Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 327–342

www.elsevier.com/locate/csl

Unsupervised crosslingual adaptation of tokenisers for spoken language recognition☆

Raymond W.M. Ng*, Mauro Nicolao, Thomas Hain

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Sheffield, England S1 4DP

Received 17 October 2016; received in revised form 7 May 2017; accepted 10 May 2017
Available online 19 May 2017

Abstract

Phone tokenisers are used in spoken language recognition (SLR) to obtain elementary phonetic information. We present a study on the use of deep neural network tokenisers. Unsupervised crosslingual adaptation was performed to adapt the baseline tokeniser trained on English conversational telephone speech data to different languages. Two training and adaptation approaches, namely cross-entropy adaptation and state-level minimum Bayes risk adaptation, were tested in a bottleneck i-vector and a phonotactic SLR system. The SLR systems using the tokenisers adapted to different languages were combined using score fusion, giving 7–18% reduction in minimum detection cost function (minDCF) compared with the baseline configurations without adapted tokenisers. Analysis of results showed that the ensemble tokenisers gave diverse representation of phonemes, thus bringing complementary effects when SLR systems with different tokenisers were combined. SLR performance was also shown to be related to the quality of the adapted tokenisers.

Keywords: Language recognition; Unsupervised adaptation; Crosslingual adaptation; Tokenisation; Phonotactic SLR

1. Introduction

In a spoken language recognition (SLR) task, an automatic system is used to infer the language identity of the given acoustic signal (Muthusamy et al., 1994). Different types of information from a speech signal can be used to identify the language spoken in an audio sample. Standard SLR methods can be categorised by the features they use. The two most popular SLR approaches are the acoustic-phonetic and the phonotactic approaches (Zissman, 1996; Ambikairajah et al., 2011; Li et al., 2013).

In the acoustic-phonetic approach, low-level acoustic features such as mel-frequency cepstral coefficients (MFCC)(Davis and Mermelstein, 1980) or shifted-delta cepstral coefficients (SDC)(Torres-Carrasquillo et al., 2002) are extracted, and statistical models such as Gaussian mixture models are trained on these features to match target languages (Zissman, 1993; Singer et al., 2003).

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail address: wm.ng@sheffield.ac.uk (R.W.M. Ng).

http://dx.doi.org/10.1016/j.csl.2017.05.002

In the phonotactic approach, an audio tokeniser converts the speech signal into graphs (lattices) of discrete tokens. Tokenisers are often implemented by phoneme recognisers whose output consists of phonemic-related units. A phoneme tokeniser trained on multilingual data would cover the phonetic variety across languages and would identify tokens with higher accuracy than monolingual ones (Schultz, 2002). High-quality multilingual tokenisers are rarely available due to limited language resources. However, regardless of the quality of the tokeniser, when the tokeniser is applied on multilingual data, the occurrence patterns of the output tokens differ from one language to another significantly. This allows for modelling and language classification (Zissman, 1996; Singer et al., 2003; Hazen and Zue, 1997; Navrátil, 2006; Glembek et al., 2008).

Since 2012, there has been an increasing use of multilayer perceptrons or deep neural networks (DNN) in SLR. Multilayer perceptron features (such as Temporal-Pattern Discrete Cosine Transform, TRAP-DCT (Schwarz, 2009)) have been employed, alone or concatenated to conventional PLP and F0 features, as SLR system input (BenZeghiba et al., 2012). In more recent years, DNNs have been used to generate bottleneck or posterior statistics related to a designated phoneme inventory, which may or may not match the target languages. On top of the DNN, i-vector SLR systems have been built to perform the language detection (Richardson et al., 2015; Ferrer et al., 2016). This study focused on the deep neural networks that process the elementary phonemic information for SLR modelling. One of the major problems of this approach arose from the use of mismatched phone tokenisers trained on languages different from target data. This issue was addressed by incorporating alternative hypotheses in the tokenisation process. Soft counts from phone lattices (Gauvain et al., 2004), phone posteriograms (D'Haro et al., 2012) or multilayer perceptron features (BenZeghiba et al., 2012) allowed to generate smooth statistics and mitigate the effect of wrong tokenisation in SLR.

The use of multilingual (Anderson and Dalsgaard, 1997) or parallel tokenisers (Zissman, 1996; D'Haro et al., 2014) has been a useful technique to boost SLR performance for many years. Regardless whether the 1-best phone sequence or a phone posteriogram vector is used, in parallel tokenisers, speech data is decoded in different manners to give diverse phonetic representations for subsequent SLR modelling. The combination of information from different tokenisers often provides gain in performance. Previous studies also showed improvements when DNNs are trained on multilingual data (Fék et al., 2015). In BenZeghiba et al. (2012), different MLPs were trained on English, French and Spanish for use in a phonotactic system similar to the classical SLR configuration with "parallel phone recognition followed by language dependent modelling" (PPRLM) (Zissman, 1996). In Ferrer et al. (2016), a DNN, initially trained on English data, was adapted on Mandarin, Spanish, and Egyptian Arabic data to provide multiple SLR systems that were combined with a fusion process.

The National Institute of Standards and Technology (NIST) has conducted a number of evaluations of automatic language recognition technology. Recent NIST language recognition evaluations (LRE) were held in 2011 and 2015 LRE (2011); (2015). These focused on languages that are similar to each other and are frequently mutually intelligible, e.g. dialectal variants LRE (2015). In NIST LRE 2015, a new requirement on fixed training data for all components in the SLR system was introduced to reflect a more realistic scenario in which linguistic resources for SLR are limited. In the tokeniser training, 300 h of transcribed Switchboard data (Godfrey and Holliman, 1993) were allowed to be used. This data consisted of monolingual (English) conversational telephone speech only. There was no transcribed multilingual data for tokeniser training. Language, channel and style mismatches between this data and the actual LRE training and test data created extra challenges in building the system.

This study compares the use of multiple tokenisers derived from different multilingual data through unsupervised training and adaptation to capture diverse linguistic information for the SLR modelling. The performance on phoneme recognition of the newly trained/adapted speech recognisers do not necessarily need to be optimal, but rather they only serve to provide different representations of the acoustic data in terms of its tokenisation, and allow complementary effects to appear in the late SLR system fusion. The tokeniser ensemble can reduce the undesirable loss of information due to wrong decoding in the tokenisation step. Besides serving as a mitigation measure for the absence of transcribed data, adaptation of tokenisers gives rise to multiple SLR systems that contain compatible information and allow effective system fusion. To the best of our knowledge, there is no work in the literature that addresses training data constraints and the use of untranscribed and multilingual data in the tokenisation for SLR. DNN adaptation has not been studied thoroughly before, particularly in an unsupervised manner for SLR application.

Inspired by the use different tokenisers trained in multiple languages (D'Haro et al., 2014), we extend our previous work on this topic (Ng et al., 2016a) where unsupervised DNN adaptation was used in the tokenisers for a

phonotactic SLR system. In the work reported here, the method has been further extended to the state-of-the-art bottleneck i-vector SLR system, and different ways of training and adapting the DNN tokenisers have been investigated.

2. Unsupervised adaptation of speech tokeniser

Unsupervised training could be a solution to acoustic modelling on a low-resource settings. These include scenarios where only a limited lexicon and acoustic data are available for model training. In most cases, there are no ground truth reference. The target phoneme-state reference is generated by automatic transcripts. Confidence thresholding is applied for data selection and iterations of training are conducted to improve the quality of a new model. This technique has been applied successfully in crosslingual acoustic model bootstrapping in different applications including ASR (Lööf et al., 2009; Vu et al., 2011) and keyword spotting (Knill et al., 2014).

Neural network adaptation for acoustic model is an area under active research and can be conducted by different means. Related work on adaptation can be found in Xue et al. (2014), where the DNN was adapted to different speakers using a sequence training criterion. In Suzuki et al. (2016), cross-entropy and sequence training adaptation on convolutional neural network (CNN) were compared.

The use of acoustic tokenisers was referred to as an "indirect" SLR method (Richardson et al., 2015). Under this approach, SLR becomes a staged task. In stage one, a DNN model (acoustic tokeniser) solves a different problem (phoneme recognition) from SLR. The idea of using adapted acoustic models for SLR has a subtle difference from those works for ASR and other speech related technologies. The acoustic model can be adapted towards acoustic model precision, or towards other criteria related to SLR. The quality of the adapted model is believed to correlate with SLR performance, but only in a loose sense. An adapted acoustic model giving suboptimal phoneme recognition results can provide complementary information to the original model. Ultimately, SLR performance improvement can be achieved by system combination. In this study, a breath-first approach is taken. Instead of implementing a sophisticated recipe of iterative unsupervised adaptation and focusing on the phoneme recognition quality of a single adapted tokeniser, multiple DNN acoustic models were constructed by implementing common model adaptation by different means and to different target languages.

The tokeniser used in the SLR system in this study is a DNN phone tokeniser implemented in a feed-forward hybrid setting. Assuming a dataset with a total of U utterances, each indexed u, $u \in [1, 2, \ldots, U]$, each utterance has different duration which is given by the number of frames ($[T_1, T_2, \ldots, T_u, \ldots, T_U]$). A feature at time t in the uth utterance is denoted by $\boldsymbol{o}_{ut}$. Triphone Hidden Markov Model (HMM) states, s_{ut}, were generated by automatic methods such as decision tree clustering and were used as the training targets. In supervised training, s_{ut} can be derived from ground truth reference, normally by word-to-state conversion through a dictionary and an automatic alignment procedure with a seed model. A primary DNN tokeniser was trained in a supervised manner with the cross-entropy (CE) criterion at frame level. $y_{ut}(s_{ut})$ denotes the DNN output posterior probability estimate at time t for utterance u, which corresponds to the reference target state s_{ut}. Cross-entropy training minimises

$$\mathcal{F}_{\mathrm{CE}} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log y_{ut}(s_{ut}). \tag{1}$$

Unsupervised adaptation of a tokeniser can be performed on new data when the ground truth transcript is not available. Automatic state-level transcripts are derived from the primary DNN tokeniser. The model is then further trained or adapted. In this paper, unsupervised adaptation is implemented by fine-tuning of weights in a primary DNN tokeniser using unlabelled multilingual data. In a crosslingual adaptation setting, the automatic state labels may lie in a different space compared with the ground truth state labels. This is potentially problematic as high error rates of target sequences may lead to divergence. In other SLR studies, language-mismatched phonemic models with unknown phoneme prediction quality were used (Zissman, 1996; BenZeghiba et al., 2012; Ferrer et al., 2016). Despite the mismatch, the combinatorial use of phoneme model sets were shown to help SLR. The purpose of adaptation is to augment the phonetic coverage of the set of models by adapting each model in the set to different data. Unsupervised adaptation of the network on untranscribed multilingual data may yield different networks, generating the variety, thus the complementary information needed for good SLR performance.

The primary tokeniser was applied to the multilingual LR training data to generate a first-best sequence $\widehat{S}_u = (\widehat{s}_{u0}, \ldots, \widehat{s}_{ut}, \ldots, \widehat{s}_{uT})$. Optionally, a decoding lattice $\mathcal{L}_u$ can be generated for discriminative training, where $\widehat{S}_u$ is one of the possible paths in $\mathcal{L}_u$. There is no constraint on the type of objective function the adapted tokeniser should be optimised with. In the following sections, two adaptation approaches on cross-entropy fine-tuning and uncertainty reweighting were tested.

2.1. DNN adaptation with cross-entropy training

In this approach, the primary DNN is used as a seed model and unsupervised crosslingual adaptation is performed. For each crosslingual adaptation setting, data from only one language is selected. Further DNN fine-tuning is carried out by means of extra iterations of cross-entropy training with the following objective function,

$$\mathcal{F}_{\text{CEadapt}} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log y_{ut}(\widehat{s}_{ut}). \tag{2}$$

Eq. (2) is similar to Eq. (1). The only difference is on $y_{ut}(\widehat{s}_{ut})$, which now corresponds to the posterior probability of the one-best decoded state with the primary DNN. With N target languages, unsupervised adaptation was performed in N independent runs, resulting N adapted networks.

2.2. DNN adaptation by uncertainty reweighting

In the second approach, an identical approach as described in Section 2.1 is taken to adapt the primary DNN to N directions. Here, DNN discriminative training is performed where the training process also considers alternative hypotheses in the decoding lattice $\mathcal{L}_u$. Minimisation of state-level Minimum Bayes Risk (sMBR) was chosen to be the objective function (Gibson and Hain, 2006), which is defined as,

$$\mathcal{F}_{\text{sMBR}} = -\sum_{u}\sum_{S} p(S|\boldsymbol{O}_u)A(S, \widehat{S}_u) \tag{3}$$

$A(S, \widehat{S}_u)$ is an accuracy term to compute the number of correct state labels corresponding to the state sequence S with respect to the first pass hypothesis $\widehat{S}$. The objective function aims to minimise Bayes risk at state-level (sMBR). The posterior probability was computed at the utterance level (Veselý et al., 2013). This objective function tries to incorporate uncertainty in previous decoding in order to achieve robust estimation .

2.3. Comparison of adaptation strategies

In this study, multiple independent adaptation trials were conducted for different target languages. Two LR system settings (bottleneck i-vector and phonotactic) were tested. To maintain a controllable scope of experiment, the CE adaptation technique is coupled with the bottleneck-ivector system. The sMBR adaptation technique is applied on a phonotactic system.

Both CE and sMBR cost functions were shown to give comparable, if not identical, performance on neural network adaptation (Suzuki et al., 2016). CE is a maximum likelihood objective and is primarily frame-based. It has a lower computational cost without the need of generating denominator lattices. sMBR is a discriminative training algorithm which takes into account of the uncertainties in the decoding process.

Considering the two SLR systems, an i-vector system takes bottleneck features which is a soft DNN output. It also prefers smooth statistics as the accurate inference of UBM and total variability matrix training depends on the normal distribution assumption. For this reason, the CE adaptation technique is thought to be more suitable for the bottleneck i−vector SLR system.

The phonotactic SLR system in this study models TF-IDF features which were derived from the n-gram statistics of the one-best DNN tokeniser output. A hard decision is made by the DNN before information is further propagated to the next SLR modelling stage. For this reason, we chose to use sMBR adaptation with the phonotactic SLR system.

A control experiment will be conducted where both adaptation techniques are applied on one of the target languages and the SLR performance is verified.

3. Data

Training and development data come from four corpora, namely Switchboard 1, Switchboard Cellular Part 2 (SWB_CELL) and two multilingual datasets, LDC2015E87 and LDC2015E88. 90% of the speakers in Switchboard 1 are selected to form a training set for both tokeniser and SLR system training, which is referred to as (SWB1_R2). SWB1_R2 and SWB_CELL are monolingual English data with a duration of 302 h and 43 h, respectively. The size of lexicon is 39 k. The phoneme set is a US-English phone set with a size of 39. LDC2015E87 comprises conversational telephone speech from the CallHome and CallFriend corpora in Egyptian Arabic, Standard Mandarin and US English. LDC2015E88 comprises data covering the seventeen other target languages in LRE 2015. The amount of data for different languages in LDC2015E87 and LDC2015E88 varies from 0.4 to 159 h. The corresponding data amounts in different languages are shown in Table 1. The multilingual data was provided without segmentation. Voice activity detection and resegmentation described in Ng et al. (2016b) were performed to derive segments in comparable length (3-seconds, 10-seconds and 30-seconds) of the test data. 80% of the segmented data are selected in each language and they are collectively referred to as LR2015-ML-TRAIN in the rest of the paper. LR2015-ML-TRAIN has a total duration of 891 h.

Several development sets were defined for internal testing purposes. TEST-SWB was created by randomly selecting 10% of the speakers (24 female and 28 male) as a held-out set from SWB1_R2 and was used to test the DNN phoneme tokeniser performance. LR2015-DEV was created by selecting 20% of the LDC2015E87 and LDC2015E88 as a development set for the tuning of system fusion weights. To relate to the previous study, LR2015-DEV is a combination of the V3DEV and V3HELDOUT data detailed in Ng et al. (2016b). 10% of the LR2015-ML-TRAIN was selected for the cross validation set in cross-entropy training of DNN tokenisers.

Test data used in this work is the official NIST LRE 2015 eval data (LR2015-EVAL). The total number of utterances is 164,334. In this work, test results are reported for 3 independent sets of roughly equal size with different nominal durations. The 3-second eval data set contains audio shorter than 7.5 s. The 10-second set contains audio equal or longer than 7.5 s and shorter than 20 s. The 30-second set contains data longer than 20 s. The number of trials with 3-second, 10-second and 30-second nominal durations is 49,981, 53,306 and 61,047, respectively.

4. DNN tokenisers

4.1. Primary DNN tokeniser (SWB)

The primary DNN tokeniser uses a feedforward DNN with 6 hidden layers in a hybrid setting. Each hidden layer contains 2048 neurons, the bottleneck layer has 64 neurons and the output layer has 3815 neurons. The DNN topology is represented in Fig. 1. The input features to the DNN were Mel−frequency cepstral coefficient (MFCC) features with delta, delta-delta and utterance-level mean normalisation. Further follow-on processing included global feature transform with linear discriminant analysis (LDA), a maximum likelihood linear transform (MLLT) and feature splicing with 5 contextual frames to the left and the right of the centre frame. The training targets were the tied triphone states obtained by alignment with a constrained maximum likelihood linear regression (CMLLR) adapted,

Table 1
Target languages and raw amount of training data in NIST LRE 2015.

Cluster	Target languages	Total length
Arabic	Egyptian (ara-arz, 159 h), Iraqi (ara-acm, 57 h), Levantine (ara-apc, 63 h), Maghrebi (ara-ary, 57 h), Modern Standard (ara-arb, 3 h)	339 h
English	British (eng-gbr, 0.4 h), General American (eng-usg, 159 h), Indian (eng-sas, 3 h)	163 h
French	West African (fre-waf, 6 h), Haitian Creole (fre-hat, 2 h)	8 h
Slavic	Polish (qsl-pol, 26 h), Russian (qsl-rus, 5 h)	31 h
Iberian	Caribbean Spanish (spa-car, 44 h), European Spanish (spa-eur, 7 h), Latin American Spanish (spa-lac, 6 h), Brazilian Portuguese (por-brz, 0.7 h)	58 h
Chinese	Cantonese (zho-yue, 4 h), Mandarin (zho-cmn, 107 h), Min (zho-cdo, 7 h), Wu (zho-wuu, 7 h)	125 h

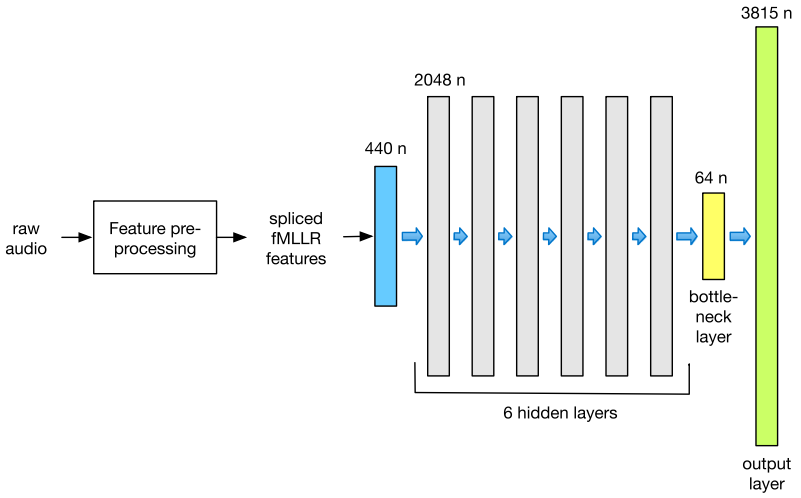

Fig. 1. DNN topology shared by all the neural networks used in the LR frameworks. Number of neurons (n) is displayed for each layer. In training stage, several DNNs have been created using different input datasets and fine-tuning strategies.

maximum mutual information(MMI)-optimised acoustic model set. The primary DNN was trained on SWB1_R2, and is abbreviated as "SWB" to indicate the single set of training data being used.

4.2. DNN adaptation

Two DNN adaptation approaches (Sections 2.1 and 2.2) were tested in this study. In the first approach, the primary tokeniser, SWB, was adapted to different languages using LR2015-ML-TRAIN. The primary SWB DNN was further trained on the data of one language subject to the cross-entropy criterion (Eq. (2)). Without extra care on the adaptation algorithm and its implementation, DNN adaptation on small amounts of data may not give sufficient and robust statistics, and overfitting may occur. For this reason, only 8 target languages in LR2015-ML-TRAIN, where training duration exceeds 10 h, were selected for the cross-entropy DNN adaptation. These languages include 4 Arabic languages, American English, Polish, Caribbean Spanish and Mandarin. The training duration in different languages is detailed in Table 1. The eight adapted DNNs are abbreviated as "CE(ara-acm), CE (ara-apc), CE(ara-ary), CE(ara-arz), CE(eng-usg), CE(qsl-pol), CE(spa-car), CE(zho-cmn)".

The second adaptation approach used the state-level MBR training technique described in Section 2.2. It was used to adapt the primary DNN (SWB) to the identical set of 8 different languages as in CE(·) above, using LR2015-ML-TRAIN. The resultant DNNs are abbreviated as "sMBR(ara-acm), sMBR(ara-apc), sMBR(ara-ary), sMBR(ara-arz), sMBR(eng-usg), sMBR(qsl-pol), sMBR(spa-car), sMBR(zho-cmn)". Models of sMBR(·) were duration-dependent. For each of the three nominal durations (3-second, 10-second, 30-second), model adaptation was conducted on the training segments of the required duration in a particular language.

The adapted DNN tokenisers use different adaptation data but share the same network topology (shown in Fig. 1) as the primary DNN tokeniser. sMBR(·) and CE(·) are derived in a similar method with different training objective functions, given by Eqs. (2) and (3).

5. Language recognition system

In this section, the detailed implementation of the bottleneck i-vector and phonotactic SLR systems is explained. Fig. 2 outlines the complete work flow of the SLR systems. The bottleneck i-vector LR system used the bottleneck-layer features from a DNN tokeniser. I-vectors were extracted, based on which logistic regression was performed for language classification. This LR technology is hereinafter abbreviated as "BNIV-LR", and systems using the BNIV-LR technology are referred to with the suffix "-B".

The phonotactic LR system modelled language identity through the occurrence statistics of phoneme *n*-grams. Outputs from the final layer of the DNN tokenisers were used to derive phoneme labels, from which phoneme

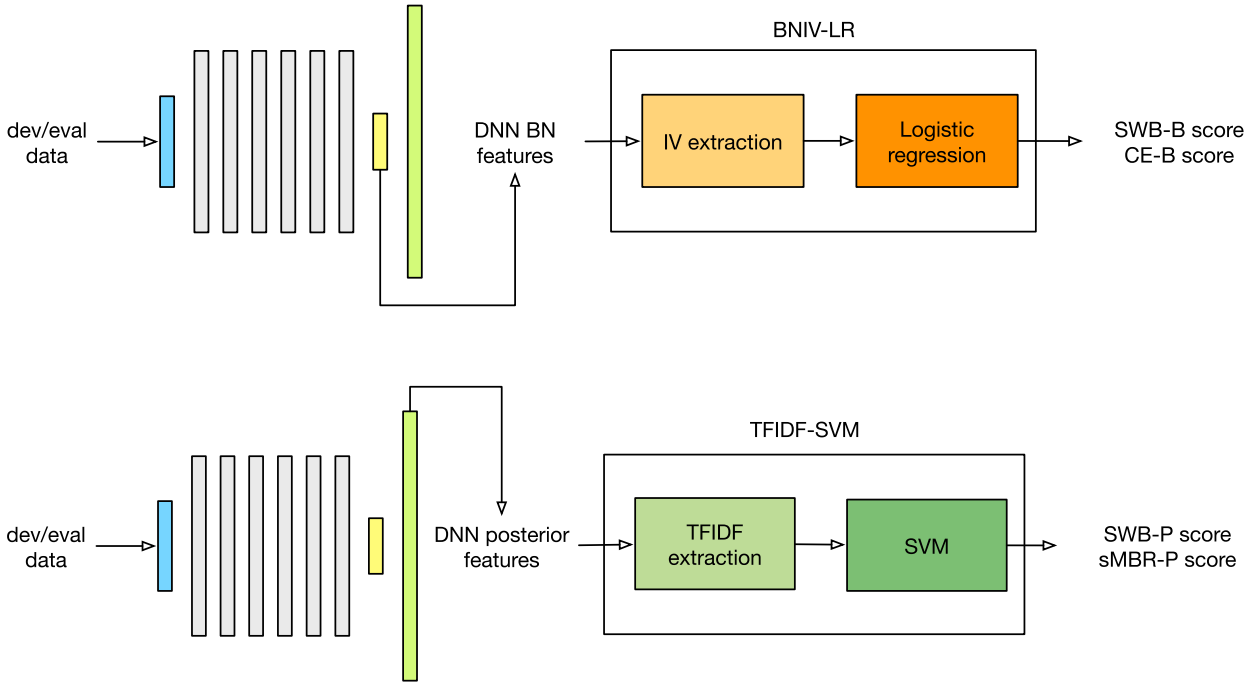

Fig. 2. Diagrams of the bottleneck i−vector (upper, BNIV-LR) and the phonotactic (lower, TFIDF-SVM) systems. The frontend DNNs were trained on SWB data and adapted to specific language data.

TF-IDF statistics were computed and support vector machine was used for language classification. This LR technology is hereinafter abbreviated as "TFIDF-SVM", and systems using the TFIDF-SVM technology are referred to with the suffix "-P" (phonotactic).

Details of the SLR systems and their parameter settings are described below.

5.1. Bottleneck i-vector (BNIV-LR / -B) system

In the bottleneck i-vector SLR systems, frame-based 64-dimension bottleneck features were extracted from LR2015-ML-TRAIN. For each DNN configuration, a universal background model (UBM) and total variability matrix were trained. Frame-level VAD based on thresholding the log mel energy was first performed to select only the bottleneck features corresponding to voiced frames. The UBM model was a 2048-component full-covariance Gaussian mixture model (GMM). For each training utterance, MAP adaptation of the UBM was performed to derive an utterance-specific GMM. The GMM means were concatenated to form supervectors of dimension $64 \times 2048 = 131,072$. Finally, a total variability matrix was trained to project the supervectors to a reduced space with 600 dimensions (Dehak et al., 2011).

The task of NIST LRE 2015 is to distinguish similar target languages within each language cluster shown in Table 1. With the extracted i-vectors, six logistic regression models were trained. These classifiers were constructed by selecting only the in-class i-vectors from the training utterances belonging to the target language and the out-of-class i-vectors from the training utterances spoken in other languages within the language cluster.

5.2. Phonotactic (TFIDF-SVM / -P) system

The phonotactic SLR systems are based on vector space modelling with TF-IDF vectors (Li et al., 2007). DNN tokenisers were applied on the multilingual training data. A phoneme bigram language model was coupled with the DNN and WFST decoding was conducted to derive phoneme transcripts. Utterance-based phoneme trigram occurrence statistics was then computed. Term frequency (TF) represents the utterance-level occurrence statistics of 171 position-dependent phonemes, their bigrams and trigrams (5 M units in total). These statistics are normalised to the

length of the sentence. Inverse document frequency (IDF) were computed separately for 3-second, 10-second and 30-second data. They are computed in a language and language-cluster agnostic way by pooling all training utterances from LR2015-ML-TRAIN. The number of documents a certain term occurred was counted and the inverse value was computed. A single TF-IDF vector was constructed for each utterance. The sparsity ratio of the TF-IDF vector is over 99%. Higher ratio is noted for utterance with shorter duration. A couple of previous studies built on the TD-IDF philosophy and improved phonotactic SLR capabilities by exploiting the most frequent and discriminative terms between languages (Corboda et al., 2007; Caraballo et al., 2010). We followed the standard TF-IDF term weighting scheme as in Ma and Li (2006), noting that further enhancements described above may improve SLR system performance.

20 binary classifiers were trained on the TF-IDF vectors for the detection of 20 languages. These classifiers were operating within-cluster. That means negative training vectors were selected only from the training utterances within the same language cluster. Classifiers were implemented by support vector machine with linear kernels (Joachims, 1999). During testing, IDF derived from the training data was used.

Language recognition scores were calibrated using a Gaussian backend (BenZeghiba et al., 2009). Within-cluster SLR trials were run on LR2015-ML-TRAIN. Multi-dimensional score vectors were derived to represent scores of all languages within the cluster. For each language, a Gaussian mixture model with 4 components was trained on the multi-dimensional score vectors. The models were used to calibrate SLR scores during test time.

5.3. Score calibration and fusion

SLR systems were implemented with three tokeniser configurations – SWB, CE, sMBR – and two SLR technologies – -B, -P.

Table 2 shows the full combination of tokenisers and SLR technologies tested in the experiment. Focusing on the non-fusion single systems, SWB-B and SWB-P refer to the use of the primary DNN tokeniser in the baseline bottleneck i–vector LR system and the baseline phonotactic LR system respectively. Following the adaptation strategy discussed in Section 2.3, eight CE adapted DNNs, which are adapted to eight different languages, were coupled with the BNIV-LR language recognisers to form 8 CE-B systems. Similarly, eight sMBR adapted DNNs were coupled with the TFIDF-SVM language recognisers to form 8 sMBR-P systems. Together with the baseline systems (SWB-B, SWB-P), there are a total of $1 + 1 + 8 + 8 = 18$ single systems and they are illustrated in Figs. 3 and 4.

To test the combinatorial effects of unsupervised tokenisers in SLR, pairwise system fusion was performed between the baseline LR systems and each of the CE-B and sMBR-P systems (to form SWB-P + sMBR-P and SWB-B + CE-B systems). Extending the combination set, 8-CE-B was formed by combining all CE-B systems, and

Table 2
Tokeniser and SLR system combination tested in the experiments.

	Tokeniser description			
	SWB	Primary	Adapted	Fusion
[Phonotactic (TFIDF-SVM) systems]				
SWB-P	✓	✓		
sMBR-P			✓[a]	
SWB-P+sMBR-P	✓		✓[a]	✓[b]
8-sMBR-P			✓	✓
SWB-P+8-sMBR-P	✓		✓	✓
[Bottleneck i–vector (BNIV-LR) systems]				
SWB-B	✓	✓		
CE-B			✓[a]	
SWB-B+CE-B	✓		✓[a]	✓[b]
8-CE-B			✓	✓
SWB-B+8-CE-B	✓		✓	✓
[Combination (BNIV-LR+TFIDF-SVM) systems]				
8-CE-B+8-sMBR-P			✓	✓
SWB-B+SWB-P+8-CE-B+8-sMBR-P	✓		✓	✓

[a] 8 systems with different adaptations specific to different languages.
[b] 8 pairwise system fusion runs between SWB and adapted systems.

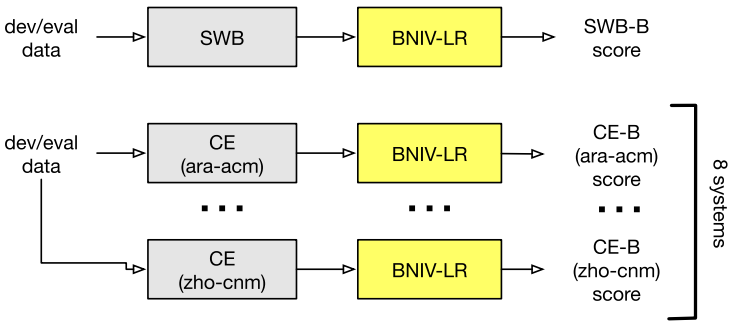

Fig. 3. Illustration of different BNIV-LR bottleneck i−vector system processes.

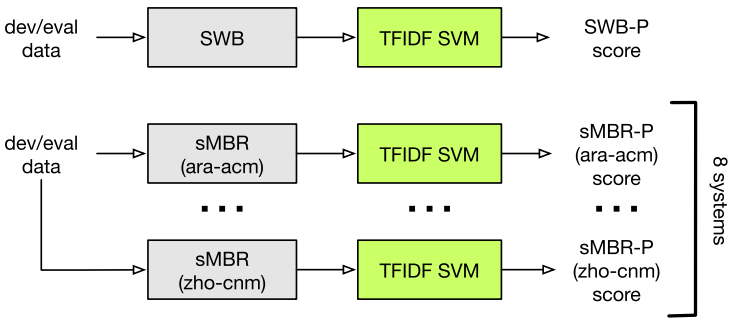

Fig. 4. Illustration of different TFIDF-SVM phonotactic system processes.

8-sMBR-P was formed by combining all sMBR-P systems. At last, system fusion was carried out to combine the primary with eight adapted systems altogether, forming SWB-P + 8-sMBR-P and SWB-B + 8-CE-B systems. Fusion between phonotactic and LR systems were also conducted. In total, the number of tested system is 2 (primary) + 16 (adapted to 1 language) + 16 (adapted fused with primary) + 6 (multiple fusion) = 40.

In each fusion trial, single system scores were converted to log likelihood ratios. LR2015-DEV was used to derive the weights for system combination subject to the minimum detection cost function, using the FoCal toolkit (Brummer, 2010). All fusion trials were carried out independently for the six language clusters. Following the work in Ng et al. (2016a), for TFIDF-SVM systems fusion was performed independently on the 3-second, 10-second and 30-second nominal duration data sets. For BNIV-LR systems, fusions were performed on mixed-duration development data sets.

6. Language recognition results

In this section, the results of SLR experiments on LR2015-EVAL are reported. Evaluation results are organised into nominal 3-second, 10-second and 30-second data. For bottleneck i−vector (BNIV-LR) systems, training data of different duration was pooled in training and fusion. For phonotactic (TFIDF-SVM) systems, following the experiment routine in Ng et al. (2016a); (2016b) duration-specific SLR model training and fusion were applied to 3-second, 10-second and 30-second data independently.

SLR results are reported in minimum Detection Cost Function (minDCF) LRE (2015). It was found that within the “French cluster” unexpected channel difference and wide range of formality in Creole Haitian exist (Torres-Carrasquillo et al., 2016). In the presentation of experimental results below, the overall system performance will be presented in minDCF computed with a global detection threshold across 18 languages without French. To visualise the SLR system performance on different language clusters, minDCF with language-dependent detection thresholds for the 6 language clusters will also be reported. For the baseline SWB and the best systems, an overall 20-language minDCF is reported.

6.1. Phonotactic LR system results

The results for TFIDF-SVM phonotactic SLR systems are summarised in the upper block in Table 3. SWB-P is the baseline phonotactic system. The minDCF scores omitting French, on 3-second, 10-second and 30-second conditions, are 40.36%, 33.51% and 29.97%, respectively. The sMBR(·) adapted tokenisers alone gave a lower average minDCF on 3-second and 30-second data (38.48%, 27.66%) but not on 10-second data (34.85%). Pairwise system fusion with SWB-P and sMBR-P(·) gave 4–8% relative reduction of minDCF for data across different nominal durations. With a fusion system comprising all eight sMBR-P(·) systems together with the baseline SWB-P system, the corresponding minDCF reduction is 10.2%, 11.3% and 17.9% for 3-second, 10-second and 30-second data, respectively.

Looking into the language specific performance for the SLR system with adapted sMBR-P(·) tokenisers, the eight SLR systems with different adaptation languages showed similar trends. Adapting the tokeniser towards one specific language did not incur additional SLR performance improvements in that language. Compared with the baseline with sMBR, all sMBR-P(·) systems gave significantly worse results on English SLR. On average there is a 10.7% relative higher minDCF for 3-second, 32.1% for 10-second and 13.6% for 30-second. This phenomenon reflects the level of SLR error increase attributed to the worsened quality of the unsupervised adapted DNN tokenisers.

Further insights into language specific performance can be derived from Fig. 5, which plots the results of SWB-P, the 8-sMBR-P fusion system and the 1+8 fusion system. In the 3-second and 30-seconds cases Iberian and Chinese SLR trials contributed to the major reduction of overall minDCF. In 10-second, the major contribution to minDCF reduction was from the Chinese SLR trials. Result degradation in English SLR trials with the adapted 8-sMBR-P system was observed in Fig. 5. These trends were consistent with what were observed in eight pairwise fusion systems with sMBR-P. More details of the phonotactic SLR systems can be found in Ng et al. (2016a). The reported minDCF in this work has a lower number compared with Ng et al. (2016a) due to the enlarged development set (LR2015-DEV, 20%) and minDCF here was computed without French.

Compared with the baseline system (SWB-P), fusion of 8-sMBR-P and the SWB-P systems gave relative minDCF reduction 10.2%, 11.3% and 17.9% on 3-second, 10-second and 30-second data. The overall 20-language minDCF for the SWB-P system was 37.57%. For the SWB-P+8-sMBR-P fusion system, the overall 20-language minDCF was 33.97%.

Table 3
Summary of minDCF (18-language, global threshold) for different SLR systems on LR2015-EVAL. sMBR-P, CE-B, SWB-P + sMBR-P and SWB-B + CE-B refer to a set of systems, each using a DNN tokeniser adapted to a distinct language. Under each system category multiple min DCF scores are computed and the average is reported.

	Min DCF (%)		
System	3-second	10-second	30-second
[Phonotactic (TFIDF-SVM) systems]			
SWB-P	40.36	33.51	29.97
sMBR-P[a]	38.48	34.85	27.66
SWB-P+sMBR-P[a]	37.21	31.08	28.68
8-sMBR-P	36.98	32.96	23.78
SWB-P+8-sMBR-P	36.23	29.72	24.62
[Bottleneck i–vector (BNIV-LR) systems]			
SWB-B	28.47	20.69	16.87
CE-B[a]	28.49	20.54	16.96
SWB-B+CE-B[a]	27.54	19.47	15.95
8-CE-B	26.57	18.46	15.22
SWB-B+8-CE-B	26.55	18.44	15.18
[Combination (BNIV-LR + TFIDF-SVM) systems]			
8-CE-B+8-sMBR-P	26.33	18.68	16.29
SWB-P+SWB-B+8-CE-B+8-sMBR-P	25.97	19.08	15.96

[a] Average min DCF from eight systems with DNN's adapted to different languages.

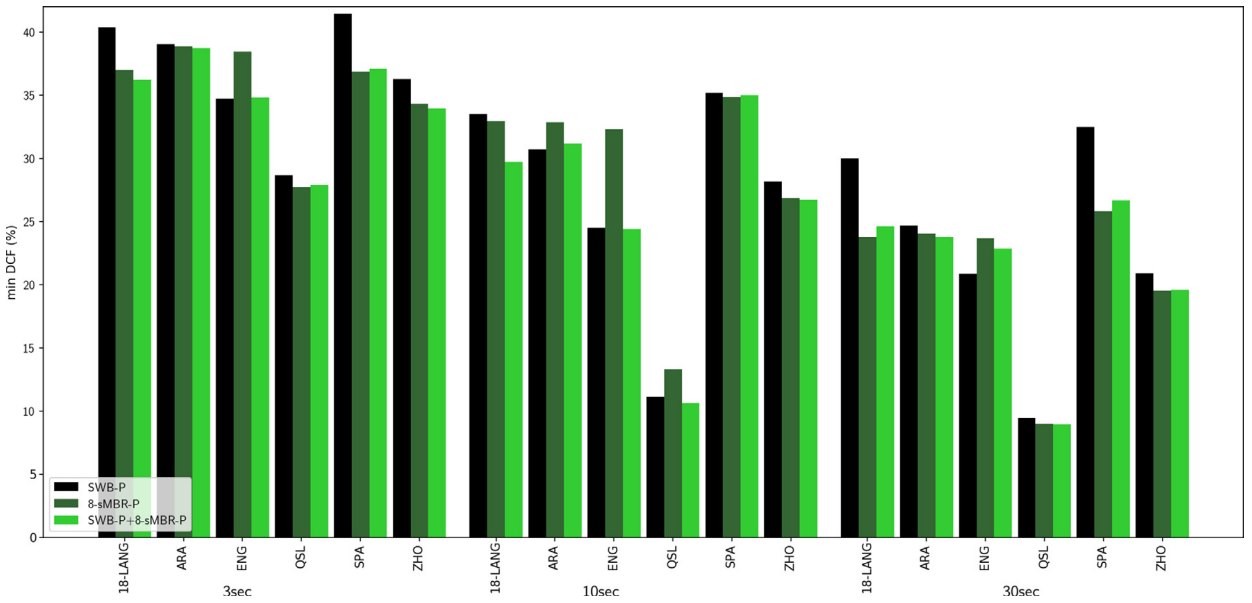

Fig. 5. SLR results of the 3 principal TFIDF-SVM systems on LR2015-EVAL across different language groups. "18-LANG" indicates the overall SLR system performance in minDCF computed with a global detection threshold across 18 languages without French. Results on the five language clusters (Arabic (ARA), English (ENG), Slavic (QSL), Iberian (SPA) and Chinese (ZHO)) were computed with language-dependent detection thresholds .

6.2. Bottleneck i-vector system results

The second block of Table 3 summarises the SLR results for all BNIV-LR systems. The minDCF omitting French languages with the SWB-B system was 28.47%, 20.69% and 16.87% for 3-second, 10-second and 30-second data, respectively. For the CE-B(·) adapted systems their average performance on 3-second, 10-second and 30-second data were 28.49%, 20.54% and 16.96%, respectively.

DNN adaptation for BNIV-LR systems used cross-entropy (CE) training, which is different from the sMBR training used for phonotactic systems. A control experiment was set up to test the difference between using the CE- and the sMBR-adapted DNN. The Egyptian Arabic language was chosen as the adaptation target. Outputs from the sMBR-adapted DNN and the CE-adapted DNN were separately taken and passed to the BNIV-LR systems trained on corresponding data. In this particular example, the CE-adapted DNN gave 0.3% absolutely lower minDCF compared to the sMBR DNN. sMBR adaptation was performed on 10-second training segments only (57 h) and CE adaptation was performed on training segments of all durations (192 h). Taking this into account, one can conclude that the SLR performance with CE and sMBR adapted DNN give results which are qualitatively similar.

Language specific performance is plotted in Fig. 6. Looking into the CE-B systems with adapted DNNs in different language clusters, systems with adapted CE-B were not found to give significant degradation in any language clusters. This is in contrast to what was observed in the TFIDF-SVM systems, where adapted DNN gave worse results for the English trials (Section 6.1). In Fig. 7, a comparison is drawn between different CE-B systems with different adapted DNN's. The system with CE-B(eng-usg) gives around 0.5% lower minDCF compared with other CE-B systems. This reflects the better quality of DNN when it is adapted to English data.

The third row in the second block of Table 3 show the pairwise BNIV-LR system fusion results. Fusion between SWB-B and any CE-B(·) systems gave relative minDCF improvements of 3.3%, 5.8% and 5.5% for 3-second, 10-second and 30-second data, respectively. Similar to what was observed in the phonotactic systems, the CE-B(·) DNN adapted to one particular language did not benefit SLR of the language in a significant way.

Fig. 6 shows the improved SLR improvements with 8-CE-B. Compared with the baseline system (SWB-B), the fusion of 8-CE-B and the SWB-B system gave relative minDCF reduction 6.7%, 10.9% and 10.0% on 3-second, 10-second and 30-second data. The overall 20-language minDCF for the SWB-B system was 27.08%. For the SWB-B+8-CE-B fusion system, the overall 20-language minDCF was 25.42%.

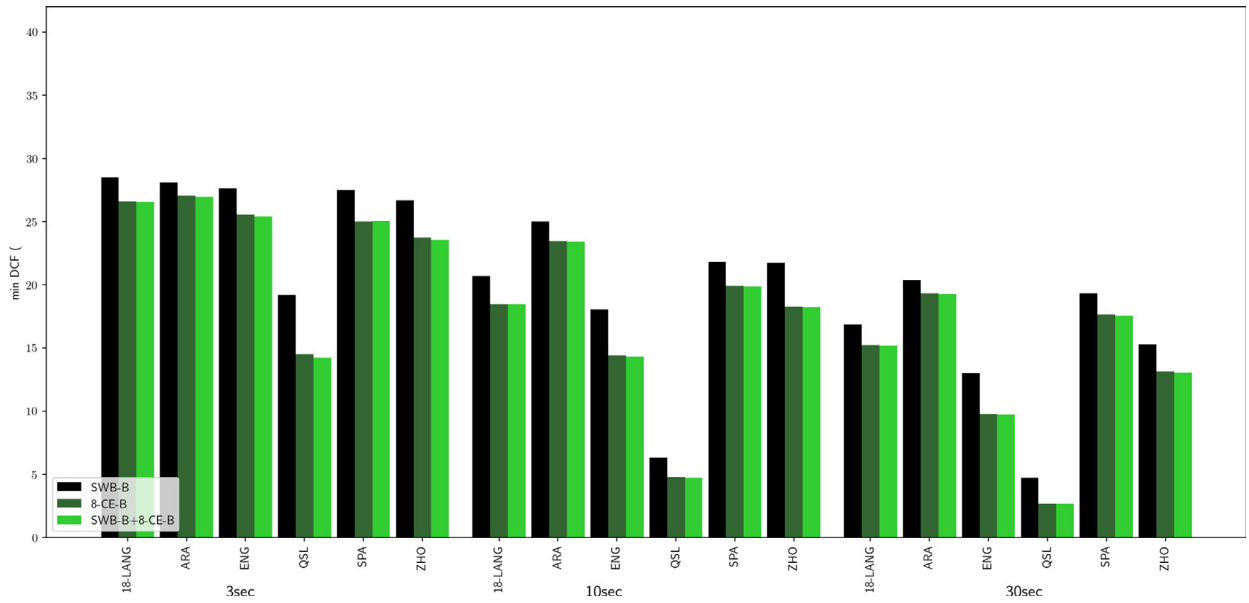

Fig. 6. SLR results of the 3 principal BNIV-LR systems on LR2015-EVAL across different language groups. "18-LANG" indicates the overall SLR system performance in minDCF computed with a global detection threshold across 18 languages without French. Results on the five language clusters (Arabic (ARA), English (ENG), Slavic (QSL), Iberian (SPA) and Chinese (ZHO)) were computed with language-dependent detection thresholds.

6.3. Overall system fusion

Overall system fusion was performed between bottleneck i-vector systems and phonotactic systems, all with adapted DNN's. The minDCF for 3-second, 10-second and 30-second data is 26.33%, 18.68% and 16.29%, respectively. Final system fusion combining all SLR systems, including the two baseline systems (SWB-B, SWB-P), was performed. The minDCF for 3-second, 10-second and 30-second data is 25.97%, 19.08% and 15.96%, respectively.

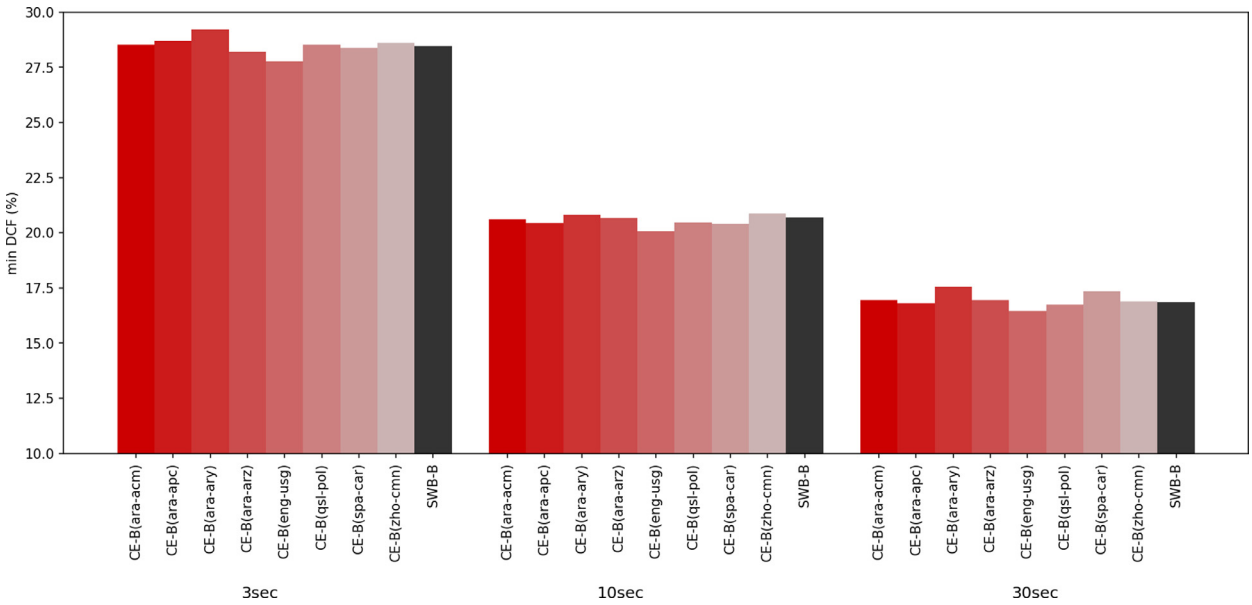

Fig. 7. Comparison of overall SLR performance on LR2015-EVAL among different CE(·) systems.

There is not significant improvement by combining the phonotactic system and the bottleneck i-vector system. This is probably due to the large performance gap between the two SLR technologies.

7. Analysis on adapted DNN outputs

In the following we aim to verify the hypothesis that adapted DNN tokenisers create diverse representations of the multilingual data for SLR. For this purpose two experiments were conducted. First, the triphone state distributions of different tokeniser outputs were analysed. Second, within-cluster language classification results from different tokenisers were compared.

In the analysis of triphone state distribution, decoding results were obtained by applying the primary and CE-adapted DNNs on different language subsets in LR2015-ML-TRAIN. Then the triphone state histograms were computed for every phoneme. Manual inspection of the histogram outputs was conducted. It was found that the phonemes generated by primary and adapted tokenisers have similar but not identical distribution. To give a specific example, Fig. 8 shows the three histograms of "schwa" with the primary tokeniser (SWB-B) and two adapted tokenisers (CE-B(ara-arz) and CE-B(eng-usg)) applied on on LR2015-ML-TRAIN(ara-arz) data. Comparing across these histograms, modest difference in the distribution can be observed. The relative difference in frequency among any pair of states (difference of y values between two states) stays roughly similar across the three tokeniser setting, but the normalised frequency values (y value) for some states fluctuate more than others.

For each of the ten systems (the primary SWB-B system, eight CE-B systems and the SWB-B + 8-CE-B fusion system), within-cluster language classification decisions were derived according to the maximum likelihood score from language detection trials. The test was performed on LR-2015-EVAL. The cluster identity of every trials was assumed to be known and the French cluster was ignored. With equal weight across five language clusters, the percentage of trials where language classification decision differs between any two systems was computed. Table 4 gives an overview on pairwise system comparison. On all but the last row in the Table, pairs of single systems differ by 18 or 19% in terms of the number of trials where decisions differ. This show that tokeniser adaptation brought about diversity across LR system systems. On the bottom row in Table 4, the SWB-B + 8-CE-B fusion system was shown to differ from its component systems by 12–14%. With only 11.7% trials having different language decision,

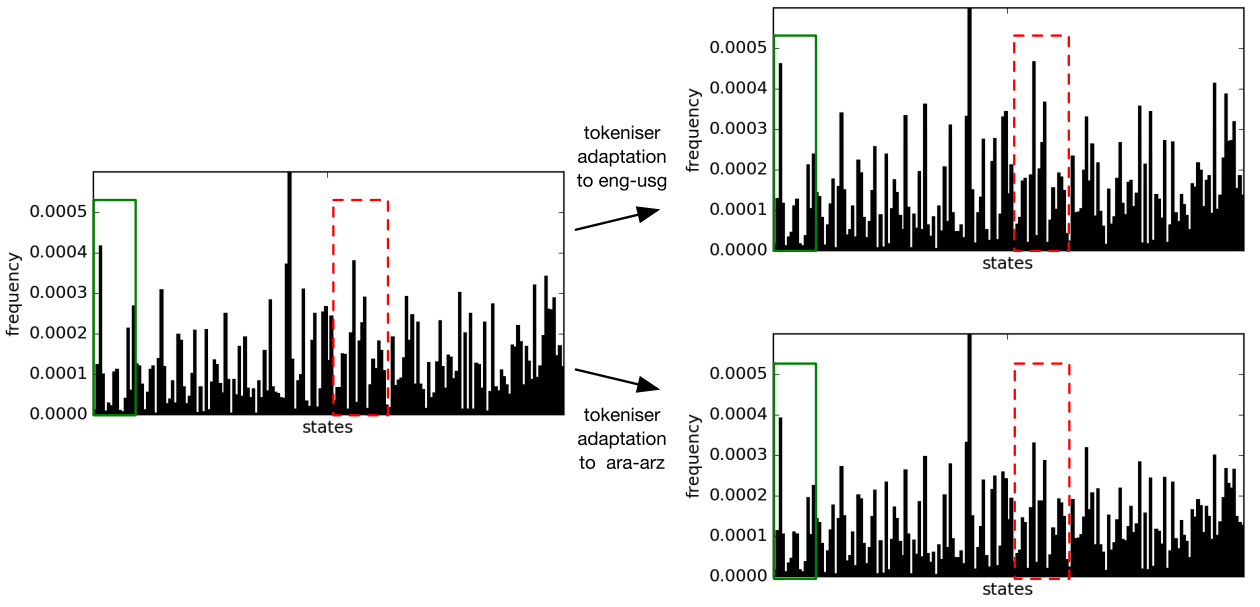

Fig. 8. Example of how the triphone states frequency distributions change after the tokeniser was adapted to different languages. The X-axis represents different triphone states in the "schwa" phoneme; The Y-axis is the normalised occurrence frequency of the state. The dash-line boxes identify areas of the three plots in which the distributions differ. States in the continuous-line boxes show similar distributions.

Table 4
Pairwise system result comparison among the baseline bottleneck i-vector system (SWB-B), 8 adapted systems with DNN tokenisers adapted to 8 different languages (CE-B), and the 8-CE-B fusion system. Pairwise system difference is represented by percentage of trials (language-cluster-balanced, French cluster excluded) where language classification results differ.

	SWB-B (%)	ara-acm (%)	ara-apc (%)	ara-ary (%)	ara-arz (%)	eng-usg (%)	qsl-pol (%)	spa-car (%)	zho-cmn (%)	8-CE-B (%)
SWB-B	0.0									
ara-acm	18.5	0.0								
ara-apc	18.5	18.0	0.0							
ara-ary	18.7	18.7	18.6	0.0						
ara-arz	19.0	18.2	18.1	18.6	0.0					
eng-usg	18.5	18.1	17.8	18.5	18.4	0.0				
qsl-pol	19.1	18.5	18.7	19.2	18.6	18.5	0.0			
spa-car	18.6	18.4	17.9	18.8	18.1	17.9	18.7	0.0		
zho-cmn	18.7	18.8	18.0	18.4	18.7	18.2	18.7	18.3	0.0	
8-CE-B	13.7	13.7	12.5	14.5	12.4	11.7	13.3	12.3	12.7	0.0

the SWB-B(eng-usg) system is "closest" to the optimal fusion system. This agrees with the quality analysis in Section 6.2 and Fig. 7.

In other studies, rich information such as confidence and domain anomalies were shown to be deducible from neural network output (Hermansky et al., 2015). Such analysis can be extended to shed light on cross-lingual tokeniser adaptation for spoken language recognition.

8. Conclusion

In this paper, we described the strategies in unsupervised crosslingual adaptation of DNN phoneme recognisers, using cross-entropy (CE) and state-level minimum Bayes risk (sMBR) objective functions, in spoken language recognition systems. CE- and sMBR-adapted DNN's were used in bottleneck i-vector and phonotactic SLR systems, giving rise to respectively 7−11% and 10−18% relative reduction in minimum detection cost function (DCF). The SLR performance depends on the diversity of the output from the ensemble tokenisers. Adapted DNN's were shown to give diverse representations of the acoustic signal. Noticeable performance improvements were observed in pairwise system fusion between an SLR system with the primary DNN and one with the adapted DNN. The quality of the adapted DNN tokenisers correlates with SLR performance. A performance degradation was observed in the phonotactic systems on English SLR test trials when the primary DNN, trained in a supervised manner, was replaced by an unsupervised adapted DNN. For the bottleneck i-vector systems, unsupervised DNN adaptation to English created a DNN with a better quality in contrast to unsupervised adaptation to other languages, which was believed to be a more difficult task. The assumption of diverse representations from adapted tokenisers was validated by a phonetic analysis on the training data, and a comparison of SLR results across different systems. Future study would focus on improving the quality of the adapted DNN, the use of different adaptation strategies and different ways of system combination such as early fusion.

Acknowledgments

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology) and Google.

References

Ambikairajah, E., Li, H., Wang, L., Yin, B., Sethu, V., 2011. Language identification: a tutorial. IEEE Circuits Syst. Mag. 11 (2), 82–108. doi: 10.1109/MCAS.2011.941081.

Anderson, O., Dalsgaard, P., 1997. Language identification based on cross-language acoustic models and optimised information combination. In: Proceedings of Eurospeech, pp. 67–70.

BenZeghiba, M.F., Gauvain, J.-L., Lamel, L., 2009. Language score calibration using adapted Gaussian backend. In: Proceeedings ofs Interspeech.

BenZeghiba, M.F., Gauvain, J.-L., Lamel, L., 2012. Phonotactic language recognition using MLP features. In: Proceeedings of Interspeech.

Brummer, N., 2010. FoCal toolkit for evaluation, fusion and calibration of statistical pattern recognisers. Available: https://sites.google.com/site/nikobrummer/focal.

Caraballo, M.A., D'Haro, L.F., Cordoba, R., San-Segundo, R., Pardo, J.M., 2010. A discriminative text categorization technique for language identification built into a PPRLM system. In: Proc. FALA, pp. 193–196.

Corboda, R., D'Haro, L.F., Fernandez-Martinez, F., Macias-Guarasa, J., Ferreiros, J., 2007. Language identification based on n-gram frequency ranking. In: Proceeedings of Interspeech.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R., 2011. Language recognition via I-vectors and dimensionality reduction. In: Proceeedings of Interspeech.

D'Haro, L.F., Corboda, R., Salamea, C., Echeverry, J.D., 2014. Extended phone log-likelihood ratio features and acoustic-based I-vectors for language recognition. In: Proceeedings of the International Conference on Acoustics, Speech and Signal Processing.

D'Haro, L.F., Glembek, O., Plchot, O., Matejka, P., Soufifar, M., Cordoba, R., Černocký, J., 2012. Phonotactic language recognition using i-vectors and phoneme posteriogram counts. In: Proceeedings of Interspeech.

Fék, R., Matějka, P., Grézl, F., Plchot, O., Černocký, J., 2015. Multilingual bottleneck features for language recognition. In: Proceeedings of Interspeech, pp. 389–393.

Ferrer, L., Lei, Y., McLaren, M., Scheffer, N., 2016. Study of senone-based deep neural network approaches for spoken language recognition. Proc. IEEE/ACM Trans. Audio Speech Lang. 24 (1), 105–116.

Gauvain, J.L., Messaoudi, A., Schwenk, H., 2004. Language recognition using phone lattices. In: Proceeedings of the International Conference on Spoken Language Processing.

Gibson, M., Hain, T., 2006. Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition. In: Proceeedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH).

Glembek, O., Matějka, P., Burget, L., Mikolov, T., 2008. Advances in phonotactic language recognition. In: Proceeedings of Interspeech.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1, ICASSP'92, San Francisco, California, IEEE Computer Society, Washington, DC, USA, 4, 517–520. http://dl.acm.org/citation.cfm?id=1895550.1895693.

Hazen, T.J., Zue, V.W., 1997. Segment-based automatic language identification. J. Acoust. Soc. Am. 101 (4), 2324–2331.

Hermansky, H., Burget, L., Cohen, J., Dupoux, E., Feldman, N., Godfrey, J., Khudanpur, S., Maciejewski, M., Mallidi, S.H., Menon, A., Ogawa, T., Peddinti, V., Rose, R., Stern, R., Wiesner, M., Veselý, K., 2015. Towards machines that know when they do not know: summary of work done at 2014 Frederick Jelinek memorial workshop. In: Proceeedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5009–5013. doi: 10.1109/ICASSP.2015.7178924.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA, pp. 169–184. Chapter 11.

Knill, K.M., Gales, M.J.F., Ragni, A., Rath, S.P., 2014. Language independent and unsupervised acoustic models for speech recognition and keyword spotting. In: Proceeedings of Interspeech, pp. 16–20.

Li, H., Ma, B., Lee, C.-H., 2007. A vector space modeling approach to spoken language identification. IEEE Trans. Audio Speech Lang. Process. 15 (1), 271–284.

Li, H., Ma, B., Lee, K.A., 2013. Spoken language recognition: from fundamentals to practice. Proc. IEEE 101 (5), 1136–1159.

Lööf, J., Golland, C., Ney, H., 2009. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system. In: Proceeedings of Interspeech, pp. 88–91.

Ma, B., Li, H., 2006. A comparative study of four language identification systems. Comput. Linguist. Chin. Lang. Process. 11 (2), 159–182.

Muthusamy, Y.K., Barnard, E., Cole, R.A., 1994. Reviewing automatic language identification. IEEE Signal Process. Mag. 11 (4), 33–41.

Navrátil, J., 2006. Recent advances in phonotactic language recognition using binary-decision trees. In: Proceeedings of Interspeech.

Ng, R.W.M., Chettri, B., Hain, T., 2016a. Combining weak tokenisers for phonotactic language recognition in a resource-constrained setting. In: Proceeedings of Interspeech.

Ng, R.W.M., Nicolao, M., Saz, O., Hasan, M., Chettri, B., Doulaty, M., Lee, T., Hain, T., 2016b. The Sheffield language recognition system in NIST LRE 2015. In: Proceeedings of the Speaker Odyssey.

Richardson, F., Reynolds, D., Dehak, N., 2015. Deep neural network approaches to speaker and language recognition. IEEE Signal Process. Lett. 22 (10), 1671–1675.

Schultz, T., 2002. Globalphone: a multilingual text and speech database developed at Karlsruhe University. In: Proceeedings of Interspeech, pp. 345–348.

Schwarz, P., 2009. Phoneme Recognition Based on Long Temporal Context. Brno University of Technology Ph.D. thesis.

Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A., 2003. Acoustic, phonetic and discriminative approaches to automatic language recognition. In: Proceeedings of Eurospeech, pp. 1345–1348.

Suzuki, M., Tachibana, R., Thomas, S., Ramabhadran, B., Saon, G., 2016. Domain adaptation of CNN based acoustic models under limited resource settings. In: Proceeedings of Interspeech, pp. 1588–1592.

The 2011 NIST language recognition evaluation plan, 2011 [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf.

The 2015 NIST language recognition evaluation plan, 2015 [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/LRE15_EvalPlan_v22-3.pdf.

Torres-Carrasquillo, P., Dehak, N., Godoy, E., Reynolds, D., Richardson, F., Shum, S., Singer, E., Sturim, D., 2016. The MITLL NIST LRE 2015 language recognition system. In: Proceeedings of the Speaker Odyssey.

Torres-Carrasquillo, P.A., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D., Deller, J.J.R., 2002. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In: Proceeedings of the International Conference on Spoken Language Processing.

Veselý, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proceeedings of Interspeech.

Vu, N.T., Kraus, F., Schultz, T., 2011. Cross-language bootstrapping based on completely unsupervised training using multilingual a-stabil. In: Proceeedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5000–5003. doi: 10.1109/ICASSP.2011.5947479.

Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L., Liu, Q., 2014. Fast adaptation of deep neural network based on discriminant codes for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1713–1725.

Zissman, M.A., 1993. Automatic language identification using Gaussian mixture and hidden MArkov models. In: Proceeedings of the International Conference on Acoustics, Speech and Signal Processing, vol.II, pp. 399–402.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44.