ are more fronted than /u/ and /a/ is more fronted than /A/.

Since many languages do not have such a distinction between vowel pairs that are associated with one grapheme, but have different realisations such /a/ and /A/, /e/ and /E/, /o/ and /O/, /i/ and /I/, /u/ and /y/, /u/ and /Y/ and /y/ and /Y/, L2 learners tend to produce attempts at pronouncing either of the two vowels in a pair, for instance /a/ or /A/, that often fall in between. Depending on the amount of deviation from the target sound these attempts will be classified as either /A/ or /a/. Problems arise when the amount of deviation is such that an attempt at producing /A/ is perceived as /a/ or vice versa, because in such cases another word will be pronounced than the intended one, for instance /maan/ (moon), instead of /man/ (man).

Given the difficulties posed by the above-mentioned vocalic contrasts to Dutch L2 learners, we set out to investigate whether it is possible to develop specific measures that achieve high accuracy in identifying the resulting pronunciation errors.

III. Method

A. Material

The speech material for our experiments was taken from the Spoken Dutch Corpus (CGN), a large corpus of Dutch as spoken in the Netherlands and Flanders by adult native speakers. CGN contains about 9 million words and a great variety of speakers of different age, gender, and region of origin, recorded in various socio-situational settings [20].

The speech material was extracted from the Northern Dutch part of CGN, and stems from 4 different components of CGN: read speech, and different broadcast speech material components that can be subsumed under the label 'broadcast monologues'. The RS material was recorded from trained speakers who read aloud novels in a studio environment, while the BM fragments were produced by speakers who were accustomed to speaking in public. These components are among the most formal in CGN, and reflect well the types of

speech that will be encountered in the final application. We used the RS material as our training set and the BM material as our test set.

CGN is a corpus of native speech and as such it does not contain the pronunciation errors L2 learners usually make. Although there are databases of non-native speech, these were considered to be too small for the purpose of this research. Given that the vocalic errors we wanted to investigate in this study concern phonemic substitutions, these can be easily simulated by artificially introducing them in a native corpus. In previous research we have used this procedure [6][21] and have seen that it works properly, as long as the simulated errors reflect errors that are actually made by L2 learners.

Errors that are often made by L2 learners are substitutions of the phonemes mentioned in Table 1 (see, e.g., [13]). Based on this information on how Dutch phones are frequently mispronounced by L2 learners, the CGN material was manipulated in such a way that realistic L2 errors were introduced. For instance, in order to train and evaluate the classification of /a/, all occurrences of /A/ in the transcriptions were replaced by /a/; and analogously for the other vowels. For more details on the procedure and on results showing that the classifiers obtained in this way show similar performance for real errors in non-native speech the reader is referred to [21]. Frequencies of the vowels under investigation in our material are shown in Table II.

TABLE I
SUBSTITUTIONS OFTEN MADE BY L2 LEARNERS. EACH ROW CONTAINS PHONEMES THAT ARE OFTEN CONFUSED, TOGETHER WITH AN EXAMPLE OF A DUTCH WORD IN WHICH THEY APPEAR (AND AN ENGLISH TRANSLATION).

/a/ maan (moon), /A/ man (man)
/i/ liep (walked), /I/ lip (lip)
/e/ leeg (empty), /E/ leg (put)
/o/ boon (bean), /O/ bon (ticket)
/u/ boek (book), /y/ vuur (fire), /Y/ bus (bus)

B. Feature Calculation

First, segmentations of the material were obtained through forced alignment. The segmentations were subsequently used to calculate a number of features. Details on the calculation of these features are provided below.

1) ASR-based Features

As our baseline we employed the widely used segmental confidence measure (CM) introduced in [8] which is the average frame-based posterior probability (AFBPP) of a forced aligned phone given the acoustic observations.

The AFBPP of a phone *ph* is calculated as:

$$afbpp(ph) = \frac{1}{t_e - t_b + 1} \sum_{t=t_b}^{t_e} \log(p(s_t^i | x_t))$$

where $p(s^i_t|x_t)$ is the frame based posterior probability of the forced aligned state s^i at time t given the observation vector x_t. $p(s^i_t|x_t)$ is calculated as:

$$p(s_t^i | x_t) = \frac{p(x_t | s_t^i)\, p(s_t^i)}{\sum_j^N p(x_t | s_t^j)\, p(s_t^j)}$$

where the summation in the denominator ranges over all N states of all triphone models. We will refer to this confidence measure as CM_{seg}. The HMM models for the automatic phone alignment were trained with SPRAAK [22]. As training material we used the RS material from the CGN corpus.

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 C_0, and their first and second order derivatives were calculated and cepstral mean subtraction (CMS) was applied. 47 3-state Gaussian Mixture Models (GMM) were trained: 46 phones and 1 silence model. In total 11,660 triphones are created, using 32,738 Gaussians.

Apart from averaging the frame-based probability over the whole segment, we also averaged over the three consecutive hidden states to model vowel onset/offset dynamics, hereby obtaining three state-based confidence measures. To this set of three features will be referred to as CM_{state}.

2) MFCCs

The 13 MFCCs, and their first and second order derivatives (as described above), were included in our feature set. We extracted MFCC-based features at three points in time within the segment, i.e. the windows closest to 25%, 50% and 75% of the length of the vowel. This makes a total of 117 (3x3x13) features referred to as *MFCCs.*

3) Phonetic Features

Using PRAAT [23], the first three formants (F1,F2 and F3) and F2-F1 were measured at the same three points in time (25%, 50% and 75%). In addition to these 12 features the mean pitch (F0) and intensity of the segments were also calculated. Since these measures can show considerable variation between speakers we carried out a normalization at the speaker level. [24] compared different vowel normalization procedures, and the best results were obtained with Lobanov's Z-score transformation [25]. Therefore, we also applied Lobanov's Z-score transformation to our data. These 14 normalized features will be referred to here as Spectral. Apart from spectral measures, we also extracted the raw segment durations from the automatically generated segmentation.

The durations of the three hidden states were also included. Apart from the 4 raw durations, we also included durations normalized for the articulation rate in the utterance, making a total of 8 duration features, referred to as *Duration.*

C. Classification: training and evaluation

For classification, we utilised support vector machines (SVM) with a linear kernel function using the LibSVM package [26]. The reason for choosing a linear kernel was that it performed as well as several non-linear kernels, i.e. Radial Basis Function (RBF) and polynomial kernels, and requires

considerably less CPU time. For each vowel, a different classifier was trained after cost parameters had been optimised through 10-fold cross-validation on the training set.

TABLE II
FREQUENCIES OF VOWELS IN TRAINING AND TEST SET

Phone	Training set	Test set
/a/	7988	4193
/A/	11092	5895
/i/	5411	3328
/I/	6967	3848
/e/	6689	3867
/E/	8242	4195
/o/	5620	3100
/O/	6359	3586
/u/	2127	1078
/y/	957	574
/Y/	1600	824
Total	63052	34488

First, the individual performance of all feature sets was examined. Afterwards, feature sets were combined. We evaluated the performance of each classifier with the Equal Error Rate (EER) on the Receiver Operating Characteristic (ROC) curve. Furthermore, 95% confidence intervals were calculated to test whether differences between performance were significant.

IV. RESULTS

In Table III it is shown how the different feature sets perform (as EER) for the different vowels. On the whole, the results for MFCCs are somewhat better than those for the CMs. For /a/-/A/, /i/-/I/, /e/-/E/ and /y/ better results are obtained for MFCCs. The results for CM_{seg} and CM_{state} do not differ much: for /a/ and /o/ significantly better results are obtained with CM_{state}, for /A/ CM_{seg} performs significantly better. The phonetic feature sets *Spectral* and *Duration* alone achieve about 60-80% correct.

In Table IV the performance for combinations of different feature sets is shown. Significant performance gains can be obtained by adding *Duration* to MFCCs and CMs for /a/, /A/, /o/ and /O/. The combinations of *Spectral* and *Duration* perform equally or better than the two sets individually, but worse than the combination of MFCCs and *Duration*. Adding CMs to the latter combination helps to lower the error rate for almost all phones, except /I/. Differences between combinations with CM_{seg} or CM_{state} are not significant.

V. DISCUSSION

Within each subset of vowels, the results are based on the same tokens. For instance, for the /a/ classification results, all occurrences of /A/ in the transcriptions are replaced by /a/, and for the /A/ results it is just the other way around. Thus it may be surprising to see that the results for the long and the short vowels are not the same. The reason for this discrepancy is that for the /a/ classification the acoustic model for /a/ was used to obtain the automatic segmentations, while for the /A/ classification the same tokens were automatically segmented by using the acoustic model for /A/. This is also how it will be done in the application. Inspection of the segmentations indeed revealed that the begin and end times do vary. The smallest differences are observed for the /o/ vs. /O/ pair, while the largest ones pertain to the /u/ vs. /y/ and /Y/ distinction. This explains the large performance differences within the latter group.

TABLE III
EQUAL ERROR RATES FOR INDIVIDUAL FEATURE SETS: CMS_{EG}, CMS_{TATE}, MFCCS, SPECTRAL AND DURATION. ASTERIKS (*) INDICATE THE BEST PERFORMING FEATURE SETS.

Target	CM_{seg}	CM_{state}	*MFCCs*	*Spectral*	*Duration*
/a/	17.0	15.9	13.8*	29.8	19.6
/A/	22.9	24.7	14.1*	30.3	25.1
/i/	18.7	19.0	13.4*	24.4	30.3
/I/	22.9	22.2	13.9*	22.3	40.8
/e/	11.4	10.7	9.7*	17.7	17.7
/E/	13.3	13.6	9.6*	17.6	32.9
/o/	26.5	24.8*	25.4	38.1	26.7
/O/	24.7*	25.2	26.1	36.9	31.0
/u/	5.0*	5.1	7.5	23.4	18.7
/y/	11.9	12.8	11.8*	22.0	27.2
/Y/	14.6	14.4*	15.1	29.6	40.7
Overall	18.9	18.9	15.0*	26.8	27.7

Note that the results presented in the current paper concern difficult cases. For instance, if we had tried to classify vowels that are acoustically more different from each other (such as /i/, /a/, and /u/), results would probably have been better. However, the latter are not the kind of substitution errors that are frequently made by language learners. For a CALL application it is important to be able to detect the errors that are frequently made by language learners.

Therefore, we first studied what frequent errors are (see Table I) [14], and we tried to develop classifiers for these frequent errors. Here we present results for many tokens present in different components of a standard general purpose corpus (CGN), e.g. in 'relatively uncontrolled material', in which different factors may have a negative effect on the performance of our classifiers.

First of all, there is a training-test mismatch. For training read speech was used, while for testing we used broadcast speech: there is a mismatch in speech style, recording

channels, etc. Furthermore, we used all tokens without using a selection procedure (e.g. for context, place of words in the utterances, prosodic effects, etc.). In the final CALL application we have more control over many of these factors: we know who the speaker is (and adapt to that speaker in various ways), what the recording channel is, and we can choose the material (the stimuli and the prompts) ourselves in such a way that we can focus on those problematic sounds that can be reliably detected. Even within the speech of natives there will be a large variation in the realisation of (distinct) vowels, and it is known that realisations of distinct vowels often overlap. By providing feedback only on clear mispronunciations, we can minimise the number of times that a correct realisation of a phoneme is classified as a mispronunciation (false rejections).

TABLE IV
EQUAL ERROR RATES FOR COMBINED FEATURE SETS.

Target	*MFCC+ Duration*	*Spectral+ Duration*	*MFCC+ Duration+ CM_{seg}*	*MFCC+ Duration+ CM_{state}*
/a/	12.5	18.5	11.3	11.1
/A/	13.0	13.8	11.6	11.7
/i/	13.3	22.2	12.5	12.6
/I/	13.7	22.4	14.0	13.7
/e/	9.1	13.0	7.9	7.8
/E/	9.7	15.7	8.4	8.4
/o/	20.8	26.9	19.3	19.2
/O/	23.9	30.3	19.5	19.7
/u/	7.2	17.8	4.7	4.6
/y/	13.0	19.9	9.7	9.7
/Y/	14.9	22.4	9.9	9.8
Overall	13.9	19.6	12.3	12.3

For the /o/ vs. /O/ distinction classification performance turns out to be lower than for all other combinations. This may partly be explained by the higher acoustic similarity between /o/ and /O/ as compared to the other vowel sets that are studied here. Shown in Table V are average frequency values of the formants (F1 and F2) for 50 males in columns 2 and 3 (taken from [27] and for 16 female speakers in columns 4 and 5 (taken from [28]). Although there are differences in the values, which was expected because columns 2-3 concern males and columns 4-5 females, it is clear that the differences between /o/ and /O/ are smaller compared to those within the other vowel sets. In order to obtain better performance for /o/ vs. /O/ we might need to look in more detail to the (phonetic) differences between these vowels, for instance the fact that /o/ often shows a considerable degree of diphthongisation.

Table V also contains information on the average durations of phonemes: the average values in column 6 are taken from [29]. The differences between the average durations in column 6 of Table V reflect the performance of the classifiers using duration alone (see column 6 of Table III). For instance, the smallest difference in duration is observed for /i/ vs. /I/, and classification with duration alone also shows the highest error rates for these vowels. On the other extreme: the largest differences in duration are observed for /a/ vs. /A/, and the best (average) classification results are also found for this vowel pair. Our classification results are thus in line with the phonetic observations.

TABLE V
AVERAGE VALUES FOR THE PHONEMES IN CLUMN 1. COLUMNS 2-5 CONTAIN AVERGE FORMANT (F1 & F2) VALUES (IN HZ) TAKEN FROM [27] AND [28] RESP. COLUMN 6 CONTAINS AVERAGE VALUES FOR THE DURATION OF THE PHONEMES (IN MSEC.), TAKEN FROM [29].

Phon.	F1	F2	F1	F2	Dur
/a/	795	1301	948	1644	186
/A/	679	1051	859	1321	103
/i/	-	-	346	2401	105
/I/	388	2003	442	2452	91
/e/	407	2017	438	2443	176
/E/	583	1725	638	2123	107
/o/	487	911	525	1033	162
/O/	523	866	581	1079	99
/u/	339	810	400	893	111
/y/	305	1730	354	2070	140
/Y/	438	1498	482	1832	98

We are aware that there is a considerable overlap between feature sets. For instance, CMs, MFCCs, and *Spectral* are all spectrally based, and thus it is not surprising to observe that there are similarities in the results. Furthermore, there is a large variation in the number of features in the sets used here: 1 for CM_{seg}, 3 for CM_{state}, 117 for MFCCs, 14 for *Spectral*, and 4 for *Duration*. It is interesting to observe that a classifier based on 1 feature (CM_{seg}) performs almost as well as the one based on 117 MFCCs. The different feature sets have some pros and cons.

The advantage of the CMs compared to the MFCCs is that the number of features is much smaller. However, more important in the final application is probably the CPU time required. The fact that MFCCs and *Duration* are part of the standard ASR procedure, i.e. they do not require a large computational overhead, might thus be appealing.

On the other hand, phonetic features (*Spectral* and *Duration*) have the advantage that they can be more easily interpreted. If formant values and durations are too low or high, feedback based on these observations can be given to the learner (i.e. the position of your tongue is too high (if F1 is too low), or the vowel should be made shorter) and to the teacher (for monitoring the learner). Clearly, the latter can be very useful in a language learning application.

Although in the current paper we studied classifiers for mispronunciation detection of Dutch vowels, the methods used are generic and can easily be ported to other languages and other sounds. First, the results presented are relevant for other languages that contain sets of vocalic phonemes that are

very similar in acoustic and articulatory terms (i.e. English, German and Swedish) and as such pose problems to L2 learners. In addition, similar classifiers can be developed for different vowel combinations, but also for consonants, as we have done in our previous research [9,10]. The relative importance of some features will differ between languages. For instance, the importance of duration in the detection of vowels will be smaller in languages in which duration is not such an important factor in the vowel system, such as Italian and Spanish. However, in the latter two languages duration plays a more important role in recognizing consonants (compared to, e.g., Dutch). Classifiers thus have to be optimized for each language, but the procedures used to develop the classifiers can be very similar.

Furthermore, in the current research the classifiers are used to detect pronunciation errors made by language learners. In doing this we focused on substitutions often made by language learners, because this is most important for the application in our language learning project. Thus we trained and tested the classifiers for certain combinations of vowels, e.g. /a/ vs. /A/. However, it is also possible to optimize the classifiers for other purposes: other combinations of sounds, or – very general - to detect whether a given sound is indeed the intended sound (e.g. /a/ or not). The cases we studied here are very difficult ones, since the vowels we try to discern are acoustically very similar. For other sound combinations the task will generally will be easier and thus performance is likely to be higher.

VI. Acknowledgment

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

References

[1]M. Eskenazi, “An overview of Spoken Language Technology for Education,” *Speech Communication*, 2009.

[2]A. Dlaska and C. Krekeler, “Self-assessment of pronunciation,” *System*, 36, pp. 506-516, 2008.

[3]S.M. Witt and S.J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning”, *Speech Communication* 30, 95–108, 2000.

[4]G. Kawai, and K. Hirose “Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology,” *Speech Communication,* 30 (2), pp. 131-143.

[5]B. Mak, M. Siu, M. Ng, Y-C, Tam, Y.-C. Chan, K.-W., Chan, Kin-Wah Chan, K-Y. Leung, S. Ho, F-H. Chong, J. Wong, and J. Loet “PLASER: Pronunciation Learning via Automatic Speech Recognition,” in Proc. *HLT-NAACL 2003 Workshop on Building Educational Applications using Natural Language Processing*, Edmonton, Canada, 23-29.

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

[7]J. Flege, “Second-language speech learning: Findings and problems,” In *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*, Winifred Strange (ed.), Timonium, MD: York Press Inc, pp. 233–273, 1995.

[8]H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, “Combination of Machine Scores for Automatic Grading of Pronunciation Quality,” *Speech Communication*, 30:121-130, 2000.

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

[10]H. Strik, K. Truong, F. de Wet and C. Cucchiarini, “Comparing different approaches for automatic pronunciation error detection,” *Speech Communication*, 2009.

[11]W. Menzel, D. Herron, R. Morton, D. Pezzotta, P. Bonaventura, P. Howarth “Interactive pronunciation training” *ReCALL*, 13(1), pp. 67-78, 2001.

[12]Y. Tsubota, M. Dantsuji, and T. Kawahara, “An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors,” *ReCALL*, 16(1), pp. 173-188, 2004.

[13]Y. Kim, H. Franco, and L. Neumeyer, “Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction,” In Proc. *Eurospeech*, pp. 645-648, Vol. 2, Rhodes, Greece, 1997.

[14]A. Neri, C. Cucchiarini, and H. Strik “Selecting segmental errors in L2 Dutch for optimal pronunciation training”, *International Review of Applied Linguistics*, 44, 357-404, 2006.

[15]B. Lindblöm, Phonetic universals in vowel systems, in *Experimental Phonology*, John J. Ohala, and Jeri J. Jaeger (eds.), 13–44, Orlando, FL, Academic Press, 1986.

[16]I. Maddieson, *Patterns of Sounds*. Cambridge, Cambridge University Press, 1984.

[17]C. Gussenhoven, Dutch, in *Handbook of the International Phonetic Association*, Part II, Illustrations of the IPA, 74–77. Cambridge, Cambridge University Press, 1999.

[18]G. Booij, *The Phonology of Dutch*. Oxford, Clarendon Press, 1995.

[19]T. Rietveld, and V.J. Van Heuven, *Algemene Fonetiek*. Bussum: Coutinho, 2001.

[20]N.H.J. Oostdijk, “The Spoken Dutch Corpus. Outline and first evaluation”, in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece, Vol. 2: pp. 887-894, 2000.

[21]S. Kanters, C. Cucchiarini, and H. Strik, “The Goodness of Pronunciation Algorithm: a Detailed Performance Study,” in *Proceedings of SLATE*, 2009.

[22]K. Demuynck, J. Roelens, D. V. Compernolle, and P. Wambacq, “SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit,” in *Proceedings of ICSLP*, p. 495., 2008.

[23]P. Boersma and D. Weenink, *Praat: doing phonetics by computer (Version 5.1.10) [Computer program]*. Retrieved July 8, 2009, from http://www.praat.org/.

[24]P. Adank, *Vowel Normalization: a perceptual-acoustic study of Dutch vowels*. Doctoral dissertation, Radboud University Nijmegen, The Netherlands, 2003.

[25]Lobanov, B. M., “Classification of Russian vowels spoken by different speakers,” *Journal of the Acoustical Society of America*, Vol. 49, Issue 2B, pp. 606-608, 1971.

[26]C.-C. Chang, C.-J. Lin, “LIBSVM: a library for support vector machines,” 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[27]L.C.W. Pols, H.R.C. Tromp and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," J.Acoust.Soc.Am. 53, 1093-110, 1973.

[28]R. Van Hout, P. Adank, & V.J. van Heuven “Akoestische metingen van Nederlandse klinkers in algemeen Nederlands en in Zuid-Limburg.”, *Taal en Tongval*, 52, pp.151-162, 2000.

[29]H. Strik and E. Konst, "A duration model for phonetic units in isolated Dutch words", AFN-Proceedings, University of Nijmegen, Vol. 15, pp. 71-78, 1992.

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

[31]http://lands.let.ru.nl/~strik/research/ST-AAP.html

[32]http://lands.let.ru.nl/~strik/research/Dutch-CAPT/

[33]Repetitor: http://lands.let.ru.nl/literature/heuvel.2008.6.pdf

[34]J.C. Wells, “SAMPA computer readable phonetic alphabet”. In *Handbook of Standards and Resources for Spoken Language Systems,* Gibbon, D., Moore, R. and Winski, R. (ed.)*,* Berlin and New York: Mouton de Gruyter. Part IV, section B, 1997.