Acoustic model optimisation for a call routing system

Neil Kleynhans*, Raymond Molapo* and Febe de Wet*†
*Human Language Technologies Research Group Meraka Institute, CSIR, South Africa
*†Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
Email: {ntkleynhans,rmolapo,fdwet}@csir.co.za

***Abstract*—The paper presents work aimed at optimising acoustic models for the AutoSecretary call routing system. To develop the optimised acoustic models: (1) an appropriate phone set was selected and used to create a pronunciation dictionary, (2) various cepstral normalization techniques were investigated, (3) three South African corpora and multiple training data combinations were used to train the acoustic models, and, (4) model-space transformations were applied. Using an independent testing corpus, which contained proper names and South African language names, a named-language recognition accuracy of 95.11 % and proper name recognition accuracy of 93.31% were obtained.**

I. Introduction

Interactive voice response (IVR) systems are widely used by companies to automatically assist their clients. The automation of services can greatly reduce company costs and in certain instances can be used by company staff to improve their productivity. Through Dual Tone Multi-Frequency (DTMF) keypads and Automatic Speech Recognition (ASR), IVR systems can capture digit information (such as account numbers) and more sophisticated information via a person's speech (e.g. person's name and surname). Unfortunately, DTMF input has an innately low information carrying capacity which is largely limited to digit-centric information. To overcome DTMF shortcomings, adding a natural spoken input and ASR information extraction capability can greatly increase the versatility of an IVR system.

A typical IVR application that makes use of speech processing capabilities is a call routing service, i.e. a system that routes incoming calls automatically to appropriate services or individuals. One such system is the AutoSecretary system introduced by Modipa *et. al.* [1], which routes incoming calls to a person based on a spoken name.

In this paper we describe the development of acoustic models for the AutoSecretary IVR application. Specifically, we focused on acoustic model optimisations which would:

- enable the system to route calls to an operator based on the callers language preference, and,
- allow new names to be added to the system relatively easily.

The next Section II describes the AutoSecretary system and provides background on some application-specific ASR issues. Section III details the ASR development effort as well as corpus selection and design. Our experiments are described in Section IV and results and a discussion are presented in Section V. Lastly, the conclusion and possible future work appear in Section VI.

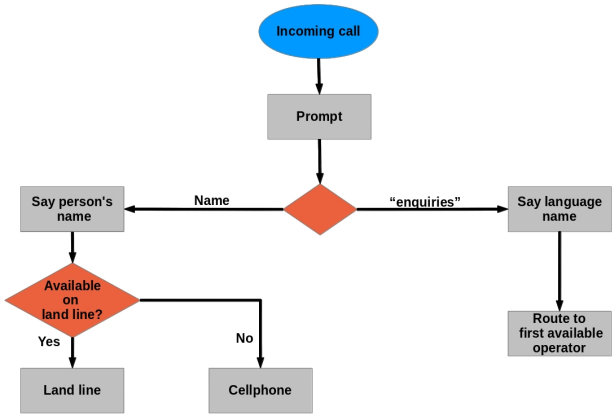

Fig. 1. High level AutoSecretary call flow.

II. Background

A. AutoSecretary IVR System

Figure 1 shows the high level call flow of the AutoSecretary call routing system. At the beginning of a session the system prompts the caller to say the name of person they are looking for or the word "*enquiries*". Following a valid name request the system will route the caller to the registered land line number. In addition, the system has the ability to route to a mobile number if it could not make a connection via the land line. If the word "*enquiries*" was spoken instead of a name, the system prompts the user for a language option - any of the eleven official South African languages - which allows the system to route the call to an operator who speaks the requested language.

The simple confidence scoring method implemented by ATK [2], is used to make a decision to either accept the recognition output if the confidence score is high or re-prompt the user to repeat their request if the confidence score is too low. Following two successive re-prompts, the system will automatically route the caller to a default operator. Figure 2

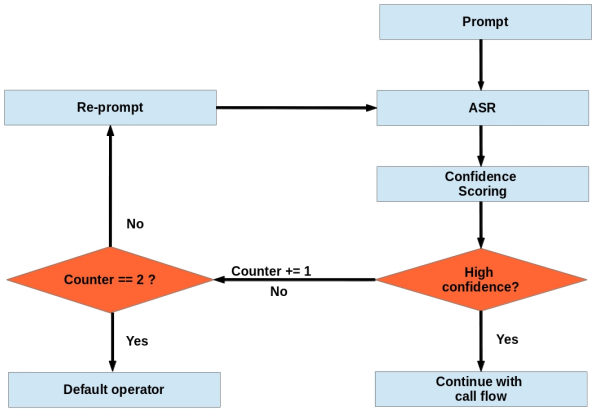

Fig. 2. AutoSecretary confidence scoring mechanism.

shows the AutoSecretary confidence scoring mechanism and how its application to the call flow.

On all successful recognitions, the system will parrot back the name to verify the selection. The caller may interrupt the transfer if the system selection is incorrect by saying "*stop*". The AutoSecretary system previously described by Modipa *et. al.* [1] had a similar call flow but did not include the functionality to route a user to an operator that spoke a particular language.

B. AutoSecretary ASR

The main issue in developing robust acoustic models for the AutoSecretary system is accurate proper name recognition. This type of problem has been encountered previously in directory assistance systems [3] and voice-navigation systems [4].

The first challenge in achieving accurate proper name recognition is robust pronunciation modelling. Initially, a phone set that can effectively represent the speech acoustic space must be chosen. This becomes an important aspect when dealing with multilingual environments which, in general, contain many sound classes and require careful phone set selection. Another major problem is that the phonemic representation of a word and the way it is pronounced vary greatly [4]. A possible cause of this mismatch is that people are altering the way in which they are pronouncing the proper name [5] based on what they think the word should sound like. This generally happens when an unknown or foreign word has to be spoken and the speaker has no prior knowledge. In multilingual environments this problem increases and becomes more difficult to solve as more languages are added. A way in which to partly overcome this problem, is to add multiple pronunciation variants to the pronunciation dictionary [1]. Adding pronunciation variants is a manually intensive task but affords greater accuracy compared to automatic methods. Automatic methods, such as G2P, have been shown to work well for common words but extracting rules for proper names still proves to be difficult [6]. Accurate proper name prediction is made difficult because proper names do not have set ways of pronouncing them [4], which makes robust rule extraction hard to accomplish. Also, predicting foreign words adds to incorrect pronunciations [5].

The second challenge is to develop robust acoustic models. In the standard HMM paradigm, creating word-based Hidden Markov Models is infeasible. As reported in [7], in the United States alone there were an estimated 1.5 million surnames with a third of these being unique. In multilingual environments, such as South Africa, these numbers would increase drastically. Another point of failure for word-based HMMs is the effort required to add new names to the system. Thus, a better approach would be to follow a large vocabulary ASR system development cycle. Here, development corpus selection is important as one would require large amounts of data to train robust acoustic models. A benefit of large vocabulary ASR systems is that they allow easier modification of the recognition grammar - for instance adding language name recognition - which adds flexibility to the system. Collecting a corpus of names per application [1] would be impractical as this would not in general produce robust acoustic models. In addition, if one would require the system to be re-deployed, a time consuming audio data collection process would have to be run before the system can be reliably operated in a new environment.

Corpus Name	# utterances	duration in hours
Lwazi English	5843	5.03
Lwazi English plus Lwazi language prompts	7770	5.57
NCHLT English	106018	76.97
AST English (5 dialects)	51745	29.80

TABLE I
THE NUMBER OF TRAINING UTTERANCES AND DURATION FOR EACH DEVELOPMENT CORPUS.

III. ASR DEVELOPMENT

In this section we describe the speech corpora used for acoustic model development, the phone set selection and pronunciation dictionary creation, the feature extraction process, acoustic model development as well as the recognition grammar and concept mapping that were used during system evaluation.

A. Training Corpora

To enable robust acoustic model development in a multilingual South African context we focused on three South African corpora. Table I shows the number of training utterances per corpus and indicates the duration in hours.

1) Lwazi: The Lwazi corpus contains annotated telephony speech data covering eleven South African languages [8], [9]. Each language-specific corpus was produced by collecting read and elicited speech data from approximately 200 speakers; with each speaker contributing roughly 30 utterances [9]. A portion of the utterances were randomly selected from a phonetically balanced corpus and the remainder are words or short phrases. Importantly, each corpus contains utterances

which captured the response of the speakers when queried about their first language.

2) NCHLT: The NCHLT ASR corpus contains annotated high-bandwidth speech data collected for eleven South African languages [10]. The individual corpora contain a minimum of 50 hours of speech data collected from 200 speakers (gender-balanced) with each speaker contributing in the order of 500 utterances. The volume of collected data improves triphone coverage and should make it easier to add new names or short phrases to the recognition grammar.

3) AST: The African Speech Technology (AST) corpus contains annotated telephony speech data for five South African languages [11]. The speech data was collected from 300 - 400 speakers and the prompts were chosen to support information retrieval, transactional teleservices and hotel booking applications. Given the prompt design, the corpus contains a large proportion of proper names and a good coverage of language prompts. Additionally, the English corpus contains data collected from five common South African English accents which should add to the robustness of the acoustic models. In the current investigation the same train, development and test sets were used as those described in Kamper *et. al.* [12].

B. Testing Corpus

A testing corpus was developed by, firstly, expanding the recognition grammar to create text prompts and then collecting speech data from a variety of speakers. The testing corpus contained speech data from approximately 20 unique speakers with each speaker contributing 22 names- and 46 language-specific utterances. The data was collected from both land line and mobile handsets which represents a close approximation to the proper testing environment. After manual validation, the final utterance count was 555 names-related and 1003 language-related utterances. The duration of the testing audio at this point was 1.42 hours. To increase the testing data size further, we included a previously collected name-surname corpus which contained 31 unique name-surname pairs. The final testing corpus contained 2.13 hours of audio data, 1480 names-related and 1003 language-related utterances.

C. Phone Set and Pronunciation Dictionary

The initial phone set was a union of all the phones found in the Lwazi corpus [8] and consisted of 87 unique phones. These were then mapped to a simplified set of 62 phones where affricates were split (*e.g.* [tS] $\rightarrow$ [t] [S], [d_0Z] $\rightarrow$ [d] [Z]) and clicks and subtle phone distinctions merged (*e.g* [h\] became [h], etc.). The motivation for simplifying the phone set is that multilingual speakers will probably not pronounce the distinctions correctly, thus removing them from the start would be better.

The corpora-specific pronunciation dictionaries were mapped to the simplified 62 phone set. As the majority of the training corpora used in our investigation were South African English (SAE) the final phone set only contained 41 South African English phones. The reduction in the number of phones, is due to English not containing phones which occur in other languages. As a final phase, foreign words had phonemic representation generated manually using the closest English phones.

The recognition pronunciation dictionary or AutoSecretary dictionary contained 158 unique entries which included multilingual person and South African language names as well as a few English honorifics (ms, mr, mrs, dr). With pronunciation variants this count increased to 415. The 41 phones in the English set were used to manually create all the pronunciations.

D. Feature Extraction

39 (13 static, 13 delta and 13 delta-delta) dimensional Mel Frequency Cepstral Coefficient (MFCC) features were generated using the Application Toolkit (ATK) [2] and the Hidden Markov Model Toolkit (HTK) [13]. These MFCCs were extracted every 10 ms from a 25 ms speech frame. The frequency bandwidth was limited to 150-3600 Hz and is applied by HTK independent of sampling rate.

Channel normalisation was performed by means of cepstral mean normalization (CMN). Four different options were considered, i.e. *no CMN*, *HTK CMN*, *Global CMN*, and *ATK CMN*.

HTK CMN is implemented by estimating a cepstral mean vector on a per utterance basis and removing the cepstral vectors' offset [13]. The *Global CMN* method estimates a cepstral mean vector from the entire training data set and then uses the vector to normalize the training and testing cepstral vectors. *ATK CMN* is implemented by first loading an initial mean vector which, for our experiments, was a global mean cepstral vector estimated on the training data [2]. This cepstral mean is updated on every *speech* frame according to the formula:

$$\boldsymbol{\mu'} = \alpha(\boldsymbol{\mu} - \boldsymbol{x})\boldsymbol{x}, \tag{1}$$

where $\boldsymbol{\mu'}$ is the updated cepstral mean, α is the time constant set to 0.995, and $\boldsymbol{x}$ the input cepstral vector. For each utterance, ATK resets the mean cepstral vector to the initial mean vector $\boldsymbol{\mu_0}$. To determine whether a frame is speech, ATK uses the first 40 frames of each utterance to train a silence detector and performs a speech / non-speech analysis on each frame. The first 10 frames of an utterance are not used to update the mean cepstral vector.

When experimenting with a specific CMN approach both the training and testing data were normalized using the same CMN technique.

E. Acoustic modelling

A standard acoustic model development strategy was used as detailed in HTK book [13]. The acoustic models were tied-state context-dependent (triphone) Hidden Markov Models (HMMs), using a three state left-to-right topology. Question-based tying was used to create the tied-state models. Eight Gaussian mixtures per HMM state were used to model the cepstral densities. Different sets of acoustic models were

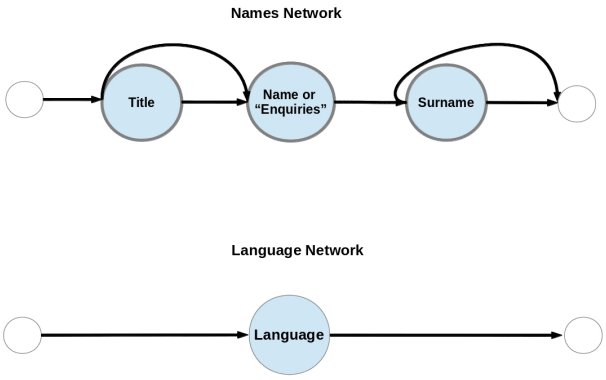

Fig. 3. The AutoSecretary name and language recognition networks.

created using the corpora described in Section III-A as well as using combinations of some of the corpora.

F. Recognition Grammar and Concept Mapping

The test set vocabulary contained 40 unique name-surname pairs and 11 unique language options. The recognition networks for names and languages are shown in Figure 3.

Expanding the recognition network provides 156 name and 46 language possibilities. The name network also contained the following words: "*enquiries*", "*switchboard*", "*reception*", and "*stop*".

During normal AutoSecretary operation the application is only aware of the unique name and language options and maps the expanded ASR text output. For example, the following mappings would be performed:

- "*Mr John Doe*" or "*John Doe*" mapped to "*John Doe*"
- "*Sesotho sa leboa*" or "*Northern Sotho*" mapped to "*Sepedi*"

For system performance evaluation we defined the various unique names and language options as "concepts" and performed "concept mapping" which reduced the expanded ASR recognition output to their unique name and language equivalents. When reporting the system results we report on "concept" accuracies unless otherwise stated.

IV. Experiments

A. Training data combinations and cepstral normalization

In addition to the English sub-corpora of the Lwazi, AST and NCHLT corpora, different combinations of the various corpora were used as training data. We defined the data combinations as follows:

1) *Lwazi English + Langs*: Training data pooled from the English sub-corpus of Lwazi and all language prompts from the remaining 10 language-specific corpora.
2) *Lwazi English + Langs + AST*: Training data pooled from (1) and the five AST English dialects.
3) *Lwazi English + Langs + NCHLT*: NCHLT English sub-corpus added to (1).
4) *Lwazi English + Langs then AST*: (1) was used to train single mixture tied-state HMMs. Then, the five AST English dialects data was added to the training data and used during mixture incrementing.
5) *Lwazi English + Langs then NCHLT*: Similar to (4) except that the NCHLT English sub-corpus was used instead of the AST data.

The four different options for channel normalization described in Section III-D ("*No CMN*", "*Global CMN*", "*HTK CMN*" and "*ATK CMN*") were tested in combination with each of these training sets.

B. Semi-tied versus Constrained MLLR

HMM-based large vocabulary ASR systems generally use diagonal covariance matrices to reduce the number of model parameters. Full covariances matrices, however, are able to model the non-Gaussian nature of data which could potentially provide an increase in accuracy. Semi-tied transformations [14] transform diagonal matrices into full covariance matrices but instead of estimating state-specific transformations, estimate class-specific transforms. These classes are usually defined by a regression class tree which groups similar HMM states together [13]. In this way the parameter count may be kept relatively low which prevents excessive recognition times. The semi-tied transform is defined as:

$$\boldsymbol{\Sigma}^{(m)} = \boldsymbol{H}^{(r)}\boldsymbol{\Sigma}^{(m)}_{diag}\boldsymbol{H}^{(r)T}, \tag{2}$$

where $\boldsymbol{\Sigma}^{(m)}$ is the component-specific diagonal covariance matrix and $\boldsymbol{H}^{(r)}$ is the class-specific semi-tied transform. Unfortunately, ATK does not implement semi-tied transforms but does support constrained maximum likelihood linear regression (CMLLR) transforms [15], [13]. Constrained MLLR is typically used for speaker and channel adaptation and performs the adaptation by transforming the mean and covariance components in the HMM set. If one compares the semi-tied and the CMLLR transforms, the forms are quite similar. The CMLLR transform is defined as:

$$\hat{\boldsymbol{\mu}}^{(m)} = \boldsymbol{H}^{(r)}\boldsymbol{\mu}^{(m)} + \boldsymbol{b}^{(r)}, \; \hat{\boldsymbol{\Sigma}}^{(m)} = \boldsymbol{H}^{(r)}\boldsymbol{\Sigma}^{(m)}_{diag}\boldsymbol{H}^{(r)T}, \tag{3}$$

where $\hat{\boldsymbol{\mu}}^{(m)}$ and $\hat{\boldsymbol{\Sigma}}^{(m)}$ are the transformed component-specific mean and covariance matrices, $\boldsymbol{\mu}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$ are the original component-specific mean and covariance matrices and $\boldsymbol{H}^{(r)}$ and $\boldsymbol{b}^{(r)}$ are the class-specific CMLLR transforms.

Both methods iteratively solve for the transform parameters by optimising a modified Expectation-Maximization auxiliary function. The auxiliary functions, found in [14] and [13], highlight the differences in the equations which change the iterative optimisation equations. As a final experiment we wanted to determine whether CMLLR transforms could perform comparably to semi-tied transforms.

		System B	
		# Correct	# Incorrect
System A	# Correct	w	x
	# Incorrect	y	z

TABLE II
A 2X2 CONTINGENCY TABLE USED IN A MCNEMAR'S TEST.

C. McNemar's Test

McNemar's test can be used to establish whether the differences in error-rates, produced by two systems, are statistically significant [16]. This test requires that the errors produced by the system are independent events and in terms of speech recongition, can be used to test isolated-word recognition results [16]. The first step in performing a McNemar's test is to create a 2x2 contingency table as shown in Table II.

From this, we define the *null* and *alternative* hypotheses as,

$$\begin{aligned} H_0 : p_x &= p_y \\ H_1 : p_x &\neq p_y \end{aligned}$$

The test statistic is a one degree of freedom chi-squared distribution (χ^2) with Yates's correction for continuity [17] and is given by,

$$\chi^2 = \frac{(\mid x - y \mid -0.5)^2}{x + y}. \tag{4}$$

The null hypothesis can be rejected or accepted by calculating the two-sided P-value of the χ^2 distribution and comparing it to standard significance levels of 0.05, 0.01 or 0.001.

V. RESULTS AND DISCUSSION

In this section we present results on various training data combinations and cepstral normalization techniques which were used to perform acoustic model optimisations for the AutoSecretary system. We also show results around our hypothesis that CMLLR can be used as an approximate replacement for semi-tied transforms. Throughout this section the tables show concept accuracies (refer to Section III-F for a description) unless otherwise stated. Training data combinations and their labels are detailed in Section IV-A and cepstral normalization techniques are described in Section III-D.

A. Training data combinations and cepstral normalization

Table III shows language and name concept accuracies for different training data combinations, various training schemes and cepstral normalization techniques. Focusing on the cepstral normalization techniques (compare results within rows) we can see that some normalization methods produced surprising results. The "*HTK CMN*" produced the worst results which indicates that for this type of application utterance-based normalization is not ideal. This may be due to the short testing utterances which are often less than a second in duration and long-term biases are not estimated properly.

The "*No CMN*" and "*Global CMN*" results are quite similar which indicates that "*Global CMN*" did not perform effective normalization. In the majority of cases (except for language experiments using "*Lwazi Eng + Langs*" and "*Lwazi Eng + Langs then NCHLT*" data combinations) the ATK normalization proved to be the best cepstral normalization approach. The "*ATK CMN*" normalization method begins with the same initial mean cepstral vector as the "*Global CMN*" normalization but adapts the mean cepstral vector as it progresses through the utterance and only updates on speech frames. This selective adaptation seems to provide a good normalization mechanism. Previously it was observed by Modipa *et. al.* that there was a large discrepancy between the off-line and online ASR accuracies. A possible cause could be the differences in HTK and ATK cepstral normalization procedures.

Turning to the language recognition results and considering only our best normalization method (compare results within the *ATK CMN* column), we see that adding the Lwazi language prompts gave quite a large boost in performance, which was to be expected. Surprisingly, the *AST* and *NCHLT* only experiments produce rather poor results. In the case of *NCHLT*, this may be put down to a channel mismatch as the *NCHLT* corpus contains high-bandwidth audio data. More investigation is needed to establish why the *AST* data performed so badly. Combining data ("*Lwazi Eng + Langs + AST*", "*Lwazi Eng + Langs + NCHLT*") resulted in a slight increase in performance when adding the *AST* data but did not achieve any increase in accuracy when adding the *NCHLT* data. Training a system on the "*Lwazi English + Langs*" then adding *AST* for mixture incrementing produced the best results. It is interesting that state-tying on the smaller "*Lwazi English + Langs*" corpus resulted in an increase in performance. Further investigation is needed to determine why state-tying on a smaller corpus produced such an increase and to establish whether such a gain would be seen if the testing vocabulary was much larger. The last experiment, where *NCHLT* was used for mixture incrementing manage to achieve a slight increase in accuracy.

For name recognition (compare results within the *ATK CMN* column), the *AST* data and combinations with the AST data produced the top results with the "*Lwazi Eng + Langs then AST*" producing the best names recognition performance. The "*Lwazi Eng + Langs then NCHLT*" produced the best result out of the non-*AST* experiments but other *NCHLT* combinations performed marginally better or worse than the "*Lwazi Eng + Langs*" data combination.

B. Semi-tied versus CMLLR

In Section IV-B we speculated if it were possible to use CMLLR as a semi-tied replacement since ATK does not support semi-tied transformations. Table IV shows name and language concept recognition accuracies for various training data combinations and using either no, semi-tied or CMLLR transformation. ATK cepstral normalization was used for all the experiments.

If one compares the semi-tied and CMLLR columns of Table IV, for both language and name recognition results, we can see that the semi-tied approach outperforms CMLLR technique in the vast majority of the experiments (12 out of

	No CMN	HTK CMN	Global CMN	ATK CMN
Lwazi Eng	85.33 / 78.72	60.88 / 70.14	85.33 / 78.92	87.13 / 83.18
Lwazi Eng + Langs	93.11 / 80.27	68.66 / 75.88	93.01 / 80.27	92.81 / 84.19
AST	58.78 / 60.14	64.57 / 72.70	60.58 / 60.07	75.25 / 77.43
NCHLT	79.84 / 78.85	67.07 / 74.73	79.54 / 78.65	80.44 / 81.28
Lwazi Eng + Langs + AST	90.12 / 77.84	69.36 / 77.64	89.52 / 77.77	93.71 / 87.91
Lwazi Eng + Langs + NCHLT	89.92 / 82.36	73.35 / 78.85	90.82 / 82.16	92.81 / 84.46
Lwazi Eng + Langs then AST	86.53 / 90.27	75.85 / 84.80	85.83 / 89.93	**95.11 / 93.31**
Lwazi Eng + Langs then NCHLT	91.82 / 87.70	77.84 / 83.92	92.22 / 87.70	91.92 / 89.46

TABLE III
Language AND *Name* CONCEPT ACCURACIES (%) FOR VARIOUS TRAINING DATA COMBINATIONS AND CEPSTRAL NORMALISATION TECHNIQUES. THE RESULTS ARE PRESENT IN PAIRS - LANGUAGE ACCURACY % / NAME ACCURACY %.

	None	Semi-tied	CMLLR
Lwazi Eng	87.13 / 83.18	86.43 / 83.65	85.83 / 81.82
Lwazi Eng + Langs	92.81 / 84.19	93.21 / 84.26	92.81 / 83.72
AST	75.25 / 77.43	78.64 / 81.42	78.54 / 78.78
NCHLT	80.44 / 81.28	78.34 / 81.28	80.64 / 79.19
Lwazi Eng + Langs + AST	93.71 / 87.91	93.01 / 89.32	94.01 / 87.23
Lwazi Eng + Langs + NCHLT	92.81 / 84.46	93.71 / 84.32	93.51 / 83.31
Lwazi Eng + Langs then AST	**95.11** / 93.31	93.71 / **94.32**	94.91 / 93.65
Lwazi Eng + Langs then NCHLT	91.92 / 89.46	91.22 / 89.46	91.32 / 88.85

TABLE IV
Language AND *Name* CONCEPT ACCURACIES (%) FOR VARIOUS TRAINING DATA COMBINATIONS AND SEMI-TIED AND CMLLR TRANSFORMATION TECHNIQUES. THE RESULTS ARE PRESENT IN PAIRS - LANGUAGE ACCURACY % / NAME ACCURACY %.

16), however, the differences in accuracies are relatively small. To investigate whether there was any significant difference between the semi-tied and CMLLR results, McNemar's test was used to analyse the recognition outputs. Refering to the fourth column of Table V, we can see that only the "*Lwazi Eng*", "*AST*" and "*Lwazi Eng + Langs + AST*" experiments produced a significant difference in the results, if one chooses a significance level of 0.05. At a stricter significance level, 0.001, all the null hypothesis would be accepted, which implies that the semi-tied and CMLLR are quite similar.

Comparing the McNemar's test P-values, calculated between semi-tied and no transform (column two Table V) and CMLLR and no-transform (column three Table V), we can see that only a few experiments produced a significant difference between the results. These are semi-tied and CMLLR "*AST*" experiments, CMLLR "*Lwazi Eng*" experiment and CMLLR "*NCHLT*" experiment. The remaining results (12 of 16) allow us to accept the null hypothesis and conclude, for these experiments, that using semi-tied or CMLLR transforms does not produce a significant increase or decrease in accuracy, as compared to a ASR system that does not implement these transforms.

To investigate further we performed a few experiments with the *Timit* and *NTimit* corpora. The results are presented in Table VI and indicate word accuracies in percent. The standard ASR system was developed (see Section III-E) and a flat recognition grammar was used which only contained words from the testing vocabulary. ATK cepstral normalization was utilized.

The results in Table VI show that semi-tied transforms provide an increase in accuracy for within corpus experiments but

Training Corpus	Testing Corpus	
	Timit	NTimit
Timit with semi-tied	60	10.11
Timit	56.60	19.44
NTimit with semi-tied	16.92	46.50
NTimit	23.38	43.46

TABLE VI
WORD ACCURACIES (%) WHEN BASELINE AND SEMI-TIED TRANSFORMS SYSTEM ON THE *Timit-NTimit* CORPORA.

for cross-corpus experiments applying semi-tied transforms reduced the ASR accuracy. The semi-tied transforms seem to amplify the data mismatch and thus decrease performance. This might explain why semi-tied transforms did not produce an average gain in performance for the AutoSecretary ASR models since there are slight mismatches between the training and testing environments which were amplified by the transform.

VI. CONCLUSION AND FUTURE WORK

The paper presented work aimed at optimising acoustic models for the AutoSecretary call routing system. The optimised acoustic models were developed by:

- creating a modified South African English phone set and an appropriate pronunciation dictionary,
- investigating various cepstral normalization techniques,
- experimenting with three South African corpora and training data combinations, and,
- applying model-space transformations.

The pronunciation dictionary contained a simplified South African English phone set which was used to robustly represent the acoustic sounds found in the South African multilin-

	None & Semi-tied	None & CMLLR	Semit-tied & CMLLR
Lwazi Eng	1.00000	0.00626	0.01383
Lwazi Eng + Langs	0.71830	0.52480	0.28980
AST	1.2e-08	0.00017	0.00347
NCHLT	0.19390	0.02830	0.62410
Lwazi Eng + Langs + AST	0.21100	0.50500	0.04658
Lwazi Eng + Langs + NCHLT	0.60010	0.38650	0.16210
Lwazi Eng + Langs then AST	1.00000	0.82200	0.90350
Lwazi Eng + Langs then NCHLT	0.59440	0.15520	0.51570

TABLE V
P-values CALCULATED USING MCNEMAR'S TEST, FOR VARIOUS TRANSFORMATION COMBINATIONS (NONE, SEMI-TIED AND CMLLR).

gual acoustic space. Each name and language entry contained multiple pronunciation variants to cope with the variability found in proper name pronunciation. For future work an investigation into automatically generating proper name pronunciations and variants should be performed to reduce the amount of manual intervention required during dictionary development. An automatic pronunciation-prediction method will also help to rapidly customize the AutoSecretary application.

The choice of cepstral normalization technique is important since the approach used to normalize the training and testing data does affect the results produced by the ASR system, as was shown by the results captured in Section V-A. The ATK normalization method proved to be the best approach while the generally used utterance-based normalized performed poorly.

Our data combination experiments showed that the best training corpus was a combination of "*Lwazi English, Lwazi language prompts and AST*". Specifically, by developing a tied-state ASR system on the Lwazi English and Lwazi language prompts, then adding the *AST* data for mixture incrementing we managed to achieve a language recognition accuracy of 95.11% and a name recognition accuracy of 93.31% on an independent test corpus. These optimised acoustic models should:

- with high accuracy be able to detect a spoken South African language name which the system can use to route a caller based on language preference, and,
- accurately recognize new names provided that an adequate number of accurate pronunciations and relevant variants are included in the pronunciation dictionary.

Surprisingly, the larger *NCHLT* corpus did not provide substantial gains in accuracy and in some cases no gains were achieved. The most likely explanation is that the data mismatch hindered its effectiveness due to the corpus containing high-bandwidth recordings instead of telephony recordings which make up the AST corpus.

In Section IV-B we postulated that CMLLR could be used as an approximate replacement for semi-tied transforms. Our results in Section V-B showed that overall the accuracies produced by both methods are quite similar and only three experiments showed statistical significant results. Furthermore, when comparing the results between systems that did not implement semi-tied or CMLLR trasforms, to those that did, the vast majority of experiments failed to produce statistcally significant improvements.

Our results indicated that although semi-tied transforms can increase the ASR system performance when the training and testing data are relatively matched, care should be taken when applying the transform when there is a data mismatch as this could degrade the system performance.

ACKNOWLEDGEMENTS

We would like to thank the following people:

- The sprint team: Marelie Davel, Charl van Heerden, Nic de Vries, Thipe Modipa, Willem Basson
- Bryan Mcalister for helping us with the testing data collection
- Herman Kamper and Thomas Niesler for sharing their AST experimental set-up with us.

REFERENCES

[1] P. Modipa, F. de Wet, and M. Davel, "ASR performance analysis of an experimental call routing system," in *20th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Stellenbosch, South Africa, Nov. 2009, pp. 127–130.

[2] S. Young. (2012) ATK Manual. [Online]. Available: http://mi.eng.cam.ac.uk/research/dialogue/ATK_Manual.pdf

[3] F. Bechet, R. de Mori, and G. Subsol, "Very large vocabulary proper name recognition for directory assistance," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01).*, Madonna di Campiglio, Italy, Dec. 2001, pp. 222–225.

[4] B. Rveil, J.-P. Martens, and H. van den Heuvel, "Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010, pp. 2149–2154.

[5] A. F. Llitjos and A. W. Black, "Knowledge of language origin improves pronunciation accuracy of proper names," in *Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sep. 2001, pp. 1919–1922.

[6] O. Giwa, M. Davel, and E. Barnard, "A Southern African corpus for multilingual name pronunciation," in *22th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Vanderbijlpark, South Africa, Nov. 2011, pp. 49–53.

[7] M. Marx and C. Schmandt, "Putting people first: Specifying proper names in speech interfaces," in *Proceedings of the 7th annual ACM symposium on User interface software and technology*, Marina del Rey, CA, USA, Nov. 1994, pp. 29–37.

[8] Meraka Institute. (2012) Lwazi ASR corpus. [Online]. Available: http://www.meraka.org.za/lwazi

[9] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages." in *10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, Sep. 2009, pp. 2847–2850.

[10] N. J. de Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. de Waal, "Woefzela - an open-source platform for ASR data collection in the developing world," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, Aug. 2011, pp. 3176–3179.

[11] J. C. Roux, P. H. Louw, and T. R. Niesler, "The African Speech Technology project: an assessment," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004, pp. 93–96.

[12] H. Kamper, F. J. M. Mukanya, and T. R. Niesler, "Multi-accent acoustic modelling of South African English," *Speech Communication*, vol. 54, pp. 801–813, Feb. 2012.

[13] 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. (2012) HTK book. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[14] M. J. F. Gales, "Semi-tied covariance matrices for Hidden Markov Models," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 272–281, May 1999.

[15] V. Digalakis, D. Rtischev, L. Neumeyer, and E. Sa, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 357–366, Sep. 1995.

[16] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Glasgow, Scotland, May 1989, pp. 532–535.

[17] F. Yates, "Contingency tables involving small numbers and the χ^2 test," *Supplement to the Journal of the Royal Statistical Society*, vol. 1, no. 2, pp. 217–235, 1934.