Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems

AUTHORS:
Febe de Wet[1]
Neil Kleynhans[2]
Dirk van Compernolle[3]
Reza Sahraeian[3]

AFFILIATIONS:
[1]Human Language Technologies Research Group, Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
[3]Center for Processing Speech and Images, Department of Electrical Engineering, University of Leuven, Leuven, Belgium

CORRESPONDENCE TO:
Febe de Wet

EMAIL:
fdwet@csir.co.za

DATES:
Received: 04 Feb. 2016
Revised: 31 May 2016
Accepted: 24 Aug. 2016

KEYWORDS:
acoustic modelling; Afrikaans; Flemish; automatic speech recognition

HOW TO CITE:
De Wet F, Kleynhans N, Van Compernolle D, Sahraeian R. Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems. S Afr J Sci. 2017;113(1/2), Art. #2016-0038, 9 pages. http://dx.doi.org/10.17159/sajs.2017/20160038

ARTICLE INCLUDES:
✓ Supplementary material
× Data set

FUNDING:
Fund for Scientific Research of Flanders; National Research Foundation (South Africa); South African Department of Arts and Culture: Programme of Collaboration on HLT.

For purposes of automated speech recognition in under-resourced environments, techniques used to share acoustic data between closely related or similar languages become important. Donor languages with abundant resources can potentially be used to increase the recognition accuracy of speech systems developed in the resource poor target language. The assumption is that adding more data will increase the robustness of the statistical estimations captured by the acoustic models. In this study we investigated data sharing between Afrikaans and Flemish – an under-resourced and well-resourced language, respectively. Our approach was focused on the exploration of model adaptation and refinement techniques associated with hidden Markov model based speech recognition systems to improve the benefit of sharing data. Specifically, we focused on the use of currently available techniques, some possible combinations and the exact utilisation of the techniques during the acoustic model development process. Our findings show that simply using normal approaches to adaptation and refinement does not result in any benefits when adding Flemish data to the Afrikaans training pool. The only observed improvement was achieved when developing acoustic models on all available data but estimating model refinements and adaptations on the target data only.

Significance:

- Acoustic modelling for under-resourced languages
- Automatic speech recognition for Afrikaans
- Data sharing between Flemish and Afrikaans to improve acoustic modelling for Afrikaans

Introduction

Speech interfaces to different types of technology are becoming increasingly more common. Users can use their voice to search the Internet, control the volume of their car radio or dictate. However, this possibility is only available to users if the required technology exists in the language they speak. Automatic speech recognition (ASR) technology already exists and is regularly used by speakers of American English, British English, German, Japanese, etc. The development of ASR systems requires substantial amounts of speech and text data. While such resources are readily available for a number of languages, the majority of the languages that are spoken in the world can be classified as under-resourced, i.e. the resources required to create technologies like ASR do not exist or exist only to a limited degree. Researchers in the field of speech technology development for under-resourced languages are investigating various possibilities to address this challenge and to establish resources and technologies in as many languages as possible.

One of the strategies that has been explored is to fast-track progress in under-resourced languages by borrowing as much as possible – in terms of both data and technology – from well-resourced languages. Here we report on an investigation on data sharing between Afrikaans – an under-resourced language – and Flemish – a well-resourced language. The approach was focused on the exploration of model adaptation and refinement techniques associated with hidden Markov model (HMM) based speech recognition systems to improve the benefit of sharing data. The focus was specifically on the use of currently available techniques, some possible combinations and the exact utilisation of the techniques during the acoustic model development process.

Most of the techniques that are used in language and speech technologies are based on statistical methods. These methods require substantial amounts of data for a reliable estimation of the statistical parameters that are used to model the language, either in its written or spoken form. The required amounts often exceed what is available for resource-scarce languages.[1] The restricted resources that are available for these languages can be supplemented with resources from other languages, especially from those for which extensive resources are available. We investigated different possibilities to improve acoustic modelling in an under-resourced language, Afrikaans, by using data from a well-resourced language, Flemish. The techniques that were investigated include bootstrapping Afrikaans models using Flemish data as well as individual and combined model adaptation techniques.

Specifically, our aim throughout was to improve the performance of Afrikaans acoustic models by adding the Flemish data using various model adaptation and refinement approaches. As we focused on the model level, we utilised maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation as well as a combination of these adaptation techniques. In addition, heteroscedastic linear discriminant analysis (HLDA) and speaker adaptive training (SAT) acoustic model refinements were investigated in terms of sharing acoustic data. The purpose of investigating these techniques – described in later sections – is to determine whether these methods are sufficient in our data sharing scenario.

Background

Some of the approaches to data combination that have been reported on in the literature include cross-language transfer[2], cross-language adaptation[3], data pooling[2,4] as well as bootstrapping[5]. However, results as well as

conclusions vary between studies and seem to be highly dependent on the amount of data that is used and the specific modelling task investigated. Some studies report small gains under very specific conditions.

In a study by Adda-Decker et al.[6] in which no acoustic data were available for the target language (Luxembourgish), English, French and German data sets were used to train a multilingual as well as three monolingual ASR systems. Baseline models for Luxembourgish were subsequently obtained by using the International Phonetic Alphabet associations between the Luxembourgish phone inventory and the English, French and German phone sets. (A phone is the smallest discrete segment of sound in a stream of speech). Results showed that the language identity of the acoustic models has a strong influence on system performance with the German models yielding much better performance than the French or English ones. The acoustic data that were available for Luxembourgish were not enough to train a baseline system. It was therefore not possible to compare the performance of the German models with models trained on the target language.

Positive results were reported for multilingual acoustic modelling when only a small amount of training data was available for Dari, Farsi and Pashto.[7] MAP adaptation of the multilingual models to the individual target languages yielded a 3% relative improvement in word error rate compared to the corresponding monolingual models. However, as more data were added during training for the individual languages, the monolingual models overtook their multilingual counterpart very quickly in terms of recognition performance – given equal amounts of training data and the same number of model parameters.

Van Heerden et al.[4] found that simply pooling data for closely related languages resulted in improvements in ASR phone accuracies. They grouped languages according to expert knowledge of language families – Nguni and Sotho. The generally observed trend was that adding one to two languages gave slight improvements in accuracy – however, this trend was not observed for Sepedi. In addition, for the majority of cases, adding a third language to the training pool resulted in a decrease in accuracy (except for isiZulu). On average, each language contained about 7 h of audio training data, thus 14 h and 21 h of training data indicated improvement.

Niesler[8] investigated the possibility of combining speech data from different languages spoken in a multilingual environment to improve the performance of ASR systems for the individual languages. The systems were all HMM based. The recognition performances of language-specific systems for Afrikaans, South African English, isiXhosa and isiZulu were compared with that of a multilingual system based on data pooling as well as data sharing by means of decision-tree clustering. The clustering process was modified to allow for language-specific questions. Data from different languages could therefore be shared at HMM state level. The results of the study showed that the multilingual acoustic models obtained using this data sharing strategy achieved a small but consistent improvement over the systems that were developed for the languages individually or by just pooling the data.

Kamper et al.[9] performed several data sharing experiments on accented English audio data collected in South Africa. They specifically considered the accents of South African English defined in the literature: Afrikaans English, Black South African English, Cape Flats English, White South African English and Indian South African English. Overall they found that their multi-accent modelling approach outperformed accent-specific and accent-independent acoustic models. To create the multi-accent acoustic models, a modified decision-tree state cluster approach was used when accent-specific questions could be asked, which allowed the sharing of data across accents at the HMM state level. This approach is similar to that of Niesler[8] except accent questions were used instead of language-specific questions. Of interest, was the analysis of the proportions of data shared at the state level. It was found that the optimal phone and word operating points were different and that the amount of data shared at these points also differed – 33% and 44%, respectively.

A current popular trend for data sharing is to make use of deep neural networks (DNNs) for robust feature extraction, for which gains have been observed even for unrelated languages. Approaches mainly focus on bottleneck features with different network architectures and optimisations. Some examples of the bottleneck feature approach are described in Veselý et al.[10] (language-independent bottleneck features), Zhang et al.[11] (multilingual stacked bottleneck features), Nguyen et al.[12] (multilingual shifting deep bottleneck features) and Vu et al.[13] (multilingual DNNs cross-language transfer). Once the features are extracted they are fed through to a Gaussian mixture model (GMM)/HMM or Kullback–Leibler divergence based HMM (KL-HMM) system, where normal ASR techniques are applied. It is difficult to interpret how exactly the DNNs are combining the different data and what effective operation is being applied to the data, but it does seem that the DNNs are applying a necessary feature normalisation.[14] In line with this feature processing, there is great scope for improvement at the feature level as shown in intrinsic spectral analysis combination investigation.[15]

Monolingual acoustic modelling for Afrikaans has been investigated previously using a conventional Mel frequency cepstral coefficient (MFCC) based HMM system and broadcast news data[16] as well as using intrinsic spectral analysis in combination with a broadband, monolingual Afrikaans corpus[15].

In a study on resource and technology transfer between closely related languages, a case study was conducted for Dutch and Afrikaans. The distance between Afrikaans and other West Germanic languages and dialects was quantified in terms of acoustically weighted Levenshtein distances.[17] The results identified Dutch and Flemish as well-resourced, donor languages for the development of language and speech technology in Afrikaans, especially in terms of supplying background data for acoustic modelling (cf. Box 1). These results were confirmed by a series of experiments that investigated the possibility of improving acoustic modelling for Afrikaans by using Dutch, Swiss German and British English as background data in Tandem and KL-HMM ASR systems. The best results were obtained when Dutch was used as out-of-language background data.[18]

Box 1: Closeness

> In the context of statistical modelling 'closeness' is defined in terms of the acoustic distances between the languages. Phonetic and lexical overlap can also be taken into consideration to determine 'closeness'. Historical and linguistic considerations may be related to but are not always reflected in objective measures such as acoustic distance.

We report on an attempt to improve acoustic modelling for Afrikaans (as an example of a resource-scarce language) by borrowing data from Flemish (as an example of a well-resourced language). Flemish was chosen as the donor language because we had access to previously developed ASR systems for Flemish as well as the relevant data. It was also decided to start with Flemish rather than a combination of Flemish and Dutch as previous studies have shown that the two languages have distinctive acoustic properties and that better recognition results are obtained if they are first modelled separately and then combined.[19]

A previous study on this topic investigated the use of multilayer perceptrons, KL-HMMs and subspace Gaussian mixture models (SGMMs) and used Dutch as a donor language.[20] The systems based on SGMMs achieved the best monolingual as well as multi-lingual performance. When the models were trained on Dutch data and adapted using the Afrikaans data, the SGMM systems also yielded the best results. Overall, the results showed that Dutch/Afrikaans multilingual systems yield a 12% relative improvement in comparison with a conventional HMM/GMM system trained only on Afrikaans.

The literature review sketches a domain in which many approaches have been explored to enable speech recognition performance gains for under-resourced languages through data sharing, but the results are quite varied. In summary, our research reported here investigates the possibility of combining Flemish and Afrikaans data at the model level using model adaptation (MLLR and MAP) and refinement (HLDA and SAT) techniques as well as combinations thereof. Although the DNN and intrinsic spectral analysis feature approaches have yielded success, this investigation will not focus on these.

Data

In this study, Flemish was used as an example of a well-resourced language and Afrikaans as an example of a closely related but under-resourced language. The Flemish and Afrikaans speech data and pronunciation dictionaries are described in this section.

Box 2: Standard and 'less standard' varieties

The data sets that were used in this study were designed to include the standard varieties of the relevant languages. For most languages it is difficult – sometimes to the point of being controversial – to define exactly what a 'standard variety' is. The Flemish data correspond to radio news bulletins. Extreme varieties of a language are usually not used for news broadcasts, although we did not confirm this supposition in terms of internationally accepted news broadcasting standards. The National Centre for Human Language Technology Afrikaans data set has a 70:30 ratio of urban versus rural accents. The 'less standard' varieties of the language are usually spoken in rural rather than urban areas. Although 'less standard' varieties could therefore be present in the data, their properties are bound to be dominated by those of the more standard variety which constitutes the majority of the data.

Flemish resources

The Spoken Dutch Corpus – Corpus Gesproken Nederlands (CGN)[21] – is a standard Dutch database (cf. Box 2) that includes speech data collected from adults in the Netherlands and Flanders. The corpus consists of 13 components that correspond to different socio-situational settings. In this study only Flemish data from component 'O' were used. This component of the database contains phonetically aligned read speech. These data were chosen for the development of the Flemish acoustic models because read speech is carefully articulated and the corresponding phone models present a 'best case scenario' of the acoustics in the language. For instance, words and phones are not affected by the co-articulation effects that typically occur in more spontaneous speech. Component 'O' includes about 38 h of speech data recorded at 16 KHz and produced by 150 speakers.

For the purposes of the current investigation the data set was divided into training and test sets as follows: 8 (4 male, 4 female) speakers were randomly chosen for the evaluation set, corresponding to about 2 h of audio data. From the remaining 36 h, 10 h of training data were randomly selected. The training set was selected to match the size of the set of unique Afrikaans prompts described in the next section. Matching training sets were used to avoid CGN data from dominating the acoustic models.

The CGN dictionary uses 48 phones, including silence. In the cross-lingual experiments, the set was reduced to 38 phonemes using knowledge-based phonetic mapping. The mapping that was used is provided in Appendix 1 of the supplementary material. Nomenclature is given in Appendix 2 of the supplementary material.

Afrikaans resources

The Afrikaans speech data that were used in this study were taken from the National Centre for Human Language Technology (NCHLT) speech corpus.[22] The development of the corpus was funded by the South African Department of Arts and Culture with the aim of collecting 50–60 h of transcribed speech for each of the 11 official South African languages. The Afrikaans set contains data collected from 210 (103 male, 107 female) speakers. The set includes about 52 h of training data and a predefined test set of almost 3 h.

During data selection for this study, an analysis was made of the type (i.e. the unique set of words) and token (i.e. the set of words) counts in the Afrikaans data set. The values for the training and test sets are summarised in the first row of Table 1. These values indicate that only 20% of the recorded utterances in the training set are unique. This figure relates to about 10 h of unique training data and 2.2 h of unique evaluation data. The unique data subset statistics are shown in the second row of Table 1 (Type frequency 1).

If each unique token is allowed to occur a maximum of five times, the training set size increases to 37.1 h and the evaluation set to 2.7 h. Row 3 in Table 1 (Type frequency 5) shows the data subset statistics for this data selection criterion.

Table 1: Summary of the National Centre for Human Language Technology Afrikaans data

	Training set			Test set		
	Types	Tokens	Duration	Types	Tokens	Duration
All data	12 274	61 413	52.2 h	2513	3002	2.7 h
Type frequency 1	12 274	12 274	10.6 h	2513	2513	2.2 h
Type frequency 5	12 274	44 538	37.1 h	2513	3002	2.7 h

From Table 1, we observed quite a large drop in training data amount when limiting the data by uniqueness or frequency of occurrence. Subsequently, the effect on ASR performance was investigated given the various training data subsets. The ASR systems were set up according to a standard configuration – MFCCs, first- and second-order derivatives, tristate left-to-right triphone models – and were built using the hidden Markov toolkit (HTK).[23] Cepstral mean and variance normalisation was applied at the speaker level.

The ASR systems were evaluated using the predefined NCHLT evaluation set as well as two additional Afrikaans corpora. The first corpus was a text-to-speech data set while the second was a broadcast news-style data set created by recording radio news broadcasts from *Radio Sonder Grense*, a local Afrikaans radio station.[16] System performance was measured in terms of phone recognition accuracy, defined as:

$$Accuracy = 100 - \left(\frac{S+D+I}{N} \times 100\right)\%, \qquad \text{Equation 1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the total number of phones in the reference.

The results of the various evaluations are summarised in Table 2. As expected, the ASR performance drops as less data are used to develop the acoustic models. Based on the NCHLT and radio broadcast data, even though there is about a 10% absolute drop in accuracy (on average) between the unique and all data sets, the ASR performance is still quite high for the unique data set given that only 20% of the training data were used. This result probably means that the full and unique data sets represent more or less the same data properties.

The text-to-speech results show very little variation for the three different sets of acoustic models. This result may be because of the nature of the corpus: it contains speech from a single speaker and the sentences are phonetically balanced. As a consequence, the data do not contain as much variation as a multi-speaker corpus such as the radio broadcast data. The specific set of training data does not seem to influence the match between the acoustic models and the single speaker in the text-to-speech corpus.

Table 2: Phone accuracy results for different sets of training data

	NCHLT	Text to speech	Radio
All data	86.24	75.39	65.81
Type frequency 1	75.04	75.19	57.87
Type frequency 5	85.21	75.28	61.04

NCHLT, National Centre for Human Language Technology

Method

Several techniques related to model adaptation and refinement and the application to data sharing were used: MLLR, MAP, SAT and HLDA. The application of the techniques is discussed in terms of data sharing.

Maximum likelihood linear regression

Maximum likelihood linear regression (MLLR), proposed by Leggetter and Woodland[24] for speaker adaptation, provides a means to update acoustic models without having to retrain the parameters directly. The technique estimates a set of linear-regression matrix transforms that are applied to the mean vectors of the acoustic models. Their initial speaker adaptation implementation performed mean-only adaptation.

Gales and Woodland[25] extended the framework to include variance adaptation. Generally, a cascaded approach is used, in which mean adaptation is applied first and then the variance transformation is applied. Another form of the MLLR transformation is the constrained MLLR transformation (CMLLR). In this approach, a joint transform is estimated in which the aim is to transform the mean and variance simultaneously. To do so, the transform is applied directly to the data vectors and not to the means and variances.

The MLLR adaptation technique utilises a regression class tree to ensure robust transformation parameter estimation. The regression class tree defines a set of classes that contain similar acoustic models that allow data to be shared amongst similar acoustic classes. The tree is developed by using a centroid splitting algorithm[23] that can be used to automatically create the user-specified number of classes, but in this study only a single class or phone-specific classes were defined. This limitation was introduced by the HTK HLDA implementation that makes use of a single class. In terms of data sharing, the adaptation process can be used to adapt acoustic models to better fit a specific language. Here we view the languages as different speakers or channels. In this scenario, we could pool the data to increase the training data amount and then utilise MLLR to adapt these models to statistically fit the target language better.

Maximum a posteriori

Gauvain and Lee[26] proposed the use of a MAP measure to perform parameter smoothing and model adaptation. The MAP technique differs from maximum likelihood estimation by including an informative prior to aid in HMM parameter adaptation. The results for speaker adaptation showed that MAP successfully adapted speaker-independent models with relatively small amounts of adaptation data compared to the maximum likelihood estimation techniques. However, as more adaptation data became available, MAP and maximum likelihood estimation yielded the same performance. In this adaptation scenario, the speaker-independent models served as the informative priors, whereas in the experiments conducted in this study, the donor language will serve as the informative prior. Similar to the MLLR data sharing scenario, MAP can be used to adapt the acoustic models to a target language. The acoustic models trained on the pooled data serve as the prior.

Acoustic model adaptation

Under certain circumstances, as shown in Van Heerden et al.[4], simply pooling speech data (combining language resources such as data and dictionaries) into a larger training set can lead to an improvement in the results. There is no guarantee, however, that an improvement in the system accuracies will be observed and if the data amounts for the target language are small, then the donor language could possibly dominate the acoustic space. Therefore, in a resource-constrained environment, a better approach may be to adapt, using a relatively small amount of data.

MAP and MLLR are commonly used to perform speaker and environment adaptation and it is fairly simple to make use of these to perform language or dialect adaptation. It has been shown previously that simply applying MLLR and MAP to data sharing does not yield improvements.[20] However, there are many points in the acoustic model development pipeline at which these techniques can be inserted and they can be used either in isolation or in certain combinations. Thus one focus of the experimental investigation is to establish which combination of adaptation techniques could produce an improvement in overall ASR accuracy and at what point during the acoustic model development it should be applied.

Acoustic model refinement

Most current ASR systems make use of HLDA and SAT to improve the overall accuracies, which in the HTK-style development cycle are applied during the last stage of model refinement. HLDA estimates a transform that reduces the dimension of the feature vectors while trying to improve class separation. The main purpose of SAT is to produce a canonical acoustic model set by using transforms to absorb speaker differences and thus create a better speaker independent model.

As these techniques are applied as last stage refinements, there are a few possibilities that can be investigated with respect to data sharing. In terms of HLDA, a donor language can be used to develop acoustic models and the target language data used to estimate the feature dimension reduction transform. For SAT, as the transforms are absorbing speaker differences, and the language or dialect used creates acoustic differences, this approach could help create an acoustic model set better suited for the target language.

Experimental set-up

For all experiments we used 10 h of randomly selected CGN data and 10 h of NCHLT data for acoustic model development and transformation estimation. The NCHLT data correspond to the set of unique utterances described above. The developed ASR systems are evaluated on the corresponding 2.2-h subset of the NCHLT evaluation data (see Tables 1 and 2). Our aim throughout was to improve the performance of NCHLT acoustic models by adding the CGN data using various model adaptation and refinement approaches.

Baseline speech recognition system

The baseline speech recognition system was developed using a standard HTK recipe. The audio files were parameterised into 39 dimensional MFCC features – 13 static, 13 delta and 13 delta-delta. These include the MFCC 0th coefficient. Cepstral mean normalisation was applied. The acoustic models were systematically developed, starting from mono phone models, expanding the mono phone models to context-dependent triphone models and finally consolidating this model set to tied-state triphone models. A three state left-to-right HMM topology was used for each acoustic model set. A phone-based question state-tying scheme was employed to develop the tied-state models. Lastly, a mixture incrementing phase was performed to better model the state distributions – eight mixture Gaussian mixture models were used for each HMM state.

Acoustic model adaptation

The first set of experiments focused on MLLR and MAP adaptation. Block diagrams illustrating the different experimental set-ups are provided in Figures 1 to 5. The following experiments were performed:

- **Baseline NCHLT:** Baseline NCHLT acoustic models were developed on the 10-h Afrikaans NCHLT data. No adaptations were applied.
- **Language CMLLR transforms:** Starting from the baseline NCHLT system, two language-based (Afrikaans on NCHLT and Flemish on CGN) CMLLR transforms were estimated using the baseline acoustic models and the separate 10-h NCHLT and 10-h CGN data. Phone-specific transforms were estimated using the phone-defined regression class tree. Once the corpus-specific transforms were estimated, the baseline acoustic models were updated using two iterations of maximum likelihood training. Both 10-h training sets were used for this update but the specific language CMLLRs were applied to the corresponding training set. The NCHLT CMLLR was applied during evaluation.
- **Retrain using language CMLLR transforms:** The language CMLLR transform generated by the 'Language CMLLR transforms' experiment was used to develop a new acoustic model set using both the 10-h NCHLT and 10-h CGN. The normal baseline training procedure was modified to incorporate the CMLLR transforms

which were used throughout the training cycle. This meant that, at each model estimation iteration, the language-specific CMLLRs were applied when updating with the corresponding training data set. During evaluation, the estimated NCHLT CMLLR transform was applied.

- **Retrain using language CMLLR transforms with MAP:** Starting with the system developed in the 'Retrain using CMLLR transforms' experiment, one final step was added to the acoustic model development cycle: two iterations of MAP adaptation were performed using the 10-h NCHLT data only. The NCHLT CMLLR transform was applied during evaluation.
- **AutoDac training approach:** For this approach, acoustic models were developed using the best method described in Kleynhans et al.[27] Initially, only the 10-h NCHLT data were used to develop the acoustic models until the state-tying phase. Then, for the last phase, mixture incrementing, the 10-h CGN data were added to the training data pool and the Gaussian densities were estimated on all the data. No CMLLR transforms or MAP adaptation were used.

Acoustic model refinement

In this experimental set-up, two additional steps were added to the acoustic model development training cycle: HLDA and SAT. Both the HLDA and SAT use a global regression tree (all states pooled into a single node). Note that no language-dependent MAP or MLLR adaptation was applied. The HLDA ASR systems appended 13 delta-delta-delta coefficients to the baseline MFCCs, which increased the feature dimension to 52. An HLDA transform was estimated using a global transform, which was then used to transform the 52-dimensional feature vectors to 39 dimensions. For SAT, a global CMLLR transform was used to model each speaker's characteristics. The following acoustic model refinement experiments were defined:

- **NCHLT HLDA-SAT:** Baseline acoustic models were developed using the 10-h NCHLT, followed by HLDA and SAT model refinements.
- **NCHLT+CGN HLDA-SAT:** Baseline acoustic models were developed using both the 10-h NCHLT and 10-h CGN data sets, and then applying the HLDA and SAT model refinements using all the training data.
- **NCHLT+CGN+NCHLT HLDA-SAT:** For this training set-up, baseline acoustic models were developed on both the 10-h NCHLT and 10-h CGN training data sets. The HLDA and SAT transformations were estimated using the 10-h NCHLT training data only.

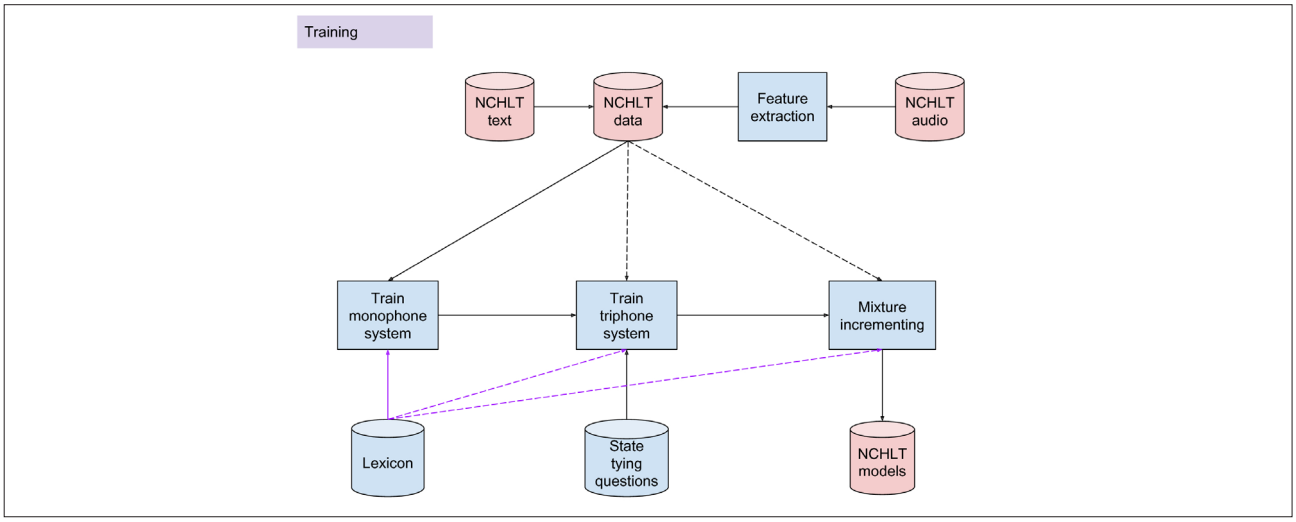
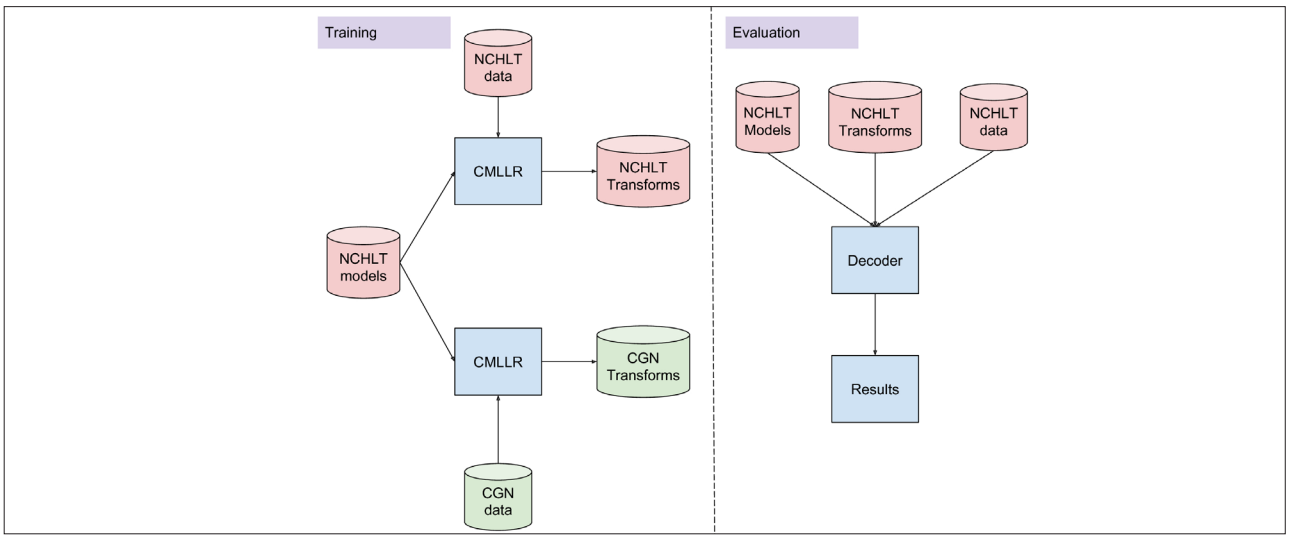
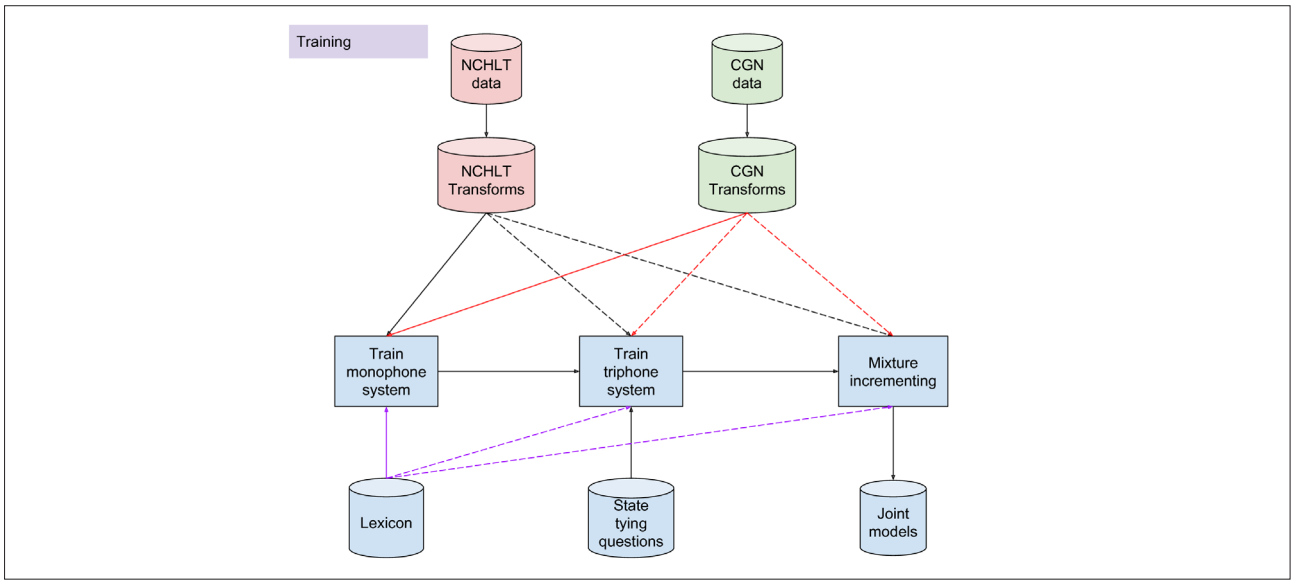


Figure 1: Baseline National Centre for Human Language Technology (NCHLT) training scheme.



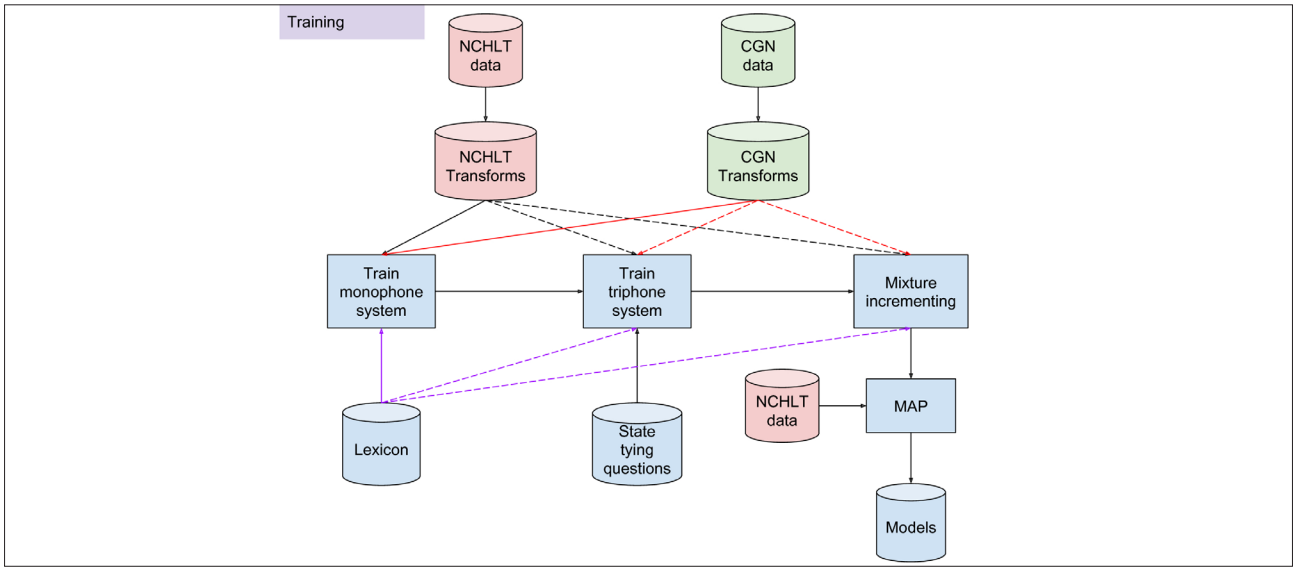
NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 2: Language constrained maximum likelihood linear regression (CMLLR) training scheme.



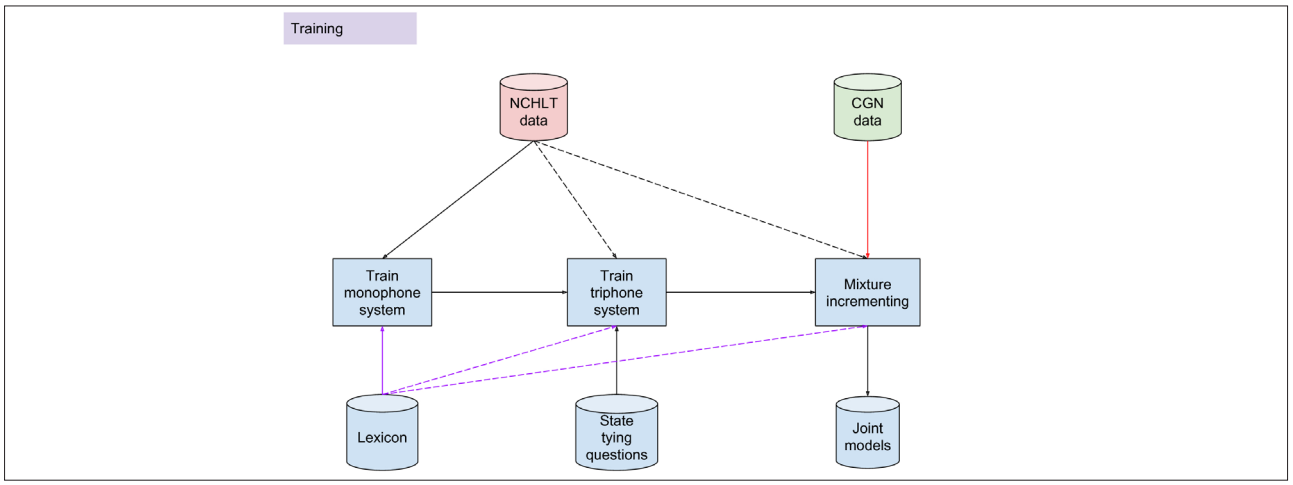
NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 3: Retrain using language constrained maximum likelihood linear regression transform training scheme.



NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 4: Retrain using language constrained maximum likelihood linear regression transforms with maximum a posteriori (MAP) training scheme.



NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 5: AutoDac training scheme.

Metrics

The ability of the different system configurations to model the training data accurately was measured in terms of the accuracy with which the test data could be decoded. Phone recognition accuracy was calculated according to Equation 1 and correctness values were derived as follows:

$$Correctness = \left(\frac{C}{N} \times 100\right)\%, \quad \text{Equation 2}$$

where C is the number of correctly recognised phones and N is the total number of phones in the reference.

Results

Experimental results are presented for CMLLR and MAP adaptation as well as HLDA plus SAT combinations. System performance is quantified in terms of phone recognition accuracy and correctness.

Acoustic model adaptation

Table 3 provides an overview of the results that were obtained using different data sets and model adaptation combinations. The first row in the table represents the performance of the baseline system without any data sharing or model adaptation.

Table 3: Correctness and accuracy results for various automatic speech recognition data sharing set-ups

	Correctness (%)	Accuracy (%)
Baseline NCHLT	78.77	71.17
Language CMLLR transforms	75.81	68.83
Retrain using language CMLLR transforms	78.02	71.82
Retrain using language CMLLR transforms with MAP	78.87	71.31
AutoDac training approach	75.69	68.41

NCHLT, National Centre for Human Language Technology; CMLLR; constrained maximum likelihood linear regression; MAP, maximum a posteriori

Unfortunately, the results in Table 3 show that none of the adaptation and training schemes provide an improvement in ASR performance, when adding CGN data to the NCHLT training data. This is in line with the results reported by Imseng et al.[20] for a similar experiment using a smaller corpus of telephone data. It would seem that both CMLLR and MAP provide insufficient mechanisms to effectively combine data from different sources in the context of cross-language data sharing.

Acoustic model refinement

The performance of the systems in which the models were refined by applying HLDA and SAT is captured in Table 4. Comparing the first row in Table 4 with the corresponding row in Table 3 shows that the application of HLDA and SAT results in a substantial improvement in both phone accuracy and correctness. When the CGN data are added to the training data, the performance decreases. However, the best result is obtained when the acoustic model set is developed on the combined data but the HLDA and SAT are estimated on the 10-h NCHLT data only. This finding may suggest that these transforms are sensitive to language-specific data. The HLDA in effect estimates a projection from a higher dimensional space to a lower one. Thus, a better projection, in terms of class separation, might be estimated on the target data only – in this case, the NCHLT data. For SAT, the single global CMLLR transforms may be insufficient to fully absorb the speaker and channel characteristics; therefore the acoustic model set is not in the best canonical form. Further tests on HTK are not possible as this is a software limitation.

Table 4: Correctness and accuracy results for heteroscedastic linear discriminant analysis (HLDA)- and speaker adaptive training (SAT)-based data sharing automatic speech recognition set-ups

	Correctness (%)	Accuracy (%)
NCHLT HLDA-SAT	85.71	79.66
NCHLT+CGN HLDA-SAT	84.37	78.33
NCHLT+CGN+NCHLT HLDA-SAT	86.89	81.07

NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Discussion

To investigate why only a single improvement was observed over the different experiments, the state-tying process was analysed as this process determines the manner in which acoustic data are shared. HTK makes use of the question-based tying scheme described by Young et al.[28]: initially all acoustic states are grouped into a single root class and then a process to split the nodes is run by 'asking' left and right context questions – all triphones that have the same left or right phone are removed from the pool and the change in pool log-likelihood is captured. The question that results in the greatest change in score is selected and a new node is created that contains all the triphones described by the question. The pre-split node contains all other triphones. The process is continued until a user-defined stopping criterion is met.

Tracking which question is used to split the data pools (create nodes) can give an indication of when the data between the two languages are shared: if language-specific questions are used to split the nodes early on in the state-tying process then no real cross-language data sharing is occurring. To perform the state-tying tracking, a modified, but similar, version of the HTK implementation was developed in which language-specific questions could be used to split the acoustic data pools. Table 5 shows the level at which a language question was used to split the data.

Table 5: The percentage of phones for which the language question was used to split the data during state tying

	State 2	State 3	State 4
First question	69.44	91.67	63.89
First or second question	86.11	91.67	77.78

The values in Table 5 show that, for the majority of cases, the best reduction in overall data pool log-likelihood can be achieved by splitting the data into language-dependent paths. The central context makes use of the language split question to partition the data, in over 91% of the cases for the very first question. This finding is significant as the central context state generally consumes the majority of the speech frames when compared to the start and end states. This result shows that minimal data sharing would occur if the system had a choice and may point to a data artefact – such as channel or environment – which prevents data sharing between the CGN and NCHLT corpora. Further investigation is needed to establish the mechanisms that are inhibiting data sharing and their relative contributions. Possible sharing prevention mechanisms could be: grammar, channel and environment. As positive pooling results were reported by Van Heerden et al.[4] and all experiments were conducted on the same corpus, channel may be a strong candidate. In this instance 'channel' refers to all the factors that could influence the acoustic properties of the speech signals, e.g. the acoustic environment in which the data were recorded and the recording equipment.

Table 5 shows that cross-language data sharing is clearly not taking place to the same extent as reported by Mandal et al.[7] The low data sharing rates are also in contrast to the results presented

by Kamper et al.[9], in which 33% and 44% sharing was seen across accents for phone and word optimal results, respectively, and by Niesler[8] where 20% sharing was measured across language at optimal system performance. For these investigations, data sharing resulted in improved system performance but it is not clear if a positive correlation exists between the percentage of data shared among clusters and the eventual ASR performance.

It could be argued that the acoustic differences between Afrikaans and Flemish are bigger than those observed between the various English accents investigated in the Kamper et al.[9] study. However, the majority of the sounds could be expected to differ to at least the same extent as the languages studied by Niesler because they are from the same language families, as are Afrikaans and Flemish. They are also similar from an acoustic point of view, as are the languages that were investigated in this study. It should be kept in mind that both Kamper et al.[9] and Niesler conducted experiments within the same corpus. Acoustic factors – other than those caused by differences between accents and languages, such as channel and environment effects – could therefore not have influenced their results. This strengthens the possibility that the lack of data sharing in the present study could probably be a result of cross-corpus rather than cross-language artefacts.

Imseng et al.[18] showed that a systematic improvement in phone performances was observed for in-domain phones that had relatively small data amounts. Thus, it would seem that we should rather target states that may need out-of-language data to improve the distribution modelling.

Conclusion

While the idea of data sharing makes sense intuitively – increase the amount of training data for robust density estimation – realising a performance gain in ASR accuracy is difficult to achieve within the context of HMM-based ASR. From the experimental results obtained in this study, using standard MAP and MLLR techniques to enable data sharing did not provide phonetic recognition performance gains. These MAP and MLLR results are in line with those presented by Imseng et al.[20] In addition, the various alternative training strategies also failed. Thus, the standard MAP, MLLR and our various training strategies are not sufficient for data sharing when simply pooling the data.

Surprisingly, the NCHLT+CGN+NCHLT HLDA-SAT experiment managed to achieve a better phone error rate; however, the baseline NCHLT+CGN HLDA-SAT did not yield a gain. The improved result may imply that the combined data are useful but the Afrikaans-specific HLDA projection and SAT acoustic model adjustment are required. This has similarities to some DNN data sharing approaches in which pre-training is performed on many languages but final network parameter optimisations are performed on the target language only.

Recent results from SGMM and DNN experiments show much more potential for data sharing between languages and should be pursued rather than MAP and MLLR. One possible line of research would be to use SGMM for data sharing but rather than pooling all the data, only include data for low occurrence phones, as suggested by results reported in Imseng et al.[18]

Acknowledgements

This research was supported by the South African National Research Foundation (grant no. UID73933), the Fund for Scientific Research of Flanders (FWO) under project AMODA (GA122.10N) as well as a grant from the joint Programme of Collaboration on HLT funded by the Nederlandse Taalunie and the South African Department of Arts and Culture.

Authors' contributions

F.D.W. and D.V.C. conceptualised and led the project on acoustic modelling for under-resourced languages; F.D.W., D.V.C., R.S. and N.K. were responsible for conceptual contributions and experimental design; F.D.W. and D.V.C. designed the phone mapping between Flemish and Afrikaans; R.S. and N.K. performed the experiments; D.F.W. and N.K. prepared the manuscript; D.V.C. is R.S.'s PhD promotor.

References

1. Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech Commun. 2014;56:85–100. http://dx.doi.org/10.1016/j.specom.2013.07.008
2. Schultz T, Waibel A. Multilingual and cross lingual speech recognition. In: DARPA Workshop 1998: Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding; 1998 February 08–11; Lansdowne, VA, USA. Lansdowne, VA: NIST; 1998. p. 259–262.
3. Schultz T, Waibel A. Language independent and language adaptive large vocabulary speech recognition. In: ICSLP 1998: Proceedings of the 5th International Conference on Spoken Language Processing; 1998 November 30 – December 04; Sydney, Australia. p. 1819–1822.
4. Van Heerden C, Kleynhans N, Barnard E, Davel, M. Pooling ASR data for closely related languages. In: SLTU 2010: Proceedings of the 2nd Workshop on Spoken Languages Technologies for Under-resourced languages; 2010 May 03–0; Penang, Malaysia. Penang: SLTU; 2010. p. 17–23.
5. Schultz T, Waibel A. Language-independent and language-adaptive acoustic modelling for speech recognition. Speech Commun. 2001;35(1):31–51. http://dx.doi.org/10.1016/S0167-6393(00)00094-7
6. Adda-Decker M, Lamel L, Adda G. A first LVCSR system for Luxembourgish, an under-resourced European language. In: LTC 2011: Proceedings of 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics; 2011 November 25–27; Poznań, Poland. Poznań: LTC; 2011. p. 47–50.
7. Mandal A, Vergyri D, Akbacak M, Richey C, Kathol A. Acoustic data sharing for Afghan and Persian languages. In: ICASSP 2011: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2011 May 22–27; Prague, Czech Republic. IEEE; 2011. p. 4996–4999. http://dx.doi.org/10.1109/ICASSP.2011.5947478
8. Niesler T. Language-dependent state clustering for multilingual acoustic modelling. Speech Commun. 2007;49(6):453–463. http://dx.doi.org/10.1016/j.specom.2007.04.001
9. Kamper H, Mukanya FJM, Niesler T. Multi-accent acoustic modelling of South African English. Speech Commun. 2012;54(6):801–813. http://dx.doi.org/10.1016/j.specom.2012.01.008
10. Veselý K, Karafiát M, Grézl F, Janda M, Egorova E. The language-independent bottleneck features. In: SLT 2012: Proceedings of the Spoken Language Technology Workshop; 2012 December 02–05; Miami, FL, USA. Miami, FL: SLT; 2012. p. 336–341. http://dx.doi.org/10.1109/slt.2012.6424246
11. Zhang Yu, Chuangsuwanich E, Glass J. Language ID-based training of multilingual stacked bottleneck features. In: Interspeech 2014: Proceedings of the International Speech Communication Association; 2014 September 14–18; Singapore. p. 1–5.
12. Nguyen QB, Gehring J, Muller M, Stuker S, Waibel A. Multilingual shifting deep bottleneck features for low-resource ASR. In: ICASSP 2014: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2014 May 04–09; Florence, Italy. p. 5607–5611. http://dx.doi.org/10.1109/ICASSP.2014.6854676
13. Vu NT, Imseng D, Povey D, Motlicek P, Schultz T, Bourlard H. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: ICASSP 2014: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2014 May 04–09; Florence, Italy. p. 7639–7643. http://dx.doi.org/10.1109/ICASSP.2014.6855086
14. Sahraeian R, Van Compernolle D, De Wet F. Under-resourced speech recognition based on the speech manifold. In: Interspeech 2015: Proceedings of the International Speech Communication Association; 2015 September 06–10; Dresden, Germany. p. 1255–1259.
15. Sahraeian R, Van Compernolle D, De Wet F. On using intrinsic spectral analysis for low-resource languages. In: SLTU 2014: Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages; 2014 May 14–16; St Petersburg, Russia. p. 61–65.
16. De Wet F, De Waal A, Van Huyssteen GB. Developing a broadband automatic speech recognition system for Afrikaans. In: Interspeech 2011: Proceedings of International Speech Communication Association; 2011 August 27–31; Florence, Italy. p. 3185–3188.
17. Heeringa W, De Wet F, Van Huyssteen GB. Afrikaans and Dutch as closely-related languages: A comparison to West Germanic languages and Dutch dialects. Stellenbosch Papers in Linguistics Plus. 2015;47:1–18. http://dx.doi.org/10.5842/47-0-649

18. Imseng D, Bourlard H, Garner PN. Boosting under-resourced speech recognizers by exploiting out of language data – Case study on Afrikaans. In: SLTU 2012: Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages; 2012 May 07–09; Cape Town, South Africa. Cape Town: SLTU; 2012. p. 60–67.

19. Despres J, Fousek P, Gauvain J, Gay S, Josse Y, Lamel L, et al. Modeling northern and southern varieties of Dutch for STT. In: Interspeech 2009: Proceedings of the International Speech Communication Association. 2009 September 06–10; Brighton, UK. p. 96–99.

20. Imseng D, Motlicek P, Bourlard H, Garner PN. Using out-of-language data to improve an under-resourced speech recognizer. Speech Commun. 2014;56:142–151. http://dx.doi.org/10.1016/j.specom.2013.01.007

21. Oostdijk N. The spoken Dutch corpus: Overview and first evaluation. In: LREC 2000: Proceedings of the Second International Conference on Language Resources and Evaluation; 2000 May 31 – June 02; Athens, Greece. p. 887–894.

22. Barnard E, Davel MH, Van Heerden C, De Wet F, Badenhorst J. The NCHLT speech corpus of the South African languages. In: SLTU 2014: Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages; 2014 May 14–16; St Petersburg, Russia. p. 194–200.

23. Young S, Evermann G, Gales M. The HTK book [document on the Internet]. c2009 [cited 2016 Jan 12]. Available from: http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf

24. Leggetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang. 1995;9(2):171–185. http://dx.doi.org/10.1006/csla.1995.0010

25. Gales MJF, Woodland PC. Mean and variance adaptation within the MLLR framework. Comput Speech Lang.1996;10(4):249–264. http://dx.doi.org/10.1006/csla.1996.0013

26. Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans Speech Audio Process. 1994;2(2):291–298. http://dx.doi.org/10.1109/89.279278

27. Kleynhans N, Molapo R, De Wet F. Acoustic model optimisation for a call routing system. In: PRASA 2012: Proceedings of the 23rd Meeting of the Pattern Recognition Association of South Africa; 2012 November 29–30; Pretoria, South Africa. p. 165–172.

28. Young SJ, Odell JJ, Woodland PC. Tree-based state tying for high accuracy acoustic modelling. In: HLT 1994: Proceedings of the Workshop on Human Language Technology; 1994 March 08–11; Plainsboro, NJ, USA. p. 307–312. http://dx.doi.org/10.3115/1075812.1075885