Optimisation of acoustic models for a target accent using decision-tree state clustering

Herman Kamper and Thomas Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
kamperh@sun.ac.za, trn@sun.ac.za

***Abstract*—In this paper we extend the decision-tree state clustering algorithm normally used to construct tied-state hidden Markov models to allow for the explicit optimisation on a particular target accent. Although the traditional algorithm guarantees overall likelihood improvements when clustering states from multiple accents, per-accent improvements are not guaranteed. We develop a tractable formulation of the targeted optimisation strategy by basing the decision-tree cluster splitting criterion on a likelihood calculated exclusively on the target accent. We find that this approach leads to deterioration compared to the traditional modelling approaches. However, when combining targeted and non-targeted approaches by linear weighting, small but consistent improvements over the traditional approaches are observed.**

I. Introduction

Accented speech is often prevalent in multilingual societies. The processing of such speech is therefore a necessary but challenging task. In previous work [1] we considered different approaches for modelling the five accents of South African English (SAE). In particular, we considered multi-accent acoustic modelling which allows selective data sharing between accents. This is achieved by including accent-based questions in the decision-tree state clustering process normally used to construct tied-state hidden Markov models (HMMs).

Although multi-accent acoustic modelling enables selective sharing, the likelihood criterion used during the decision-tree state clustering process is calculated on data from all accents. The process therefore guarantees an overall likelihood improvement, but not per-accent improvements. In some practical scenarios it might, however, be desirable to obtain the best possible acoustic model set for a particular accent. This leads to the question of whether the multi-accent decision-tree state clustering approach can be extended to optimise the likelihood on a particular target accent. Selective sharing would still be allowed across accents, but data will only be shared if it is advantageous for the target accent. In this paper we develop, evaluate and analyse such techniques.

We base our investigation on databases for the five accents of SAE identified in the literature [1], [2]. The acoustic modelling approaches developed in [1] will serve as baselines in the evaluation of the proposed targeted modelling approaches.

II. Related Research

Several studies have considered acoustic modelling of different accents of the same language. One approach is to simply train separate accent-specific models that allow no sharing between accents [3]. An alternative is to pool data from all accents considered, resulting in a single accent-independent acoustic model set [4]. Adaptation techniques in which models trained on one accent are adapted using data from another accent have also been considered [5], [6].

Recently, selective data sharing across accents through the use of appropriate decision-tree state clustering algorithms has received some attention [1], [7]. In these studies the multilingual modelling approach first proposed by Schultz and Waibel [8] was extended to apply to multiple accents of the same language. In this paper we extend the multi-accent acoustic modelling approach to allow targeted optimisation on an individual accent from the set of accents considered.

III. General Experimental Methodology

A. Training and test sets

Our experiments were based on the African Speech Technology (AST) databases [9]. These consist of annotated telephone speech recorded over fixed and mobile telephone networks and contain a mix of read and spontaneous speech. As part of the AST Project, five English accented speech databases were compiled corresponding to the five South African accents of English identified in the literature [2]: Afrikaans English (AE), Black South African English (BE), Cape Flats English (CE), White South African English (EE) and Indian South African English (IE). These databases were transcribed both phonetically, using a common IPA-based phone set consisting of 50 phones, as well as orthographically.

Each of the five databases was divided into training, development and evaluation sets. As indicated in Tables I and II, the training sets each contain between 5.5 and 7 hours of speech from approximately 250 speakers while the evaluation sets contain approximately 25 minutes from 20 speakers for each accent. The development sets were used only for the optimisation of the recognition parameters before final testing on the evaluation data. For the development and evaluation sets the ratio of male to female speakers is approximately equal and all sets contain utterances from both land-line and mobile phones. There is no speaker-overlap between any of the sets. The average length of an utterance is approximately 2 seconds.

B. General acoustic modelling procedure

Speech recognition systems were developed using the HTK tools [10]. Speech audio data was parametrised as 13 Mel-

TABLE I
TRAINING SETS FOR EACH ACCENT.

Accent	Speech (h)	No. of utterances	No. of speakers	Phone tokens
AE	7.02	11 344	276	199 336
BE	5.45	7779	193	140 331
CE	6.15	10 004	231	174 068
EE	5.95	9878	245	178 954
IE	7.21	15 073	295	218 372
Total	31.78	54 078	1240	911 061

TABLE II
EVALUATION SETS FOR EACH ACCENT.

Accent	Speech (min)	No. of utterances	No. of speakers	Phone tokens
AE	24.16	689	21	10 708
BE	25.77	745	20	11 219
CE	23.83	709	20	11 180
EE	23.96	702	18	11 304
IE	25.41	865	20	12 684
Total	123.13	3710	99	57 095

frequency cepstral coefficients (MFCCs) with their first and second order derivatives to obtain 39 dimensional observation vectors. Cepstral mean normalisation was applied on a per-utterance basis. The parametrised training sets were used to obtain three-state left-to-right single-mixture monophone HMMs with diagonal covariance matrices using embedded Baum-Welch re-estimation. These monophone models were then cloned and re-estimated to obtain initial cross-word triphone models which were subsequently subjected to decision-tree state clustering. This was followed by five iterations of re-estimation. Finally, the number of Gaussian mixtures per state was gradually increased, each increase being followed by a further five iterations of re-estimation. This yielded diagonal-covariance cross-word tied-state triphone HMMs with three states per model and eight Gaussian mixtures per state.

As part of the research presented here, several different acoustic model sets were developed following this general training procedure. For each modelling approach a different variant of the decision-tree state clustering algorithm was applied. Since decision-tree state clustering is central to this study, the standard algorithm is described briefly in Section IV. Variants of the algorithm are subsequently described in Sections V and VI.

C. Language models

Comparison of recognition performance was based on phone recognition experiments. Using the SRILM toolkit [11], backoff bigram phone language models were trained for each accent individually from the corresponding training set phone transcriptions. Absolute discounting was used for the estimation of language model probabilities [12]. The development sets were used to optimise the word insertion penalty (WIP) and language model scaling factor (LMS) used during recognition. Because optimal WIP and LMS values showed almost no variation between accents, the same WIP and LMS settings were used for all experiments.

Since the presented work considers only the effect of the acoustic models, it was assumed that during testing the accent of each utterance was known. In order to isolate acoustic modelling effects, evaluation therefore involved presenting each test utterance only to a system employing an acoustic and language model matching the accent of that utterance.

IV. DECISION-TREE STATE CLUSTERING

The standard decision-tree state clustering algorithm that is used to construct tied-state triphone HMMs (Section III-B) is reviewed in this section. The content is based on [13] and [14].

A. Overview

The clustering process begins by pooling into a single cluster the data of corresponding states from all triphones with the same basephone. This is done for all triphones observed in the training set. A set of linguistically-motivated questions is then used to split these clusters. Such questions may, for example, ask whether the left context of a particular triphone is a vowel or whether the right context is a silence. There are, in general, many such questions and each potential question results in a split which subsequently results in an increase in training set likelihood. For each cluster the optimal question (leading to the largest likelihood increase) is determined. In this way clusters are subdivided repeatedly until either the increase in likelihood or the number of observation vectors associated with a resulting cluster (the cluster occupancy count) falls below a certain predefined threshold.

The result is a phonetically-motivated binary decision-tree where the leaf nodes represent clusters of triphone HMM states which are to be tied by pooling data. This ensures that model parameters are estimated on a sufficient amount of training data. Furthermore, each state of a triphone not seen in the training set can be associated with a leaf node in the decision-trees. This allows the synthesis of triphones that are required during recognition but are not present in the training set.

B. Details of decision-tree construction

Suppose question q splits the cluster with states $\mathbb{S}$ into two clusters with states $\mathbb{S}_1(q)$ and $\mathbb{S}_2(q)$, respectively. The increase in log likelihood resulting from the split can be calculated as

$$\Delta L_q = L(\mathbb{S}_1(q)) + L(\mathbb{S}_2(q)) - L(\mathbb{S}) \tag{1}$$

where $L(\mathbb{S})$ denotes the log likelihood of the training observation vectors assigned to the states in $\mathbb{S}$. The question q^* which maximises (1) is selected as the optimal question to split the cluster. In order to compute (1), however, the calculation of the likelihood of an arbitrary cluster of states must be tractable.

Let $\mathbb{S}$ denote an arbitrary set of HMM states and let $L(\mathbb{S})$ be the log likelihood of the training observation vectors assigned to the states in $\mathbb{S}$ under the assumption that all states in $\mathbb{S}$ share a common mean $\boldsymbol{\mu}(\mathbb{S})$ and covariance matrix $\boldsymbol{\Sigma}(\mathbb{S})$. We also assume that the transition probabilities have a negligible effect on the log likelihood and can therefore be ignored [14]. The

log likelihood that the observation vectors were generated by the states in $\mathbb{S}$ can then be calculated as

$$L(\mathbb{S}) = \log \prod_{f \in \mathbb{F}} p(\mathbf{o}_f|\mathbb{S}) = \sum_{f \in \mathbb{F}} \log \left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] \quad (2)$$

where $\mathbf{o}_f$ is the observation vector associated with frame f and $\mathbb{F}$ is the set of training frames for which the observation vectors are associated with the states in $\mathbb{S}$, i.e. $\mathbb{F} = \{f : \mathbf{o}_f \text{ is generated by states in } \mathbb{S}\}$. The observation probability density functions (PDFs) are single-mixture Gaussian PDFs.

The direct calculation of $L(\mathbb{S})$ using (2) requires direct recourse to the observation vectors $\mathbf{o}_f$. This is computationally intractable since datasets are large and the likelihood calculation will have to be repeated several times. Fortunately it can be shown (Appendix A) that [13]:

$$L(\mathbb{S}) = -\frac{1}{2}\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] + n\right\} \sum_{s \in \mathbb{S}} \sum_{f \in \mathbb{F}} \gamma_s(\mathbf{o}_f) \quad (3)$$

where n is the dimensionality of the observation vectors and $\gamma_s(\mathbf{o}_f)$ is the posterior probability that the observation vector $\mathbf{o}_f$ is generated by HMM state s. The log likelihood of a cluster of states is therefore only dependent on the shared covariance matrix $\boldsymbol{\Sigma}(\mathbb{S})$ and the total state occupancy of the cluster $\sum_{s \in \mathbb{S}} \sum_{f \in \mathbb{F}} \gamma_s(\mathbf{o}_f)$. It can be shown that the former can be calculated from the means and covariance matrices of the states in the cluster [13]. The state occupancy counts are determined during the Baum-Welch re-estimation procedure which precedes clustering. Thus, $L(\mathbb{S})$ can be calculated without recourse to the observation vectors and the decision-tree construction process becomes computationally tractable.

V. Traditional Modelling Approaches

The following gives an overview of acoustic modelling approaches considered in previous work [1] and summarises relevant results. These results are the baselines for Section VI.

A. Accent-specific and accent-independent acoustic modelling

As described in Section II, *accent-specific acoustic models* are obtained by not allowing any sharing of data between accents. By growing separate decision-trees for the different accents, triphone HMM states are clustered separately. Only questions relating to phonetic context are employed, resulting in completely distinct sets of acoustic models for each accent.

In contrast, *accent-independent models* are obtained by blindly pooling accent-specific data across accents for phones with the same IPA symbol, resulting in a single accent-independent model set. A single set of decision-trees is constructed across all accents and the clustering process employs only questions relating to phonetic context, resulting in a single accent-independent set of triphone HMMs for all accents.

These two approaches were applied to the training sets of the five accents of SAE described in Section III-A. For each accent, the decision-tree likelihood improvement threshold was optimised separately on its corresponding development set. This approach was followed for all experiments presented in this paper since the purpose here is to achieve best performance on a particular target accent and not to optimise average performance over all accents, as was the case in [1].

The first two entries in Table III show the phone recognition performance measured on the evaluation sets for the accent-specific and accent-independent modelling approaches. Accent-independent models perform better than the accent-specific models for all accents except BE. The average accuracy of the accent-independent models is also better by approximately 0.76% absolute. This improvement has been calculated to be statistically significant at the 99.9% level using bootstrap confidence interval estimation at the utterance level with 10^4 bootstrap replications over all five accents [15].

B. Multi-accent acoustic modelling

The third and final acoustic modelling approach considered in [1] is similar to accent-independent modelling. Again, the state clustering process begins by pooling corresponding states from all triphones with the same basephone. However, in this case the set of decision-tree questions take into account not only the phonetic character of the left and right contexts but also the accent of the basephone. The HMM states of two triphones with the same IPA symbol but from different accents can therefore be kept separate if there is a significant acoustic difference or can be tied if there is not. We refer to such models as *multi-accent acoustic models*. Figure 1 shows an example in which the centre state of the triphone [t]-[iy]+[ng] is tied across the AE and EE accents while the first and last states are modelled separately.

The third entry in Table III indicates the performance when using multi-accent acoustic models. For AE and IE, improved performance over the first two acoustic model sets is observed. For CE and EE, deterioration is seen relative to the accent-independent models. For BE, deterioration is seen relative to the accent-specific models. Nevertheless, the multi-accent models show a very small improvement in average accuracy over the accent-independent models. This improvement is statistically significant only at the 60% level.

To obtain some indication of what happens in the decision-tree clustering process, the type of questions most frequently asked during clustering can be considered. Figure 2 analyses the decision-trees of the multi-accent acoustic models giving optimal performance on the AE development set. The figure

TABLE III
Per-accent and average (avg.) phone recognition accuracies (%) measured on the evaluation set. The different acoustic model sets are described throughout the paper.

Acoustic model set	AE	BE	CE	EE	IE	Avg.
Accent-specific	64.80	56.77	64.59	72.97	64.27	64.68
Accent-independent	65.97	55.98	66.51	74.45	64.40	65.44
Multi-accent	66.20	56.56	66.31	73.94	64.60	65.50
Targeted multi-accent	64.60	55.17	64.11	72.65	64.44	64.21
Weighted targeted	66.74	56.56	66.13	73.94	64.96	65.65
Weight w_t used above	0.51	0.5	0.53	0.5	0.54	

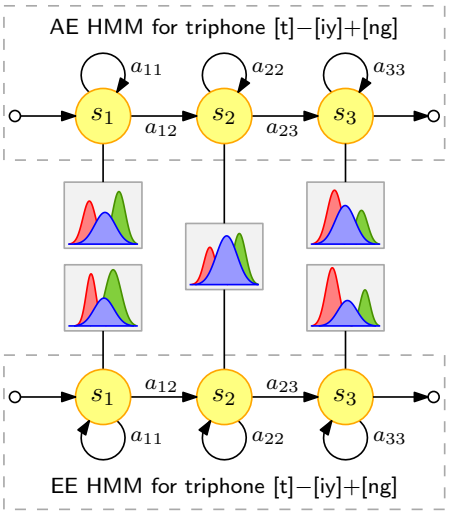


Fig. 1. Multi-accent HMMs for corresponding AE and EE triphones.

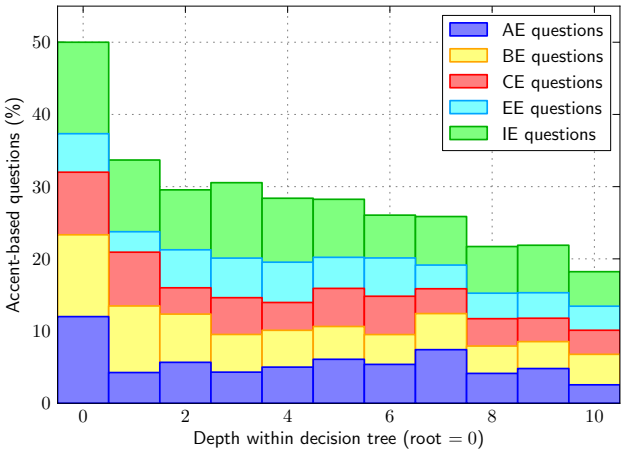


Fig. 2. The percentage of questions that relate to specific accents at various depths within the decision-trees for the multi-accent acoustic model set with optimal recognition performance on the AE development set.

shows that about 50% of all questions at the root nodes are accent-based and that this proportion drops to 34% and 30% for the roots' children and grandchildren respectively. Of the 12 970 resulting clusters (the leaf nodes) in the decision-trees, 13.2% are AE-only, 22.2% share AE with some other accent(s) and 64.7% are non-AE. These statistics and the analysis in Figure 2 are used for comparison in the next sections.

VI. Targeted Modelling Approaches

This section describes new extensions which we have made to the multi-accent acoustic modelling approach (Section V-B). We treat the results presented in Section V as baselines.

A. Motivation and overview

When clustering triphone states from several accents, the log likelihood $L(\mathbb{S})$ used as splitting criterion in the decision-tree clustering process is calculated over *all* accents. Although a particular cluster split guarantees an overall improvement in likelihood, improvements on a per-accent basis are not guaranteed. This raises the question whether the algorithm can be altered to optimise the likelihood on a particular target accent. In such an approach, a specific phonetic or accent-based question would be applied only when it is advantageous for the models of the selected target accent to do so.

B. Targeted multi-accent acoustic modelling

Suppose we have a cluster of states $\mathbb{S} = \mathbb{S}_x \cup \mathbb{S}_t$ with the states $\mathbb{S}_x$ generating observation vectors for frames $\mathbb{F}_x$ and $\mathbb{S}_t$ generating observation vectors for frames $\mathbb{F}_t$. Our aim is to optimise performance on the target states $\mathbb{S}_t$. In the traditional decision-tree state clustering procedure, the log likelihood of this cluster $\mathbb{S}$ generating the observation vectors for frames $\mathbb{F} = \mathbb{F}_x \cup \mathbb{F}_t$ would be calculated according to (3) and the optimisation criterion would be based upon this figure. We propose to determine instead the log likelihood of the target states $\mathbb{S}_t$ generating the observation vectors for frames $\mathbb{F}_t$. While all states in $\mathbb{S}$ still share a common mean $\boldsymbol{\mu}(\mathbb{S})$ and covariance matrix $\boldsymbol{\Sigma}(\mathbb{S})$, we base the cluster splitting criterion on this alternative log likelihood. By doing so, parameter estimation is still based on data from all frames $\mathbb{F} = \mathbb{F}_x \cup \mathbb{F}_t$ but the likelihood optimised is restricted to a set of target states $\mathbb{S}_t$ and no longer based on all the states $\mathbb{S}$.

The log likelihood of states $\mathbb{S}_t$ generating the associated observation vectors for frames $\mathbb{F}_t$ can be calculated as

$$\begin{aligned} L_t(\mathbb{S}) &= \log \prod_{f \in \mathbb{F}_t} p(\mathbf{o}_f|\mathbb{S}) \\ &= \sum_{f \in \mathbb{F}_t} \log \left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] \end{aligned} \tag{4}$$

This log likelihood is still dependent on all the states $\mathbb{S}$ since $\boldsymbol{\mu}(\mathbb{S})$ and $\boldsymbol{\Sigma}(\mathbb{S})$ are based on data from all the states.

As was the case in (2), the direct calculation of (4) is computationally intractable since it requires recourse to the observation vectors. However, we can again show (Appendix B) that this amended log likelihood can be calculated from the means, covariance matrices and state occupancy counts of the states in $\mathbb{S}$:

$$\begin{aligned} L_t(\mathbb{S}) = &-\frac{1}{2} N_t \left\{\log[(2\pi)^n |\boldsymbol{\Sigma}(\mathbb{S})|]\right\} - \frac{1}{2} n(N_x + N_t) \\ &+ \frac{1}{2} \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbb{S}) N_x [\boldsymbol{\Sigma}(\mathbb{S}_x) \\ &+ (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathsf{T}}]\} \end{aligned} \tag{5}$$

with

$$N_t = \sum_{s \in \mathbb{S}_t} \sum_{f \in \mathbb{F}} \gamma_s(\mathbf{o}_f) \text{ and } N_x = \sum_{s \in \mathbb{S}_x} \sum_{f \in \mathbb{F}} \gamma_s(\mathbf{o}_f) \tag{6}$$

Since $\boldsymbol{\mu}(\mathbb{S}_x)$, $\boldsymbol{\mu}(\mathbb{S})$, $\boldsymbol{\Sigma}(\mathbb{S}_x)$ and $\boldsymbol{\Sigma}(\mathbb{S})$ are only the means and covariance matrices of the states in the corresponding clusters, the calculation of $L_t(\mathbb{S})$ as in (5) is computationally tractable.

C. Evaluation and analysis: targeted modelling

By considering each of the SAE accents in turn as the target accent, the *targeted multi-accent acoustic modelling* approach

was applied to the five training sets described in Section III-A. Phone recognition performance is shown in the fourth entry of Table III. The targeted multi-accent models are outperformed by all other models, yielding the lowest average accuracy of 64.21%. Worse performance is also achieved on a per-accent basis for all accents except for IE, for which a slight improvement over the accent-specific models is observed.

Figure 3 analyses the decision-trees of the targeted multi-accent acoustic models giving optimal performance on the AE development set. A striking feature is that the only accent-based question ever employed by the trees relate to the target accent AE. In fact, it is possible to show (Appendix C) that the target-accent-question will always be asked rather than a non-target-accent-question. Figure 3 shows that 53% of all questions at the root nodes relate to AE and that this proportion drops to 27% and 18% for the roots' children and grand-children, respectively. Of the 5718 resulting clusters in the decision-trees, 84.7% are AE-only, 5.3% combine data from all five accents, and 10% combine data from all the accents apart from AE. This last group of clusters was consequently not used during recognition.

In comparison with the analysis of the multi-accent decision-trees in Figure 2, slightly more accent-based questions are asked at the root nodes and the proportion of accent-based questions tapers off much more quickly in the targeted case. This indicates that earlier separation of the AE accent occurs in the AE-targeted multi-accent decision-trees. Increased separation of AE is also observed when comparing the resulting cluster statistics in the targeted case to those of the non-targeted case (final paragraph, Section V-B); for the former, only 301 clusters (5.3% of 5718 clusters) share data from AE with data from any of the other accents while, for the latter, this figure is 2876 clusters (22.2% of 12 970 clusters).

Even though most clusters model AE separately, some sharing does occur in the targeted case. However, by comparing the results of the accent-specific and targeted multi-accent acoustic model sets in Table III, this small degree of sharing seems to lead to a deterioration compared to the case where accents are clustered separately from the outset.

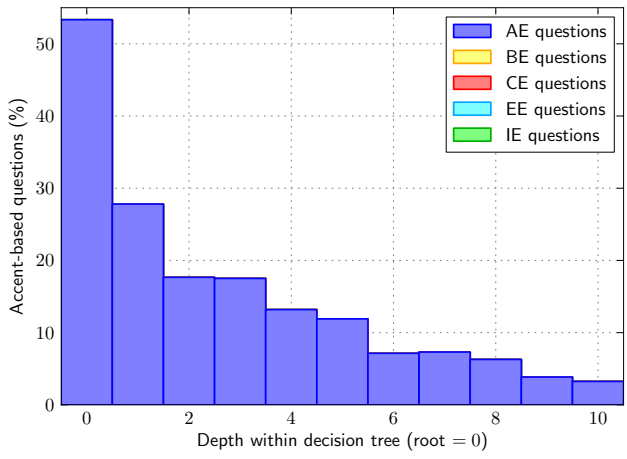


Fig. 3. The percentage of questions that relate to specific accents at various depths within the decision-trees for the targeted multi-accent acoustic model set with optimal recognition performance on the AE development set.

Although the comparative analysis presented in this section was described for the AE accent, the same trends were observed for the other four accents. Empirically we have therefore shown that the decision-trees constructed during targeted multi-accent acoustic modelling tend to model the target accent separately. However, this leads to deteriorated performance compared to simple accent-specific acoustic modelling.

D. Weighted targeted multi-accent acoustic modelling

The preceding section showed that targeted multi-accent decision-trees tend strongly towards the separation of the target accent. In this section we propose a further variant of the standard decision-tree state clustering algorithm (as applied in multi-accent modelling) in order to counteract this tendency.

Suppose again that we have a cluster of states $\mathbb{S} = \mathbb{S}_x \cup \mathbb{S}_t$ with the states $\mathbb{S}_x$ generating observation vectors for frames $\mathbb{F}_x$ and $\mathbb{S}_t$ generating observation vectors for frames $\mathbb{F}_t$. We propose that, instead of basing our cluster splitting criterion solely on the log likelihood $L_t(\mathbb{S})$ on the target states $\mathbb{S}_t$, we also assign some weight to the log likelihood $L_x(\mathbb{S})$ of the non-target states $\mathbb{S}_x$ generating the observation vectors $\mathbb{F}_x$. We calculate this alternative log likelihood as

$$L_w(\mathbb{S}) = w_t L_t(\mathbb{S}) + w_x L_x(\mathbb{S}) \tag{7}$$

with $w_t > 0$, $w_x > 0$ and $w_x = 1 - w_t$. The likelihood $L_t(\mathbb{S})$ is calculated according to (5) and, analogously, $L_x(\mathbb{S})$ is calculated as

$$\begin{aligned} L_x(\mathbb{S}) = -\frac{1}{2} N_x \{\log[(2\pi)^n |\boldsymbol{\Sigma}(\mathbb{S})|]\} - \frac{1}{2} n(N_t + N_x) \\ + \frac{1}{2} \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbb{S}) N_t [\boldsymbol{\Sigma}(\mathbb{S}_t) \\ + (\boldsymbol{\mu}(\mathbb{S}_t) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_t) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}]\} \end{aligned} \tag{8}$$

In this last equation the roles of the target and non-target states are simply reversed from the case presented in (5).

In Appendix D we show that when $w_t = w_x = 1/2$, this weighted targeted log likelihood reduces to $L_w(\mathbb{S}) = 1/2\, L(\mathbb{S})$ with $L(\mathbb{S})$ the overall log likelihood as in (2) and (3). Thus, when using equal weights, this new cluster splitting criterion is equivalent to that used for multi-accent acoustic modelling as described in Section V-B. When $w_t = 1$ and $w_x = 0$, we have $L_w(\mathbb{S}) = L_t(\mathbb{S})$, which is the unweighted targeted case presented in Section VI-B. Both multi-accent acoustic modelling and targeted multi-accent acoustic modelling are therefore special cases of this *weighted targeted multi-accent acoustic modelling* approach.

E. Evaluation and analysis: weighted targeted modelling

We again considered each of the SAE accents in turn as the target accent and applied the weighted targeted approach to the five training sets. Phone recognition performance is shown as the fifth entry in Table III. For each accent the target weight w_t was optimised on its development set. These weights are indicated in the final line of Table III.

The weighted targeted multi-accent model set achieves improved performance for AE and IE. Although multi-accent modelling is a special case of the weighted targeted approach, poorer performance might still occur since the weights are optimised on a development set. This is illustrated by the performance on CE, for instance, where accuracy deteriorates from 66.31% to 66.13%. For BE and EE, the target weight was determined to be 0.5 and the performance of the multi-accent models is therefore achieved: 56.56% and 73.94% respectively. The average performance of the weighted targeted approach is better than that achieved by any of the other approaches. The improvements in average accuracy of the weighted targeted multi-accent models (65.65%, Table III) over the accent-independent (65.44%) and multi-accent models (64.50%) are both statistically significant at the 80% level.

Figure 4 analyses the decision-trees of the weighted targeted multi-accent acoustic models giving optimal performance on the AE development set. Since the weight assigned to the target is small (0.51), the decision-trees are very similar to the non-targeted case shown in Figure 2. Of the 12 823 resulting clusters in the weighted targeted decision-trees, 13.6% are AE-only, 22.2% share AE with some other accent(s) and 64.2% are non-AE. The AE-only clusters are therefore slightly more here than in the trees analysed in Figure 2 where 13.2% of the 12 970 clusters were AE-only (final paragraph, Section V-B).

Although the improvements of the weighted targeted multi-accent acoustic modelling approach over the other approaches are relatively small, they do indicate that some gain can be obtained by targeting the decision-tree likelihood optimisation on a specific accent in this manner.

VII. Summary and Conclusions

We have described new techniques that extend the standard decision-tree state clustering algorithm used to construct tied-state hidden Markov models to allow explicit optimisation on a target accent. Using databases for the five accents of South African English, we compared these new techniques to the accent-specific, accent-independent and multi-accent acoustic modelling approaches developed in previous work.

We showed that it is possible to derive expressions that allow the tractable implementation of the new clustering methods. In a first approach, the decision-tree state clustering process was altered so that the likelihood criterion used during decision-tree construction is calculated only on a target accent. Phonetic or accent-based questions are then asked only when it is advantageous for the target accent. However, both per-accent and overall average phone recognition performance indicated that this approach leads to poorer models compared to those obtained previously. Further analysis indicated that this is mostly due to the tendency of the targeted decision-trees to separate out the target accent into isolated clusters.

In order to alleviate this tendency towards separate modelling, we implemented a further extension to the algorithm in which the likelihood criterion also assigns some weight to the likelihood on non-target accents. By weighting the likelihoods on the target and non-target accents, the amount of separation could be controlled. Using this weighted targeted multi-accent modelling approach, very small average improvements ($\sim$0.2% absolute) were obtained over all other approaches.

In future work the proposed techniques should be compared to classical adaptation approaches. Clustering is also performed fairly early on in the complete acoustic model training process and is performed on the training set; changes in state-tying do not guarantee improvements for the final higher-mixture acoustic models. This warrants further investigation.

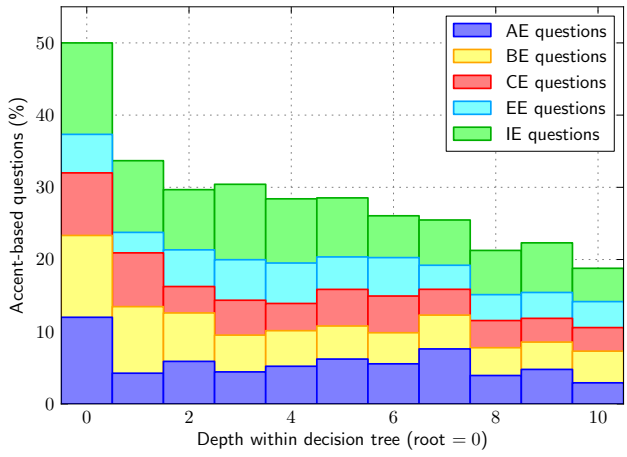


Fig. 4. The percentage of questions that relate to specific accents at various depths within the decision-trees for the weighted targeted multi-accent acoustic model set with optimal recognition performance on the AE development set ($w_t = 0.51$).

Appendix A
Log Likelihood of a Cluster of States

The log likelihood that the observation vectors were generated by the states in $\mathbb{S}$ can be calculated as

$$L(\mathbb{S}) = \log \prod_{f \in \mathbb{F}} p(\mathbf{o}_f|\mathbb{S}) = \sum_{f \in \mathbb{F}} \log \left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] \quad \text{(A.9)}$$

where the observation PDFs are assumed to be single-mixture Gaussian PDFs:

$$\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S})) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}(\mathbb{S})|}} e^{\left\{-\frac{1}{2}(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))\right\}} \quad \text{(A.10)}$$

From (A.10), equation (A.9) can then be written as

$$L(\mathbb{S}) = -\frac{1}{2} \sum_{f \in \mathbb{F}} \log[(2\pi)^n |\boldsymbol{\Sigma}(\mathbb{S})|] - \frac{1}{2} \sum_{f \in \mathbb{F}} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \quad \text{(A.11)}$$

The covariance matrix of the cluster of states $\mathbb{S}$ can be calculated as

$$\boldsymbol{\Sigma}(\mathbb{S}) = \frac{1}{N} \sum_{f \in \mathbb{F}} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \quad \text{(A.12)}$$

where N is the number of frames in $\mathbb{F}$ and given by

$$N = \sum_{s\in\mathbb{S}} \sum_{f\in\mathbb{F}} \gamma_s(\mathbf{o}_f) \tag{A.13}$$

with $\gamma_s(\mathbf{o}_f)$ the posterior probability that the observation vector $\mathbf{o}_f$ is generated by HMM state s. By cross-multiplication, equation (A.12) becomes

$$N\,\mathbf{I} = \sum_{f\in\mathbb{F}} \boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \tag{A.14}$$

In [16, p. 62] the matrix identity

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^{\mathrm{T}}) \tag{A.15}$$

is given, where $\mathbf{x}$ is an $n\times 1$ vector, $\mathbf{A}$ is an $n\times n$ matrix and tr denotes the trace of a matrix. By taking the trace of both sides of (A.14) and then applying (A.15) we obtain

$$\begin{aligned} nN &= \mathrm{tr}\left[\sum_{f\in\mathbb{F}} \boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \\ &= \sum_{f\in\mathbb{F}} \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \\ &= \sum_{f\in\mathbb{F}} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \end{aligned} \tag{A.16}$$

where n is the dimensionality of the observation vectors.

By substituting (A.16) into (A.11) we obtain the result:

$$\begin{aligned} L(\mathbb{S}) &= -\frac{1}{2}\sum_{f\in\mathbb{F}} \log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] - \frac{1}{2}nN \\ &= -\frac{1}{2}\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] + n\right\} N \\ &= -\frac{1}{2}\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] + n\right\} \sum_{s\in\mathbb{S}}\sum_{f\in\mathbb{F}} \gamma_s(\mathbf{o}_f) \end{aligned} \tag{A.17}$$

APPENDIX B
LOG LIKELIHOOD OF A TARGETED SUBSET OF STATES

The log likelihood of states $\mathbb{S}_t$ generating the associated observation vectors for frames $\mathbb{F}_t$ can be calculated as

$$\begin{aligned} L_t(\mathbb{S}) &= \sum_{f\in\mathbb{F}_t} \log\left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] \\ &= -\frac{1}{2}\sum_{f\in\mathbb{F}_t} \log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] \\ &\quad - \frac{1}{2}\sum_{f\in\mathbb{F}_t} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \\ &= -\frac{1}{2}N_t\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} \\ &\quad - \frac{1}{2}\sum_{f\in\mathbb{F}_t} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \end{aligned} \tag{B.18}$$

where

$$N_t = \sum_{s\in\mathbb{S}_t}\sum_{f\in\mathbb{F}} \gamma_s(\mathbf{o}_f) \text{ and } N_x = \sum_{s\in\mathbb{S}_x}\sum_{f\in\mathbb{F}} \gamma_s(\mathbf{o}_f) \tag{B.19}$$

Calculation of the second term in (B.18) is slightly involved and we derive an expression for this term as follows.

The covariance matrix of the PDF of the cluster $\mathbb{S}$ is

$$\begin{aligned} \Sigma(\mathbb{S}) &= \frac{1}{N_x + N_t}\sum_{f\in\mathbb{F}} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \\ &= \frac{1}{N_x + N_t}\left[\sum_{f\in\mathbb{F}_x} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right. \\ &\qquad \left. + \sum_{f\in\mathbb{F}_t} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \end{aligned} \tag{B.20}$$

which leads to

$$\begin{aligned} \boldsymbol{\Sigma}(\mathbb{S})\,(N_x + N_t) &= \sum_{f\in\mathbb{F}_x} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \\ &\quad + \sum_{f\in\mathbb{F}_t} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \end{aligned} \tag{B.21}$$

An expression for the first term on the right hand side of (B.21) can be obtained as follows:

$$\begin{aligned} &\sum_{f\in\mathbb{F}_x} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \\ &= \sum_{f\in\mathbb{F}_x} ((\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x)) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S})))\times \\ &\qquad ((\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x)) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S})))^{\mathrm{T}} \\ &= \sum_{f\in\mathbb{F}_x} \left[(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))^{\mathrm{T}}\right. \\ &\qquad + (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \\ &\qquad + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))^{\mathrm{T}} \\ &\qquad \left. + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \\ &= N_x\,\boldsymbol{\Sigma}(\mathbb{S}_x) + \sum_{f\in\mathbb{F}_x} (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \\ &= N_x\left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \end{aligned} \tag{B.22}$$

where, in the third step, we used the definitions:

$$\boldsymbol{\mu}(\mathbb{S}_x) = \frac{1}{N_x}\sum_{f\in\mathbb{F}_x} \mathbf{o}_f \tag{B.23}$$

and

$$\boldsymbol{\Sigma}(\mathbb{S}_x) = \frac{1}{N_x}\sum_{f\in\mathbb{F}_x} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}_x))^{\mathrm{T}} \tag{B.24}$$

By substituting (B.22) into (B.21), it follows that

$$\begin{aligned} &\boldsymbol{\Sigma}(\mathbb{S})\,(N_x + N_t) = \\ &\quad N_x\left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right] \\ &\quad + \sum_{f\in\mathbb{F}_t} (\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}} \end{aligned} \tag{B.25}$$

Multiply (B.25) with $\boldsymbol{\Sigma}^{-1}(\mathbb{S})$ and take the trace:

$$\begin{aligned} &n(N_x + N_t) = \\ &\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbb{S})N_x\left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right]\right\} \\ &+ \sum_{f\in\mathbb{F}_t} \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right\} \end{aligned} \tag{B.26}$$

and use the identity in (A.15):

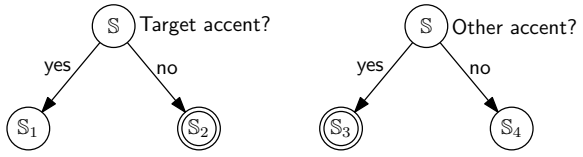
$$n(N_x + N_t) = \\ \text{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbb{S})N_x\left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right]\right\} \\ + \sum_{f\in\mathbb{F}_t}(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \quad \text{(B.27)}$$

The last term on the right hand side of (B.27) is the required second term in (B.18). We thus obtain the result:

$$\begin{aligned} L_t(\mathbb{S}) = &-\frac{1}{2}N_t\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} - \frac{1}{2}n(N_x + N_t) \\ &+ \frac{1}{2}\text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbb{S})N_x[\boldsymbol{\Sigma}(\mathbb{S}_x) \\ &+ (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}]\} \quad \text{(B.28)} \end{aligned}$$

APPENDIX C
ACCENT-BASED QUESTIONS IN TARGETED MULTI-ACCENT DECISION-TREES

Consider the two possible cluster splits illustrated in Figure 5. Assume we are using $L_t(\mathbb{S})$ as splitting criterion. In (a) the question relates to the target accent, e.g. "is the accent AE?" (assuming we optimise AE). In (b) the question relates to some non-target accent, e.g. "is the accent EE?". We show that case (a) will always occur rather than case (b).



(a) Question relates to the target accent.

(b) Question relates to some non-target accent.

Fig. 5. Two potential questions split the cluster $\mathbb{S}$. Mathematically it can be shown that (a) will always occur rather than (b).

$\mathbb{S}_2$ will contain no states from the target accent and $L_t(\mathbb{S}_2) = 0$; this cluster would therefore be a leaf node. Similarly, $\mathbb{S}_3$ will contain no states from the target accent and $L_t(\mathbb{S}_3) = 0$; again resulting in a leaf node. What distinguishes the likelihood improvement in the two cases is therefore $L_t(\mathbb{S}_1)$ and $L_t(\mathbb{S}_4)$. The former is given by

$$L_t(\mathbb{S}_1) = L_t(\mathbb{S}_t) = \sum_{f\in\mathbb{F}_t}\log\left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}_t), \boldsymbol{\Sigma}(\mathbb{S}_t))\right] \quad \text{(C.29)}$$

in accordance with (4). This log likelihood is the one maximised when performing maximum likelihood estimation of $\boldsymbol{\mu}(\mathbb{S}_t)$ and $\boldsymbol{\Sigma}(\mathbb{S}_t)$ on frames $\mathbb{F}_t$. For case (b) we have

$$L_t(\mathbb{S}_4) = \sum_{f\in\mathbb{F}_t}\log\left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}_4), \boldsymbol{\Sigma}(\mathbb{S}_4))\right] \quad \text{(C.30)}$$

with $\mathbb{S}_t \subset \mathbb{S}_4$. In this case $\boldsymbol{\mu}(\mathbb{S}_4)$ and $\boldsymbol{\Sigma}(\mathbb{S}_4)$ are obtained by maximising the log likelihood on all the frames $\mathbb{F}_4$ associated with $\mathbb{S}_4$, which is different to the calculation in (C.30) since $\mathbb{F}_t \subset \mathbb{F}_4$. It thus follows that $L_t(\mathbb{S}_4) < L_t(\mathbb{S}_1)$. The target-accent-question (a) will therefore always be asked in favour of a non-target-accent-question (b).

APPENDIX D
EQUAL WEIGHT TARGETED MODELLING

Using the form of (B.18) for both $L_t(\mathbb{S})$ and $L_x(\mathbb{S})$, we obtain the following result when $w_t = w_x = 1/2$:

$$\begin{aligned} L_w(\mathbb{S}) &= \frac{1}{2}L_t(\mathbb{S}) + \frac{1}{2}L_x(\mathbb{S}) \\ &= -\frac{1}{4}(N_x + N_t)\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} \\ &\quad - \frac{1}{4}\sum_{f\in\mathbb{F}}(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbb{S})(\mathbf{o}_f - \boldsymbol{\mu}(\mathbb{S})) \\ &= -\frac{1}{4}(N_x + N_t)\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} - \frac{1}{4}n(N_x + N_t) \\ &= -\frac{1}{4}\left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|] + n\right\}N = \frac{1}{2}L(\mathbb{S}) \quad \text{(D.31)} \end{aligned}$$

where $N = N_x + N_t$ and we used (A.16) in the third line.

ACKNOWLEDGEMENTS

Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University.

REFERENCES

[1] H. Kamper, F. J. Muamba Mukanya, and T. R. Niesler, "Multi-accent acoustic modelling of South African English," *Speech Communication*, vol. 54, no. 6, pp. 801–813, 2012.

[2] E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., *A Handbook of Varieties of English*. Berlin, Germany: Mouton de Gruyter, 2004.

[3] V. Fischer, Y. Gao, and E. Janke, "Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 787–790.

[4] R. Chengalvarayan, "Accent-independent universal HMM-based speech recognizer for American, Australian and British English," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2733–2736.

[5] K. Kirchhoff and D. Vergyri, "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition," *Speech Commun.*, vol. 46, no. 1, pp. 37–51, 2005.

[6] J. Despres, P. Fousek, J. L. Gauvain, S. Gay, Y. Josse, L. Lamel, and A. Messaoudi, "Modeling Northern and Southern varieties of Dutch for STT," in *Proc. Interspeech*, Brighton, 2009, pp. 96–99.

[7] M. Caballero, A. Moreno, and A. Nogueiras, "Multidialectal Spanish acoustic modeling for speech recognition," *Speech Commun.*, vol. 51, pp. 217–229, 2009.

[8] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.

[9] J. C. Roux, P. H. Louw, and T. R. Niesler, "The African Speech Technology project: An assessment," in *Proc. LREC*, Lisbon, Portugal, 2004, pp. 93–96.

[10] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. L. Moore, J. J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.

[11] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, CO, 2002, pp. 901–904.

[12] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, pp. 359–394, 1999.

[13] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop Human Lang. Technol.*, Plainsboro, NJ, 1994, pp. 307–312.

[14] J. J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 1995.

[15] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. ICASSP*, Montreal, Quebec, Canada, 2004, pp. 409–412.

[16] H. A. Engelbrecht, "Automatic phoneme recognition of South African English," Master's thesis, Stellenbosch University, 2004.