The limitations of data perturbation for ASR of learner data in under-resourced languages

Jaco Badenhorst
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
jbadenhorst@csir.co.za

Febe de Wet
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
febe.dewet@gmail.com

***Abstract*—This paper reports on the recognition of second language (L2) isiXhosa speech produced by beginner level adult language learners. The speech samples were produced and recorded during the development of a Mobile Assisted Language Learning (MALL) application. The application aimed to provide a means for students to practise their oral skills and improve their pronunciation of isiXhosa. Automatically derived proficiency indicators can enhance MALL applications by enabling Computer Assisted Pronunciation Training (CAPT) and monitoring students' progress. However, the automatic recognition of low-proficient, non-native speech is a particularly challenging task, especially for under-resourced languages. Data augmentation strategies aim to increase the quantity of training data, improve model robustness and avoid overfitting. In this study we investigated whether directly adjusting the speed of raw audio signals (simulating additional training speakers) improved phone recognition accuracy for learner data. We present results for subspace Gaussian mixture models (SGMMs) and deep neural networks (DNNs) implemented using Kaldi. The under-resourced system's tendency to overfit on within-corpus test data is clearly illustrated and contrasted with cross-corpus results for non-native data. Compared to first language data, the speech rate of most language learners is considerably slower. Our results indicate that adjusting the speed of the learner data improves phone recognition accuracy.**

Index Terms: speech data perturbation, speech recognition for under-resourced languages, non-native speech recognition, low-proficient learner speech, phone recognition accuracy, speech rate analysis

I. Introduction

South Africa is a multi-lingual country. It's constitution recognises 11 official languages. In urban areas English is most often used as the *lingua franca*, while the other official languages tend to dominate in more rural areas. Professionals who receive their training in English or one of the other dominant languages and who start their careers in rural areas therefore often work in communities where people communicate in a language that they are not proficient in.

The most prominent example of this phenomenon are the medical professionals who receive their training in English or Afrikaans, but who then have to perform community service at remote hospitals and clinics where people are not fluent in either of the two languages. Language differences therefore pose a significant challenge to successful clinical communication. Many universities in the country are trying to address this situation by including at least one course in an additional language in medical curricula.

For instance, the undergraduate programmes offered by the Faculty of Medicine and Health Sciences at Stellenbosch University all include clinical communication modules in either Afrikaans or isiXhosa[1]. The data that was used in this study was collected during a project aimed at developing a Mobile Assisted Language Learning (MALL) application to supplement the isiXhosa course. The application was designed to provide an opportunity for students to practise their pronunciation in their own time and at their own pace. It was also foreseen that the application would be extended to incorporate Computer Assisted Pronunciation Training (CAPT).

The implementation of almost all measures that are used in CAPT rely on accurately aligned learner data which depends on the availability of speech recognition technology in the target language [1], [2]. IsiXhosa, like all the official languages of South Africa, is a severely under-resourced language [3]. Very little speech and text data is available to enable language technology development. The aim of this study was therefore to test the ability of new state-of-the-art modeling and data perturbation techniques to yield an automatic speech recognition system for isiXhosa that can align learner data accurately enough to enable the derivation of automatic proficiency indicators from speech data.

II. Background

Baseline isiXhosa automatic speech recognition (ASR) systems have been developed before, but have always been evaluated on test data that was collected in the same manner as the training data [3]. Initial experiments with learner speech (L2) indicated that the performance of a system based on Hidden Markov Models (HMMs) trained with first language (L1) isiXhosa data degraded substantially for out-of-corpus, L2 test data [4]. In addition to the phenomena that are typically associated with L2 speech, an analysis of the results showed that the acoustic differences between the training and test data accounted for much of the observed drop in performance. Applying cepstral mean and variance normalisation (CMVN) at speaker level did improve the results to some extent, but the phone recognition accuracy observed for the L2 data was still much lower than the corresponding values for the within-corpus L1 data.

In the current study the HMM-based ASR system was replaced with systems based on subspace Gaussian mixture (SGM) and Deep Neural Network (DNN) models. In addition, the possibility of using data perturbation to improve recognition accuracy for L1 as well as L2 data was explored. Data augmentation has been proposed as a means to improve recognition accuracy for large vocabulary ASR tasks [5]. It's potential to enhance ASR performance has also been illustrated for low-resourced languages [6], [7], [8], [9]. One of the aims of this study was to determine to what extent the reported improvements in system performance generalise to other data sets.

We did not implement any specific proficiency indicators or error detection methods. Instead, the focus was on the optimisation of phone recognition accuracy for low-proficient, adult learners whose speech is known to be difficult to process accurately using ASR [10]. This choice was motivated by the fact that most of the measures of

[1] These are the dominant languages in the Western Cape province where the university is located.

oral proficiency that can be derived automatically rely on an accurate alignment of speech.

III. Data

Table I gives an overview of the L1 and L2 data sets that were used in this study. The L1 data was selected from the isiXhosa component of the NCHLT speech database [3] and the L2 data was collected from students at Stellenbosch University's medical campus.

A. L1 data

First language isiXhosa data was selected from the NCHLT database of South African languages. The isiXhosa component of the database comprises around 56 hours of data and includes speech produced by 209 native speakers of the language. The data set is balanced in terms of gender and the associated transcriptions include 29 130 unique types and 136 904 tokens. A pre-defined test set was released with the NCHLT corpus. It includes 4 male and 4 female speakers. We used the same development set as in a previous study [4], also consisting of 4 male and 4 female speakers.

After the selection of the development set 40 873 utterances remained in the training data pool. The transcriptions of only 14 590 utterances within this set are unique types. The transcriptions of the unique set occur multiple times for different speakers.

Data set	# Utterances	Duration (h:m)
NCHLT Train	40 873	49:23
NCHLT Development	3 008	03:47
NCHLT Test	2 770	03:06
L2A	2 167	02:05
L2B	986	01:07

TABLE I: Number of utterances and duration of each data set

To enable a comparison between the effects of adding more data to the training set and data perturbation, we selected a subset of the data that included the unique types for which examples produced by at least three speakers appear in the corpus. As is shown in Table II, selecting three speakers for each unique type at random provided 12 927 recordings for the *Train_3* data set. *Train_3* contains recordings from 189 speakers, 92 males and 97 females, with 5 498 and 7 429 recordings respectively.

One of the three examples of each type was randomly chosen to compile a unique training set, *Train_1*. *Train_1* includes 4 309 utterances produced by 91 male and 96 female speakers.

B. L2 data

Two sets of learner data were collected. In each instance 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 course the students were enrolled for.

The first set of learner data (*L2A*) was collected from students who had completed the basic isiXhosa semester module. The second L2 data set (*L2B*) was collected during a new group of students' first semester of isiXhosa. These students represent adult beginners, similar to the user group described in [10]. Each student read 15 target utterances as part of three different simulated usage sessions during the course of the semester.

The L2A data set was collected under controlled conditions and a technical supervisor was present during all the recording sessions. The L2B set was compiled by means of simulated application usage. Although the students were still asked to read target phrases, they were free to move around and use the mobile devices without supervision. An analysis of the L2 data indicated that the L2B set contained many more empty recordings, background speech, reading errors, laughing, whispering, etc. [4]. The L2B set is probably a better representation of real world data, but also more difficult to process automatically.

C. Data perturbation

1) L1 data: Table II provides an overview of the data sets that were created by applying data perturbation to the L1 training data. To test the effect of data augmentation, the limited *Train_1* data set was expanded to the same number of utterances as *Train_3*. Perturbing the speed of each utterance using the `Sox`[2] utility, we simulated two new utterances for each unique type in *Train_1*, resulting in the *Perturbed_1* data set. The speed adjustment not only changes the duration of each utterance, but also the spectral frequencies, effectively simulating additional speakers [5]. We re-sampled the signal using the `Sox` *speed* function at speed factors 0.9 and 1.1 (90% and 110% of the original rate).

Data set	# Utterances	Duration (h:m)
Train_1	4 309	05:15
Perturbed_1	12 927	15:51
Train_3	12 927	15:35
Perturbed_3	38 781	47:36
Train_All	40 873	49:23
Perturbed_3_All	66 727	80.86

TABLE II: Number of utterances and duration of training sets with and without data augmentation

Applying the same procedure to *Train_3* resulted in six additional utterances per unique type which were used to create the *Perturbed_3* data set. *Train_All* corresponds to the complete NCHLT training set in Table I and lastly we created a *Perturbed_3_All* set by adding only the speed adjusted utterances of the *Perturbed_3* data set to *Train_All*.

2) L2 data: Low-proficient learner speakers tend to articulate at a slower rate than L1 speakers. Increasing the speed of these slower L2 utterances can simulate a faster L1 rate of speech (ROS) to some extent. `Sox` provides a *tempo* command to achieve this kind of perturbation, ensuring that the pitch and spectral envelope of the signal does not change [5]. We used it to adjust the tempo of slower utterances with speed factors ranging from 1.05 to 1.3 (105% to 130% original rate) in steps of 0.05.

IV. Experimental design

A. ASR system

We trained phone recognition systems using the open source Kaldi toolkit and followed a training recipe based on the Wall Street Journal and TIMIT example recipes [11]. In particular, we used a setup of position independent phones, converting the training transcriptions to a phone level representation so that each word label directly maps to a single monophone label before training commences. We created an ergodic phone loop by constructing a flat ARPA language model consisting of equiprobable 1-grams.

We used a standard front-end, applying a 25ms Hamming window with a 10ms shift between frames. The sample-frequency was set to 16KHz. Performing a linear prediction coefficient (LPC) analysis of default order 12, 13 cepstra were extracted (which includes C0). Mean and variance normalisation were applied per speaker for each data set (training as well as L1 and L2 test sets). Delta and double delta coefficients were added.

[2] http://sox.sourceforge.net/

The training features were used to estimate 3-state left to right HMM triphone models. Input alignments for SGMM training were derived from triphone models after incorporating linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT) training and speaker adaptive training (SAT).

The Kaldi *nnet2* setup was used to train DNN-HMM hybrid models. As introduced in [12], the DNNs were trained using the p-norm generalised maxout unit. Standard parameters were kept as is. The parameter p was always set to two, training four hidden layers. The initial and final learning rates were kept at the default values of 0.02 and 0.004, respectively.

B. Measuring recognition accuracy

Word error rate (WER) is a typical measure of recognition performance given the one-best ASR hypothesis and human transcriptions. Kaldi estimates this metric as a minimum edit distance [13] between word labels in the ASR output and reference transcription. Three edit operations, *substitution*, *deletion* and *insertion* transform one set of labels into the other.

To compare our results with previous work performed using the HTK toolkit [14] we calculated phone recognition accuracy as:

$$ACC \quad = \quad \frac{H-I}{N} \times 100\% \tag{1}$$

where $H = N - D - S$ refers to the number of correct labels, S to the number of substitution, D deletion and I insertion operations. N is the total number of labels in the defining transcriptions. We used the *HResults* tool to estimate H, S, D and I and calculated N as $N = H + S + D$. Estimation of all accuracy measures used speech phone labels only, ignoring silence labels. Since the Kaldi systems in this study were based on phone rather than word labels, accuracy was determined at phone level.

C. Optimising recognition performance

Adjusting either the acoustic or language-scale in Kaldi is effective to search through different ratios of acoustic and language model contributions. Setting the acoustic-scale (which is used during decoding) close to the inverse of the language-scale setting provides good results in general.

To find the best ratio between the acoustic and language model contributions for the L1 data set, we varied the language-scale parameter during scoring (integer values in the range of 1-20) and kept the acoustic-scale parameter at the default (value of 0.1) setting. In each experiment the language-scale adjustments were performed on the L1 development set data only. The language-scale value yielding the best accuracy was selected and kept constant for all the measurements on the L1 and L2 test data.

In [15] it was pointed out that more insertion errors tend to occur in slow speech segments. Therefore it should be possible to optimise results further by adjusting the insertion penalty during decoding. For the results presented in Section VI-C, we verified that reducing the number of insertion errors further (by tweaking the language-scale parameter) did not lead to any significant improvements of the results at any speed factor threshold.

V. Rate Of Speech analysis

ROS is often used as an indicator of oral proficiency. In this study ROS was calculated in the same way as in [4]: the ratio between the number of the speech phones in an utterance and its total duration. The ROS values were derived to compare L1 and L2 data. The results of the comparison informed our data perturbation method for the L2 data.

A. Phone alignment

ROS estimation requires time alignments at phone level. To produce these alignments for the L1 and L2 data sets, we force aligned the decoded phone labels. With the exception of the L1 training data, a training graph containing only this single decoded sequence of labels for each utterance was used. Alignment of the L1 training data was accomplished with the reference (more accurate) training graphs. A final decode then generated the required lattices from which time alignments were extracted (using the *ali-to-phones* Kaldi implementation).

B. Distribution of utterances

The *Train_3* histogram in Figure 1 depicts the distribution of the per utterance ROS values derived from the training data ($\bar{L1}_{ros} = 8.46$, $\sigma = 1.78$). ROS histograms for the L2 utterances are shown in the same figure (L2A in Figure 1(b) and L2B in Figure 1(c)). The histograms show that the distributions of the L2A and L2B data sets differ. The L2A histogram ($\bar{L2A}_{ros} = 7.88$, $\sigma = 1.31$) resembles the *Train_3* histogram more closely than the L2B distribution ($\bar{L2B}_{ros} = 6.92$, $\sigma = 1.40$).

C. L2 speed factors

The data in Figure 1 shows that many of the L2 utterances were articulated at a slower rate than the L1 training data, especially in the L2B data set. As described in section III-C2, the tempo of utterances could be perturbed to match the observed speech rate of the training data better. To resemble $\bar{L1}_{ros}$, a number of between 2 068 (DNN) and 2 297 (SGMMs) utterances required speed factors greater than 1.0, given the total of 3 144 L2 utterances. The ROS of about 988 utterances was faster than $\bar{L1}_{ros}$.

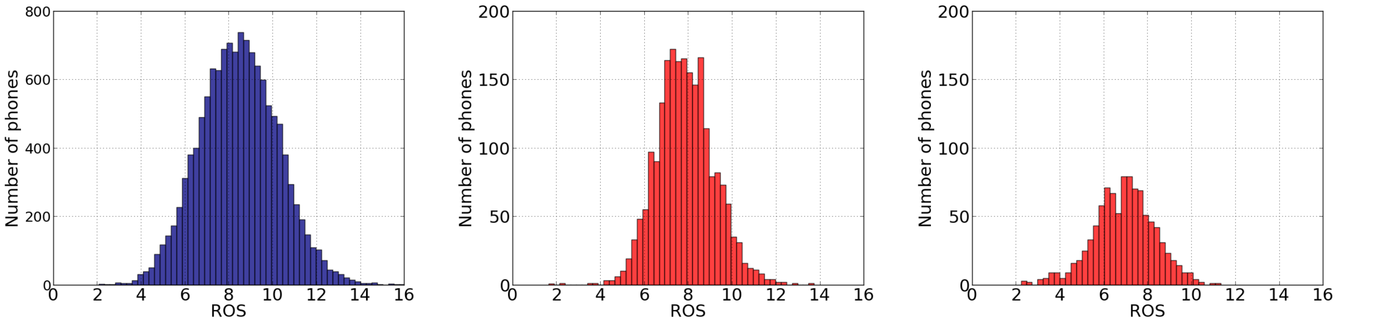
To investigate the merit of creating versions of the test utterances with speech rates closer to $\bar{L1}_{ros}$, we applied different selection thresholds. Perturbed versions of utterances were only selected when the estimated ROS more closely represented $\bar{L1}_{ros}$ and without exceeding the set threshold to control the size of the allowed speed factors. In this manner the speed of slower utterances was increased gradually and the effect measured for each speed factor step size of 0.5.

VI. Results

Section VI-A contains a complete overview of phone recognition accuracies for SGMM and DNN systems trained on different data sets. In Section VI-B the effect of the difference in recording quality between the L2A and L2B data sets is highlighted and Section VI-C reports on the improvement in recognition performance as a result of altering the speed of the L2 data.

A. Training data perturbation

As described in Section III-C, perturbing the speed of individual training utterances effectively simulates additional speakers. To investigate the quality and utility of the new utterances as training material, we compared phone recognition accuracies for the *Train_1*, *Perturbed_1*, *Train_3* and *Perturbed_3* data sets defined in Sections III-A and III-C. The recognition accuracies obtained by systems trained on the *Train_All* and *Perturbed_3_All* data sets are also reported. The results for an initial comparison, based on systems with a flat 1-gram language model, are shown in Table III. The table provides a system description (System) in terms of the training data set and acoustic model type (in brackets) and lists the phone recognition accuracy (ACC) for both L1 and L2 test sets.



(a) L1 (Train_3)

(b) L2A

(c) L2B

Fig. 1: *Per utterance ROS values for L1 and L2 data*

System	L1 (ACC)	L2 (ACC)
Flat 1-gram ARPA		
Train_1 (SGMM)	77.26	64.67
Train_1 (DNN)	79.28	62.86
Perturbed_1 (SGMM)	77.77	65.30
Perturbed_1 (DNN)	79.22	61.52
Train_3 (SGMM)	79.33	67.35
Train_3 (DNN)	82.61	67.33
Perturbed_3 (SGMM)	79.58	67.45
Perturbed_3 (DNN)	82.60	65.56
Train_All (SGMM)	80.12	67.23
Train_All (DNN)	84.47	69.41
Perturbed_3_All (SGMM)	79.10	65.67
Perturbed_3_All (DNN)	83.24	68.52

TABLE III: Comparing phone recognition accuracies (ACC) for different systems

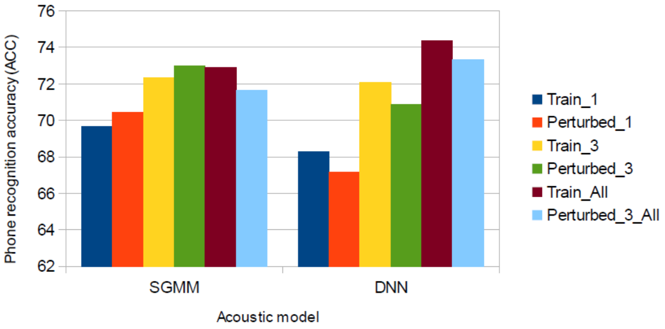


Fig. 2: Phone recognition accuracies for SGMM and DNN systems on the L2A data set

Table III shows that results did improve for the SGMM system as a result of data perturbation (compare *Train_1* SGMM and *Perturbed_1* SGMM). In addition, the absolute cross-corpus (L2) accuracy increased from 64.67% to 65.30%. The same trend was not observed for the DNN system. With only 5 hours of training data (see Table II), the DNN system did not outperform its SGMM counterpart on L2. L1 modelling with the *Perturbed_1* training data did not improve with a DNN estimator and a degradation for the L2 test set was observed compared to the *Train_1* results.

The data perturbation technique clearly did not achieve a gain of similar magnitude than real training data. Experiments *Train_3* (SGMM) and *Train_3* (DNN) yielded much better results than the *Perturbed_1* system. For this training set the DNN system continued to outperform (82.61%) the SGMM system on the L1 test set. The performance gap between these systems on L2 also disappeared.

The *Perturbed_3* experiments revealed a much smaller effect of data perturbation on the results. The performance increase of the SGMMs on L2 data diminished: 0.1% absolute difference and DNN recognition accuracy of L2 data degraded.

Training on all the available data (*Train_All*), which include many examples of some unique types, did not result in much gain over the *Train_3* SGMM systems. Using a DNN system provided the best results (84.47 and 69.41 for L1 and L2 respectively).

Given the training data, it is possible to learn and incorporate information on regularly observed phone sequences using n-gram language models. To verify the utility that a stronger phone sequence predictor might have, we repeated the *Train_3* and *Perturbed_3* experiments with a 2-gram model based on phone transcriptions of the *Train_1* data set. Table IV presents the corresponding results. In comparison to the flat 1-gram model, trends remained very similar. With the 2-gram model phone recognition accuracy for the L2 data improved to 70.37%.

System	L1 (ACC)	L2 (ACC)
2-gram ARPA		
Train_3 (SGMM)	83.06	70.34
Train_3 (DNN)	84.70	70.37
Perturbed_3 (SGMM)	83.34	70.34
Perturbed_3 (DNN)	84.67	69.84

TABLE IV: Comparing phone recognition accuracies (ACC) for 2-gram language models

B. *Speech quality*

The observed differences in system performance between the L1 and L2 data could be ascribed to acoustic differences between the two data sets as well as the quality of the L2 recordings. To investigate the role of this effect more closely, we compared phone accuracies for the L2A and L2B data separately. Figure 2 depicts the phone recognition accuracies for 2 166 utterances of the L2A data set decoded with each of the systems (using a flat 1-gram model) presented in Table III. The same results are depicted for 978 L2B utterances in Figure 3.

The results show that the recognition rate for the more proficient learners in the L2A data set is better than the accuracies measured for the beginner level learners in L2B. The figure also indicates that data perturbation improved results for the SGMM system evaluated on the L2A data with the *Perturbed_3* system achieving a similar

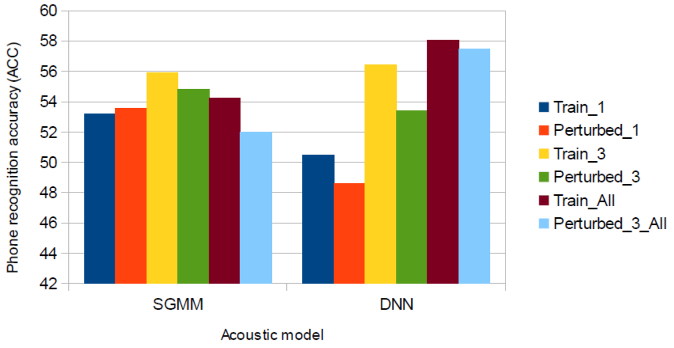


Fig. 3: Phone recognition accuracies for SGMM and DNN systems on the L2B data set

result to training on all the NCHLT training data. In contrast, the L2B *Perturbed_3* result for the SGMM system degraded.

With a DNN estimator no *Perturbed* systems yielded an improvement but, with sufficient training data, the DNN systems outperform all of the SGMM systems.

C. Speed factor analysis

Figure 4 depicts phone recognition accuracies achieved for speech factor thresholds up to the point where some utterances were played 30% faster than the original. The figure shows results for SGMM and DNN systems trained on the *Train_3* set using the flat 1-gram language model only. The results show that, in general, increasing the speed of slow L2 utterances proved beneficial.

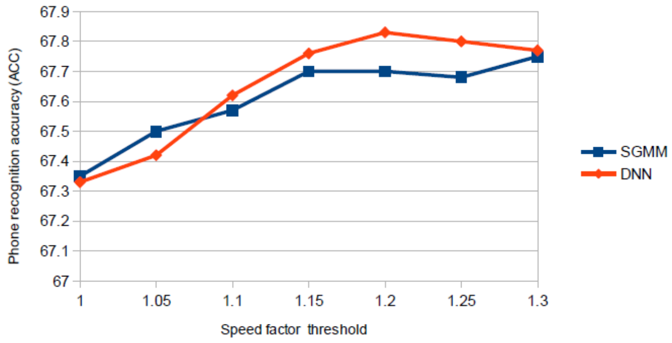


Fig. 4: L2 data: Phone recognition accuracies for SGMM and DNN systems with different speed factor thresholds

At a speed factor threshold of 1.0 no perturbed utterances were selected and the respective phone accuracies are therefore equal to the values in Table III. We found that the DNN system marginally outperformed the SGMM system at a speed factor threshold of 1.2, resulting in an absolute improvement of 0.5%.

With regard to the difference between the L2A and L2B data, Figures 5 and 6 illustrate the contribution of each set to the change in recognition accuracy. The observed improvement is higher for the L2B data than for the L2A data. This result could be expected given the distributions in Figure 1a: there are more slower utterances in the L2B set that stand to gain from the speed adjustment. Furthermore, the DNN predictor outperformed the SGMM predictor on the L2B data rather than the L2A data.

VII. Discussion

Limited training data imposes unique constraints on ASR development. The ability of systems to generalise well, given the complexity of the speech signal, reduces with smaller training sets. Furthermore, L2 speakers, and in particular low-proficient learner speakers, introduce additional variability into the data. The aim of our investigation was therefore to explore state-of-the-art modelling techniques as applied to under-resourced ASR. In particular we attempted to determine which methods generalise well to the L2 space given these conditions, before exploring pronunciation differences between the L1 and L2 speech.

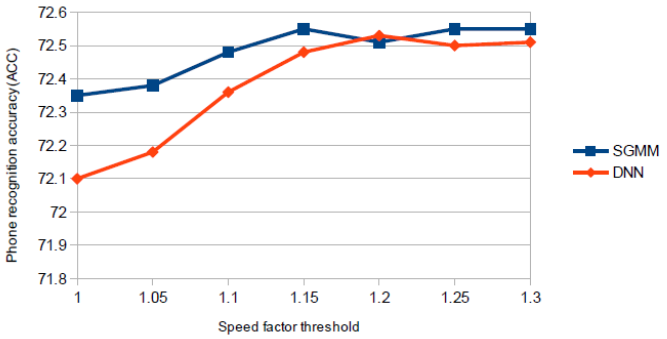


Fig. 5: L2A data: Phone recognition accuracies for SGMM and DNN systems with different speed factor thresholds.

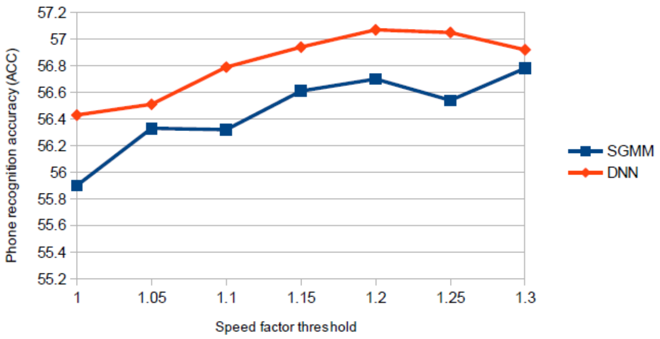


Fig. 6: L2B data: Phone recognition accuracies for SGMM and DNN systems with different speed factor thresholds.

Substantial pronunciation differences between L1 and L2 speech are more difficult to recognise or predict with systems trained on limited sets of L1 data. Recently, ASR systems trained on hundreds of hours of training data have been shown to benefit from data perturbation techniques. Not only did these techniques create more training data, but it has been claimed that the robustness of such systems also improved [5].

It may not be obvious why a particular data perturbation technique produces an improved ASR system [6]. In terms of phone examples, apart from phone confusion mistakes, it might make sense to separately consider two types of variability that could both be expected to affect the acoustic match for an L1 system recognising L2 speech: (1) inter-speaker variability and (2) any remaining intra-phone variability.

With only 5 hours of training data, the *Train_1* system results confirmed that data perturbation (15 hours) did indeed yield an SGMM system that generalised better to L2. We also confirmed that this amount of data was simply too limited to attempt DNN training with standard parameters.

The situation improved much with 15 hours of real training data (*Train_3* systems). Although some overfitting of the DNN on L1 data clearly remained, in general the SGMM and DNN systems generalised equally well to the L2 domain. As could be expected, simply perturbing the speed of the *Train_1* utterances was not nearly as effective as adding real data samples.

Comparing the *Perturbed_3* and *Train_3* system results, it was clear that simulating additional phone examples from more speakers (*Perturbed_3*) had very little effect on with-in corpus results and generalisation to L2 data degraded. The effect of intra-phone variability seemed to dominate at this point. This observation seems to indicate that more phone examples are required to improve results, rather than more examples from different speakers. In [16] a similar insensitivity with regard to the utility of more speakers to increase training data was observed. Results for the *Train_All* (SGMM) system confirms this prediction. Generalisation to L2 remains similar to the *Train_3* system. Also, given the results for *Train_All* and *Perturbed_3_All* the fact that no further gain could be obtained from data perturbation is not surprising. The combination of a stronger DNN predictor and more training examples of the same unique types significantly improved results.

One reason why it is challenging to recognise L2 learner data is that such data probably has more intra-phone variability. Mispronunciation and alternative speech rates which, in turn, may lead to even greater pronunciation differences, occur more commonly for learner speakers. Accordingly, the results for the learner data in Section VI-B showed that perturbed system performance is worst for beginner level learners. In addition, the improvement in recognition accuracy observed for the DNN system trained on the *Train_All* data set was less significant for the L2B test data than the L2A data.

We attempted to limit the observed variability further by speeding up slower utterances so that the speech of the perturbed test utterances would have an ROS matching the training data more closely. This strategy did result in some recognition improvement.

Separate investigation of the L2A and L2B test sets revealed interesting differences. For the L2B data the DNN results seemed to keep outperforming the SGMM, while this observation is reversed for the L2A data. Compared to the SGMM, the DNN predictor seems to generalise better to intra-phone variability within the L2B data. This effect also benefits slightly from the type of limitation our data perturbation technique imposed on the data. It seemed that adjusting the ROS of the L2B data to the observed ROS for the phone examples of the training data further reduced the inter-speaker differences, while a large component of intra-phone variability remained.

VIII. Conclusions

Optimising an under-resourced ASR system to recognise low-proficient learner data more accurately presents unique challenges. In this study we performed an analysis of recognition accuracy for state-of-the-art acoustic modelling and investigated the extent to which data perturbation may reduce phone mismatch of L1 train and L2 test data. Our results showed that speed perturbation of the training data can improve phone recognition accuracy of L2 data under specific conditions.

The speed perturbation techniques generate a particular type of inter-speaker variability. Therefore, the results also showed that the quality of real phone examples surpass the quality of the simulated data. With sufficient examples, the DNN system provides improved generalisation to both the L1 and L2 data of low quality. This finding has not been reported for the NCHLT data to date.

L2 phone examples likely contain more intra-phone variability than L1 data. With data perturbation it is possible to limit the difference in speech rates observed for L1 and L2 data. Increasing the speech rate of slower L2 utterances yielded improved recognition accuracy. Future work on L2 recognition will focus on intra-phone variability.

References

[1] M. Eskenazi, A. Alwan, and H. Strik, Eds., *Speech Communication: Special Issue on Spoken Language Technology for Education Spoken Language*. Elsevier, 2009.

[2] O. Engwall, Ed., *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*. KTH, Stockholm, Sweden, 2012.

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

[4] J. Badenhorst, A. Tshoane, and F. de Wet, "What does learner speech sound like? A case study on adult learners of isiXhosa," in *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016*. IEEE, 2016, pp. 1–6.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, September 2015, pp. 3586–3589.

[6] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proceedings of Interspeech*, San Francisco, USA, September 2016, pp. 2378–2382.

[7] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages." in *Proceedings of Interspeech*, Singapore, September 2014, pp. 810–814.

[8] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages." in *Proceedings of Interspeech*, Singapore, September 2014, pp. 1420–1424.

[9] E. Gauthier, L. Besacier, and S. Voisin, "Speed perturbation and vowel duration modeling for ASR in Hausa an Wolof languages," in *Proceedings of Interspeech*, San-Francisco, United States, September 2016, pp. 3529–3533.

[10] S. Robertson, C. Munteanu, and G. Penn, "Pronunciation error detection for new language learners," in *Proceedings of Interspeech*, San Francisco, USA, September 2016, pp. 2691–2695.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on Automatic Speech Recognition and Understanding (ASRU)*, no. EPFL-CONF-192584, Hilton Waikoloa Village, Big Island, Hawaii, December 2011.

[12] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalised maxout networks," in *Proceedings of IEEE ICASSP*, Florence, Italy, May 2014, pp. 215–219.

[13] D. Jurafsky and J. Martin, *Speech & language processing*. Prentice Hall, 2000.

[14] 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//.

[15] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Transactions on speech and audio processing*, vol. 12, no. 4, pp. 391–400, 2004.

[16] J. Badenhorst, C. V. Heerden, M. Davel, and E. Barnard, "Collecting and evaluating speech recognition corpora for 11 South African languages," *Language Resources and Evaluation*, vol. 3, no. 45, pp. 289–309, 2011.