What does learner speech sound like? A case study on adult learners of isiXhosa.

Jaco Badenhorst, Alfred Tshoane & Febe de Wet
Human Language Technology Research Group, CSIR Meraka Institute, Pretoria, South Africa
Email: {JBadenhorst, ATshoane, fdwet}@csir.co.za

***Abstract*—This paper reports on an analysis of isiXhosa speech produced by adult language learners. The learners whose speech was recorded were all acquiring isiXhosa as an additional language and the majority of the students had beginner level oral proficiency skills. The speech samples were produced and recorded during the development of a Mobile Assisted Language Learning (MALL) application to support clinical communication skills training at Stellenbosch University's Faculty of Medicine and Health Sciences. The aim of the application was to provide a means for students to practise their oral skills and improve their pronunciation in isiXhosa. The speech data was processed manually as well as automatically and the results reveal that 30% of the recordings do not contain suitable audio. It was also found that, on average, absolute differences between first language speakers and additional language learners are not good indicators of proficiency. However, automatically derived proficiency measures for the majority of the learners improved during the course of a semester.**

I. Introduction

Clinical communication skills training is part of all the undergraduate programmes offered by the Faculty of Medicine and Health Sciences at Stellenbosch University. The Human Language Technology research group at the CSIR's Meraka Institue collaborated with the isiXhosa language tutors at the faculty on the development of a Mobile Assisted Language Learning (MALL) application that could supplement isiXhosa lectures and course material. The main aim of the application was to provide an opportunity for students to improve their oral proficiency in isiXhosa.

Lecture time for clinical communication skills training is limited and does not allow for extensive pronunciation training. In addition, many students are hesitant to speak isiXhosa (especially in a full classroom) because they find the pronunciation difficult. It was therefore decided to develop an application that students could use to practise their pronunciation in their own time and at their own pace.

Some language learning applications prompt students to record their own voice and compare their recording to a target pronunciation of the utterance. However, in a preparatory study on user preference, students indicated that they did not like listening to recordings of their own voices [1]. Moreover, by comparing the two recordings students cannot always tell where the errors in their own pronunciations are or what they should do to improve their pronunciation to sound more like the target speaker.

These drawbacks are addressed by incorporating Automatic Speech Recognition (ASR) into the language learning environment [2]. Phone recognition is used to analyse the properties of the student's speech. Some of the signal properties can be used to derive pronunciation scores and these scores, in turn, can be used to evaluate the speech and to provide feedback on pronunciation [3], [4]. Signal properties that are commonly used for this purpose include Rate of Speech (ROS) and the acoustic match between the utterance and previously trained models of the target speech [5]. This match is quantified in terms of a likelihood score, e.g. the Goodness of Pronunciation (GOP) score [6].

However, it is well-known that the performance of ASR systems trained on speech produced by first language (*L1*) speakers deteriorates when they are used by non-native (*L2*) speakers [7], [8], [9]. In addition, meaningful pronunciation scores can only be derived from utterances that contain speech and in which the speech can clearly be discerned from background noise. Recordings should therefore be pre-processed before pronunciation scores can be extracted.

During the development of the ASR-enabled MALL application, speech data from the target user group was collected. This paper reports on an analysis of the collected data. The aim of the analysis was to determine how many of the recordings are suitable for the derivation of meaningful pronunciation scores. In addition, ROS and GOP values were extracted from the recorded data and compared to scores derived from isiXhosa speech produced by L1 speakers.

II. Background

South Africa's official languages can all still be classified as under-resourced and, as a consequence, the development of language and speech technology is not at an advanced level. However, a number of resource development projects have been successful in providing speech data for technology development, e.g. [10], [11], [12], [13], [14]. While some of these corpora include examples of accented speech, none contain examples of learner speech. Examples of speech produced by learners therefore had to be elicited and recorded for the purposes of this study.

Experience has shown that performing basic quality checks during data collection can enhance the quality of the collected data substantially [15]. Although the collection of the isiXhosa learner speech was not an extensive resource development project, similar automatic quality measurements were used to identify recordings that did not contain suitable data.

978-1-5090-3334-5/16/$31.00

For many languages speech processing technology has advanced to a level where it can support various applications. One such an application is ASR-enabled Computer Assisted Language Learning (CALL) systems and, more recently, ASR-enabled MALL applications. Proficiency indicators derived from ASR output can be used as a means to measure different aspects of oral proficiency automatically [16], [17]. In Computer Assisted Pronunciation Training (CAPT) systems, automatically derived proficiency indicators are used to provide real time feedback on pronunciation [18], [19], [20], [21], [22]. The aim of the current study was to collect samples of learner data that could be used to develop an ASR-enhanced MALL application [23]. This paper reports on an analysis of the collected data and compares the pronunciation indicators that were derived from useable recordings with similar measures derived from native data.

The data sets that were used in this study are described in the next section. Section IV provides an overview of the pre-processing that was performed on the data to identify suitable utterances. Experiments and results are presented in Sections V and VI, followed by a discussion in Section VII.

III. Data

During the current investigation, L1 speech as well as L2 (learner) speech were used. Table I gives an overview of the different data sets.

A. First language isiXhosa data

The isiXhosa component of the NCHLT speech corpus was used as an example of L1 speech in this study [13]. This data set includes speech produced by 209 native speakers (balanced in terms of gender) of the isiXhosa language, with a total duration of just more than 56 hours. Associated transcriptions include 29 130 unique types and 136 904 tokens. A pre-defined test set that was released with the corpus includes 4 male and 4 female speakers. For the purposes of the current study a development set, also consisting of 4 male and 4 female speakers, was selected from the training data (Row 3 in Table I).

Not all the utterances in the NCHLT speech corpus are unique. The isiXhosa training set contains 40 873 utterances, of which 15 500 are unique. Only the subset of unique utterances were used for acoustic model development during the current investigation. (Row 2 in Table I)

B. Learner data

The L2 isiXhosa data was collected in two phases. During both phases the students were asked to read target utterances and their responses were captured using a data collection tool on a mobile telephone. The target utterances were derived from the lecture notes of the isiXhosa clinical communication skills module the students were enrolled for.

1) Tygerberg 2014: The first batch of data was collected from students who had already completed one semester of clinical communication skills training in isiXhosa. All the recordings were evaluated by a first language speaker of isiXhosa as either being an intelligible version of the target utterance, or not.

2) Tygerberg 2015: The second data collection phase coincided with a new group of students' first semester of isiXhosa clinical communication skills training. Each student was requested to read 15 target utterances (1) at the beginning of the semester, (2) after five weeks had passed and (3) during the last week of the semester. This data was recorded to identify changes (if any) in the acoustic properties of the speech produced by the students.

Data set	# Utterances	Duration (h:m:s)
NCHLT Train	40 873	49:22:31
NCHLT Train Unique	15 500	20:54:29
NCHLT Development	3 008	03:46:35
NCHLT Test	2 770	03:06:29
Tygerberg 2014	2 167	02:04:43
Tygerberg 2015	986	01:06:32

TABLE I
Number of utterances and duration of different data sets.

IV. Data pre-processing

A. Automatic pre-processing

Automatic measurements of data quality that were developed during a previous project were used to identify utterances that do not contain useable audio [15]. A Root-Mean-Square based duration value of less than 0.5 seconds and failed forced alignment were used to identify 20 empty files in the learner data (3 in Tygerberg 2014 and 17 in Tygerberg 2015).

B. Manual pre-processing

The speech samples that were collected during the second data collection phase (Tygerberg 2015) were manually labeled. A number of factors that could have a significant impact on the acoustic properties of the speech signals were identified. The eight event categories that each file was checked for are as follows:

- Empty: The audio file is empty.
- Device/Handling noise: Noise caused as a result of the device being moved during recording or by a sound/beep that results from the press of a button and an obstruction of the device microphone.
- Low volume: Speech is too soft to understand what is being said.
- Whispering: Speaker whispers during recording.
- Laughter: Speaker laughs while recording a prompt.
- Background noise: Any non-speech noise that occurs during the recording that is loud enough and may have an impact on recognition results.
- Background speech: Any speech that occurs in the background during the recording.
- Transcription mismatch: Speaker omits and/or inserts word(s) or does not read the entire prompt.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

V. Experimental set-up

A. Baseline ASR system

HTK-based acoustic models were trained using the unique sub-set of the NCHLT isiXhosa data (see Section III-A) [24]. Standard 3-state left-to-right HMMs were trained both with and without semi-tied transforms. Speaker level cepstral mean and variance normalisation (CMVN) were applied to the 13 mel cepstral coefficients that were derived from each speech frame. *All* the utterances in the NCHLT isiXhosa training set were used to derive the CMVN transform, not only the data in the unique sub-set. Delta and delta-delta features were also included in the feature vectors.

B. Adapted ASR system

Given that the L1 and L2 data was not collected under exactly the same acoustic conditions, some form of acoustic mismatch between the two data sets can be expected. To prevent channel and other effects from influencing the proficiency measurements, the models derived from the L1 data should be adapted using the L2 data.

Model adaptation was approximated using a feature transform. By re-estimating the CMVN feature transformations from a combination of NCHLT and Tygerberg data before training the acoustic models on the unique sub-set of the data, the channel differences are diminished [25]. The individual speakers in the two data sets were not taken into consideration for the derivation of the CMVN transform. In stead, the male and female data were clustered to simulate only two different speakers in the adaptation data.

C. Speech proficiency indicators

Two metrics that can be used as indicators of pronunciation were derived from the output of the ASR systems: Rate of Speech (ROS) a Goodness of Pronunciation (GOP) score. ROS was calculated as proposed in [16]:

$$ROS = \frac{N_p}{T_{sp}} \tag{1}$$

where N_p denotes the number of speech phones in an utterance and T_{sp} is the total duration of speech in the utterance, excluding pauses. GOP corresponds to the likelihood ratio defined in [6] as:

$$GOP(q_i) = \frac{|log(P(q_i|O))|}{NF(O)} \tag{2}$$

where $NF(O)$ corresponds to the number of frames in acoustic segment O. A GOP score was determined for each phone q_i in an utterance and utterance level scores were subsequently obtained by taking the average of all the phone scores in the utterance.

VI. Results

A. Phone recognition accuracy

Table II shows a summary of phone recognition accuracy [24] values for different experiments. The accuracy values give an indication of the acoustic match between the baseline and adapted models and the L1 Test and L2 Tygerberg data. The table shows the insertion penalty (IP) value for each experiment, and provides the recognition accuracy (ACC) values for models with and without semi-tied transforms applied. Insertion penalty values were optimised using the evaluation data described in Section III-A and the rows labelled "Dev: NCHLT" and "Dev: NCHLT (Norm)" show the corresponding recognition accuracy.

Evaluation Test	IP	ACC	IP	ACC (semi-tied)
Dev: NCHLT	-32	80.91	-27	80.56
NCHLT	-32	80.90	-27	81.39
Tygerberg	-32	12.79	-27	15.26
Dev: NCHLT (Norm)	-28	74.69	-27	74.31
NCHLT (Norm)	-28	75.81	-27	76.37
Tygerberg (Norm)	-28	50.80	-27	50.14
Tygerberg (2014)	-28	55.27	-27	54.22
Tygerberg (2015)	-28	38.92	-27	38.73

TABLE II
Phone recognition accuracy for different models and data sets.

The initial L1 acoustic model ("NCHLT") generated a best phone recognition accuracy of 81.39%, but using this model to evaluate L2 data performs poorly: the corresponding result for the Tygerberg data is only 15.26%.

It is possible to reduce feature mismatch by pooling all the L1 training and L2 evaluation data and computing CMVN vectors for two classes of gender (see Section V-B). Repeating the initial evaluation tests with these normalised features ("Norm") resulted in much better recognition performance (50.80%) of the L2 data. It is clear that not having speaker specific, but gender specific feature normalisation and adding the L2 data reduced the ability of the L1 models to recognise the L1 test data: a reduction of about 5% absolute for the "NCHLT (Norm)" tests can be observed in Table II.

The last two rows in Table II show the recognition accuracy of the first and second data collections separately. Much better recognition accuracy is achieved for the "Tygerberg 2014" data set than for the "Tygerberg 2015" data set.

Recognition accuracy for the 2015 data set is much lower than expected. The data was therefore inspected manually to determine the cause of the deterioration observed in the results (see Section IV-B). To select a set of "clean" utterances from the 2015 data set, the recognition accuracy corresponding to four selection options was determined. Each option consisted of a set of manual pre-processing categories. If any of these categories were marked for a particular utterance, the utterance was discarded. The event categories were combined as follows:

- *Option 1:* Empty, Whispering, Laughter, Background speech, Transcription mismatch
- *Option 2:* Empty, Low volume, Whispering, Laughter, Background speech, Transcription mismatch
- *Option 3:* Empty, Device/Handling noise, Low volume, Whispering, Laughter, Background speech, Transcription mismatch
- *Option 4:* Empty, Low volume, Whispering, Laughter, Background noise, Background speech, Transcription

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

mismatch

Table III lists the recognition accuracy values corresponding to the four options. The results in the table are ordered according to the number of utterances ("#Utts") that were not discarded.

Option	ACC	ACC (semi-tied)	#Utts
1	44.73	45.11	678
2	46.66	47.41	616
3	45.60	46.54	484
4	47.18	47.62	434

TABLE III
Phone recognition accuracy for manually pre-processed Tygerberg 2015 utterances for different combinations of manual pre-processing criteria.

Event categories such as Empty, Whispering, Laughter, Background speech and Transcription mismatch are expected to cause more severe recognition errors than the other categories. All utterances with noise events included in Option 1 were therefore removed to select a reasonably clean data set. For the remaining 678 utterances a recognition accuracy of 45.11% was measured. Also removing low volume utterances from this set (Option 2) improved recognition accuracy further by 2.20%, while a total of 616 utterances remained. The counts for Options 3 and 4 show that removing Device/Handling and Background noise resulted in much smaller data sets (484 and 434 utterances), with only 0.2% improvement in phone recognition accuracy for Option 4.

Given the results in Table III it was decided to conduct further experiments with the data set selected using Option 1. The recognition results are similar to those obtained for the Tygerberg 2014 data set and the biggest number of utterances are available for proficiency analysis. Furthermore, the phone recognition accuracy values in Tables II and III show that the results obtained for the models for which semi-tied transforms were applied are consistently better than the results for models without these transforms. Only results for models with semi-tied transforms are therefore reported in subsequent sections.

B. Global measurements

Two speech proficiency indicators, ROS and GOP, were used to conduct further analyses on the L2 data. Since both these measures are derived from recognition output, it is of value to compare their global values with the recognition experiments described in Section VI-A. Table IV shows the ROS and GOP values corresponding to the semi-tied results for the test sets in Table II. The values in the table correspond to the mean ROS and GOP values (standard deviation in brackets) of the utterances in each test set and the number of utterances for each set (#Utts) is also indicated in the table.

Table IV shows that the mean ROS value for L1 speech (NCHLT) is 8.34 phones per second. Similarly, a low GOP estimate of 1.09 indicates good orthographic alignment of transcriptions and acoustic models for the L1 data. The mean GOP value (1.43) for the updated acoustic model (pooling both L1 and L2 speech data during feature transformation) is slightly higher than the corresponding value for the L1 only model, while ROS remains constant. The differences observed between the Tygerberg and Tygerberg (Norm) results clearly indicate that reducing feature mismatch has dramatic implications for L2 data. Much higher mean ROS and GOP values resulted from poor recognition of L2 data for the unmatched model. While ROS for the matched model decreased slightly, the mean GOP estimate for L2 data is almost twice (3.09) that of the L1 estimate (1.43).

Evaluation test	ROS	GOP	#Utts
NCHLT	8.34 (1.64)	1.09 (0.93)	2770
Tygerberg	11.54 (3.37)	6.89 (2.01)	3133
NCHLT (Norm)	8.34 (1.69)	1.43 (1.20)	2770
Tygerberg (Norm)	7.98 (1.82)	3.09 (1.81)	3133
Tygerberg (2014)	8.18 (1.75)	3.02 (1.70)	2255
Tygerberg (2015)	7.47 (1.88)	3.29 (2.04)	878
Tygerberg (Option 1)	7.46 (1.54)	2.83 (1.48)	615

TABLE IV
Mean ROS and GOP values for different data sets.

The last three rows in Table IV show mean ROS and GOP values for different subsets of the Tygerberg (Norm) evaluation test. The mean ROS for the 2014 L2 data is close to the corresponding L1 NCHLT value and the 2015 data set deviates most from the L1 estimates, as could be expected. A higher GOP of 3.29 confirms the poor recognition of the corresponding experiment in Table II. Surprisingly, this data set has a significantly lower mean ROS (7.47). The last entry in Table IV confirms this low ROS value and a decrease in mean GOP to a value of 2.83.

C. Comparing first and second language speakers

The mean ROS and GOP values presented in the previous section were used to determine how the acoustic properties of different data sets compare in terms of these measurements. While the general trends provide vital information with regard to channel mismatches, it is also important to keep in mind that these metrics are highly speaker specific. Tables V and VI present per speaker estimates of the L1 and L2 (Tygerberg 2014) speakers. The values in these tables clearly show speaker differences. In Table V, the values correspond to the eight speakers of the L1 test set, ordered according to the mean ROS estimate for each speaker. The mean GOP estimate (standard deviation in brackets) and the number of utterances per speaker are also shown in the table. Finally, the last two rows in the table correspond to the results obtained for two L1 speakers who participated in the 2014 data collection.

The values in Table V indicate that the ROS values for L1 speakers ranged from as low as 6 to as high as about 10 phones per second. GOP estimates seem to remain lower than a value of 1.50 for the NCHLT test data. Table VI shows that the ROS values corresponding to the Tygerberg 2014 L2 data are in the same range as the values that were observed for the L1 test data. However, the table also shows that the per speaker GOP estimate ranges between 2.50 and 4.50.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

Speaker	ROS	GOP	#Utts
504f	6.35 (1.39)	1.47 (0.89)	305
505m	7.57 (1.30)	0.42 (0.68)	312
502f	7.88 (1.41)	1.42 (0.99)	322
503f	8.08 (1.54)	1.35 (0.88)	365
507m	8.56 (1.18)	0.46 (0.58)	358
500m	8.98 (1.22)	1.43 (0.95)	365
506f	9.26 (1.43)	0.81 (0.74)	360
501m	9.50 (1.28)	1.31 (0.86)	383
027	8.08 (1.11)	1.90 (1.35)	128
005	10.41 (1.70)	3.45 (2.02)	164

TABLE V
ROS and GOP estimates for L1 speakers.

In terms of GOP, the value of 1.90 that was measured for one of the Tygerberg 2014 L1 speakers (027) was within the expected range. This is not the case for the second L1 speaker (mean GOP value of 3.45). Listening to a random sample of utterances generated by speaker 005 revealed fast speech where other pronunciation effects such as vowel deletion is apparent. Also, for a significant fraction of utterances the audio is cut off, resulting in transcription errors. These two observations explain the high mean GOP for this speaker.

Speaker	ROS	GOP	#utts
018	6.21	3.07	169
015	6.26	2.65	144
044	6.95	3.21	138
010	7.40	3.64	31
011	7.49	3.91	30
028	8.01	2.63	73
012	8.08	2.45	166
033	8.10	4.57	139
031	8.20	2.07	217
004	8.34	3.65	167
034	8.51	4.68	94
022	8.44	2.73	278
017	8.54	2.91	30
008	8.97	2.71	136
038	9.58	2.85	73

TABLE VI
ROS and GOP estimates L2 speakers of the Tygerberg 2014 data.

1) Language training: A total of nine speakers in the Tygerberg 2015 data set recorded test prompts before and after a semester of clinical communication skills training (see Section III-B). Speech proficiency indicators derived from the two data sets were used to analyse the changes in the acoustic properties of these recordings. Mean ROS and GOP values for each speaker are shown in Tables VII and VIII respectively. For each indicator, the tables show the values corresponding to the beginning and the end of the semester as well as the number of utterances remaining after manual pre-processing according to Option 1 described in Section IV-B.

From Table VII it is clear that seven of the speakers initially speak significantly slower than after a semester of training. For one speaker (013) the average ROS decreases. Applying Option 1 criteria during data selection results in all the data associated with speaker 007 being discarded.

Speaker	ROS (before)	#Utts	ROS (after)	#Utts
041	5.32 (1.07)	4	7.62 (1.44)	13
045	6.29 (0.97)	13	7.70 (1.83)	9
040	6.34 (0.41)	15	7.51 (0.84)	14
014	6.39 (1.20)	16	8.86 (1.40)	17
003	6.93 (1.37)	13	7.52 (1.07)	12
016	7.02 (0.86)	15	8.51 (1.37)	13
042	7.59 (1.49)	8	8.19 (1.25)	14
013	7.19 (0.80)	10	6.65 (0.82)	10
007	-	-	-	0

TABLE VII
ROS estimates for eight speakers that recorded test utterances before and after a semester of clinical communication skills training.

Speaker	GOP (before)	#Utts	GOP (after)	#Utts
040	2.86 (1.26)	15	2.47 (1.23)	14
013	3.62 (1.77)	10	3.35 (1.43)	10
041	3.84 (2.10)	4	3.75 (1.35)	13
042	3.26 (1.94)	8	2.72 (1.29)	14
045	2.08 (0.76)	13	3.34 (1.88)	9
014	2.38 (1.02)	16	2.71 (1.07)	17
016	2.82 (1.16)	15	4.18 (2.06)	13
003	2.52 (0.81)	13	3.41 (2.16)	12
007	-	-	-	0

TABLE VIII
GOP estimates for eight speakers that recorded test utterances before and after a semester of clinical communication skills training.

Table VIII shows that even after removing the 263 utterances corresponding to Option 1 (see Table III), per speaker GOP estimates remain unstable. Half of the speakers obtain slightly lower GOP values after a complete semester of language training, while the GOP deteriorates for the other speakers.

VII. Discussion

The results presented in the previous section indicate that a combination of the speech proficiency indicators (ROS and GOP metrics) is sufficient to distinguish between L1 and different L2 speech proficiency levels under designed acoustic conditions. In general a significant offset between L1 and L2 speakers in terms of mean GOP estimates were established that quantify the pronunciation difference between L1 and L2 speech. Similarly the ROS metric was found to be instrumental to determine fast speech (affecting standard pronunciation) and to track the progress of L2 learners.

In order to make reliable measurements, properly adapted ASR models are particularly important. Acoustic mismatch has an adverse effect on ROS and GOP estimates. The values in Table IV for the "Tygerberg" evaluation test show just how much these values may increase artificially. Acoustic mismatch has to be minimised before L1 analysis of L2 data can be attempted.

A simple CMVN feature normalisation was found to sufficiently reduce the acoustic mismatch for the purposes of this paper, but more sophisticated adaptation methods are certainly possible. Further work on speaker adaptation methods could be key to develop L2 evaluation systems using limited data.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

The adapted ASR models also proved successful in finding reasonable ROS and GOP estimates for an L1 speaker within the Tygerberg 2014 data set, while inspection of a second L1 speaker confirms acoustic properties different from the NCHLT L1 data.

A key difference between Tygerberg 2014 and Tygerberg 2015 is that the first data collection was conducted under much more controlled circumstances than the second. For Tygerberg 2014 a technician was present during all the recording session. In 2015, on the other hand, students were given a recording device and asked to make recordings on their own. This difference in recording conditions clearly reflects in the low recognition accuracy (below 40%) for the Tygerberg (2015) test (Table II).

Manual pre-processing of the Tygerberg 2015 data revealed a list of eight event categories that could affect recognition accuracy and consequently, speech proficiency indicators. The removal of utterances pertaining to most of these categories result in recognition accuracies comparable to the Tygerberg 2014 data. This finding shows the importance of data pre-processing in real-world application of automatic evaluation systems. Future research should address these problems.

Learner recordings made before and after a whole semester of training suggest that, while learners become more proficient with regard to speaking rate, pronunciation does not seem to improve much during this time frame. The latter finding seems to hold for the larger Tygerberg 2014 data set, where global GOP estimates remained substantially higher than for L1 speakers.

VIII. Conclusion

The results of this study clearly indicate that the pre-processing of speech data is required before reliable indicators of pronunciation proficiency can be derived from the data. Different sources of acoustic variation should be taken into account, e.g. recording conditions, speaker proficiency in the target language, speaker idiosyncrasies, etc. The differences between the two sets of learner data show that data collected during a formal data collection campaign and during simulated usage are quite different and that a substantial portion of usage data (up to 30%) does not contain useable speech data. This restriction should be kept in mind during application design as well as data analysis.

References

[1] I. Wilken, "A Language Application for Health Science Students: a study on user experience," Master's thesis, Department of African Languages, Faculty of Humanities, University of Pretoria, South Africa, 2016.

[2] C. Cucchiarini, A. Neri, and H. Strik, "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863, 2009.

[3] H. Strik, K. P. Truong, F. De Wet, and C. Cucchiarini, "Comparing classifiers for pronunciation error detection." in *Interspeech*, 2007, pp. 1837–1840.

[4] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, vol. 6, 2012.

[5] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," *Speech Communication*, vol. 30, no. 2, pp. 109–119, 2000.

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

[7] D. Van Compernolle, "Recognizing speech of goats, wolves, sheep and non-natives," *Speech Communication*, vol. 35, no. 1, pp. 71–79, 2001.

[8] L. M. Tomokiyo, "Handling non-native speech in lvcsr: A preliminary study," in *Proceedings of the EUROCALL/CALICO/ISCA workshop on Integrating Speech Technology in (Language) Learning (InSTIL)*, 2000.

[9] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, p. 1, 2009.

[10] J. C. Roux, P. H. Louw, and T. Niesler, "The African Speech Technology project: An assessment." in *LREC*, 2004.

[11] J. Badenhorst, C. van Heerden, M. Davel, and E. Barnard, "Collecting and evaluating speech recognition corpora for 11 south african languages," *Language resources and evaluation*, vol. 45, no. 3, pp. 289–309, 2011.

[12] M. H. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of internet audio for resource-scarce ASR," in *Proc. INTERSPEECH*, 2011, pp. 3153–3156.

[13] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proccedings of the 4^{th} Workshop on Spoken Language Technologies for Under-resourced Languages*, St Peterburg, Russia, May 2014, pp. 194–200.

[14] F. de Wet, J. Badenhorst, and T. Modipa, "Developing speech resources from parliamentary data for south african english," *Procedia Computer Science*, vol. 81, pp. 45–52, 2016.

[15] J. Badenhorst, A. de Waal, and F. de Wet, "Quality measurements for mobile data collection in the developing world," in *Proccedings of the 3^{rd} Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, May 2012.

[16] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[17] F. de Wet, C. Van der Walt, and T. Niesler, "Automatic assessment of oral language proficiency and listening comprehension," *Speech Communication*, vol. 51, no. 10, pp. 864–874, 2009.

[18] N. Stenson, B. Downing, J. Smith, and K. Smith, "The effectiveness of computer-assisted pronunciation training," *Calico Journal*, pp. 5–19, 1992.

[19] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Computer assisted language learning*, vol. 15, no. 5, pp. 441–467, 2002.

[20] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[21] W. K. Lo, S. Zhang, and H. M. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system." in *INTERSPEECH*, 2010, pp. 765–768.

[22] B. Luo, "Evaluating a computer-assisted pronunciation training (capt) technique for efficient classroom instruction," *Computer Assisted Language Learning*, vol. 29, no. 3, pp. 451–476, 2016.

[23] M. Carranza, C. Cucchiarini, P. Burgos, and H. Strik, "Non-native speech corpora for the development of computer assisted pronunciation training systems," *Edulearn 2014 Proceedings*, pp. 7–9, 2014.

[24] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book. revised for HTK version 3.4," March 2009, http://htk.eng.cam.ac.uk//.

[25] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, p. 133147, August 1998.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE