The Effect of Language Identification Accuracy on Speech Recognition Accuracy of Proper Names

Oluwapelumi Giwa[1,2] and Marelie H. Davel[1,2]
[1]Multilingual Speech Technologies, North-West University, South Africa.
[2]CAIR, CSIR Meraka, South Africa.
oluwapelumi.giwa@gmail.com, marelie.davel@gmail.com

***Abstract*— Utilising the known language of origin of a name can be useful when predicting the pronunciation of the name. When this language is not known, automatic language identification (LID) can be used to influence which language-specific grapheme-to-phoneme (G2P) predictor is triggered to produce a pronunciation for the name. We investigate the implications when both the LID system and the G2P system generate errors: what influence does this have on a resulting speech recognition system? We experiment with different approaches to LID-based dictionary creation and report on results in four South African languages: Afrikaans, English, Sesotho and isiZulu.**

I. Introduction

Speech recognition applications that utilise names, such as directory enquiry systems, require the ability to predict the pronunciation of these names accurately. Proper names (names of people, places or things) form a very large set of words, and their pronunciation can be difficult to predict.

One of the factors that has been shown to improve modelling of proper names is the ability to identify their language of origin [1]. Many personal names originate from other languages [2], and their pronunciation can be influenced by the pronunciation rules of the original language. In South Africa, the difficulty of identifying the language of origin of a name is exacerbated by two factors: the scarcity of resources for model training and the co-existence of multiple languages. Many names are in fact multilingual. For example, an Afrikaans name, such as 'Paul' (*/ p @u l /*, using SAMPA[1] notation) will be pronounced differently from the English name 'Paul' (*/ p O: l /*).

In this paper, we are interested in the effect of language identification (LID) accuracy on automatic speech recognition (ASR) accuracy of proper names in the South African context, and experiment with four languages: Afrikaans, English, Sesotho and isiZulu. We use mono- and multilingual LID systems to predict the language tags of proper names. We then use language-specific grapheme-to-phoneme (G2P) systems to generate pronunciations based on the different LID tags and analyse the result. Using the predicted pronunciations, we develop ASR systems and investigate the implications of LID of proper names on ASR accuracy. Specifically, we would like to know whether more accurate LID tags influence recognition accuracy.

[1]The 'Speech Assessment Methods Phonetic Alphabet' is a standard computer-readable notation for phoneme descriptions. See [3].

The paper is structured as follows: Section II provides background on related LID studies of proper names. Section III presents the data used in subsequent experiments, and Section IV provides an overview of the experimental design. Section V describes results in terms of LID, G2P and ASR performance. Finally, Section VI summarises the main findings.

II. Background

Various studies of natural speech synthesis have found that knowledge of the source language of a proper name is important to determine the correct pronunciation of that name [4], [1]. In the South African context, Kgampe and Davel [5] investigated the pronunciation of Afrikaans, English, Setswana and isiZulu names, and showed that the linguistic origin of proper names and the mother tongue of the respondent have a significant effect on the pronunciation of such names.

Data-driven G2P prediction or even fully graphemic systems perform very well for standard words [6]. However, these techniques become less accurate for words that do not follow the standard pronunciation rules of the language. Names can be of diverse etymological origin and are sometimes borrowed from another language without following the process of assimilation to the phonological pattern of the new language [7]. Pronunciation prediction of names therefore typically still rely on large dictionaries and semi-automatic processing (a combination of manual and automatic processing) [8].

To address the complications associated with proper name pronunciation prediction, a number of authors [1], [9], [10] propose a combination of two lexical modelling approaches, consisting of (1) G2P conversion based on language-specific rules, followed by (2) phoneme-to-phoneme (P2P) conversion. The language-specific G2P conversion approach makes use of the source language of the proper name in context before applying the language-specific G2P rules to predict its pronunciation. For example, Réveil *et al.* [9] performed a study on how the source language of a word affects the ASR performance. In their experiment, they made use of a language-specific G2P converter, mono- and multilingual acoustic model and language-specific P2P converter. They observed that when pronouncing foreign words, speakers tend to use the G2P rules of their own (mother tongue)

language rather than the G2P rules of the true language of origin.

LID of proper names have been studied by authors such as Bhargava and Kondrak [11], who experimented with two multilingual name corpora, namely the *Transfermarkt* corpus containing European soccer players' names in 13 possible languages, and the *Chinese-English-Japanese (CEJ)* corpus containing first and last names in these languages. They used support vector machines (SVM) and *n-gram* counts as classifier and features respectively. In [12], Joint Sequence Models (JSMs) were found to provide competitive results when compared with n-gram SVMs. In related work, JSMs were also used to classify names as multilingual [13].

III. Data

Two data sets are used to evaluate the effect of LID on ASR accuracy: the South African Directory Enquiries (SADE) corpus [14], and the Multipron corpus [15]. The SADE corpus was developed to support the development of directory enquiry applications in South Africa. This corpus contains audio samples from multilingual speakers producing proper names typically encountered in a directory enquiries system. SADE encompasses all 11 official South African languages, from which a subset of four languages is selected.

Multipron was designed to contain samples of the diverse proper names that occur in South Africa. The corpus captures a balance between own-language pronunciations and imitation of foreign-language pronunciation, as well as the different styles of imitation that occur, which are also of broader interest. Table I lists the number of unique words, average word length and character count per language, with 'afr', 'eng', 'sot' and 'zul' representing Afrikaans, English, Sesotho and isiZulu, respectively. Considering the statistics in Table I, SADE data set is almost exclusively bilingual; where English and Afrikaans languages dominate. Also, we observe that the majority of the words are monolingual with a smaller percentage of 9.3% identified as multilingual.

TABLE I
SADE vs Multipron corpus: language distribution and word statistics.

	Word count		**Avg. word length**		**Character count**	
Lang	*SADE*	*Multipron*	*SADE*	*Multipron*	*SADE*	*Multipron*
afr	1 050	252	6.9	6.9	7 308	1 743
eng	6 634	522	7.4	6.2	48 733	3 232
sot	465	264	7.8	7.4	3 612	1 965
zul	458	254	8.1	7.2	3 689	1 839

In addition, the *NCHLT-inlang* dictionaries [16] are used as training data when creating both LID and G2P models. These dictionaries were developed as part of the NCHLT speech corpora [17], and contain 15,000 word-pronunciation pairs in each of South Africa's 11 official languages.

A. Corpus reference dictionary

The dictionaries for the SADE and Multipron corpora were obtained in different ways: The original Multipron corpus [15] constitutes a combination of name-surname pairs as a single word with their corresponding pronunciations, which were recorded and then manually transcribed per audio clip. In [18], name pairs were split into separate entities (compensating for boundary effects), resulting in a dictionary with more pronunciation variants per word. The latter split version is the dictionary utilised here.

The SADE pronunciation dictionary, on the other hand, was obtained semi-automatically by combining manual verification and correction with G2P prediction. Initial name pronunciations were obtained using G2P rules extracted from already existing resources. In order to identify incorrect pronunciations, a phoneme-based dynamic programming score (PDP) [19] was used. These scores are based on speech recogniser output, and are calculated using a data-driven matrix (assigning variable weight across phoneme substitutions). PDP output was also used to generate cross-lingual pronunciation variants. A final round of verification was carried out on audio/transcription pairs to flag wrong transcriptions and mark these for manual correction [18]. We observe that less pronunciation variants per word exist in this dictionary as compared to the Multipron dictionary.

IV. Experimental design

A. Overview

To analyse the effect of LID on ASR accuracy, we develop four ASR systems that only differ in the pronunciation dictionaries used during training and testing. These dictionaries are generated using a combination of LID and G2P prediction. The same set of language-specific G2P predictors are used for all dictionaries, but different LID options are evaluated (causing different predictors to be triggered per word). The overall process is depicted in Fig. 1, and described in the remainder of this section.

B. LID tag prediction

Each corpus includes a LID tag per word, which can be used to obtain an oracle result (*Ref-LID*). Three additional cases are considered where JSM models are used for LID prediction:

- Single-language tags: the LID technique is forced to classify each word as monolingual.
- Multi-language tags: words are classified as multilingual. Each word may therefore have more than one LID tag.
- All-four-language tags: we assume that all words originate from all four target languages by tagging each word with all the target languages.

One question that needs addressing is the selection of the training data on which the LID system should be trained. To avoid bias towards any of the evaluation corpora employed, especially during later ASR experiments, a model trained on the *NCHLT-inlang* data set is used to predict the language of origin of proper names. A 10 000 word subset of the original 15 000 word list is used per language. This list contains only unique words that are monolingual. The JSM-based LID model developed is described in [13]. In summary, the JSM model parameters used include initialisation with counts,

unconstrained contexts, full Expectation Maximisation, and discounting (optimised using a 10% hold-out set, and folded back post training). Models are trained up to an order of 8. To select the final word LID across sets, *'forced pron'* voting is applied. See [13] for more detail.

Once LID prediction has been completed, there are therefore four versions of the word list, each version associated with a different set of LID tags.

C. Phoneme mapping

As different phoneme transcription conventions are used for the corpora involved, these must be reconciled prior to G2P training or testing. The goal of creating a phoneme mapping is to use common phonemes from all available corpora to obtain a final phoneme set that has the same symbol per phoneme. Cases where corpora share the same symbol or use a different symbol for the same phoneme, are straightforward to handle. However, there are cases where two distinct phonemes in one corpus are transcribed as a single phoneme in another, resulting in an absence of certain phonemes. This occurs since phonemes contained in *NCHLT-inlang* dictionaries capture distinctions that are linguistically important, compared to the phoneme set used to obtain the two reference dictionaries, which capture distinctions humans were able to make reliably during cross-lingual transcriptions.

For example, in our reference dictionaries, for languages in which duration is considered important, a differentiation is made between long and short vowels (*/u:/* and */u/*, */i:/* and */i/*) and speakers tend to produce distinct samples that are either lengthened or not. On the other hand, for languages that do not put emphasis on duration, speakers may arbitrarily produce */i:/* or */i/* or a duration in between – this is then very difficult to transcribe consistently. In cross-lingual transcriptions, such language-specific transitions are therefore sometimes omitted [20].

To obtain a common phoneme set that reconciles differences across the three corpora, two phoneme set mappings are defined:

- One that makes the sets consistent but retains all phoneme distinctions, referred to as 'detailed'.
- Another that applies a merging approach where not all phoneme distinctions are utilised, referred to as 'combined'.

The 'detailed' phoneme set will therefore produce more conservative results than the 'combined' set.

D. G2P dictionary prediction

To obtain phonemic transcriptions and generate a dictionary for the different word lists, G2P conversion is performed on each list using Default&Refine [21]. The algorithm is trained on language-specific generic text to extract G2P rules, which are then language-dependent. Each word translation is based on language-specific G2P rules. To generate G2P models, the full set of the *NCHLT-inlang* data (15 000 words per language) is used as training data. Using these models, we generate four dictionaries:

- Ref-LID dictionary: G2P-based dictionary obtained from the manually tagged word list.
- Single LID dictionary: G2P-based dictionary obtained from the JSM-based single-language tags.
- Multi LID dictionary: G2P-based dictionary obtained from the JSM-based multi-language tags.
- All-four languages dictionary: G2P-based dictionary obtained when assuming all four languages apply.

Figure 1 shows the different steps necessary to obtain a phonemic transcription for each G2P-based dictionary.

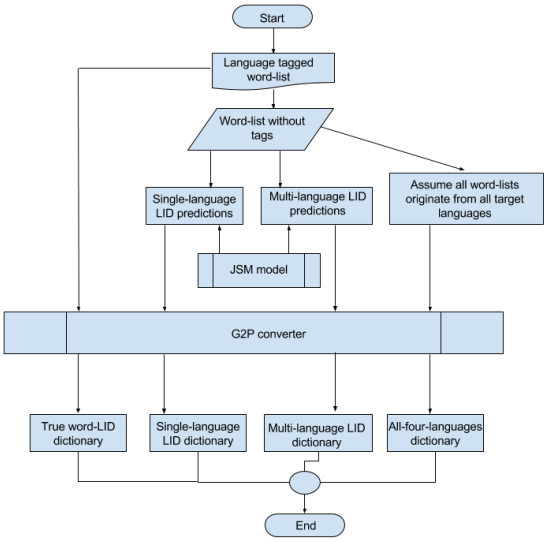

Fig. 1. Different processing step used to generate G2P-based dictionaries. [22]

As discussed in Section IV-C, the resulting dictionaries can only be used when mapped to a reconciled phoneme set. The mappings introduced in Section IV-C are applied here.

E. ASR system

In order to build a baseline ASR system, we use the entire SADE corpus [14]. The SADE corpus contains 13 hours 56 minutes and 9 seconds of speech from 40 speakers, with 500 prompts per speaker. The data corpus is balanced across gender with 20 male and 20 female speakers. We partition the entire corpus into a 65% training and 35% testing set. No development set is used, and all parameters are set on the training data directly.

The baseline system employs a standard Kaldi-based system using a recipe similar to the Babel recipes [23]. We build a context-dependent crossword HMM-based phone recogniser with triphone models and Gaussian mixture models (GMMs). For speaker-specific transforms, feature-space maximum likelihood linear regression (FMLLR) and a maximum likelihood linear transform (MLLT) is used per speaker. To perform speaker-specific normalisation, both Cepstral Mean Normalisation (CMN) and Cepstral Variance Normalisation (CVN) are used. Features are obtained by splicing

together seven frames 13-dimensional Mel-frequency cepstral coefficients (MFCCs) each. The feature dimensionality is later reduced using linear discriminant analysis (LDA) to 40. These models generate the alignments used to initialise a standard 3-layer deep neural network (DNN).

The test vocabulary and LM used has a significant effect on the recognition accuracy. We therefore analyse two cases: (1) in order to minimise the effect of the LM, we utilise a flat LM knowing that recognition accuracy will be poor; (2) we also use a trained n-gram-based LM to verify that our system is sufficiently accurate (that is, to evaluate the overall system development process). To obtain a complete pronunciation lexicon necessary to train an ASR system and avoid out-of-vocabulary (OOV) tokens, we:

- identify all words not included in Section IV-D, and extract their corresponding pronunciations (either single or multiple) from the original SADE transcribed dictionary, referred to as the 'supplemental pronunciation dictionary', and
- combine each dictionary being analysed (the case-specific dictionaries from Section IV-D) to obtain a complete pronunciation dictionary for ASR purposes.

Note that the supplemental pronunciation dictionary retains variants (single or multiple) as contained in the reference SADE dictionary. All phonemes are mapped to the 'combined' phoneme set using the mapping approach discussed in Section IV-C . Also, in the SADE corpus, words with two or less characters, spelled-out or words outside the four target languages are removed from the reference and hypothesised dictionaries for proper comparative analysis.

F. Performance measure

ASR accuracy is evaluated in terms of word error rate while LID performance is calculated using standard precision, recall, and F-measure.

No single evident measure exists for calculating G2P accuracy for variant-based dictionaries. In this analysis, we use four G2P metrics defined in [24], namely variant-based phoneme accuracy (V-PA), variant-based word accuracy (V-WA), single-best phoneme accuracy (S-PA) and single-best word accuracy (S-WA). Variant-based G2P accuracy is computed per word, by evaluating each reference pronunciation against all hypothesised pronunciations (for that specific word) to obtain the best-matching pronunciation, as well as the accuracy score for that reference-hypothesis pair. These scores are then averaged across all pronunciation variants of the specific word, occurring in the reference dictionary. A best-variant score of 1 for any given word means that there is a hypothesised pronunciation that matches the reference pronunciation, for every single pronunciation variant occurring in the reference dictionary.

V-PA and V-WA only differ on whether the accuracy score (that is averaged across all reference pronunciation variants) is a phoneme accuracy score (V-PA) or word accuracy score (V-WA). Both the overall V-WA and V-PA are then obtained by averaging these values over all the words in the dictionaries. S-PA and S-WA are computed by obtaining only the single best accuracy per word (from the best-matching reference and hypothesis pair), and then averaging over all words in the dictionary. Similarly, S-WA is a word-based accuracy score and S-PA a phoneme-based accuracy score.

V. Analysis and Results

A. LID Analysis

We first evaluate the LID accuracy of the three predicted LID sets (*Single, Multi* and *All-four*), where we use the true language of origin of each word as a reference to estimate the LID accuracy. Table II shows the results of the two JSM-based and *All-four* LID predictions using precision, recall and F-measure.

TABLE II
LID Precision, Recall and F-measure using different LID approaches.

Data set	Dict	Precision (%)	Recall (%)	F-measure (%)
SADE	All-four	27.11	100.00	44.66
	Multi	84.87	92.80	88.66
	Single	88.86	81.94	85.26
Multipron	All-four	26.20	100.00	41.52
	Multi	72.63	88.48	79.78
	Single	78.64	75.04	76.80

As expected, from our previous study [13], *Multi LID* outperforms *Single LID* across both data sets. Also, as expected, we observe very low precision on the *All-four* approach, where we assume all word lists originate from all target languages.

B. G2P Analysis

For each corpus, we report G2P accuracy by comparing the four hypothesised dictionaries against the reference dictionary using both phoneme sets. Tables III and IV show V-PA, V-WA, S-PA and S-WA obtained using the 'detailed' and 'combined' phoneme sets respectively, and based on different dictionary approaches. Across all performance metrics, accuracy obtained using the 'detailed' phoneme set is lower as compared to 'combined' because of the higher penalty incurred with the stricter phoneme mapping strategy employed. We observe that across both phoneme sets, accuracy obtained using the *All-four* dictionary outperforms other G2P-based dictionaries; the Multi dictionary outperforms the *Ref-LID* dictionary, and *Single* performs the worst (of the four G2P-based dictionaries).

Given the performance measure used, the higher the number of pronunciation variants, the better the G2P accuracy. This explains why the *All-four* approach seems the best dictionary approach, since we over-generate pronunciation variants across all four target languages. These results are not expected to mirror ASR accuracy, as it is known that more variants tend to introduce higher confusability in ASR systems. Columns Ref_{avg} and Hyp_{avg} represent the average number of reference and hypothesised pronunciations per word. For the 'Multipron' corpus, the number of pronunciation variants per word is more than in the SADE corpus, due to the fact that each audio clip produced by a speaker

TABLE III

V-PA, V-WA, S-PA AND S-WA ACHIEVED WITH 'DETAILED' PHONEME SET FOR DIFFERENT DICTIONARY APPROACHES ON TWO DATA SETS.

Data set	**Dict**	**V-PA**	**V-WA**	**S-PA**	**S-WA**	**Ref.$_{\mathbf{avg}}$**	**Hyp.$_{\mathbf{avg}}$**
Multipron	All-four	76.20	19.93	94.27	71.13	5.52	3.53
	Multi	70.55	15.49	90.64	59.60	5.52	1.35
	Ref-LID	68.44	13.28	89.15	53.41	5.52	1.05
	Single	66.95	13.13	88.03	52.55	5.52	1.00
SADE	All-four	83.80	39.74	84.79	42.03	1.12	3.53
	Multi	80.49	36.86	81.73	39.02	1.12	1.32
	Ref-LID	79.89	35.81	81.06	37.62	1.12	1.08
	Single	76.97	33.12	78.53	35.10	1.12	1.00

TABLE IV

V-PA, V-WA, S-PA AND S-WA ACHIEVED WITH 'COMBINED' PHONEME SET FOR DIFFERENT DICTIONARY APPROACHES ON TWO DATA SETS.

Data set	**Dict**	**V-PA**	**V-WA**	**S-PA**	**S-WA**	**Ref.$_{\mathbf{avg}}$**	**Hyp.$_{\mathbf{avg}}$**
Multipron	All-four	78.94	25.19	96.41	81.73	5.21	3.4
	Multi	73.77	20.20	93.26	70.98	5.21	1.33
	Ref-LID	72.07	18.06	92.12	66.56	5.21	1.05
	Single	70.54	17.51	90.86	64.55	5.21	1.00
SADE	All-four	89.59	59.96	90.53	62.73	1.12	3.39
	Multi	86.94	57.26	88.19	60.20	1.12	1.31
	Ref-LID	86.79	56.53	87.97	59.12	1.12	1.08
	Single	83.42	52.91	85.08	56.04	1.12	1.00

was manually transcribed. Comparing Tables III and IV, we observe that trends remain similar using either phoneme set. Hence, we use the 'combined' phoneme set from here onwards.

TABLE V

AVERAGE NUMBER OF PRONUNCIATION VARIANTS, AS WELL AS THE WER USING FLAT AND TRAINED LMS.

		WER	
Dictionary	*Average variants*	*Flat LM*	*Trained LM*
Manual	1.12	57.30	14.60
Ref-LID	1.08	62.70	16.40
Single	1.00	63.90	18.20
Multi	1.31	64.90	19.90
All-four	3.39	67.80	20.20

C. ASR Analysis

Word recognition is performed using both a flat and a trained LM. For the trained LM, we use the *4-gram* modified Kneser-Ney technique where the minimum *n-gram* order is set to 1. To determine the optimal LM weight, we perform a 2-fold cross-validation on the testing data. To understand how each of the dictionaries influences the performance of each system, we measure word error rate.

One of the most common reasons for ASR errors observed here is acoustic confusability due to homophones. Examples include spelled-out words, numerical words, dates, etc. While this ambiguity is typically resolved by the language model, a system developed with a flat language model may be unfairly penalised. To report the performance of the system, we therefore consider two cases where:

- All homophones are retained without performing any preprocessing task.
- Each word in the hypothesised string is remapped to its reference counterpart if the pronunciation of the reference word and observed word matches. See Table VI, for an example.

TABLE VI

EXAMPLE OF HOMOPHONE REMAPPING A HYPOTHESISED ('HYP') STRING TO BETTER MATCH THE REFERENCE ('REF') STRING.

Original string	REF	l-3 communications hldgs.
	HYP	l-3 communications holdings
After homophone remapping	REF	l-3 communications hldgs.
	HYP	l-3 communications hldgs.

Remapping homophones reduces the WER by only 1% on average across the target dictionaries. Trends observed before or after homophones preprocessing (realigning homophones) are the same. Homophones therefore affect the results less than we initially anticipated.

As expected, we achieve the best result on the reference dictionary. ASR performance decreases as soon as predicted dictionaries are used. The order, from best to worst performing technique, is now as follows: *Manual, Ref-LID, Single, Multi* and *All-four*. While the differences between *Ref-LID, Single* and *Multi* are small but consistent, there is a much larger performance difference between the reference dictionary and the above three, as well as between these three and the *All-four* dictionary.

In Table V, we observe that the same trends are observed across dictionary approaches (*Manual, Ref-LID, Single, Multi,* and *All-four*) regardless of the language model used. We also observe that more variants tend to result in poorer ASR performance. While this observation does not hold for the *Ref-LID* dictionary, this dictionary benefits more from prior information relating to the true source languages of words before G2P conversion. Note that the drop in performance of the recognition between *Ref-LID* and *Manual*

dictionaries correspond to errors made by the G2P converter, while the difference in WER between the two LID-based dictionaries and *Ref-LID* can be associated with the LID prediction error.

VI. Conclusion

This work focused on the implications of producing language-based pronunciation variants for proper name recognition, where different LID techniques are used to predict the most probable source language(s) of a word. Both G2P and ASR performance were analysed and compared.

To understand the implications of creating LID-based dictionaries, we considered four dictionaries that were generated using a combination of LID and G2P prediction. The same language-specific G2P predictors were used for all dictionaries but different LID options were evaluated: when the true source language is known, when a single source language is predicted, when multiple source languages are predicted, and when it is simply assumed that all words may be from all relevant source languages. These dictionaries were evaluated against a reference dictionary (developed with each of the corpora and manually corrected/verified) to measure G2P accuracy. ASR performance was evaluated by developing a full-blown ASR system and evaluating performance with both a flat and trained LM.

The extent to which LID accuracy influences ASR performance is most visible from the results in Table V. The effect of improved LID tags can be observed, with the *Single* approach performing the best of the predicted tags. While the G2P error accounts for the WER difference between *Ref-LID* and *Manual* results, the difference between the *Single* and the *Ref-LID* results provides a measure of the effect of remaining LID prediction error

During G2P analysis, it became clear that the way in which variants are dealt with during accuracy calculation has a large effect on measured performance. Using existing G2P accuracy measures, variants in the hypothesised dictionary are not penalised sufficiently, an issue we aim to address in future work.

VII. Acknowledgment

We would like to acknowledge Charl van Heerden for his assistance with the ASR experiments, as well as Ulrike Janke for her editing assistance. This work was partially supported by the National Research Foundation (NRF). Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF do not accept any liability in regard thereto.

References

[1] A. Font Llitjos and A. W. Black, "Knowledge of language origin improves pronunciation accuracy of proper names." in *Proc. INTERSPEECH*, Aalborg, Denmark, 2001, pp. 1919–1922.

[2] M. M. J. G. K. D. Spiegel, Murray F., "Synthesis of names by a demisyllable-based speech synthesizer (SPOKESMAN)," in *Proc. EUROSPEECH*, 1989, pp. 1117–1120.

[3] D. Gibbon, R. Moore, and R. Winski, *Handbook of standards and resources for spoken language systems*. Walter de Gruyter, 1997.

[4] T. Vitale, "An algorithm for high accuracy name pronunciation by parametric speech synthesizer," *Computational Linguistics*, vol. 17, no. 3, pp. 257–276, 1991.

[5] M. Kgampe and M. H. Davel, "Consistency of cross-lingual pronunciation of south african personal names." Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 2010.

[6] M. Davel, E. Barnard, C. van Heerden, W. Hartmann, D. Karakos, R. Schwartz, and S. Tsakalidis, "Exploring minimal pronunciation modeling for low resource languages," in *Proc. INTERSPEECH*, 2015, pp. 538–542.

[7] A. F. Llitjós, "Improving pronunciation accuracy of proper names with language origin classes," in *ESSLLI Student Session*. Citeseer, 2001, p. 53.

[8] M. F. Spiegel, "Proper name pronunciations for speech technology applications," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 419–427, 2003.

[9] B. Réveil, J.-P. Martens, and B. Dhoore, "How speaker tongue and name source language affect the automatic recognition of spoken names," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 2995–2998.

[10] Q. Yang, J.-P. Martens, N. Konings, and H. v. d. Heuvel, "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names," in *Proc. of the 3rd International Conference on Language Resources and Evaluation LREC06*, Genova, Italy, 2006, pp. 287–292.

[11] A. Bhargava and G. Kondrak, "Language identification of names with SVMs," in *Proc. North American Chapter of the Association of Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Los Angeles, CA, 2010, pp. 693–696.

[12] O. Giwa and M. H. Davel, "Language identification of individual words with joint sequence models," in *Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 14-18 September, Singapore, 2014, pp. 1400–1404.

[13] O. Giwa and M. Davel, "Text-based language identification of multilingual names," in *Proc. Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 166–171.

[14] Thirion, Jan W.F. and van Heerden, Charl and Giwa, Oluwapelumi and Davel, Marelie H., "The South African Directory Enquiries (SADE) corpus," in preparation.

[15] O. Giwa, M. H. Davel, and E. Barnard, "A Southern African corpus for multilingual name pronunciation," in *Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Vanderbijlpark, South Africa, 2011, pp. 49–53.

[16] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT Dictionaries: Project Report," Multilingual Speech Technologies, North-West University, Tech. Rep., May 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[17] E. Barnard, M. H. Davel, C. J. V. Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, 2014, pp. 194–200.

[18] J. W. Thirion, M. H. Davel, and E. Barnard, "Multilingual pronunciations of proper names in a southern african corpus," in *23st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2012, pp. 102–108.

[19] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *of the third International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTUÍ2)*, 2012, pp. 68–75.

[20] M. Kgampe and M. H. Davel, "Consistency of cross-lingual pronunciation of South African personal names," in *Proc. 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Nov. 2010, pp. 123–127.

[21] M. Davel and E. Barnard, "Pronunciation prediction with default & refine," *Computer Speech and Language*, vol. 22, no. 4, pp. 374–393, 2008.

[22] O. Giwa, "Language identification for proper name pronunciation," PhD thesis, North-West University (South Africa), Vaal Triangle Campus, 2016.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No.: CFP11SRW-USB.

[24] M. H. Davel, C. J. van Heerden, and E. Barnard, "G2P variant prediction techniques for ASR and STD," in *Proc. INTERSPEECH*, 2013, pp. 1831–1835.