Cross-bandwidth adaptation for ASR systems

Neil Kleynhans* and Etienne Barnard†
*Human Language Technologies Research Group Meraka Institute, CSIR, South Africa
* School of Electrical, Electronic and Computer Engineering, North-West University
† Multilingual Speech Technologies Group, North-West University
Email: {ntkleynhans,etienne.barnard}@gmail.com

***Abstract*—Mismatches between application and training data greatly reduce the performance of automatic speech recognition (ASR) systems. However, collecting suitable amounts of in-domain and application-specific data for training is resource intensive and may not be feasible for resource-scarce environments. Utilising limited amounts of in-domain data and a combination of feature normalisation and acoustic model adaptation techniques has therefore found wide use in ASR systems. Various approaches have been proposed, and it is not clear when to make use of a particular approach given a specific amount of adaptation data. In this work we investigate the use of standard feature normalisation and model adaptation techniques, for the scenario where adaptation between narrow- and wide-band environments must be performed. Our investigation focuses on the dependence of the adaptation data amount and various adaptation techniques by systematically varying the adaptation data amount and comparing the performance of various adaptation techniques. From this we establish a guideline which can be used by an ASR developer to choose the best adaptation technique given a size constraint on the adaptation data. In addition, we investigate the effectiveness of a novel channel normalisation technique and compare the performance with standard normalisation and adaptation techniques.**

I. Introduction

It is well known that speech recognition systems perform poorly when there is a mismatch between the acoustic models and testing audio data. The mismatch can manifest itself in several ways; the leading causes are environmental noise, channel differences, various speaking styles and different dialects. A system's performance can be greatly increased if the mismatch is sufficiently reduced. Using task-specific corpora could negate the acoustic mismatch, but these corpora are often difficult to come by and for resource-scarce languages the choices are severely limited. Alternatively, acoustic model adaptation and feature normalisation techniques provide a means to reduce the mismatch and play a crucial role in speech recognition system deployment. Feature normalisation improves the feature robustness by trying to remove channel or environmental distortions while acoustic model adaptation shifts the model's means and scales the variances to accommodate for the change in data statistics. In general, model-based adaptations perform better than feature normalisation approaches but require transcriptions to estimate the class-specific mismatches and apply the appropriate transforms.

In a resource-scarce environment, adaptation techniques are particularly important as one often has to develop ASR systems using available data which is not matched to the application: for instance, using high-bandwidth data to train acoustic models for an application that operates on telephone quality audio data – over the course of using the system, application-specific data can be collected and used to adapt or create new acoustic models and thereby improve the performance. Generally, given access to a language-specific corpus, it would be highly efficient to train acoustic models with the available data and then apply task-specific optimisations. When moving between different operating environments, the optimisations would have to take into consideration the data mismatch which leads to performance degradations.

Thus, we will investigate unsupervised techniques for channel normalisation, which can be applied to mismatched data applications. In addition, current ASR model adaptation techniques learn a set of transformations or update acoustic model parameters from provided adaptation data. The performance gains which are attained by the various techniques are dependent on the amount of adaptation data from which the statistics are estimated. Therefore, a comparative investigation will be performed to determine the effectiveness of current model adaptation techniques based on the amount of available adaptation data. The specific scenario that will be investigated is one in which plentiful speech data resources of either telephone (narrow) bandwidth or high bandwidth are available. We will investigate how feature normalisation and model adaptation techniques can increase the ASR system performance gains given increasing amounts of adaptation data from the less-resourced bandwidth.

II. Background

Leggetter and Woodland [1] showed in their speaker-adaptation experiments that mean-only Maximum Likelihood Linear Regression (MLLR) adaptation, using a full global regression matrix, yielded improvements after three adaptation utterances (roughly 11 seconds of speech). The performance gain saturates at about 15 utterances. To make better use of the additional adaptation data, additional regression classes were suggested.

Gauvain and Lee [2] managed to achieve significant improvements in word-error-rates (WER) using Maximum a-Posteriori (MAP) adapted models compared to Maximum Likelihood Estimation (MLE) trained models. In their experiments they used three model types: (1) speaker-dependent (SD) models trained on a specific speaker's data only, using MLE training, (2) speaker-adapted models (SA-1) which were created by MAP adapting a speaker-independent model – a model trained on data sourced from many speakers –, and, (3) a second set of speaker-adapted models (SA-2), created by MAP adapting gender-dependent models – models trained on female or male data only. Table I shows a summary of

the results using the various acoustic models MAP adapted or MLE trained at different data amounts (adapted from Gauvain and Lee [2]).

TABLE I. WERs FOR MAP-ADAPTED AND MLE-TRAINED ACOUSTIC MODELS AT DIFFERENT DATA AMOUNTS. ADAPTED FROM GAUVAIN AND LEE [2].

Model Type	0 Min	2 Min	5 Min	30 Min
SA-1	13.5 %	8.7 %	6.9%	3.4 %
SA-2	11.9 %	7.5%	6.0 %	3.5 %
SD	-	31.5	12.1%	3.5 %

Wallace *et. al.* [3] investigated various supervised and unsupervised adaptation techniques to improve automatic transcription generation using speech recognition to extract transcriptions from telephony-quality audio data. Their experiments focused on speaker-dependent adaptation. The supervised adaptation experiments used hand normalised transcriptions while the unsupervised adaptation approach used transcriptions generated by the baseline non-adapted acoustic models. The supervised techniques showed continued WER reductions for the following order of adaptation techniques: global MLLR, regression tree MLLR, MAP, cascaded global plus regression tree MLLR, and, cascaded global plus regression tree MLLR plus MAP. For the unsupervised adaptation experiments the order is somewhat different: global MLLR, MAP, regression tree MLLR, cascaded global plus regression tree MLLR, and, cascaded global plus regression tree MLLR plus MAP. The unsupervised experiments highlight the sensitivity of MAP to inaccurate transcriptions and robustness of the MLLR approach. The adaptation amount experiments, which used 10, 30 and 60 minutes of adaptation data, showed for both supervised and unsupervised adaptation, the global MLLR approach could not provide further performance gains after 30 minutes of data. The cascaded global plus regression tree MLLR and cascaded global plus regression tree MLLR plus MAP showed continued improvements as more adaptation was data added.

Wang *et. al.* [4] showed that for non-native speaker adaptation and for a fixed number of speakers, MAP adaptation consistently performed better than MLLR, independent of adaptation data amount. The only scenario where MLLR provided a performance gain over MAP was when the number of speakers found in the adaptation data were varied; however, even in that case MAP proved a better choice once the adaptation data amount exceeded 20 minutes.

For the specific ASR scenario describe by Bocchieri *et. al.* [5], MAP adaptation provided the best results for fewer than 1500 sentences, which equated to 3.5 hours of audio data and roughly 2 hours of speech data. Between 1500 and 6000 sentences, training the context-trees on in-domain data and estimating the state distributions on both in- and out-domain data resulted in the best performance – 6000 sentences corresponded to 14.4 hours of audio data and approximately 9 hours of speech data. For 6000 sentences and above, retraining the Hidden Markov Model (HMM) acoustic models on the in-domain data provided the best results. However, no data threshold was provided for the use of MLLR. In summary, Bocchieri *et al.* [5] described a strategy to port existing acoustic models to new applications. Their approach for increasing amounts of in-domain data t is (adapted from [5]):

- If $0 < t < t_{mllr}$ use out-of-domain acoustic models,
- If $t_{mllr} < t < t_{map}$ use MLLR adapted out-of-domain acoustic models.
- If $t_{map} < t < t_{cntx}$ use MAP adapted out-of-domain acoustic models.
- If $t_{cntx} < t < t_{new}$ retrain acoustic models on in-domain and out-of-domain data (building context-trees on in-domain data).
- If $t_{new} < t$ retrain acoustic models on in-domain data.

We can, in summary, state that the expected order of increasing performance gains provided by the use of various adaptation techniques is: feature normalisation, adaptation by MLLR transformation, MAP adaptation and retraining the models. We can see, however, that the boundaries where one would chose a specific adaptation technique over another are quite varied. The transition boundaries are dependent on the ASR task where parameters such as speaker number and adaptation type (environment, dialect or speaker) have an influence over the transitions. The boundary measure is usually specified by duration, generally measured in minutes or hours. This can be misleading, as adaptation performance is likely to depend on how much speech data is actually available within the audio data. For instance read, conversational or distressed speech would all have different ratios of speech to non-speech. In terms of current HMM-based ASR systems a more informative unit would be the total number of words, phones or triphones found in the adaptation data. Table II shows the duration (in seconds) of audio per 1000 phones for various corpora – as can be seen the numbers vary, thus specifying adaptation data amount in time units will result in differing training unit amounts across the corpora.

TABLE II. DURATION, IN SECONDS, OF AUDIO PER 1000 PHONES FOR VARIOUS CORPORA.

Corpus	Type	Duration per 1000 phones (seconds)
TIMIT (train)	Read	78.92
WSJ0 (train)	Read	91.82
MoneyWeb (train)	Conversational	98.21
Lwazi English	Read	90.94
NCHLT English (RAW)	Read	129.77
BN (high-fidelity speech)	Read	78.44

For the course of our experiments we make use of standard normalisation (investigating an unsupervised transfer-function filtering approach in addition) and adaptation techniques and report on additional evidence regarding the relative contributions of different adaptation methods. Some of these findings confirm facts that are already known in the literature. The amount of adaptation data will be specified by the number of triphones which provides a standard calibration unit for ASR systems. As stated, our experiments will focus on mixed narrow-bandwidth telephone-quality and high-bandwidth high-quality audio data applications and investigate how to port acoustic models starting from a narrow-bandwidth scenario and progress towards a high-bandwidth one and vice versa – a scenario of great relevance in developing-world contexts.

Lastly, from the literature survey, the cited work made use of standard acoustic model adaptation techniques to either adapted speaker-independent models to speaker-dependent models or adapted out-of-domain models using in-domain data. In all cases, however, the in- and out-domain data had similar channel and environmental characteristics and the purpose of their research was to reduce the mismatch caused by

speaker characteristics and differing triphonic content. Since our work focuses on model adaptation and data normalisation between mixed-bandwidth and differing quality audio data, we are provided with an opportunity to investigate whether the established data-related performance gains of the various adaptation techniques hold for bandwidth adaptation as well.

The format of the work is structured in the following manner: a description of the various feature normalisation and adaptation techniques used in our experiments can be found in Section (III), the corpora used, the data selection strategy, ASR system and performance gain curves are described in Section (IV), results are presented in Section (V) and concluding remarks can be found in Section (VI).

III. Method

A. Feature Normalisation

Feature normalisation techniques strive to remove biases in data statistics introduced by environmental or channel variations. This is achieved by applying a set of transforms to the feature vectors which either normalise the feature vectors to a standard set of values or transform them to the training set values.

A simple feature normalisation strategy is to band-limit the spectral content of the audio signals. Moreno and Stern [6] demonstrated the importance of matching the portion of speech bandwidth which is used to extract speech features on the Timit and NTimit corpora.

In the cepstral domain, cepstral mean normalisation (CMN) is widely used. This method estimates an average cepstral vector over a set of cepstral observations and removes the bias from each vector – the technique performs well in removing convolutional noise and constant channel distortions.

Chen and Bilmes [7] showed through their in-depth analysis that CMN worked well at removing convolution noise but performed poorly in removing additive noise. It was further shown that the effects of additive noise, depending on the noise level, can be reduced if variance normalisation was applied to the cepstral coefficients. Lastly, Chen and Bilmes re-introduced filtering of the cepstral coefficients which limits the modulation frequencies and improves the dynamic range of the cepstral trajectories by suppressing noise effects. The cepstral trajectories were filtered using a finite length auto-regression moving average (ARMA) filter. Based on the gains reported by Chen and Bilmes [7], all extracted feature vectors are normalised using their approach (referred to as *MVA*).

1) Transfer-Function Filtering: Gelbart and Morgan [8] showed that feature normalisation can be achieved by removing a long-term average log spectral estimate from spectral analysis frames. Their technique, however, required lengthy speech segments to estimate the average log spectrum and relatively longer analysis windows. Such delays would be impractical for real-time ASR systems. It was shown previously [9] that channel normalisation can be performed by inverse filtering the short-term spectra.

Starting with the basic idea we formulated a slightly different approach. If it is assumed that the discrete cosine transform (DCT) of the logarithmic short-term spectra are drawn from a multivariate Gaussian distribution, then channel normalisation can be realised by the mapping of normal distributions. The first step is to estimate the *mean* (μ) and *covariance* (Σ) statistical moments, which, using the MLE, are given by

$$E[\mathbf{X}] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i \quad (1)$$

$$\mu = E[\mathbf{X}] \quad (2)$$

$$\Sigma = E[\mathbf{X}^2] - E[\mathbf{X}]^2, \quad (3)$$

where μ is the mean value, Σ is the covariance and $E[]$ is the expected value operator. The measures are extracted from channel-specific data and require no transcriptions. The manner in which the estimates are obtained are as follows;

- Block the audio in 25 ms frames and overlap consecutive frames by 10 ms (standard values used in current ASR feature extraction).
- Firstly apply the logarithmic transform to short-term frame spectra then apply the discrete cosine transform.
- Update the mean and covariance accumulators.
- Once all frames have been processed, calculate the final mean and covariance values.

Given various mean and covariance statistics estimated from different channels, one set of feature vectors (in our case the DCT of the logarithmically mapped short-term spectra) can be normalised to another distribution by firstly normalising the feature vectors to zero mean and unit covariance distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, then applying an affine transform to the feature vectors to shift their statistics such that they will produce the target mean and covariance measures. This is achieved by applying the following steps to each feature vector:

- $\mathbf{Z}_{zero} = \mathbf{A}_{src}^{-1}(\mathbf{Z}_{src} - \mu_{src})$, where $\mathbf{A}_{src}$ is given by the Cholesky decomposition of Σ_{src}, $\mathbf{A}_{src}\mathbf{A}_{src}^T = \Sigma_{src}$.
- $\mathbf{Z}_{tgt} = \mathbf{A}_{tgt}\mathbf{Z}_{zero} + \mu_{tgt}$, where $\mathbf{A}_{tgt}$ is given by the Cholesky decomposition of Σ_{tgt}, $\mathbf{A}_{tgt}\mathbf{A}_{tgt}^T = \Sigma_{tgt}$.

After transforming a feature vector to the most likely target vector, the inverse discrete transform is applied and the values mapped by the exponential function. This channel normalised linear spectrum is sent through to the feature extraction unit for final processing. This technique is similar to the one proposed by Gelbart and Morgan [8] but differs in the following ways: (1) applied to short-term spectrum 25 ms, and, (2) extends log spectral subtraction by assuming the analysis frames are drawn from a Gaussian distribution and applies mean and variance normalisation.

B. Model Adaptation

1) Maximum Likelihood Linear Regression: The MLLR technique estimates a set of linear transforms from adaptation data, then updates the model parameters by applying the transforms to the mean and variance parameters. The technique also requires relatively small amounts of adaptation data since it uses binary regression class trees to group similar models together and thus create larger class-specific pools of adaptation data. The MLLR implementation is elegant since multiple transforms can be applied to model parameters. For instance a typical transform estimation process would initially estimate a set of mean transforms, then apply these mean transforms to

the models, estimate a set of variance transforms – forming a cascade of transforms.

For our MLLR experiments we used the following approach;

1) Estimate a 40-class regression tree.
2) Estimate 40-class-specific semi-tied transforms[10].
3) Using the semi-tied transforms as parent transforms, estimate 40-class-specific mean transforms.
4) Using the mean transforms as parent transforms, estimate 40-class-specific variance transforms.

The number of regression classes was set to 40 which correlates well to the average number of sound classes in a language. These mean and variance transforms are stored in separate files and are loaded and applied to the models during recognition.

2) Maximum A-Posteriori adaptation: MAP adaptation provides a means to adapt the model parameters without having to retrain the models from scratch. The MAP training procedure incorporates prior information which provides a parameter estimation benefit compared to standard MLE parameter estimates [2]. The effectiveness of MAP adaptation is only fully realised at relatively large data amounts as the technique updates different model components separately. Thus, the adaptation data must cover quite a large set of different training examples and each example a sufficient number of times. Gauvain and Lee [2] showed, however, that using MAP adaptation to speaker adapt existing speaker-independent models requires much less data to gain substantial improvements in the WERs (compared to retraining the models). Therefore, it does seem that MAP possesses a lower critical data limit than the limit needed to train robust acoustic models.

HTK [11] provides a mechanism to update the weights, means, variances and various combinations of the these. The MAP adaptation experiments that we performed either adapted the weights-means combination or weights-means-variances combination and used 10 adaptation iterations.

IV. Experimental Setup

A. Corpora

The various feature normalisation and model adaptation experiments were performed on pairs of American English and IsiNdebele read-speech corpora. To ensure a mismatch between corpora a narrow- and high-bandwidth version of each language were chosen, and to simulate the typical environment for low-resource languages, we experimented using within-language cross-channel adaptation.

1) Wall Street Journal: The Wall Street Journal (WSJ) Continuous Speech Recognition (CSR) corpus contains high-bandwidth American English read-speech utterances and orthographic transcriptions [12]. For our purposes we sourced the speaker-independent sub-corpus which contains a separate training and testing set with no speaker overlap. The audio was recorded with a Sennheiser microphone at a sample rate of 16 kHz and contains financially oriented content. The transcriptions contain three text subsets: a small set spoken by all the speakers, a few sentences which have limited speaker overlap and a unique sentence set. There are an equal number of male and female speakers. Table III shows the make-up of the WSJ corpus.

TABLE III. The WSJ corpus statistics for the training and testing sets.

Set Type	# utterances	# hours	# speakers
Train	12776	24.9	101
Test	1142	2.2	10

2) NTimit: NTimit (Network Timit) is a narrow-bandwidth telephone-quality read-speech corpus [13]. NTimit was created by transmitting the Timit corpus data through "local" and "long-distance" telephone networks in the United States. The purpose of the NTimit corpus was to aid in the investigation of telephone network distortions on speech. The Timit corpus is an high-bandwidth American English read-speech corpus [14]. The main corpus design criteria ensured phonetic diversity which enables the study of general speech characteristics. The data was collected across the United States and encompassed the eight main dialect regions of the country. Each speaker contributed ten sentences; two were common to all speakers and were used to investigate dialect variations, five were selected to provide phonetic diversity and the last three were sourced from the Brown corpus. Table IV shows the NTimit corpus statistics.

TABLE IV. The NTimit corpus statistics for the training and testing sets.

Set Type	# utterances	# hours	# speakers
Train	4617	3.9	462
Test	1675	1.4	168

3) NCHLT: The NCHLT corpus is high-bandwidth read-speech corpus containing audio data and transcriptions collected from eleven South African languages [15]. The audio data was recorded using mobile devices. The transcriptions contain short sentences and were derived from large text corpora in order to attain coverage of the most common triphones of the target language. For our cross-channel experiments we limited ourselves to using the IsiNdebele sub-corpus (which was the only completed sub-corpus at the initiation of our experiments). The initial corpus contained 90297 utterances collected from 209 speakers. After running pre-processing, which removed utterances that contained English words, clipped audio data and audio files containing incorrect header information, the corpus was reduced to 60687 utterances and 169 speakers. For English word detection we employed an in-house English pronunciation dictionary and created a lookup table containing a list of all the words found in the dictionary. The NCHLT corpus does not have a dedicated training and testing set; hence, five-fold cross validation was used to partition the corpus and create the desired sets. Table V shows the five-fold training/adaptation and testing corpus statistics.

4) Lwazi: The Lwazi corpus contains read and elicited speech recordings collected from eleven South African languages [16]. There are approximately 200 speakers per language and the audio data was recorded over the telephone network. Each speaker contributed thirty utterances; sixteen sentences were sourced from phonetically rich text while the remaining 14 sentences were elicited by questions that produced either short phrases or single words (e.g. yes/no answers, digits, etc...). To create a counterpart for the NCHLT corpus we chose the IsiNdebele sentences. As with the NCHLT

TABLE V. THE NCHLT-ISINDEBELE TRAINING / ADAPTATION AND TESTING CORPORA. THE CORPORA STATISTICS ARE REPORTED BY CROSS-VALIDATION FOLDS.

Training / Adaptation

Fold	# utterances	# hours	# speakers
1	47348	61.78	136
2	49206	63.31	136
3	49710	63.28	136
4	49128	64.36	136
5	48847	63.05	136

Testing

Fold	# utterances	# hours	# speakers
1	13339	16.61	33
2	11481	15.08	33
3	10977	15.11	33
4	11559	14.03	33
5	11840	15.34	33

corpus, we had to create a speaker-independent training and testing sets - we did this by partitioning the corpus into five sub-corpora. Table VI shows the sub-corpus statistics for the training/adaptation and testing sets respectively.

TABLE VI. *The Lwazi-IsiNdebele training/adaptation and testing cross-validation corpora.*

Training / Adaptation

Fold	# utterances	# hours	# speakers
1	4817	4.09	160
2	4804	4.11	160
3	4813	4.08	160
4	4807	4.11	160
5	4811	4.11	160

Testing

1	1196	1.03	40
2	1209	1.01	40
3	1200	1.05	40
4	1206	1.02	40
5	1202	1.02	40

B. Data Selection

To investigate the relationship between the amount of adaptation data and the performance of each adaptation method, we needed to devise an algorithm that would grow adaptation data pools from a given data set. Additionally, we needed to obtain an average accuracy value so we decided to repeat each experiment five times, which meant five adaptation data pools had to be created. Our simple data growing algorithm performed the following steps:

- Randomly partition the data into five sub-corpora and ensure each pool has unique speakers.
- For each sub-corpus (sub-corpora are processed independently), start at the first file and sum up the number of triphones contained in each subsequent file added to the data pool. At specified triphone counts, save the file list up to that point.
- If the desired triphone counts cannot be achieved, within a given data pool, randomly select data from the other sub-corpora until the count is reached.

It must be noted that the algorithm "grows" the adaptation pool. For example, if we would like to create two lists which contain files contributing 100 and 250 randomly selected triphone counts, then the 250 triphone count file list will contain all the files present in the 100 triphone counts file list as well as additional files which make up the difference.

C. Baseline ASR systems

The speech recognition system, was based on a standard HMM-based system [17]. Firstly, the audio data was converted to a set of standard Mel-Frequency Cepstral Coefficients (MFCC) vectors. The vectors were estimated from a 25 ms audio window and a 100 vectors per second of speech were calculated. Each vector was constructed by concatenating 13 static, 13 first derivative and 13 second derivative coefficients. Band-limiting was implemented by limiting MFCC extraction to a frequency range of 250 to 3400 Hz. MVA was applied on a per utterance basis and all coefficients were normalised. The HMMs, used to model the cross-word context-dependent triphones, were of a three state left-to-right structure and each state contained 8 mixture diagonal covariance Gaussian models. A question-based tying scheme was followed to create a tied-state data sharing system [18] - where any context-dependent triphone having the same central context could be tied together. As a last step a 40-class binary regression tree was estimated and a semi-tied transform was estimated for each class.

The pronunciation dictionaries for the NCHLT and Lwazi corpora were source from previous work as outlined in Davel and Martirosian [19]. The American English systems made use of the *CMUDict* pronunciation dictionary [20].

The performance of the various ASR systems will be measured using phone-level accuracies. Flat-phone language models were used during the decoding phase.

D. Performance Gain Curves

To create performance gain curves for the various adaptation techniques a set of cross-channel adaptation experiments were performed on both the WSJ-NTimit and the NCHLT-Lwazi corpus pairs. To generate the performance gain curves, the following procedure was used:

- An ASR system was trained on band-limited audio data sourced from one of aforementioned corpora's training set.
- A portion of adaptation data was selected from the corresponding cross-channel training corpus set. The data selection approach is outlined in Section (IV-B).
- The ASR system was adapted using the adaptation techniques and the selected adaptation data. The adaptation techniques under investigation are transfer-function filtering, MLLR and MAP. In addition, an ASR system was trained on the adaptation data without the use any adaptation techniques.
- The adapted and retrained ASR systems were used to recognise the corresponding cross-channel testing dataset.
- The process was repeated on increasing amounts of adaptation data.

The adapted and retrained ASR system performances were measured using phone-level accuracies. For a specific adaptation technique, the WSJ and NTimit experiment results were averaged over the five adapted ASR systems created at each triphone count interval. For the NCHLT and Lwazi

experiments, the results were averaged over the five folds and five adapted ASR systems created at each triphone count interval for a chosen adaptation approach.

V. Results

A. Performance Gain: WSJ - NTimit

Figure 1 shows the accuracies obtained from NTimit acoustic models trained on band-limited (250 to 3400 Hz) audio data and adapted using different adaptation techniques and various amounts of adaptation data sourced from the WSJ training data. The experiments represent a scenario where an ASR system initially uses acoustic models trained on narrow-bandwidth telephone-quality data and the application has to recognize high-bandwidth high-quality data. For all experiments MVA feature normalisation and band-limiting was utilized unless otherwise stated. In the figure the following tags appear in the legend:

- **NTIMIT_WSJ_BP** - Acoustic models trained on all the band-limited NTimit training data and recognised band-limited WSJ test data.
- **WSJ_BP** - Acoustic models trained on all the band-limited WSJ training data and recognised band-limited WSJ test data.
- **WSJ_16k** - Acoustic models trained on all the 16 kHz WSJ training data and recognised 16 kHz WSJ test data.
- **NTIMIT_WSJ_TFF** - Acoustic models trained on band-limited NTimit data which was normalised using transfer-function filtering (TFF) which uses increasing amounts of WSJ to estimate the filtering function. The test data was band-limited WSJ data.
- **NTIMIT_WSJ_MLLR_BP** - Acoustic models trained on band-limited NTimit data and then adapted using MLLR which is estimated on increasing amounts of band-limited WSJ data. The test data was band-limited WSJ data.
- **NTIMIT_WSJ_MLLR_16k** - Acoustic models trained on band-limited NTimit data and then adapted using MLLR which is estimated on increasing amounts of 16 kHz WSJ data. The test data was 16kHz WSJ data.
- **NTIMIT_WSJ_MAP_BP** - Acoustic models trained on band-limited NTimit data and then adapted using MAP for increasing amounts of band-limited WSJ data. The test data was band-limited WSJ data.
- **NTIMIT_WSJ_MAP_16k** - Acoustic models trained on band-limited NTimit data and then adapted using MAP for increasing amounts of 16 kHz WSJ data. The test data was 16kHz WSJ data.
- **WSJ_RETRAIN_16k** - Acoustic models trained on increasing amounts of 16 kHz WSJ training data and recognised 16kHz WSJ test data.

Interpreting the plots we can see at really low adaptation data levels (fewer than 400 triphone examples) the transfer-function feature normalisation gives the best performance gain. Around 400 triphone examples MLLR starts to give better performance gains as the transfer-function feature normalisation gain has saturated. MLLR continues to give the best gain until 7000 triphone examples where retraining the acoustic models with the 16 kHz WSJ data starts to deliver the best performance. The 16 kHz WSJ acoustic models performance improves considerably between 7000 and 200000 triphone training examples. Surprisingly the MAP adaptation method did not out-perform the retrained models at any stage. Even though the MAP adapted models did not give the desired performance (the expected TFF $\rightarrow$ MLLR $\rightarrow$ MAP $\rightarrow$ RETRAIN

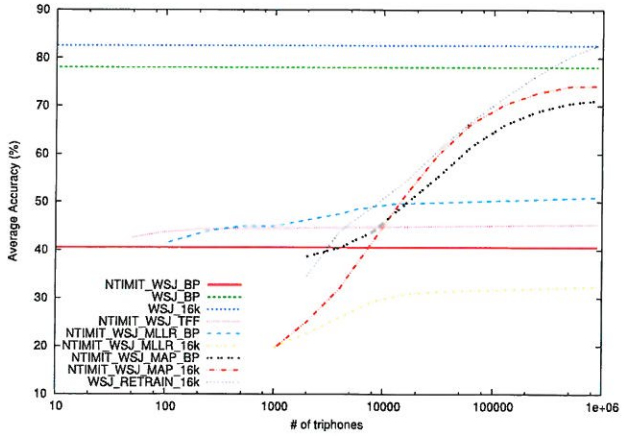

Fig. 1. A narrow-bandwidth to high-bandwidth scenario and accuracies obtained using various acoustic models and adaptation techniques.

transition), the MAP adapted band-limited acoustic models initially performed better (from about 2000 – 10000 triphone examples) compared to the MAP adapted 16k models. It is also interesting to see how quickly the MAP adaptation performance gain plateaus: the phase of linear accuracy improvements as data increases starts to end around 80000 triphones. Lastly, the MLLR adaptation using the 16 kHz WSJ data did not perform well at all – producing accuracies well below the non-adapted ASR setup NTIMIT_WSJ_BP. This shows that MLLR performs better when there is a smaller mismatch between acoustic models and adaptation data whereas MAP has a better ability to deal with large data mismatches.

Figure 2 shows accuracies obtained using WSJ acoustic models trained on band-limited data and adapted using increasing amounts of NTimit training data and various adaptation techniques. The scenario is now an ASR system initially using acoustic models trained on band-limited high-quality audio data and the application has to recognise narrow-bandwidth telephone-quality data. For these experiments, all audio was band-limited and MVA feature processing was applied. Each experiment was repeated five times to obtain average accuracy values.

As with the narrow- to high-bandwidth scenario we see regular trends. The transfer-function feature normalisation performs the best at low triphone counts. Around 100 triphone examples MLLR starts producing better gains and continues as the best option to around 35000 training examples. At this point the retrained models start delivering the best gains.

B. Performance Gain NCHLT - Lwazi

To corroborate the data dependence trends obtained with the WSJ-NTimit corpora, we repeated the cross-channel experiments on the NCHLT-Lwazi corpora. The only difference is that the transfer-function normalisation was dropped as the MLLR appears to give approximately the same performance gains at really low data amounts. As with the WSJ-NTimit experiments, each experiment was repeated five times and in addition, the experiments were run independently on each cross-validation fold. Figure 3 shows the average improvement in accuracies (across folds) as more adaptation data is used

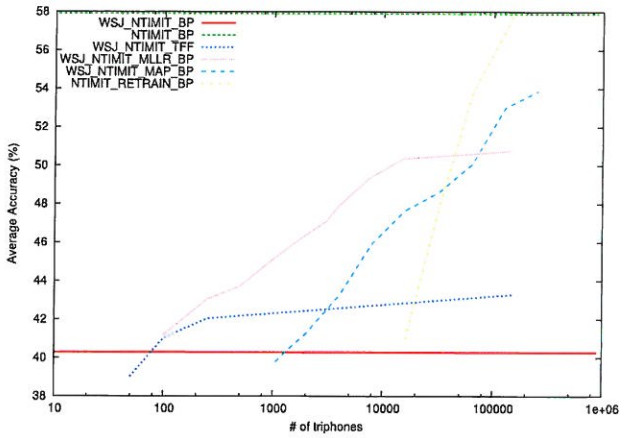

Fig. 2. A high-bandwidth to narrow-bandwidth scenario and accuracies obtained using various acoustic models and adaptation techniques.

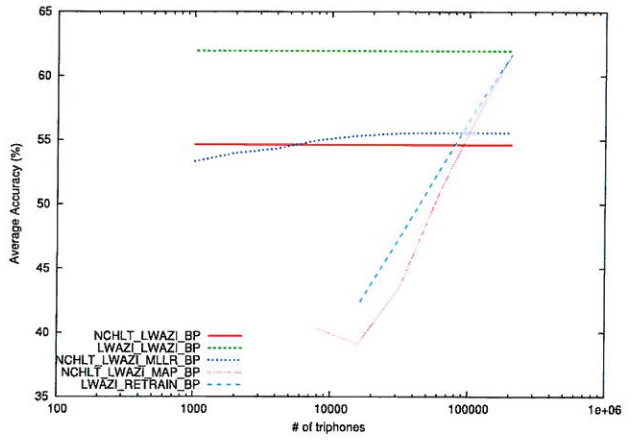

Fig. 3. The average accuracies obtained using various adaptation methods to port high-bandwidth (NCHLT) acoustic models to narrow-bandwidth (Lwazi) telephonic environment.

to adapt the high-bandwidth acoustic models to the narrow-bandwidth environment using various techniques.

As can be seen in figure 3, unexpectedly, the MLLR (mean and variance) initial performance is worse than applying no model adaptation, which implies that the limited adaptation data does not generalize well. For a triphone count between 6000 to 9000 the MLLR starts producing a performance gain but saturates relatively quickly around 12000 triphones. As with the WSJ-NTimit results the retrained acoustic models out-perform the MAP adapted models. The retrained acoustic models start to produce better results around 70000 – 80000 triphones.

Figure 4 shows the average performance gains, as the adaptation data amount is systematically increased and used to adapt the narrow-bandwidth acoustic models to high-bandwidth environment.

Figure 4 is quite similar to the NTimit to WSJ transition experiments. At 100 triphone counts, MLLR provides a gain in performance and continues to produce the best gain in

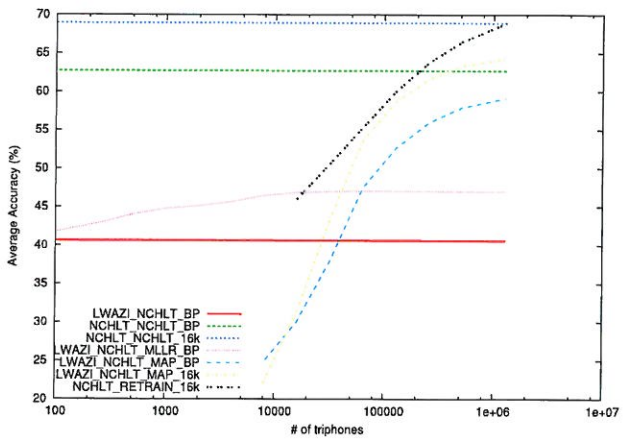

Fig. 4. The average accuracies obtained using various adaptation methods to port narrow-bandwidth (Lwazi) telephonic acoustic models to high-bandwidth (NCHLT) clean environment.

performance until a triphone count of around 18000, where the retrained models start providing the best accuracy. Again, the 16k MAP performs better than its band-limited counterpart but does not improve on the retrained models.

VI. Conclusion

We have analysed the performance gains afforded by the use of several standard feature normalisation and model adaptation techniques for adapting between narrow-band and wide-band speech corpora. The feature normalisation approaches investigated were bandwidth limiting, cepstral mean and variance normalisation with arma filtering (MVA) and a novel transfer-function filtering normalisation. Amongst the model adaptation techniques, we evaluated MLLR for mean and variance adaptation and MAP adaptation of the weights, means and variances. The main conclusions that may be drawn from the work are:

- The novel transfer-function filtering feature normalisation approach performed comparably to MLLR for low adaptation counts but the observed gains plateau quickly as more data was added. Other benefits of the transfer-function normalisation method are that it does not require transcriptions to perform the normalisation and can be applied independently of the various model-based adaptations.
- For low adaptation data amounts MLLR provides the best accuracy gain.
- MLLR works well in reducing the mismatch for bandwidth matched adaptations but failed to achieve improved ASR system accuracies when transforming band limited acoustic models to full bandwidth models (16kHz).
- As the adaptation data count approaches 10000 to 100000 triphone examples, retraining the acoustic models becomes a viable option – out-performing MLLR and MAP.

- Around the 10000 to 100000 adaptation triphone count MAP starts to perform better than MLLR but never beats the retraining the acoustic models.
- Our findings are in agreement with many results in the literature (e.g. MLLR performs better at low data amounts compared to MAP), but also in conflict with some other findings (retraining models out-performs MAP adaptation); this emphasises the fact that some of the strengths and weaknesses of the various adaptation techniques depend on the particular use case (e.g. speaker adaptation vs. dialect adaptation vs. channel adaptation). The main contribution of the this work is to arrive at a consistent picture of the behaviour that can be expected for the specific case of adaptation between narrow- and high-bandwidth applications.

We have demonstrated the efficiency of feature normalisation and model adaptation techniques to reduce the mismatch between telephone-quality and high-bandwidth speech audio. To obtain the best results for channel mismatched scenarios one should employ bandwidth matching, MVA feature normalisation, apply MLLR mean and variance transformation at relatively low adaptation data amounts and after 10000 triphone training examples, retrain the acoustic models on data sourced from the operating environment.

Similar to previously published work we have seen MLLR provide the best adaptation for low data amounts but the observed gains become saturated relatively quickly as more data is added. At this saturation point MAP adapting and retraining the acoustic models become better adaptation options. For channel and environmental adaptation, retraining the acoustic models provides better results compared to MAP adaptation. This is contrary to the speaker-adaptation task where the channel and environment characteristics are similar and the only substantial difference is the triphonic content and speaker characteristics. In this case MAP has a much greater window of data amounts where it is the best adaptation option. We believe that this picture will be particularly useful for system developers in the developing world, who are likely to be confronted with this scenario in practice.

REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, no. 2, p. 171, 1995.

[2] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[3] R. Wallace, K. Thambiratnam, and F. Seide, "Unsupervised speaker adaptation for telephone call transcription," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Taipei, Taiwan: IEEE, April 2009, pp. 4393–4396.

[4] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Hong Kong, Hong Kong: IEEE, April 2003, pp. 540–543.

[5] E. Bocchieri, M. Riley, and M. Saraclar, "Methods for task adaptation of acoustic models with limited transcribed in-domain data," in *Proceedings of INTERSPEECH*. Jeju Island, Korea: ISCA, October 2004, pp. 2953–2956.

[6] P. J. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Adelaide, Australia: IEEE, April 1994, pp. 109–112.

[7] C. P. Chen and J. A. Bilmes, "MVA processing of speech features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 257–270, 2007.

[8] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, December 2001, pp. 103–106.

[9] N. Kleynhans and E. Barnard, "A channel normalization technique for speech recognition in mismatched conditions," in *Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, November 2008, pp. 115–118.

[10] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, 1999.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book. revised for HTK version 3.4," March 2009, http://htk.eng.cam.ac.uk//.

[12] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[13] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Albuquerque, New Mexico, USA: IEEE, April 1990, pp. 109–112.

[14] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[15] N. J. de Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, E. Barnard, and A. de Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[16] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proceedings of INTERSPEECH*. Brighton, United Kingdom: ISCA, September 2009, pp. 2847–2850.

[17] S. Young, "Acoustic modelling for large vocabulary continuous speech recognition," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 169, pp. 18–39, 1999.

[18] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[19] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Interspeech 2009*, 2009, pp. 2851–2854.

[20] CMU, "The CMU Pronouncing Dictionary," 2013. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict