INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Combining weak tokenisers for phonotactic language recognition in a resource-constrained setting

Raymond W. M. Ng, Bhusan Chettri, Thomas Hain

Department of Computer Science, University of Sheffield, United Kingdom
{wm.ng,b.chettri,t.hain}@sheffield.ac.uk

Abstract

In the phonotactic approach for language recognition, a phone tokeniser is normally used to transform the audio signal into acoustic tokens. The language identity of the speech is modelled by the occurrence statistics of the decoded tokens. The performance of this approach depends heavily on the quality of the audio tokeniser. A high-quality tokeniser in matched condition is not always available for a language recognition task. This study investigated into the performance of a phonotactic language recogniser in a resource-constrained setting, following NIST LRE 2015 specification. An ensemble of phone tokenisers was constructed by applying unsupervised sequence training on different target languages followed by a score-based fusion. This method gave $5-7\%$ relative performance improvement to baseline system on LRE 2015 eval set. This gain was retained when the ensemble phonotactic system was further fused with an acoustic iVector system.

Index Terms: Language recognition, phonotactics, multilingual adaptation.

1. Introduction

In a spoken language recognition (SLR) task, an automatic system is used to infer the language identity of the given acoustic signal [1]. Different types of information from a speech signal can be used to identify languages. Standard SLR methods can be categorised by the features they use. The two most popular SLR approaches are the acoustic-phonetic and phonotactic approaches [2, 3, 4]. In the acoustic-phonetic approach, low-level acoustic features such as Mel-frequency cepstral coefficients (MFCC)[5], or shifted-delta cepstral coefficient (SDC)[6] are extracted, on which statistical models such as Gaussian mixture models are trained to model languages [7, 8]. For the phonotactic approach, an ASR-style tokeniser is required to convert the speech signal into graphs (lattices) of discrete tokens, most notably phonemes. The occurrence patterns of these tokens differs from target languages, thus allowing for modelling and language classification [2, 8, 9]. Acoustic SLR is often believed to give better results. Nevertheless, with detailed control of tokeniser quality and adequate modelling, phonotactic SLR could outperform its counterpart in certain situations [10]. Furthermore system fusion is an important necessity for language recognition, and having a second similar performing system allows further performance gain. This was shown most recently for some evaluation systems in LRE 2015 [11].

This paper focuses on the phonotactic approach in SLR. From the literature, a number of major advances in the field can be observed. For instance, clustering and adaptation were proposed to address the issue of sparse statistics of higher-order N-grams, especially for short utterances [12, 13]. Soft counts from phone lattices [14], phone posteriograms [15] or multilayer perceptron features [16] were used to generate smooth statistics for language recognition. The use of parallel tokenisers has long been a useful trick to boost performance [2, 17]. Regardless whether the 1-best phone sequence or a phone posteriogram vector is used, with parallel tokenisers the speech data is decoded in multiple different ways giving a diverse view for subsequent language recognition model training. The combination of different streams of tokeniser results often provide gains in the performance.

The National Institute of Standards and Technology (NIST) has conducted a number of evaluations of automatic language recognition technology. Recent NIST language recognition evaluations were held in 2011 and 2015 [18, 19]. These evaluations focus on languages that are similar to each other and frequently mutually intelligible, e.g. dialectal variants [19]. In NIST LRE 2015, a new requirement on Fixed Training Data for all components in the LR system was introduced that created extra challenges to system building. For tokeniser training a 300-hour Switchboard data set with phonetic transcriptions and alignments was permitted to be used. These data were monolingual (English) conversational telephone speech only. Language, channel and style mismatch between that data and the LR training and test data was well known.

Experiments for this paper investigate how to make best use of the tokeniser training data for a phonotactic language recognition task in a resource constrained setting (i.e. no other transcribed training data can be used). Without the availability of extra transcribed data, the Deep Neural Netowrk (DNN) tokeniser was adapted (fine-tuned) in an unsupervised manner using the sequence training criterion towards different languages. Adaptation was performed using multi-lingual LR training data. The multiple adapted tokenisers operated in parallel for phonotactic language recogniser training and testing. The results were ultimately combined at the score level. Despite having the single source of transcribed text, after DNN adaptations the ensemble of phonotactic LR systems demonstrated *minDCF* 's which were $5-7\%$ lower than those obtained when using a single unadapted phonotactic tokeniser. This gain was retained when the system was fused with an acoustic iVector language recognition system.

This work is related to previous work on phonotactic language recognition discussed above. We have built on the idea of using different tokenisers trained in multiple languages [17]. Here we have employed a standard phonotactic LR setup using normalised tri-gram phone count. The focus has been put on the adaptation of tokeniser from a single data source where multi-lingual transcribed data is not available. Related work on adaptation can be found in [20], where the DNN is adapted to different speakers using sequence training criterion. To the

http://dx.doi.org/10.21437/Interspeech.2016-630

best of our knowledge, there is no work in the literature that addresses training data constraints in the context of phonotactic language recognition.

2. Tokeniser adaptation

The tokeniser used in language recognition is a DNN phone tokeniser implemented in a feed-forward hybrid setting. Assume a dataset with a total of U utterances each indexed u (i.e. $u = [1, 2, \ldots, u, \ldots, U]$). Each utterance has different duration and is represented by the number of frames ($[T_1, T_2, \ldots, T_u, \ldots, T_U]$). An observation at time t in utterance u is denoted by $\boldsymbol{o}_{ut}$. With the reference state label s_{ut}, a baseline DNN tokeniser was trained on the cross-entropy criterion at the 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 $\mathcal{F}_{\text{CE}}$ where

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

To adapt a tokeniser to different target languages, the weights of the DNN was retrained (fine-tuned) on multi-lingual (out-of-domain) data in an unsupervised manner. This could potentially be considered as problematic as high error rates of target sequences will be reaffirmed. The original tokeniser was applied on the multi-lingual LR training data to generate a 1-best decode sequence $\hat{S}_u = (\hat{s}_{u0}, \cdots, \hat{s}_{ut}, \cdots, \hat{s}_{uT})$. There is no constraint on what kind of objective function the adapted tokeniser should be optimised. In this paper, minimisation of state-level Minimum Bayes Risk (sMBR) was chosen to be the objective function [21], which is defined as,

$$\mathcal{F}_{\text{sMBR}} = -\sum_{u}\sum_{S} p(S|\boldsymbol{O}_u) A(S, \hat{S}_u) \quad (2)$$

$A(S, \hat{S}_u)$ 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 $\hat{S}$. The objective function aims to minimise Bayes risk at state-level (i.e. maximised accuracy). The posterior probability was computed in the utterance level to achieve robust estimation. [22].

3. Experimental setup

This study focuses on phonotactic language recognition, which is considered as a two-stage process – phone tokenisation and language recognition. Data usage and systems are designed to conform to this two-stage system regime.

3.1. Data

The training and development data used in this study comes mainly from three corpora. The Switchboard 1 (SWB) contains transcribed English conversational telephone speech data with a total duration of 302 hours including silence. It was used for the training of the source phone tokenisers. In addition, two multi-lingual datasets (LDC2015E87, LDC2015E88) were used for language recogniser training [19]. LDC2015E87 comprises conversational telephone speech from the CallHome and CallFriend collections, in Egyptian Arabic, Standard Mandarin and US English. LDC2015E88 comprises data in seventeen further target languages as used in NIST LRE 2015. The amount of data for different languages varies from 0.4 hours to 63 hours in LDC2015E88, and the data amount in LDC2015E87 is 159 hours for Egyptian Arabic and US English respectively and 107 hours for Standard Mandarin (Table 1). In this paper, the two multi-lingual data sets were used for unsupervised adaption of tokenisers.

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

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

This study proposes to improve the poor performance of phonotactic LR systems, thus tests were focused on 3-second and 10-second data only. The language recognition systems were tested on an internal evaluation data set (HELDOUT) constructed by extracting 10% from LDC2015E87 and LDC2015E88 [23], as well as the official LRE2015 EVAL data.

3.2. Unadapted tokeniser

The baseline tokeniser followed a feedforward DNN hybrid setting with 6 hidden layers, where each contains 2048 neurons, which are followed by a bottleneck layer with 64 neurons and an output layer with 3815 neurons (as per the number of senones). The input features to the DNN were Mel–frequency cepstral coefficient (MFCC) features with differentials and mean normalisation. Further follow-on processing used by 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 senone alignment results from a constrained maximum likelihood linear regression (CMLLR) adapted, maximum mutual information(MMI)-optimised acoustic model set. This first-pass DNN is referred to as unadapted DNN, and abbreviated as "1-SWB" to indicate a single set of training data. 1-SWB models were trained using the cross-entropy criterion.

1-SWB was compared with the typically used Hungarian (HU) phone tokeniser trained on SpeechDat database [24]. The SpeechDat-HU tokeniser is known to have a rich phonetic variety for a good performance of language recognition. Nevertheless, because of the unavailability for the raw training data, a shallow neural network phone tokeniser trained on TempoRAl Pattern (TRAP) techniques was used in this study [25].

3.3. Unsupervised adaptation of tokenisers

1-SWB was adapted to 20 different languages using the training data given for each of the 20 target languages in NIST LRE 2015 as shown in Table 1[19]. This study focused on language recognition on speech with 3-second and 10-second durations. In order to remove as much non-speech as possible DNN-based voice activity detection was applied on the raw training data,

with the aim to derive speech segments of compatible durations [23]. Segments of required durations in a particular language are then used for sequence training with state-level MBR criterion (Eq(2)) to derive an adapted tokeniser. A total of 20 adapted tokenisers were constructed. These are further collectively denoted as 20-SWB'. Language recognition will be performed on each of these 20-SWB' tokenisers. The average and the standard deviation of the min Detection Cost Function (*minDCF*) will be computed.

3.4. Language recogniser setup

In this study, different tokenisers and adaptation settings are explored. They share a common language recogniser setting based on vector space modelling with tf-idf vectors [26]. In brief, the tokenisers were applied on the multi-lingual training data to derive phone transcripts. Utterance–based phone tri-gram occurrence statistics was then computed, from which term frequency (tf) and inverse document frequency (idf) was derived. A single tf-idf vector was constructed for each utterance, allowing to train 20 binary classifiers for each of the 20 target languages. The 20 target languages in NIST LRE 2015 are organised into 6 language clusters [19]. For the training for each classifier the positive training vectors were selected from training utterances belonging to the target language. Negative training vectors were selected only from the training utterances within the same language cluster. During testing, idf was inherited from the training data and the likelihood for each of the 20 languages was computed to obtain the final language recognition result.

Language recognition scores were applied to a Gaussian backend. For each system and each target language, a Gaussian mixture model with 4 components was trained on the multi-dimensional score vectors, which resulted from the decoding of the training data. During testing, the likelihood of each language-dependent GMMs was computed. The Gaussian back–end system is denoted as $\mathcal{N}(\cdot)$.

System fusion was performed among a subset of the full set of 20-SWB' systems. In each fusion trial, single system scores are converted to log likelihood ratios and 10% DEV data portion extracted from LRE2015 TRAIN data set was used to learn a linear weight for system combination, with respect to the minimum detection cost [27]. System fusion trials were carried out independently for the six language clusters and the 3-second and the 10-second nominal duration data set.

4. Results

4.1. Quality of tokenisers

Table 2 shows the results of the phonotactic language recognition system using three different (sets of) tokenisers. Across the three tokenisers, the Switchboard (SWB) tokeniser as used in the official NIST LRE 2015 evaluation gave the best LR performance in terms of the *minDCF* score.

The SpeechDat-HU tokeniser was a shallow NN tokeniser and this was believed to be the cause of inferior LR performance. On a separate internal test using TIMIT as the tokeniser training data and LRE96 as the language recognition task, the TRAP tokeniser gave 3% absolute higher (i.e. worse) *minDCF* compared to a DNN tokeniser. If the TRAP versus DNN performance difference on LRE96 can be transferred to performance on LRE2015, results in Table 2 may suggest a potential gain of phonotactic language recognition results with a better performing tokeniser.

The average LR performance of the 20 adapted tokenisers were also included in Table 2 for reference. Note that after unsupervised sequence training on the hypothesis transcripts of the twenty target languages, the LR performance of individual tokeniser settings showed worse performance. On 3-second and 10-second eval data, the *minDCF* increased by 1.6% and 2.4% absolute respectively, in comparison to the unadapted SWB tokeniser results. Nevertheless, the availability of the 20 tokeniser system was expected to benefit from the output variety, and therefore benefit the overall results with system fusion.

Table 2: LRE performance (*minDCF*) for different tokenisers

	HELDOUT		EVAL	
Tokeniser	03s	10s	03s	10s
SpeechDat-HU	37.27	36.26	42.12	40.57
1-SWB	36.37	31.82	41.33	38.20
20-SWB'	38.46	36.94	42.90	40.63
	$\pm$0.22	$\pm$0.61	$\pm$0.08	$\pm$0.28

4.2. Gaussian backend and tokeniser combination

Table 3 shows results for system fusion with a Gaussian backend as well as score-level score calibration or system fusion between multiple phonotactic LR systems. Compared with the results shown in Table 2, the Gaussian backend [$\mathcal{N}(\cdot)$] reduced the *minDCF* for both unadapted and adapted tokenisers on 10-second test data by at least 5% relative. However, the Gaussian backend only worked for the adapted tokenisers on 3-second test data. This may be due to the empirical choice of training the Gaussian components on mismatched 30-second training data for the case of unadapted tokeniser [23].

The bottom half of Table 3 shows results for logistic regression score calibration, with the LR system score from the unadapted SWB (1-SWB) tokeniser. Calibration yielded a 7% relative reduction of *minDCF* on HELDOUT data while the performance improvement on EVAL data was none or marginal. For the score fusion with adapted SWB (SWB') tokenisers, a mixed trend could be observed when a subset of tokenisers were used in fusion. When all LR systems with 20 different adapted tokenisers (20-SWB') were combined a consistent gain of at least 10% relative could be observed on HELDOUT data. The corresponding gain on EVAL data was $5 - 9\%$ relative.

Comparing between the LR system with single tokeniser, cal(1-SWB), and the fusion of LR systems with adapted tokenisers, fusion(20-SWB'), the 20-SWB' system is 14% and 11% relative better on HELDOUT data. On EVAL data it was 7% and 5% relative better.

Gaussian backend was an important process to normalise the scores from multiple LR systems for fusion. According to our experiments, excluding the Gaussian backend would eliminate all possible gain from system fusion.

4.3. Fusion with acoustic systems

In a final set of experiments the single SWB tokeniser phonotactic LR system (1-SWB) and multiple LR systems with 20 adapted-SWB tokenisers (20-SWB') were combined with an acoustic LR system. The acoustic LR system was an iVector system and results are shown in Table 4. The *minDCF* for the iVector system is much lower than those of the 1-SWB and 20-SWB' on HELDOUT data. Nevertheless, the performance gap narrows on EVAL data. All fusion settings resulted in an improvement of performance. Fusion with 20-SWB' gave bet-

Table 3: LRE performance (*minDCF*) for tokeniser combination. On EVAL data, 20-SWB' systems were $5-7\%$ relative better than 1-SWB (results underlined)

	HELDOUT		EVAL	
Tokeniser	03s	10s	03s	10s
[Gaussian backend]				
$\mathcal{N}$(1-SWB)	38.87	29.69	42.01	35.33
$\mathcal{N}$(20-SWB')	36.04	32.02	40.66	37.87
	$\pm$0.53	$\pm$1.27	$\pm$0.17	$\pm$0.77
[Score calibaration / system fusion]				
cal(1-SWB)	36.11	27.43	<u>41.63</u>	<u>36.07</u>
fusion(ara-5-SWB')	31.65	25.83	38.43	36.24
fusion(eng-3-SWB')	31.57	26.88	39.23	36.51
fusion(fre-2-SWB')	32.79	27.10	40.27	36.17
fusion(qsl-2-SWB')	31.84	28.20	39.03	37.94
fusion(spa-4-SWB')	31.80	26.06	38.95	35.25
fusion(zho-4-SWB')	31.55	27.93	39.07	37.37
fusion(20-SWB')	31.19	24.50	<u>38.55</u>	<u>34.40</u>

Table 4: LRE performance (*minDCF*) for acoustic systems

	HELDOUT		EVAL	
System$^\sharp$	03s	10s	03s	10s
iVector	22.83	17.38	36.61	34.22
iVector+1-SWB	21.33	16.68	35.58	32.42
iVector+20-SWB'	21.53	15.80	35.26	31.73
bn-iVec	18.47	14.85	32.55	29.78
bn-iVec+1-SWB	18.04	13.67	32.41	28.18
bn-iVec+20-SWB'	17.75	14.43	32.02	28.21
iVector+bn-iVec	18.68	13.70	32.50	28.24
iVector+bn-iVec+1-SWB	18.43	12.31	32.26	27.88
iVector+bn-iVec+20-SWB'	17.36	13.38	31.76	28.12

$^\sharp$ 1-SWB is score calibrated, 20-SWB' is a fusion system with 20 systems

ter or equal performance as fusion with 1-SWB. Across different data sets and different durations, fusion with 1-SWB gave a relative 3-7% reduction on *minDCF* , and fusion with 20-SWB' gave a relative 4-9% reduction on *minDCF* . Despite the marginal difference, the performance gain with the ensemble of tokenisers was carried over in the acoustic-phonotactic fusion system.

We also tried system fusion of 20-SWB' with (i) the bottleneck-iVector system [23], and (ii) a combination of the iVector and the bottleneck-iVector systems. All acoustic-phonotactic fusions demonstrated reduction of *minDCF* . However, as the performance gap between the acoustic and phonotactic counterparts was larger again, the robustness of 20-SWB' over 1-SWB was weakened. This is particularly true for 10-second test data.

5. Summary

This study investigated the use of unsupervised tokeniser adaptation to create variations of a baseline tokeniser for use in phonotactic language recogntion. The baseline Switchboard tokeniser was adapted towards a set of 20 target languages from the NIST LRE 2015. Fusion of the outputs of the 20 adapted tokenisers results in at least a 10% relative reduction in *minDCF* on the internal development data set and a $5-7\%$ relative reduction on LRE2015 EVAL data. This gain was retained when the system was fused with an acoustic iVector langauge recognition system. These results indicated the usefulness of phonotactic LR approach despite the lack of a high quality tokeniser. Future studies will focus on the robustness of phone occurrence statistics, particularly for short duration test sentences. The soft estimate with phone posteriogram or bottleneck features could be incorporated in the phonotactic approach. Adaptation of tokenisers to maximised langauge discriminability of the decoded phones may also give promising options for further exploration.

6. Acknowledgements

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

7. Data access statement

Data used in this paper was obtained from these resources: Switchboard 1, Switchboard Cellular Part 2 and multilingual training data sets for NIST LRE 2015 (LR2015E87, LR2015E88). Specific file lists used in the experiments, as well as result files can be accessed online with DOI:10.15131/shef.data.3462599

8. References

[1] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, Oct. 1994.

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

[3] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 82 –108, secondquarter 2011.

[4] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. Reynolds, and J. J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002.

[7] M. A. Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proc. ICASSP*, vol. II, 1993, pp. 399–402.

[8] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic and discriminative approaches to automatic language recognition," in *Proc. Eurospeech*, 2003, pp. 1345–1348.

[9] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2324–2331, Apr. 1997.

[10] G. Gelly, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, and A. Messaoudi, "Language recognition for dialects and closely related languages," in *Odyssey 2016*, 2016, pp. 124–131.

[11] "NIST LRE 2015 workshop," Informal communication, 2015.

[12] J. Navrátil, "Recent advances in phonotactic language recognition using binary-decision trees," in *Proc. Interspeech*, 2006.

[13] O. Glembek, P. Matějka, L. Burget, and T. Mikolov, "Advances in phonotactic language recognition," in *Proc. Interspeech*, 2008.

[14] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language recognition using phone lattices," in *Proc. ICSLP*, 2004.

[15] L. F. D'Haro, O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba, and J. Černocký, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in *Proc. Interspeech*, 2012.

[16] M. F. BenZeghiba, J.-L. Gauvain, and L. Lamel, "Phonotactic language recognition using MLP features," in *Proc. Interspeech*, 2012.

[17] L. F. D'Haro, R. Corboda, C. Salamea, and J. D. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based I-vectors for language recognition," in *Proc. ICASSP*, 2014.

[18] "The 2011 NIST language recognition evaluation plan (LRE11)," http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf, 2011.

[19] "The 2015 NIST language recognition evaluation plan (LRE15)," http://www.nist.gov/itl/iad/mig/upload/LRE15_EvalPlan_v22-3.pdf, 2015.

[20] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1713–1725, Dec 2014.

[21] M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition," in *INTERSPEECH 2006, Ninth International Conference on Spoken Language Processing*, 2006.

[22] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.

[23] R. W. M. Ng, M. Nicolao, O. Saz, M. Hasan, B. Chettri, M. Doulaty, T. Lee, and T. Hain, "The sheffield language recognition system in NIST LRE 2015," in *Proc. Speaker Odyssey*, 2016.

[24] P. Pollák, J. Černocký, J. Boudy, K. Choukri, H. van den Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, J. Sadowski, P. Staroniewicz, H. Tropf, J. Kochanina, A. Ostroukhov, M. Rusko, and M. Trnka, "SpeechDat(E) - Eastern European telephone speech databases," 1998.

[25] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Brno University of Technology, 2009.

[26] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 271–284, Jan. 2007.

[27] N. Brummer, "Focal toolkit for evaluation, fusion and calibration of statistical pattern recognisers," 2010. [Online]. Available: https://sites.google.com/site/nikobrummer/focal