7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features

K. Kyriakopoulos, M.J.F. Gales, K.M. Knill

ALTA Institute / Engineering Department
Cambridge University, UK
{kk492,mjfg,kate.knill}@eng.cam.ac.uk

Abstract

The distances between and relative movements of phones in acoustic space in language learners have been shown to be indicative of the speaker's proficiency, in a way that is compact and independent of bias-inducing voice qualities. Typically these features are based on known transcriptions, "read aloud" style tasks. This paper examines the information that can be extracted about speakers from phone distance features (PDFs) when the transcription is unknown. Here, phone distances are obtained by measuring the relative entropy between a distribution trained on the speaker's manner of pronunciation of each of the phones of the English language and distributions trained on each of the other phones. These features are extracted from untranscribed audio and so rely on automatic speech recognition (ASR) output. The ASR can have high word error rates, as spontaneous, non-native speech is being recognised. Two forms of speaker characterisation are examined using these features: first, the use of PDFs to predict the speaker's proficiency and second, their use in classifying the mother tongue (L1) of the speaker. For both tasks, recorded answers to sections of the BULATS English Speaking test were used. Using only PDFs for predicting the grade within a Gaussian Process based grader showed performance comparable to using a range of standard fluency style features. This indicates the robustness of PDFs to errors in ASR output. Additionally, the same PDF features can detect with high accuracy the L1 of the speakers from among 21 L1s using a deep neural network based classifier. Experiments on South American Spanish show that it is further possible to discriminate between the speakers' countries of origin.

1. Introduction

The process by which a language learner improves their pronunciation can be thought of as a path through acoustic space from their initial incorrect pronunciation, affected by their native language (L1) and dialect, towards a pronunciation more closely resembling native speech. It is therefore useful, in the context of Computer Aided Pronunciation Training (CAPT), to be able to automatically characterise the path that the learner is following and evaluate their position along it.

This paper investigates phone distance features (PDFs) for characterising the pronunciation of a non-native speaker of English, from recordings of un-transcribed spontaneous speech. It examines to what extent these features carry information about the speaker's starting point (L1 and country of origin) and their progress along the learning path (as measured by human-assigned proficiency scores).

Approaches for automatic assessment of pronunciation in the literature often include comparison to native speaker models [1, 2, 3], which can introduce considerable bias with regards to accent and voice quality. The most common features used are prosodic [4, 5, 6] and ASR confidence measures (both at the word and phone level) [7, 2, 1]. Recent investigations have shown promising results based on phone distance based measures [8, 9], on which this paper is based. Most existing systems rely on "read aloud" style tasks with known transcriptions. There are a few systems which can grade spontaneous speech, e.g. [10, 11, 12], however, their scope is usually limited. This paper uses phone distance features, an extension on the vowel distance approaches in [13, 9], to grade spontaneous non-native speech, using only non-native training data.

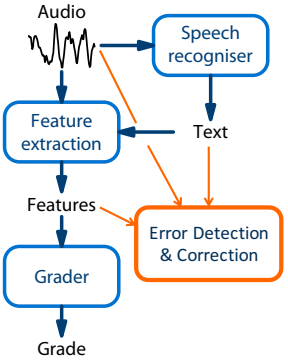

Figure 1: *Architecture of system for automatic assessment and feedback of spoken language.*

The spontaneous nature of the speech and the reliance of PDFs on an aligned phone sequence necessitates that the candidate's audio must first be passed through an automatic speech recogniser (ASR) to determine what the speaker said, and the recognised text used together with initial audio for feature extraction (Figure 1). The error rate of this ASR will always be an issue and so the robustness of the PDFs to the extent of this error is also investigated.

Approaches employed so far in the literature for L1 classification include i-vector modelling [14], GMMs trained on MFCCs [15], and prosodic [16] approaches, with varying degrees of success. The approach investigated in this paper predicts, from recordings of spontaneous speech, the speaker's native language (L1) from among 21 different languages and, in the case of Spanish speakers, their country of origin from among three countries.

2. Phone Distance Features

Pronunciation is a key predictor of speaker proficiency, and is expected to become more native, reducing strain to the listener caused by L1 effects, as the learner progresses up the CEFR lev-

10.21437/SLaTE.2017-11

els [17]. A large component of pronunciation is the manner in which the phones of the language are rendered. Extracting features to represent pronunciation of phones, however, presents a number of difficulties, particularly when dealing with spontaneous speech. First, acoustic models of the phones are not a robust predictor of proficiency, due to the large variation across speakers with different accents, voice qualities and L1s but of otherwise similar level. The forms of native pronunciation being emulated may also vary from speaker to speaker, owing to the large variation in English native speech, creating problems with using native speaker comparisons. The spontaneous nature of the speech further complicates obtaining comparable native speaker models and strengthens the need for general non-native reference approaches.

To overcome these issues, this paper employs an approach based on the distances between phones. Rather than characterising each phone by the distribution of acoustic features in its articulations, it is defined relative to the pronunciation of each of the other phones, with the full set of phone-pair distances describing the speaker's overall accent. Distances between acoustic models should be more robust to speaker variability than the models themselves. In [9] phonetic pronunciation features consisting of a set of phone-pair distances were proposed for vowels and applied to read speech. Here, the features consist of a set of phone-pair distances covering all 47 phones in English and are applied to both read and spontaneous speech. This yields 1081 distances in total.

Phone distance features should thus robustly represent the pronunciation of a speaker in samples of spontaneous, untranscribed audio, in a way that is compact and independent of the speaker's irrelevant voice qualities.

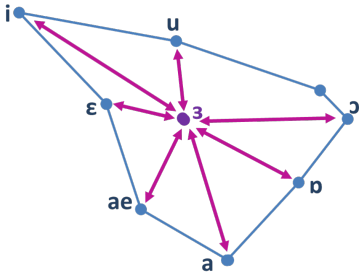

Figure 2: *Illustration of the phone distance concept*

The speaker's recorded utterances are passed through an ASR and time aligned to the most probable phone sequence given the recognised word sequence. A set of statistical models is then trained to represent the manner of pronunciation of each of the phones in the English language. For each possible phone pair, the distance between the phone models is measured by the symmetric Kullback-Leibler (K-L) divergence [18]. If the statistical models for phones ϕ_i and ϕ_j are $p(\phi_i)$ and $p(\phi_j)$, respectively, the K-L divergence between the two phones is defined as

$$\mathrm{D_{KL}}(p_i||p_j) = \int p(\phi_i) \log\left(\frac{p(\phi_i)}{p(\phi_j)}\right) d\phi_i. \qquad (1)$$

Since the K-L divergence is not symmetric and the distance measure should be invariant of the order in which the distributions are taken, one type of the symmetric K-L divergence (also known as Jensen–Shannon divergence [19]) is used, which can be written as

$$\mathrm{D_{JS}}(p_i||p_j) = \frac{1}{2}\left[\mathrm{D_{KL}}(p_j||p_i) + \mathrm{D_{KL}}(p_j||p_i)\right], \qquad (2)$$

Each phone is modeled by a single multivariate Gaussian with a mean, $\boldsymbol{\mu}$, and diagonal covariance matrix, $\boldsymbol{\Sigma}$. The input vector consists of PLP features, extracted from the speaker's audio. For each speaker, a model set is trained on all the speech from that speaker. Full recognition is run to acquire 1-best hypotheses from which time aligned phone sequences are generated. Single Gaussian models for each phone are then trained given these alignments. The K-L divergence of $\mathrm{D_{JS}}(p_i||p_j)$ is calculated as

$$\mathrm{D_{KL}}(p_i||p_j) = \frac{1}{2}\Big[\mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) - d + \ln\left(\frac{\det\boldsymbol{\Sigma}_j}{\det\boldsymbol{\Sigma}_i}\right)\Big], \qquad (3)$$

where $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ are the operators for the trace and determinant of the matrix, respectively.

If there is insufficient data to train the multivariate Gaussian of a particular phone, the PDFs corresponding to all phone-pairs containing that phone are set to -1. The resultant vector thus also contains information about which phones the speaker avoided pronouncing. This information may itself help predict speaker proficiency and L1. This approach was found to yield higher score prediction accuracy than replacing missing PDFs with the mean value based on other speakers in the training set.

This paper investigates the hypothesis that features extracted in this way are strongly representative of the speaker's accent, by evaluating how well they predict the speaker's proficiency, native language and country of origin.

3. Data

The experiments reported in this paper are based on candidate responses to the spoken component of the Business Language Testing Service (BULATS), provided by Cambridge English Language Assessment. The BULATS speaking test has five sections, all related to business scenarios [20]. Section A consists of short responses to prompted questions. Candidates read 8 sentences aloud in Section B. Sections C-E consist of spontaneous responses of several sentences in length to a series of spoken and visual prompts. Candidates are scored on a scale from 0 to 30, based on their overall proficiency, mapping to standard CEFR levels as shown in Table 1.

BULATS score range	Level description	CEFR level
29-30	Upper advanced	C2
25-29	Advanced	C1
20-25	Upper intermediate	B2
15-20	Intermediate	B1
10-15	Elementary	A2
2-10	Beginner	A1
0-2	Fail/Incomprehensible	pre-A1

Table 1: *Equivalence between BULATS scores and CEFR levels (adapted from [21])*

For the purposes of the experiments in this paper, the data is segmented into four non-overlapping data sets: BLX0, which is used to train the ASR, TRN, which is used to train the regressors and classifiers, EVL1 which is used to evaluate the classifiers, and EVL2, which is used to evaluate the regressors,classifiers and ASR.

It is important that none of the speakers in the training or evaluation sets (TRN, EVL1 or EVL2) are present in the ASR training set, so that ASR induced error in the PDFs is uniform across the data used to train and evaluate the other systems.

For each speaker in each of the four sets there is available the audio, a human assigned proficiency score and meta-data describing the candidate's L1 and country of origin. BLX0 and EVL2 additionally are accompanied by crowd-sourced transcriptions, which are used in training and evaluating the ASR. Finally, each speaker in EVL2 has also been scored by a highly qualified expert grader, who are known to have extremely high inter-annotator agreement (upwards of 0.95). Evaluation of the regressor (which is done by Pearson Correlation Coefficient - PCC - of actual to predicted grades) is therefore only performed on EVL2.

When evaluating country of origin on speakers whose L1 is known to be Spanish, subsets of TRN and EVL1 including only Spanish speakers (called TRN_S and EVL1_S) are used.

Tables 2 and 3 show the breakdown of L1 and country ofo-rigin in the latter three sets (TRN, EVL1 and EVL2) used in the course of this investigation.

L1	Set		
	TRN	EVL1	EVL2
Spanish	4502	2156	-
Tamil	1468	790	-
Gujarati	1015	230	94
Hindi	563	294	-
Telugu	462	250	-
Malayalam	395	184	-
Bengali	333	152	-
Russian	303	170	-
French	291	115	36
Polish	258	69	39
Vietnamese	245	67	37
Kannada	226	131	-
Arabic	202	51	39
Portuguese	176	78	-
Dutch	173	47	32
Thai	144	43	36
Japanese	135	68	-
Marathi	106	67	-
Italian	107	41	-
Korean	90	53	-
Oriya	65	26	-

Table 2: *Breakdown of number of speakers in each data set by native language (L1)*

Country	Set	
	TRN_S	EVL1_S
Colombia	798	296
Mexico	3208	1578
Spain	359	220

Table 3: *Breakdown of speakers in Spanish data set by country of origin*

4. Experimental Setup

4.1. ASR

Due to the incorrect pronunciations, grammar and rhythm, related to the speaker's proficiency level and first language (L1), the accuracy of standard commercial "off-the-shelf" ASR systems is too low for non-native learner English. Instead specific ASR systems are trained.

A stacked-hybrid DNN-HMM acoustic model is used for ASR. It is trained on a 108 hour (1075 speaker) Gujarati L1 BULATS data set with merged crowd-sourced transcriptions [22] (BLX0, mentioned above), using the HTK toolkit [23, 24]. The input consists of 9 consecutive frames of 40-D filterbank features with delta appended to each frame feature. A bottleneck DNN is trained on the AMI corpus [25], and 39-D BN features extracted for the BULATS data and transformed using a global semi-tied covariance matrix [26]. The transformed BN features are appended to HLDA [26] projected PLP features with CMN and CVN applied at the speaker level to yield a 78-D per frame input feature. The input to the stacked hybrid DNN-HMM is a concatenation of 9 consecutive transformed feature vectors, 702-D. The DNN structure is $702x1000^5x6000$. A Kneser-Ney trigram LM is trained on 186K words of BULATS test data and interpolated with a general English LM trained on a large broadcast news corpus, using the SRILM toolkit [27].

4.2. Regression and Classification DNNs

Both regression and classification tasks are performed using deep neural networks (DNNs), constructed using Torch. The networks each have 6 hidden layers (8 layers in total), with 1200 hidden units per hidden layer. Dropout of 50% of the units is implemented at each layer. Weight decay and stop validation are employed to further improve generalisation. The regression networks are trained for minimum MSE, while the classification networks are trained for minimum cross-entropy. The networks are trained in batches of 1000 speakers at a time, for 200 iterations. The regression networks are evaluated (on the evaluation sets described in Section 3, using the Pearson correlation coefficient (PCC) and mean square error (MSE) between predicted and actual results. The classifiers are evaluated by the percentage of speakers correctly classified.

4.3. Baseline Features

The fluency and prosodic features described in [5], plus the number and fraction of disfluencies, fraction of speech in the recording duration and vowel frequency, are used as baseline features for both the grader and the classifiers. Each system is implemented using just baseline features, just phone distance features and the two combined.

5. Results and Discussion

The results of four experiments using the systems described in the previous sections are presented and discussed in the following subsections. First, the ASR trained on BLX0 is evaluated on EVL2 to obtain indicative ASR word error rates, by profi-

ciency and language. Using phone distance features, extracted as described in Section 2, using the ASR trained on BLX0, a score predictor is trained of TRN and evaluated on EVL2, an L1 classifier is trained on TRN and evaluated on EVL1 and a country of origin classifier is trained on TRN_S and evaluated on EVL1_S.

5.1. ASR Performance

This ASR trained on BLX0 as described in Section 3.1 is evaluated using the transcriptions of EVL2. It has an overall word error rate (WER) of 47.5% and a phone error rate (PER) of 33.9%. The relatively high PER suggests a considerable amount of noise will be present in the phone distance features. Note, however, that the inherent inaccuracies and noisiness of crowdsourced transcriptions may lead error rate figures to be exaggerated.

The WER for the mixed-L1 data is further broken down by L1 and CEFR level in Table 4. As can clearly be seen, recognition error decreases with increasing proficiency of the speaker, a result that holds across all L1s. This is to be expected as higher proficiency speakers are likely to speak more intelligibly and their speech is therefore easier for the ASR to recognise.

	Spanish	Arabic	Dutch	French	Thai	Viet.
A1	69.8	69.7	78.7	55.8	65.4	65.4
A2	58.7	67.4	45.7	48.0	56.0	55.9
B1	48.6	47.2	41.3	45.0	50.7	53.5
B2	47.1	47.3	40.3	45.0	48.1	56.6
C	48.8	48.6	43.1	36.7	41.3	43.6
All	50.9	52.0	42.5	43.6	50.2	53.0

Table 4: *Word error rates (WER) of ASR on indicative Mixed-L1 data set (EVL2) broken down by L1 and CEFR level*

5.2. Score Prediction

The baseline and phone distance features are now used to build an automatic grader, which attempts to predict human-assigned scores, trained on the ordinary human grader assigned scores in TRN and evaluated on the expert-assigned scores in EVL2.

As seen in Table 6 below, PDFs outperform baseline features in both MSE and PCC when used on their own and yield considerable improvements when used in combination with them. This is particularly promising when considering that these results are in the presence of considerable ASR error and that the scores being predicted are general proficiency scores and not pronunciation-specific.

	PCC	MSE
Base	0.737	26.4
PDF	0.751	23.6
Base+PDF	0.832	15.8

Table 5: *Performance (PCC and MSE) of DNN grader described in §4.2, trained on TRN and evaluated on EVL2*

5.3. Candidate L1 Classification

Having established that PDFs are strong predictors of proficiency, a DNN classifier is now built to determine whether they can also be used to predict candidates' native languages.

Table 6 below shows the performance of the same features when used to classify the speakers' native language (L1) from among 21 candidates. The baseline features already perform significantly better than random chance, but the PDF-trained DNNs significantly outperform them, suggesting the phone distances are indeed indicative of speaker L1. Combining PDFs and baseline features degrades the accuracy slightly, suggesting most information about L1 contained in the baseline features is also captured by the PDFs.

	Accuracy (%)	
	EVL1	EVL2
Base	53.1	31.9
PDF	69.0	61.2
Base+PDF	66.5	60.0

Table 6: *Accuracy (percentage of the speakers correctly classified) for DNN L1 classifier described in §4.2, trained on TRN and evaluated on EVL1 and EVL2*

As can be seen in Table 7, L1 classification performance is highest for those languages with the most data in the training and testing sets (e.g. Gujarati, Spanish and Tamil) and lowest for those with the least data (e.g. Marathi, Italian, Korean and Oriya).

	% Correct L1	% # Speakers in TRN	Most confused
Overall	66.5	-	-
Spanish	97.7	4502	Portuguese
Tamil	76.7	1468	Telugu
Gujarati	74.5	1015	Hindi
Hindi	62.3	563	Telugu
Marathi	0.0	106	Hindi
Italian	2.4	107	Spanish
Korean	3.7	90	Spanish
Oriya	0.0	65	Hindi

Table 7: *Breakdown by L1 of accuracies for L1 classifier evaluated on EVL1, using baseline + PDF, for L1s with the most and leasts number of speakers in the training set (TRN), along with L1 that its spears are most frequently misclassified as.*

The identities of the L1s that speakers with each L1 are most frequently misclassified as confirm expectations regarding similarities between languages. Romance languages (Italian, Portuguese and French) are most frequently confused with Spanish, Indo-Aryan languages (Gujarati, Marathi, Bengali and Oriya) are most commonly misclassified as Hindi, and Dravidian languages (Telugu and Malayalam) most commonly misclassified as Tamil. This is confirmed by observing the confusion matrices in Tables 8, 9 and 10 below, which demonstrate that the majority of incorrect classifications within each group of languages are as other L1s in the same group.

Breaking down L1 classification accuracy by CEFR level (Table 11), the classifier is the least accurate (with both baseline and PDF features) for poor speakers (A1) and its performance then increases with proficiency. This is attributed to the decrease in WER with increasing proficiency. As WER falls, features are more representative of the speaker's actual pronunciation and less noisy, resulting in more powerful classification.

	Spanish	French	Portugese	Italian
Spanish	97.7	0.5	0.0	0.0
French	16.5	43.5	0.0	0.0
Portugese	21.8	24.4	29.1	0.0
Italian	36.6	26.8	0.0	2.4

Table 8: *Percentage of speakers of each Romance L1 classified as other Romance languages*

	Gujarati	Hindi	Bengali
Gujarati	74.3	10.9	1.7
Hindi	11.6	62.9	0
Bengali	10.5	55.9	20.3
Marathi	3.0	74.6	0
Oriya	3.8	73.1	0

Table 9: *Percentage of speakers of each Indo-Aryan L1 classified as other Indo-Aryan languages*

	Tamil	Telugu	Malayalam	Kannada
Tamil	76.7	8.5	0	0
Telugu	37.6	27.5	0	0
Malayalam	49.5	23.4	12.4	0
Kannada	24.4	19.1	0	0

Table 10: *Percentage of speakers of each Dravidian L1 classified as other Dravidian languages*

	%Baseline	%PDF	%Baseline+PDF
Overall	53.1	69.0	66.5
Al	41.3	60.0	56.5
A2	47.0	60.1	58.5
B1	55.2	70.0	67.8
B2	53.5	70.5	67.7
C1	56.3	71.8	67.8
C2	50.0	57.5	77.5

Table 11: *Detection rate for speaker L1 classifier, evaluated on EVL1, broken down by CEFR level*

Performance rises faster with proficiency for baseline features than for PDFs, suggesting the latter are more robust to ASR error. Performance levels out and slightly dips as proficiency enters the C levels, which can be attributed to the speakers' pronunciation becoming more similar to native speech and therefore less indicative of their L1.

5.4. Candidate Country of Origin Classification

A similar methodology to the previous section is now employed to attempt to classify the speaker's country of origin, to see whether the predictive powers of PDFs can narrow foreign accents down further than the level of L1.

As seen in Tables 12 and 13, the classifier performs very well when identifying the country of origin of speakers with a known L1. This suggests that PDFs (as well as, to some extent, the baseline features) are able to capture phonological differences in the way speakers of the same L1 with different regional dialects pronounce the phones of English. As with the L1 classifier, PDFs considerably outperform the baseline features, while adding the baseline features to the PDF only slightly increases performance.

As with L1 classification, performance increases with CEFR level (Table 14), again clearly attributable to decreasing word error rates.

6. Conclusions

Phone distance features were presented as a means of representing the relative manner in which a learner renders the phones of the English language, based only on recordings of spontaneous speech. They were shown to be a strong predictor of the speaker's proficiency (as assigned by human graders), their native language and, at least for speakers of Spanish, their country of origin, performing significantly better than baseline features at all three prediction tasks. Although they depend on accurate ASR for best results and their predictive power decreases with increasing WER, they were found to be more robust to ASR errors than the baseline features.

7. Acknowledgements

This research was funded under the ALTA Institute, Cambridge University. Thanks to Cambridge English, Cambridge University, for supporting this research and providing access to the BULATS data.

8. References

[1] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech and Language*, vol. 21, pp. 219–230, 2007.

[2] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners." in *INTERSPEECH*, 2014, pp. 1468–1472.

[3] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of english learner pronunciation using discriminative classifiers," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5351–5355.

[4] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for english as l2," *Speech Prosody*, 2010.

[5] R. van Dalen, K. Knill, and M. Gales, "Automatically Grading Learners' English Using a Gaussian Process," in *Proceedings of Workshop on Speech and Language Technology for Education (SLaTE)*, 2015.

[6] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[7] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, 2000.

[8] S. Asakawa, N. Minematsu, T. Isei-Jaakkola, and K. Hirose, "Structural representation of the non-native pronunciations." in *INTERSPEECH*, 2005, pp. 165–168.

[9] N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for CALL," in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*, 2006, pp. 126–129.

[10] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

	Accuracy (%)
Base	77.3
PDF	85.5
Base+PDF	87.0

Table 12: *Accuracy (percentage of the speakers correctly classified) for DNN Country of Origin classifier described in §4.2, trained on the Spanish speakers in TRN (aka. TRN§) and evaluated on the Spanish speakers in EVL1 (aka EVL1_S)*

	% Correct
Spain	71.5
Colombia	45.5
Mexico	97.5

Table 13: *Breakdown by country of detection rates for country classifier, using baseline + PDF features, trained and evaluated on Spanish-only data (TRN_S/EVL1_S)*

	%Baseline	%PDF	%Baseline+PDF
Overall	77.3	85.5	87.0
A1	51.1	34.4	51.1
A2	55.0	75.7	78.1
B1	79.7	88.7	90.3
B2	81.8	89.8	89.6
C1	77.3	86.3	85.8
C2	86.7	86.7	93.3

Table 14: *Detection rate for speaker country classifier, evaluated on Spanish data, broken down by CEFR level*

[11] AISpeech, 2012. [Online]. Available: http://bit.ly/2mMyxRX

[12] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proceedings of INTERSPEECH*, 2014.

[13] C. Graham, F. Nolan, A. Caines, and P. Buttery, "Automated assessment of non-native speech using vowel formant features," ALTA Institute, Phonetics Lab, DTAL - University of Cambridge.

[14] H. Behravan, V. Hautamaki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "i-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 29–41, 2016.

[15] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using gaussian mixture models," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 343–346.

[16] J. H. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 836–839.

[17] C. of Europe, *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001.

[18] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler Distance," *IEEE Transactions on Information Theory*, 2001.

[19] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information theory*, 2003.

[20] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[21] BULATS. Business language testing service. [Online]. Available: http://www.bulats.org/computer-based-tests/results

[22] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr 2015.

[23] S. Young *et al.*, *The* HTK *book (for* HTK *version 3.4.1)*. University of Cambridge, 2009. [Online]. Available: http://htk.eng.cam.ac.uk

[24] ——, *The* HTK *book (for* HTK *version 3.5)*. University of Cambridge, 2015. [Online]. Available: http://htk.eng.cam.ac.uk

[25] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 28–39.

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

[27] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.