PORTING A SPOKEN LANGUAGE IDENTIFICATION SYSTEM TO A NEW ENVIRONMENT

Marius Peché[1], Marelie Davel[2] and Etienne Barnard[2,]
mpeche@csir.co.za, mdavel@csir.co.za, ebarnard@csir.co.za

[1]Department of Electrical, Electronic and Computer Engineering, University of Pretoria.
[2]HLT Research Group, Meraka Institute, CSIR.

ABSTRACT

A speech processing system is often required to perform in a different environment than the one for which it was initially developed. In such a case, data from the new environment may be more limited in quantity and of poorer quality than the carefully selected training data used to construct the system initially. We investigate the process of porting a Spoken Language Identification (S-LID) system to a new environment and describe methods to prepare it for more effective use. Specifically we demonstrate that retraining only the classifier component of the system provides a significant improvement over an initial system developed using acoustic models channel-normalized to the new environment. We also find that the most accurate system requires retraining of both the acoustic models and the final classifier.

Index Terms — Spoken Language Identification, S-LID.

1. INTRODUCTION

Spoken Language Identification (S-LID) is the process whereby a sample of audio speech from an unknown source is classified as one of several possible languages [1]. This can be done in a number of ways, including sampling the prosodic information or processing information extracted from specified tokens, where such tokens may be phonological or syntax related [2]. In the latter case, spoken LID differs significantly from textual LID because text already consists of properly defined and accurate tokens (such as alphabetical letters) while these tokens (such as phonemes) must first be extracted from audio speech, and may not be accurate.

In addition, more accurate S-LID systems usually are more complex and require a larger amount of data to create systems with sufficient performance [3]. The popular Parallel Phone Recognition and Language Model (PPR-LM) approach [1] provides reasonably high system accuracy with acceptable data requirements, and is the approach experimented with in this paper.

In a PPR-LM system, separate phone recognizers are used to tokenize an incoming audio signal individually, and a classifier trained to identify the language spoken based on the token strings received in parallel from the various phone recognizers. Initially based on language modeling scores, various classifiers have since been used in literature, with Support Vector Machines (SVMs) achieving high accuracy [4].

Once developed for a specific environment, it is often required that a S-LID system be ported to a new environment. Data from such a new environment may be more limited in quantity and of poorer quality than the carefully selected training data used to construct the system initially.

We investigate the process of porting a PPR-LM based S-LID system to a new environment and describe methods to prepare it for more effective use. Specifically we compare the effect of re-training the classifier component with that of re-training both the acoustic modeling and classifier component and report on results.

The paper is structured as follows: In section 2 we describe the design of our baseline system. In section 3 we describe the porting process step by step, specifically focusing on data preparation, initial system adaptation, classifier adaptation, acoustic model adaptation and final analysis. Section 4 contains some concluding remarks.

2. BASELINE SYSTEM DESIGN

We develop an initial S-LID system able to identify three languages: English, French and Portuguese. English data is obtained from the Wall Street Journal corpus, and French and Portuguese data from the GlobalPhone corpus [6]. (These two corpora have similar acoustic characteristics.)

Using a PPR-LM approach, we develop three Automatic Speech Recognition (ASR) systems capable of performing phone recognition, each in one of the languages English, French or Portugues. These ASR systems utilize Hidden Markov Models (HMM) which have been trained to recognize bi-phones from Mel Frequencies Cepstral Coefficients (MFCC). The training of the HMMs as well as the extraction of the MFCCs from the audio signal are performed using the HMM Tool Kit (HTK) [7]. These phone recognition systems run in parallel with one other, each yielding a phoneme string for a given speech sample.

We use the 'Bag-of-Sounds' principle to model the frequencies of phonemes as a vector, with the frequency of each phoneme within the sample of speech representing an element of this vector. These vectors are then used to train a Support Vector Machine (SVM) using the LIBSVM [5] toolkit. In all experiments, a radial basis function kernel is used and the kernel width and misclassification cost are optimized using a grid search. Multiple classes are handled using a 1 against n-1 scheme.

Using a flat phone grammar with approximately 40 phones per language, we achieve phone recognition accuracies of 48% to 66% for the three ASR systems on an independent test set. While these accuracies seem fairly low, they are sufficient to obtain highly accurate S-LID results, as displayed in Table 1. S-LID results are obtained using the same test set as used to report on ASR accuracies, and durations of the speech samples range from 10 to 60 seconds each.

Language	Word recognition accuracy	S-LID accuracy
English	52.8%	98.9%
French	66.2%	94.9%
Portuguese	48.1%	97.7%

Table 1: Accuracies achieved by baseline system

3. PORTING THE S-LID SYSTEM

In this section we first discuss the new environment investigated and the data available from this environment, before providing detail with regard to the different aspects of our approach to porting the S-LID system, specifically consisting of (1) data preparation, (2) initial system adaptation, (3) classifier adaptation, (4) acoustic model adaptation and (5) final analysis.

3.1 Data description

In order to investigate porting of the S-LID system to a new environment, we utilize a telephone corpus of African variants of the three languages of interest (referred to from here onwards as the African corpus). The African corpus consists of approximately 45 hours of speech separated into English, French and Portuguese variants spoken on the African continent. The speech is untranscribed and no additional speaker information is available. Single calls are assumed to be from a single speaker and most calls are assumed to be from different speakers. The amount of data, in hours, is displayed in Table 2.

The initial S-LID system is therefore required to perform in a new environment with significantly different channel conditions and speech dialects. In addition, the new data contain non-speech signals as well as competing background noises.

Data from the three different languages in the new corpus are identified according to language. The new corpus is separated into a training and test corpus as indicated below, with the same test set used to report on results. Care is taken to ensure that the same speaker is not included in both the training and test set.

Language		**GlobalPhone**		**African Corpus**	
		train	test	train	test
English	Hours	20.2	4.85	16.77	4.3
	Speakers	83	19	250	25
French	Hours	21.6	5.3	8.25	2.07
	Speakers	80	21	109	26
Portuguese	Hours	14.4	3.6	11.18	2.84
	Speakers	77	25	108	28

Table 2: Training and testing data statistics.

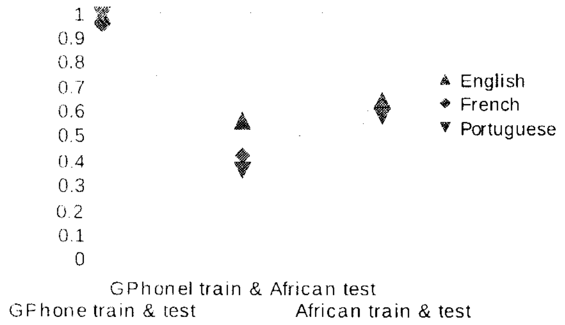

Figure 1: S-LID Accuracy when the GlobalPhone ASR system is used, but the SVM is trained on different corpora.

3.2 Data Preparation

Our first task is to pre-process the new data. We use diarization techniques to separate the different speakers and to remove any non-speech signals. We also remove long sections of silence from the audio, perform amplitude normalization and segment the new audio files into sections of no larger than one minute each.

3.3 Initial System Adaptation

Once the new data has been preprocessed (as above) it can be used to estimate the channel conditions of the new environment. The GlobalPhone corpus can now be downsampled (to 8KHz, the sampling frequency of the African corpus), amplitude normalized and channel normalized in order to better match the new environment [8].

In order to verify the new system, ASR and S-LID accuracies are calculated using the same test corpus as before. While ASR accuracies decrease with between 7% and 14% absolute, overall S-LID accuracy increases from 97.18% to 98.26%.

It should be noted that these results still refer to data from the previous (GlobalPhone) environment. Once this system is tested using the new data, an S-LID accuracy of only 47.02% is obtained.

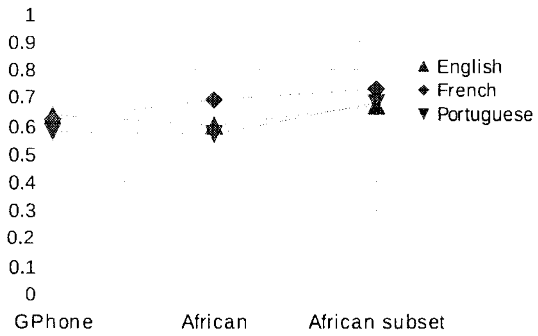

Figure 2: S-LID Accuracy with the African corpus when different ASR systems are used

3.4 Classifier Adaptation

We now adapt the classifier to the new environment: we tokenize the new training data using the normalised GlobalPhone recognizers and use these phone stings to re-train the SVM. S-LID accuracy improves dramatically, from the previous 47.02% to 62.57%. The differences in performance between an optimal system (GPhone train&test), an unported system (GPhone train & African test) and the ported system with only the SVM adapted (African train & test) are depicted in Figure 1.

3.5 Acoustic Model Adaptation

In order to further improve the performance of the system with the African corpus, we train new acoustic models for the tokenizers. We use the normalised GlobalPhone recognisers to bootstrap transcriptions for the new data (since the African corpus is not transcribed), and use these transcriptions to train new acoustic models. [9] Once new acoustic models are trained, these are used to re-tokenize the new audio data and re-train the classifier.

Initially results are disappointing as S-LID accuracy falls to 60.58%. However, when transcriptions are filtered to exclude the training and test utterances that were clearly hard to recognize (transcriptions that contain fewer than one and a half phones per second) S-LID accuracy increases to 68.88%. This is the highest accuracy obtained using the full test set. The effect of using different ASR systems on S-LID accuracy is depicted in Figure 2.

3.6 Final Analysis

While an improvement from 47.02% to 68.88% is significant, these results are still lower than anticipated and further analysis of the data set is required. When subsets of the new data set are systematically listened to by human verifiers it is noted that the new corpus contains data of highly variable quality.

While initial random testing of the corpus provided some indication of the quality of the data, a systematic analysis by human verifiers indicates that significant portions of the corpus contain the following problematic subsets:

- Data incorrectly labeled or unusable, meaning that the language spoken is neither English, French or Portuguese ('Unusable').
- Data correctly labeled, but spoken with a strong accent ('Accented').
- Data correctly labeled but consisting mostly of noise with similar spectral characteristics as speech, which the diarization system did remove (also included as 'Unusable').

Data correctly labeled and identifiable as either English, French or Portuguese are indicated as 'Correct' by the human verifiers. The number of samples evaluated that falls within each category for each of the languages is listed in Table 3. (Note that only a subset of the full corpus was evaluated.)

Language	Unusable	Accented	Correct
English	109	72	118
French	110	4	159
Portuguese	35	1	109

Table 3: Number of samples per category as verified by human verifiers

This table provides a new perspective on the results obtained (in Section 3.5). As a large percentage of the samples are in fact unusable, an S-LID accuracy of 68.38% is indeed highly encouraging. Further analysis and optimisation can now be done using the smaller "correct" subsets in order to obtain a better indication of system performance.

4. CONCLUSION

In this paper we describe the process of adapting an existing S-LID system to a new environment. We describe the different stages in such a process and provided results for each stage. We highlight the importance of verifying the quality of the data from the new environment systematically as an important step during system porting. We show that bootstrapping transcriptions from existing ASR systems, and re-training the classifier using the bootstrapped transcriptions provide a significant improvement in performance and that the most accurate system requires retraining of both the acoustic models and the final classifier, this is with a small margin only

In further work we are currently repeating some of the above experiments using only the small portions of data identified as "Correct" during human verification. We are also investigating automated mechanisms to identify problematic audio samples during system development.

5. REFERENCES

[1] Y.K. Muthusamy, E. Barnard and R.A. Cole, "Reviewing Automatic Language Recognition", IEEE Signal Processing Magazine, Oct 1994.

[2] Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li and Eng Siong Chng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification", In ICASSP-2006, Toulouse, France. pp 205-208.

[3] Marc A. Zissman, Kay M. Berkling, "Automatic language identification", Speech Communication, 35 (2001) pp 115-124.

[4] Haizhou Li, Bin Ma, "A phonotactic language model for spoken language identification," In Proc. ACL-2005, pp. 515–522.

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[6] Tanja Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University", In ICSLP-2002, Denver, Colorado, USA. pp 345-348.

[7] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, "The HTK book. revised for HTK version 3.3," September 2005. Software available at http://htk.eng.cam.ac.uk/

[8] Neil Kleinhans, "Channel normalization for speech recognition in mismatched conditions", *accepted for publication*, In PRASA-2008, Cape Town, South Africa.

[9] Marius Peché, Marelie Davel, Etienne Barnard, "Phonotactic spoken language identification with limited training data" In Interspeech-2007, Antwerp, Belguim. pp 1537-1540.