Spoken language identification system adaptation in under-resourced environments

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*—Speech technologies have matured over the past few decades and have made significant impacts in a variety of fields, from assistive technologies to personal assistants. However, speech system development is a resource intensive activity and requires language resources such as text annotated audio recordings and pronunciation dictionaries. Unfortunately, many languages found in the developing world fall into the resource-scarce category and due to this resource scarcity the deployment of Automatic Speech Recognition (ASR) systems in the developing world is severely inhibited. Given that few task-specific corpora exist and speech technology systems perform poorly when deployed in a new environment, we investigate the use of acoustic model adaptation. We propose a new blind deconvolution technique which rapidly adapts acoustic models to a new environment and increases their overall robustness. This new technique is utilized in a Spoken Language Identification (SLID) system and significantly improves the system's accuracy by 6% relative to the baseline system and achieves comparable performances when compared to relatively more computationally intensive standard adaptation techniques.**

I. Introduction

Spoken Language Identification (SLID) technology has a variety of uses [1], [2], [3] and will most probably play an increasing role in global services. The basic goal of SLID is to determine the language being spoken in an utterance. It is possible to discriminate between languages because there are differences in the phonology, morphology, syntax and prosody[2]. A good SLID system will extract these salient traits and base its decision on the best match.

The system is typically deployed in a multilingual environment where audio data obtained from a variety of languages can be classified and grouped together for some follow-up task to process – for example, for manual processing by human listeners, or to instruct a multilingual speech recognition system to load the appropriate acoustic models. Currently popular SLID systems use a combination of acoustic level and phonotactic information to classify an audio segment and by fusing these information sources achieve high performance rates when trained on large corpora, operated in clean environments and on high-bandwidth audio. It is well known, however, that the performance of speech-based systems degrade considerably when deployed in noisy environments and when tasked to process audio data that is mismatched compared with the training data. Another challenge is to develop and deploy SLID in resource-scarce environments where access to training or adaptation audio data is severely limited.

Thus, the main *aim* of our research is to investigate adaptation techniques to improve the robustness of a SLID system designed to operate in a resource-scarce real-world environment.

II. Background

Zissman [2] analysed four approaches to the SLID task which were:

- Gaussian Mixture Model (GMM) classification,
- single-language phone-recognition followed by language-dependent interpolated n-gram language modelling (PRLM),
- parallel PRLM, which uses multi-language recognisers, and,
- language-dependent parallel phone recognition (PPR).

The GMM approach trains a language-specific GMM and classifies an utterance based on the model which gives the best log-likelihood score. The PRLM system uses a single-language phone recogniser to tokenise an utterance. In a training phase, data from a new language is tokenised and a n-gram language model is trained on the token sequence. During testing, an utterance is tokenised and the n-gram language model which will best generate the observed sequence is chosen as the spoken language. For parallel PRLM, additional phone recognisers are introduced to add more independent token streams, which allows the training of more n-gram language models. For example, if there were three phone recognisers, each language would have three n-gram language models with which to make a decision. The scores generated by the separate language-specific n-gram language models are averaged to obtain a single language score. Lastly, the PPR approach utilises a bank of language-specific acoustic models to perform language identification. During the decoding of an utterance, a language-specific n-gram language model is used in conjunction with the language-specific acoustic models to generate the phone tokens. The acoustic scores associated with the selected tokens are summed and language and length normalised. The best language log-likelihood value is chosen. Out of the four candidates analysed the PPR system performed the best, but required the most extensive development effort.

Li *et. al.* [3] proposed a SLID approach which built upon the PPR system. Their method utilised the PPR system as a front-end, to produce phone sequences, and added a Support Vector Machine (SVM) classifier to build a model from the phone sequences. The SVM builds a set of language classifiers based on the biphone frequencies produced by the bank of phone recognisers. The system was evaluated against

the language model back-end and provided superior results independent of the test performed.

An alternative approach which has recently gained popularity employs Total Variability Factor Analysis on a linear model representation, called the iVector representation [4].

Although the iVector approach has been greatly beneficial for speaker-verification tasks, it is not clearly advantageous for spoken language identification: evaluations such as the recent Albayzin 2012 Language Recognition Evaluation [5] have found similar performance for PPR-SVM and iVector-based systems.

Peche *et. al.* showed the versatility of PPR-SVM SLID architecture by utilizing the system in a limited data environment [6], porting the system to operate it in a new low-bandwidth environment [7] and successfully applying the system in real-world resource-scarce environment, to identify South African languages [8]. The SLID task was more challenging since the in-domain audio had no transcriptions which added further complications.

The starting point of our investigation was to choose a SLID system which has been shown to operate effectively in a resource-scarce environment and reliably process audio data sourced from the real-world. From the various SLID systems surveyed, the SLID system presented by Peche *et. al.* [6], [7], [8] was the best candidate and served as our baseline SLID system. Briefly, the entire SLID framework consists of data filtering, phone recogniser training, classifier training and evaluation. Of these four levels our work complemented the phone recogniser training phase and focused on improving the acoustic modelling. Logically, better acoustic models would produce phone strings which were more accurate and thus would enable the classifier to be trained on more robust data and hopefully produce an improved overall system performance. In addition, to implementing standard acoustic model adaptation techniques we investigated an unsupervised channel adaptation approach and compared its effectiveness to the standard approaches used.

In Section III we present the data filtering approach, describe the SLID system, provide detail into the acoustic model adaptation and describe the performance measures. Results are presented in Section IV and concluding remarks are captured in Section V.

III. Method

A. Data Filtering

The original audio data (collected from the operating environment of a commercial client of the Meraka Institute) resembles that of low-bandwidth telephony recordings. The exact recording methodology is unknown. The received audio was packaged in the Microsoft WAVE format using a sample rate of 8 kHz, two signed bytes to represent a sample and two channels per recording. A visual inspection of the data reveals that the audio does pass through a band limiting filter with a lower cut off frequency around 250 Hz and an upper cut off frequency around 3400 Hz. The languages present in the audio recordings are English, French and Portuguese. The quality of the audio data is variable and ranges from clean to very noisy recording channels. Modem tones, DTMF tones, clicks and facsimile sounds can be found in the audio. The content of the audio is typical of conversational speech and is quite varied; it fluctuates between speakers with accents, well articulated speech, incoherent speech and speech filled with disfluencies. Only speaker and language labels are identified in the metadata provided with the audio data and no transcriptions are present. Table I shows the amount of raw audio data per language for this corpus. For convenience, we will refer to this real-world corpus as the *EFP-LID corpus*.

TABLE I. The amount of raw audio data in hours per language for the EFP-LID corpus.

Language	Duration (Hours:Minutes)
English	40:22
French	11:21
Portuguese	21:45

The raw audio data is innately heterogenous and required a data-filtering process to transform the data into a familiar ASR-style corpus. Our data-filtering approach can be defined as the automatic processing of real-world audio data by which the actions transform the raw data into a set of homogeneous parts and provides a normalisation barrier between the raw audio and the SLID system. This type of data filtering is known as *diarization*. Broadly, the diarization process followed (for our data context) was (adapted from [9])

- **Silence detection** - silence audio portions are identified and removed.
- **Audio Segmentation** - the silence free audio is segmented into chunks. The segment boundaries are calculated using the Bayesian Information Criterion (BIC) [10] which determines a boundary by comparing the statistics of adjacent audio portions and setting a boundary if the statistical difference is greater than a user-defined threshold.
- **Speech / Non-Speech classification** - audio segments are classified and assigned a class label. Typically speech, music, and non-speech events can be found in audio recordings and need to be identified to create homogeneous audio groups. Audio segments generated from the previous audio segmentation step are classified and assigned either with a Speech or Non-speech label.
- **Audio Concatenation** - audio segments which are assigned the same class labels are concatenated. The segments from a specific utterance, given the same class label, are concatenated and form the new cleaned audio utterance.

A final EFP-LID corpus was created by processing the original audio data using our data-filtering approach. Table II shows the language-specific audio data amounts, in hours, for the training and testing subsets of the final EFP-LID corpus. The training and testing datasets were selected at random. Our channel adaptation work was based on this final version of the EFP-LID corpus which contained low-bandwidth telephone-quality audio data for the English, French and Portuguese languages.

TABLE II. The audio amounts, in hours, for the training and testing data sets for the EFP-LID corpus. These values were obtained after processing the data with the diarizer.

Language	Training	Testing
English	15.13	3.71
French	11.01	2.78
Portuguese	10.71	2.93

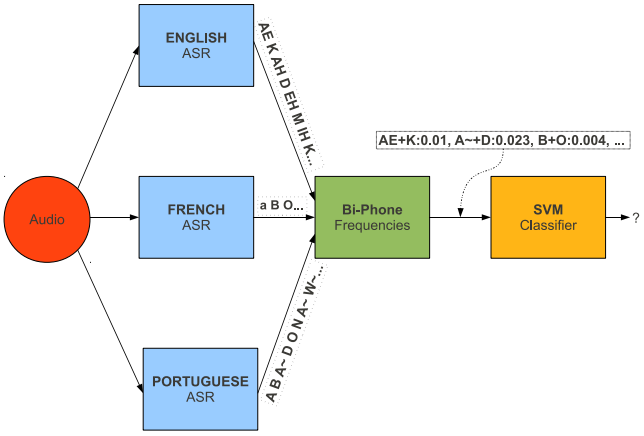

Fig. 1. The PPR-SVM system architecture.

B. Spoken Language Identification System

One of the most widely used SLID architectures is the Parallel Phone Recogniser front-end [2] and classifier back-end scheme [3], [8]. In this set-up, a bank of phone recognisers are used to generate phonetic information streams from the audio which are then fused together to form an input to a classifier which makes a final decision about the spoken language. The phonetic data streams generally exhibit a high dimensionality and large sparseness, because accurate classification requires that sequences of several phones be treated as single units [3]. This makes it difficult to choose a standard distribution form with which to model the distribution of the data. However, given the nature of the phonetic information streams, SVMs[11] are a logical choice since SVMs achieve good performance levels on sparse high-dimension datasets [3], [12]. This type of SLID system is referred to as a PPR-SVM set-up and has been shown to provide the best result compared to other PPR set-ups in a detailed comparison[3]. The basic PPR-SVM architecture used for our purposes is shown in Figure 1.

The advantage of this type of SLID system is that adding new languages to the system does not require adding language-specific phone recognisers [6]. Only the SVM parameters have to be trained on the new and existing phonetic streams to incorporate the new language.

1) Phone recognisers: Once the audio data has been processed, the speech-only audio is sent through a bank of phone recognisers. The phone recognisers generate phone strings which represent the most likely text to have been spoken in the utterance. Since English, French and Portuguese are present in the EFP-LID corpus and we had access to language-specific high-bandwidth corpora, a bank of phone recognisers were built using these three languages. The high-bandwidth French and Portuguese corpora form part of the GlobalPhone corpus [13], while the high-bandwidth English data was sourced from a portion the Wall Street Journal (WSJ) corpus [14] ("The Continuous Speech Recognition Wall Street Journal Phase I (CSR-WSJ0) Corpus"). These corpora are clean ASR corpora, with a read speech style and the audio data recorded at a sampling rate of 16 kHz. Table III shows the data amount in hours and speaker numbers for the training and testing sets for the GlobalPhone and WSJ corpora. The phone

TABLE III. GLOBALPHONE CORPUS STATISTICS FOR ENGLISH, FRENCH AND PORTUGUESE LANGUAGES

	Training		Testing	
Language	# Hours	# Speakers	# Hours	# Speakers
English	20	83	4.85	19
French	21.6	80	5.3	21
Portuguese	14.4	77	3.5	25

recognition systems employ standard Hidden Markov Models (HMM) as used in ASR. A typical phone model consists of a three state (entry and exit states not counted) left-to-right HMM which is used to model triphone contexts (including cross-word contexts). Each state was modelled using GMMs with seven components and for improved robustness state-tying was employed. A 39 dimensional Mel-frequency cepstral coefficient (MFCC) feature vector was used to encode the audio data frames. Each MFCC feature vector was constructed by appending 13 static, 13 first derivative and 13 second derivative coefficients. The audio data frames were extracted by applying a blocking process which used a 25 ms frame width and advanced the frame by 10 ms to create an overlap between adjacent frames. To reduce the mismatch between the high-bandwidth data and EFP-LID corpus data we band-limited the frequency range to 250 – 3400 Hz. Bandwidth matching has been shown previously by Moreno and Stern [15] to improve ASR performance. The frequency band range were chosen after manual verifying the average frequency cutoffs.

2) SVM training: SVM classifiers [11] are highly adept at learning near-optimal separating boundaries from high-dimensional sparse datasets. Because of this trait these classifiers can be used successfully to classify languages given the phone strings produced by the phone recognisers. As shown in [3] calculating frequency counts from the phone strings provides a salient data feature which can be used to correctly classify languages. Improved results can be achieved if bigram or trigram phone frequencies are calculated but a trade off must be struck between computational time and time for a detection. For the purposes of our SLID system, bigram frequency counts were used. The SVM was based on the C-support type and used Radial Basis Function (RBF) kernels.

To train the SVM model the following training steps were performed:

- For each language, the training audio was decoded using the English, French and Portuguese phone recognisers and the phone strings collated.
- The bigram phone frequency counts for each language were calculated.
- The SVM model was trained on the normalised language-specific biphone frequency counts. A grid search optimization process was followed to find the optimal penalty and kernel width values.

During the testing phase the first two training steps were followed and then for the last step the SVM model was used to classify bigram phone frequency counts.

C. Acoustic Model Adaptation

As the SLID system had to process low-bandwidth telephony data, the phone recognisers had to be updated accordingly. The initial acoustic models trained on GlobalPhone WSJ

high-bandwidth data would perform poorly as there is a large mismatch between the acoustic models and the audio data found in the EFP-LID corpus. The amount of training data in the EFP-LID corpus is sufficient to support retraining or adaptating the acoustic models for the various languages. The audio data, however, had no accompanying transcriptions which are needed for the retrain or adaptation process. A simple solution is, for each language, to use the high-bandwidth phone recogniser to decode the audio data and create the necessary transcriptions. Without adaptation these transcriptions would not be useful but by utilizing various adaptation techniques the acoustic model's environmental robustness can be improved and more reliable transcriptions can be generated.

To produce accurate transcriptions the mismatch between the EFP-LID audio data and the acoustic models must be reduced. A reduction can be achieved by using ASR adaptation techniques to update the acoustic models' parameters to better fit the low-bandwidth audio data. Our first step in improving the match was to reduce the bandwidth of the high-bandwidth audio data by limiting the frequency content to between 250 Hz and 3.4 kHz. These limits match well with telephone channel bandpass cut offs. Instead of filtering the audio data, which unnecessarily creates more audio data, we set the MFCC extraction process to extract features from the newly set bandwidth limits.

1) Transfer-Function Filtering: Standard ASR adaptation techniques require transcriptions for the adaptation audio in order to estimate the transforms which reduce the data-model mismatch. As the EFP-LID corpus contains no transcriptions, some transcription generation process must be run to create this missing resource. A channel adaptation technique which does not require any knowledge about the content of the audio would be invaluable in this scenario.

As shown in [16] it is possible to perform blind channel estimation and then estimate an inverse filter to remove the channel distortions. The process proposed in [16] is rather elaborate and requires the separate estimation of the long-term clean and distorted speech statistics using their custom signal decomposition methods known as the "adjustable bandwidth concept". We employ a simpler approach, estimating the long-term average of the short-term speech spectral information, of the source and target audio data, to calculate the inverse filter which can be used to transform one audio dataset to another. This is possible if we assume that the channel response remains constant (linear and time-invariant) and that the speech statistics for the data sets are similar. We have previously shown in [17] that this is indeed possible to transform two high-bandwidth corpora such that on average the short-term frequency response are similar. The EFP-LID system gives us a further opportunity to test our hypothesis on a mixed low- and high-bandwidth scenario. As the high- and low-bandwidth signals have different frequency ranges, the high-bandwidth signal was band-limited to match the low-bandwidth signal. This was achieved by applying a band-pass filter which had a frequency range of 250 – 3400 Hz. A mathematical derivation on how to estimate the inverse filter follows.

First off, we can estimate the average short-term spectral statistics by summing over all the short-term spectral data frames extracted in the feature extraction process. The average

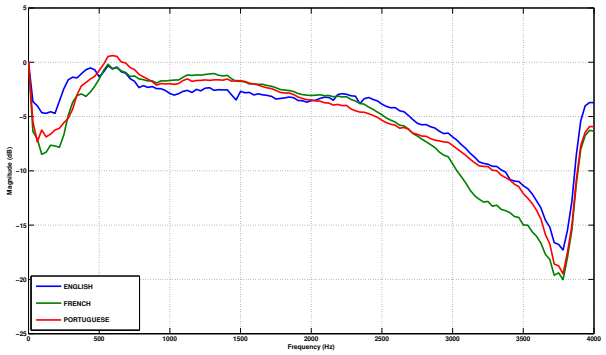

Fig. 2. Inverse filter response for English, French and Portuguese. These responses transform GlobalPhone data to the EFP-LID data.

short-term spectral estimate is given by equation(1),

$$Y_{avg}^{Channel}(f) = \frac{H(f)}{M}\sum_{n=1}^{M} X_n(f) \quad (1)$$

where $Y_{avg}(f)$ is the average short-term spectral estimate for a specific channel, $H(f)$ is the constant channel response and $X_n(f)$ is the speech spectral estimate for the frame *n*. Given two of these estimates and using the linear filtering assumption, we can estimate the inverse transforming filter by dividing the estimates. Thus, the inverse filter is calculated by equation (2),

$$H_{inv}(f) = \frac{Y_{avg}^{Channel_1}(f)}{Y_{avg}^{Channel_2}(f)} \quad (2)$$

$$H_{inv}(f) = \frac{\frac{H_1(f)}{M_1}\sum_{n=1}^{M_1} X_n(f)}{\frac{H_2(f)}{M_2}\sum_{n=1}^{M_2} X_n(f)} \quad (3)$$

where $H_{inv}(f)$ is the inverse filter response, $Y_{avg}^{Channel_1}(f)$ is the average short-term spectral estimate for $Channel_1$, $Y_{avg}^{Channel_2}(f)$ is the average short-term spectral estimate for $Channel_2$, $X_n(f)$ is the short-term speech spectral estimate, and, $H^1(f)$ and $H^2(f)$ are the channel responses of the differing channels. If we assume that the speech statistics are similar we may disregard them and the inverse filter is given by equation (4),

$$H_{inv}(f) = \frac{H^1(f)}{H^2(f)} \quad (4)$$

which shows that the inverse filter is only determined by the ratio of the channel responses. The assumption that the speech statistics are similar will not be valid for small amounts of data, but will hopefully not be far off as both corpus sizes increase. Figure 2 shows the inverse filter response which transforms the GlobalPhone data to match the EFP-LID data. As the high-bandwidth signal was band-pass filtered (range 250–3400 Hz), the inverse filter response outside the frequency range are signal estimation and normalisation artefacts.

To apply the inverse filter, we move to the log-domain and update each frame as follows,

$$\log(X_{new}) = \log(X_{old}) + \log(H_{inv}(f)) \quad (5)$$

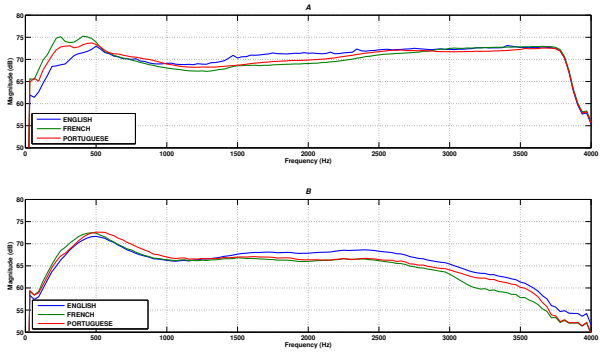

Fig. 3. The average spectrum of the GlobalPhone languages in (A) and the average spectrum of the EFP-LID languages in (B).

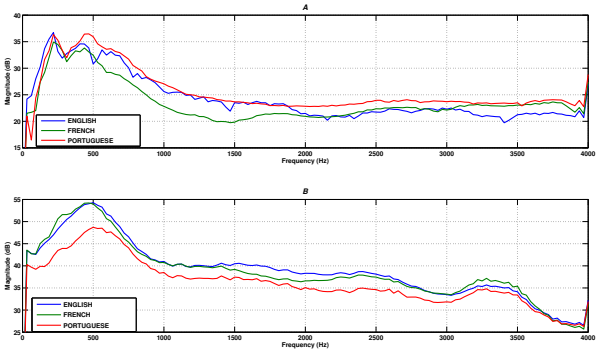

Fig. 4. The spectrum variance of the GlobalPhone languages in (A) and the spectrum variance of the EFP-LID languages in (B).

where X_{new} is the transformed short-term spectrum, X_{old} is the unmodified short-term spectrum and $H_{inv}(f)$ is the inverse channel response.

As part of the optimization phase of the channel adaptation task, we assumed that the log-domain representation of the spectral components were best described by a normal distribution. Following from this, we assumed that any changes observed in the means and variances, of two spectral estimates, were induced by channel effects. Therefore, to remove the channel effects we have to modify the spectral components such that the means and variances would match a target distribution. Using the maximum likelihood approach we can estimate the mean and variance for each frequency component by,

$$\mu_i = \frac{1}{N}\sum_{n=1}^{N} X_n(f) \quad (6)$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_i^n(f) - \mu_i)^2 \quad (7)$$

where μ_i is the mean estimate for frequency component i, σ_i^2 is the variance estimate for frequency component i and $X_i^n(f)$ is a spectral estimate for frequency component i at frame n. Diagonal covariance matrices are used throughout. To correct for the channel distortion, all we need to do is apply this simple update formula,

$$X_{new}(f) = \frac{\sigma_t(X_{old}(f) - \mu_s)}{\sigma_s} + \mu_t \quad (8)$$

where $X_{new}(f)$ is the transformed spectrum, $X_{old}(f)$ is the original spectrum, μ_s and μ_t are the source and target means, and, σ_s and σ_t are the source and target standard deviations. Figure 3 shows the average spectrum estimates for the GlobalPhone and EFP-LID languages and Figure 4 shows the spectrum variance estimates for the GlobalPhone and EFP-LID languages.

Our strategy to improve the EFP-LID transcription generation process by using the transfer-function adaptation proceeded as follows:

1) estimate the average short-term spectrum across the corpora and for all languages,
2) estimate the short-term spectrum standard deviation across the corpora and for all languages,
3) create channel adapted GlobalPhone and WSJ corpora by modifying the short-term spectrum through inverse filtering and transforming the data characteristics so that the average spectrum approximates that of the EFP-LID corpus, and,
4) create another channel adapted version of the GlobalPhone and WSJ corpora by modifying the mean and standard deviation of the short-term spectra so that they match the EFP-LID corpus statistics.

2) Broad Class Adaptation: The standard ASR adaptation techniques rely on transcriptions to estimate adapted parameters. The transcriptions are used to estimate class-specific transforms (with phone, biphone or triphone classes typically being employed) and thus improve the adaptation performance. The transfer-function method makes the assumption that the channel response is linear time-invariant; in light of the results of Reynolds et. al. [18] we know that this is a very rough approximation. The channel response does change depending on the input energy and the energy input has a strong correlation to the sound class (i.e. the phonetic identity of the sound).

To incorporate this knowledge into the adaptation approach we needed a classifier that could reliably identify different sounds within an audio file. Running cross-corpus tests using the Timit [19] and NTimit [20] corpora, it was found that a Timit trained classifier, using seven broad classes achieve an accuracy of 60% while using the normal 39 English phoneme set only achieved an accuracy of 33 %. The seven broad classes were chosen based on linguistic knowledge and were defined to be vowels, fricatives, affricates, glides, nasals, plosives and silence. The broad class classification results were encouraging, but for adaptation to new languages it is not always realistic to expect that knowledge of the most appropriate broad classes will be available.

Therefore, another means was needed to generate the broad classes. For some HTK-based adaptation tasks a regression class tree has to be built, so that data can be grouped together in order to estimate more robust transformation parameters [21]. This provided us with an alternative way to create the broad classes for the English, French and Portuguese data. The binary regression class trees are built by using a centroid-splitting algorithm which initially groups all the data, and then iteratively splits the nodes until the desired number of classes are reached. Each node contains data samples which are considered in an Euclidean sense to be similar. After some

experimentation it was decided to use six broad classes, as it was felt that this number of classes gave a good data split.

3) Maximum Likelihood Linear Regression: The Maximum Likelihood Linear Regression (MLLR) technique has become a standard approach to speaker and environment adaptation for ASR systems. Initially introduced to aid in speaker adaptation, the technique can easily be extended to cater for channel and environment normalisation [22]. The adaptation technique uses a maximum likelihood approach to estimate a set of linear transformation matrices which modify the mean vectors of the acoustic models. Gales and Woodland later modified the MLLR framework to introduce variance transformations [23]. The advantage of this technique is that is uses regression class trees to pool similar HMM mixture components together and thus form larger data groups. Using more data in this way enables better transformation matrix parameter estimates [22]. To transform GlobalPhone and WSJ acoustic models to represent the EFP-LID data space better we experimented with two MLLR adaptation types.

The first MLLR form estimates transforms involving means and variances in two steps. In the first step, mean transforms are estimated, whereas the variance transforms are estimated in the second step using the mean transforms as a parent transform. For the MLLR transform we used a two-class regression class tree: class one includes all the silence mixtures while class 2 includes all the speech mixtures. The MLLR adaptation technique requires transcriptions to perform the estimation and since the EFP-LID system contained no transcriptions we used the broad class decodes as reference transcriptions. We estimated three separate mean-variance transforms: one for each language.

A second form of the MLLR transformation is the constrained MLLR transformation (CMLLR). In this approach, a joint transform is estimated which tries to transform the mean and variance in one step. To achieve this, the transform is applied to the data vector and not the mean and variance as is the case in the first MLLR transformation form. Again, we only used two class regression class tree, estimated separate language transforms and used the broad class decodes as reference transcriptions. For lack of EFP-LID transcriptions, we could not complete the entire constrained MLLR transform estimation process which requires the acoustic models to be retrained after applying the transform to the data.

D. Performance Measures

For each adaptation technique we show two system performance measures. The first measure is the overall system accuracy percentage which was calculated by dividing the number of correct identifications for all three languages by the total number of possible correct identifications. The second measure is a table of the confusion matrix which shows the accuracy of identifying a specific language and what errors were made in trying to identify that language.

To test the results significance we employed the Pearson's chi-squared statistic with Yates's correction for continuity [24]. The chi-squared statistic is generally used to test for goodness-of-fit which for our purposes will tell us whether or not the obtained correct and erroneous classifications, produced by the alternative adaptation system, are likely to be drawn from the baseline system distribution. Leading from this we propose a null hypothesis; H_0: the expected and observed values are drawn from the *SAME* distribution, while the alternative hypotheses states; H_1: the expected and observed values are *NOT* drawn from the same distribution. For our experimentation we set the significance level at 0.01 or 1 % - the null hypothesis will not be rejected for any P-value above 0.01.

IV. Results

In this section we present results for the SLID system which was trained and tested on EFP-LID corpus data. The systems' results give us an indirect indication on how well the various adaptation techniques worked in adapting the GlobalPhone and WSJ acoustic models which were used to generate transcriptions for the EFP-LID corpus.

A. Adaptation Results

Our first set of results show the baseline system performance which is captured in Table IV. The overall system accuracy was 73.57 %. The only channel adaptation that was applied was Cepstral Mean normalisation (CMN). English classification out-performs both French and Portuguese, with Portuguese classification performing the worst around 50 %. Preliminary investigations showed the SLID system achieved an overall language classification rate greater than 99 % for high-bandwidth clean data (full frequency bandwidths used i.e. 0 – 8 kHz). The most likely explanation for the observed performance loss are the low-bandwidth recordings, lack of reliable transcriptions and variable quality of the training and testing recordings.

TABLE IV. A confusion matrix produced by the SLID system using the EFP-LID corpus and having applied CMN channel adaptation.

Language	English	French	Portuguese
English	84.17	09.35	06.47
French	14.72	76.14	09.13
Portuguese	29.41	16.04	54.54

The next set of results show the SLID performance when applying per-utterance Cepstral Mean Variance normalisation (CMVN) – the mean is normalised to zero and the variance scaled to one. The overall system accuracy was 75.15 % and confusion matrix is shown in Table V. Compared to the baseline performance, that of CMN, there was an approximately 1.6 % absolute increase in overall accuracy. The percentages in the confusion matrix show that Portuguese accuracy increased while the other languages' values remained almost constant. It was decided for further experiments CMVN would serve as the baseline feature normalisation and the other feature normalisation and model adaptations techniques would be applied in conjunction.

TABLE V. A confusion matrix produced by the SLID system using the EFP-LID corpus and applying CMVN channel adaptation.

Language	English	French	Portuguese
English	84.53	06.47	08.99
French	12.69	76.65	10.66
Portuguese	29.10	11.11	59.79

The overall accuracy of the SLID system which used inverse filtering (IF) was 77.41 % – the initial transfer-function filtering approach. This translates to an 3.5 % absolute improvement in accuracy compared to the baseline system. For this experiment, all the language training data was pooled for the different corpora and the inverse filter was calculated independently of language. Table VI contains the confusion matrix results which show a substantial increase in accuracy for Portuguese and a general decrease in the error amount.

TABLE VI. A CONFUSION MATRIX PRODUCED BY THE SLID SYSTEM USING THE EFP-LID CORPUS AND APPLYING THE INVERSE FILTERING CHANNEL ADAPTATION.

Language	English	French	Portuguese
English	84.89	07.55	07.55
French	12.18	77.16	10.66
Portuguese	24.34	08.99	66.67

The next channel adaptation technique used to improve the SLID system performance was the spectral mean and variance normalisation (SMVN) method – transfer-function filtering optimisation. As with the previous inverse filtering experiment, the corpus-specific means and variances were calculated independently of language. The overall system accuracy was 78.46 % which is 4.9 % absolute improvement compared to the baseline system. The language accuracies for French and Portuguese, as seen in Table VII, have all improved except for English which has an increased number of misclassifications.

TABLE VII. A CONFUSION MATRIX PRODUCED BY THE SLID SYSTEM USING THE EFP-LID CORPUS AND APPLYING SPECTRAL MEAN AND VARIANCE NORMALISATION CHANNEL ADAPTATION.

Language	English	French	Portuguese
English	81.65	11.15	07.19
French	10.15	83.77	06.09
Portuguese	21.16	10.58	68.25

The last spectral mean and variance normalisation experiment estimated language-specific means and variances (LS-SMVN) and updated the languages accordingly. For this adaptation approach the overall system accuracy jumped to 79.82 % compared to the baseline result – 6 % absolute improvement. This overall increase in performance can be seen in the better confusion matrix values shown in Table VIII.

TABLE VIII. A CONFUSION MATRIX PRODUCED BY THE SLID SYSTEM USING THE EFP-LID CORPUS AND APPLYING LANGUAGE-SPECIFIC SPECTRAL MEAN AND VARIANCE NORMALISATION CHANNEL ADAPTATION.

Language	English	French	Portuguese
English	86.33	07.19	06.45
French	07.61	83.25	09.14
Portuguese	24.87	08.47	66.67

The overall system accuracy for the constrained MLLR (CMLLR) adaptation approach was 76.05 % which translates to 2.5 % relative increase in performance as compared with the baseline accuracy. In Table IX we see that for French and Portuguese an accuracy increase but this is accompanied by a decrease in the English accuracy. It would be interesting to establish how much better the constrained MLLR adaptation method would have performed if the acoustic models could have been retrained while applying the transform (which is the standard approach).

TABLE IX. A CONFUSION MATRIX PRODUCED BY THE SLID SYSTEM USING THE EFP-LID CORPUS AND APPLYING CONSTRAINED MLLR CHANNEL ADAPTATION.

Language	English	French	Portuguese
English	81.65	09.35	08.99
French	06.60	82.74	10.66
Portuguese	30.69	08.47	60.85

Lastly, the results for the mean and variance MLLR (MV-MLLR) adaptation method. This approach produced an overall system accuracy of 79.64 %; 6 % absolute improvement compared with the baseline accuracy. This result is similar to the language-specific spectral mean and variance normalisation. Table X shows the language specific accuracies. Interestingly for Portuguese, the number of misclassification between English increases substantially and the French error decreases accordingly.

TABLE X. A CONFUSION MATRIX PRODUCED BY THE SLID SYSTEM USING THE EFP-LID CORPUS AND APPLYING MEAN AND VARIANCE MLLR CHANNEL ADAPTATION.

Language	English	French	Portuguese
English	85.61	04.68	09.71
French	05.58	86.29	08.12
Portuguese	30.85	04.79	64.36

B. Adaptation Results Significance

Table XI shows a summary of the overall system accuracies for different adaptation methods as well as the Pearson significance tests where CMN results are used as reference. Although the CMVN, IF and CMLLR approaches produce an increase in system performance the increases are not significant. The SMVN, LS-SMVN and MV-MLLR produce significant system accuracy increases with LS-SMVN providing the best result.

TABLE XI. SUMMARY OF OVERALL SYSTEM ACCURACIES FOR VARIOUS ADAPTATION APPROACHES AND PEARSON SIGNIFICANCE TESTS.

	Reference - CMN			
Adaptation Technique	System Accuracy (%)	χ^2	P-Value	Significant?
CMN	73.57	-	-	-
CMVN	75.15	1.99	0.85	No
IF	77.41	10.97	0.052	No
SMVN	78.46	19.85	0.00134	Yes
LS-SMVN	79.82	16.353	0.0059	Yes
CMLLR	76.05	7.375	0.1942	No
MV-MLLR	79.64	15.544	0.0083	Yes

V. CONCLUSION

In this work, we presented an SLID system which uses an PPR-SVM architecture to process low-bandwidth telephony quality recordings and determine if the language spoken in the recording was either English, French or Portuguese. The system's acoustic models were trained on in-domain audio data (EFP-LID corpus) which contained no transcriptions. We successfully generated the missing transcriptions by using high-bandwidth acoustic models and applying novel and standard adaptation techniques. The best overall system accuracy that we obtained was 79.82 %, utilising the language-specific spectral mean and variance normalisation adaptation technique. To summarize our main contributions;

- We showed that it is indeed possible to apply a new unsupervised channel normalisation technique to aid in the

bootstrapping of a spoken language identification (SLID) system to a low-bandwidth environment.

- The comparative SLID results (Section IV) show that the proposed log-domain spectral mean and variance normalisation produced a significant improvement in accuracy compared to the baseline system accuracy and performed comparably with the MLLR model adaptation technique.
- We have developed blind deconvolution techniques which do not require transcriptions for their proper implementation and have shown them to perform comparably with standard adaptation techniques.
- Inverse filtering can be successfully applied to reduce the mismatch caused by channel distortions and change in bandwidth.
- Log-domain spectral mean and variance normalisation, working under the assumption of normally distributed spectral components, can be used to reduce the channel distortion effect and improve acoustic model robustness, producing the best results (along with mean and variance MLLR) in our comparisons.

It would be interesting to investigate whether it is possible to further improve the accuracy of the SLID system by combining several of proposed channel normalisation techniques. The most likely candidates are CMVN feature normalisation coupled with the language-specific mean and variance spectral normalisation and MLLR mean and variance model adaptations. However, there is no guarantee that the combination would achieve the desired gain as shown in [17] where channel normalisation combinations, in some cases, reduced the system's performance. This work does suggest that some trial-and-error experimentation would have to take place to determine which combinations would prove to be most beneficial and to gain some insight on how to approach the channel normalisation technique combination process.

It is noteworthy that the simple unsupervised channel normalisation technique performed well on the SLID task. It would be worthwhile to see if the gains can be translated to ASR applications. In addition, we know that model-based adaptations try to estimate class-based transforms or model updates, thus it would be interesting to see if it could be possible to extend the unsupervised channel normalisation to include class information and to establish if significant gains in performance can be achieved.

REFERENCES

[1] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proceedings of INTERSPEECH*, vol. 92. Banff, Alberta, Canada: ISCA, October 1992, pp. 895–898.

[2] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 1, pp. 31–44, 1996.

[3] H. Li, B. Ma, and C. H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 271–284, 2007.

[4] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," in *Proceedings of INTERSPEECH*. Firenze, Italy: ISCA, August 2011, pp. 861–864.

[5] L. J. Rodríguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, G. Bordel, and M. Diez, "The albayzin 2012 language recognition evaluation," in *Proceedings of INTERSPEECH*. Lyon, France: ISCA, August 2013, pp. 1497–1501.

[6] M. Peché, M. H. Davel, and E. Barnard, "Phonotactic spoken language identification with limited training data," in *Proceedings of INTERSPEECH*. Antwerp, Belgium: ISCA, August 2007, pp. 1537–1540.

[7] M. Peché, M. Davel, and E. Barnard, "Porting a spoken language identification system to a new environment." in *Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, November 2008, pp. 103–107.

[8] M. Peché, M. H. Davel, and E. Barnard, "Development of a spoken language identification system for South African languages," *SAIEE Africa Research Journal*, vol. 100, no. 4, pp. 97–112, 2009.

[9] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization," in *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA: ISCA, September 2006, pp. 2178–2181.

[10] J. Zdansky, "BINSEG: An efficient speaker-based segmentation technique," in *Proceedings of INTERSPEECH*. Pittsburgh, Pennsylvania, USA: ISCA, September 2006, pp. 2182–2185.

[11] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[12] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.

[13] T. Schultz, "Globalphone: a multilingual speech and text database developed at Karlsruhe University," in *Proceedings of INTERSPEECH*. Denver, Colorado, USA: ISCA, September 2002, pp. 345–348.

[14] 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.

[15] 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.

[16] S. J. Wenndt and A. J. Noga, "Blind channel estimation for audio signals," in *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, vol. 5. Big Sky, Montana, USA: IEEE, March 2004, pp. 3144–3150.

[17] 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.

[18] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. Detroit, Michigan, USA: IEEE, May 1995, pp. 329–332.

[19] 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.

[20] 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.

[21] 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//.

[22] 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.

[23] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

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