Use of Graphemic Lexicons for Spoken Language Assessment

K.M. Knill, M.J.F. Gales, K. Kyriakopoulos, A. Ragni, Y. Wang

ALTA Institute / Engineering Department
Cambridge University, UK
{kate.knill,mjfg,kk492,ar527,yw396}@eng.cam.ac.uk

Abstract

Automatic systems for practice and exams are essential to support the growing worldwide demand for learning English as an additional language. Assessment of spontaneous spoken English is, however, currently limited in scope due to the difficulty of achieving sufficient automatic speech recognition (ASR) accuracy. "Off-the-shelf" English ASR systems cannot model the exceptionally wide variety of accents, pronunications and recording conditions found in non-native learner data. Limited training data for different first languages (L1s), across all proficiency levels, often with (at most) crowd-sourced transcriptions, limits the performance of ASR systems trained on non-native English learner speech. This paper investigates whether the effect of one source of error in the system, lexical modelling, can be mitigated by using graphemic lexicons in place of phonetic lexicons based on native speaker pronunications. Graphemic-based English ASR is typically worse than phonetic-based due to the irregularity of English spelling-to-pronunciation but here lower word error rates are consistently observed with the graphemic ASR. The effect of using graphemes on automatic assessment is assessed on different grader feature sets: audio and fluency derived features, including some phonetic level features; and phone/grapheme distance features which capture a measure of pronunciation ability.

Index Terms: graphemic speech recognition, spoken language assessment, automatic grading, non-native speakers

1. Introduction

By 2020 the number of people worldwide using or learning English as an additional language is expected to exceed 1.5 billion [1]. Automatic systems to support learners in practice and for examination are essential to handle this level of demand. There are a few systems available, however, their scope is limited e.g. [2, 3, 4]. To assess a learner's spoken communication skills requires a system that can handle spontaneous speech with all its disfluencies and non-standard grammatical content. Automatic speech recognition (ASR) is needed to determine what the learner said as input to automatic grading and feedback systems (Figure 1). 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. ASR improvements, such as the use of DNNs [5, 6], are still insufficient to achieve the required accuracy to create systems which can, e.g., provide feedback on spontaneous speech at all proficiency levels. This paper considers whether using a graphemic lexicon, in place of the standard phonetic lexicon, can yield more accurate ASR and help assessment.

There is limited non-native English ASR training data available. It cannot cover anywhere near the variations observed in testing due to L1, proficiency, spontaneous speech disfluencies and recording conditions. In addition, non-native speech can be very hard to understand, and often contains unusual names so transcription quality is compromised. Crowd-sourcing enables more transcriptions at the cost of lower inter-annotator agreement and more spelling errors [7].

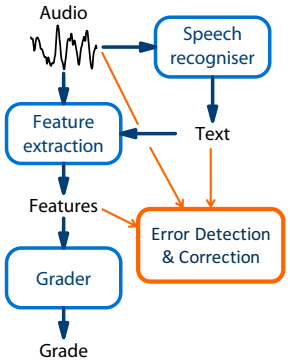

Figure 1: *Architecture of system for automatic assessment and feedback of spoken language.*

Since whole word models are not feasible for recognising large vocabularies, most ASR systems are based on phonetic subword units. A lexicon is required to map words into their phonetic sequences. Lexicons such as CMUdict [8] and Combilex [9] have undergone a lot of manual checking and provide good quality pronunciations, however, they are based on native English speaker pronunciations. Pronunciations for words not found in the lexicon have to be automatically generated using a G2P system (e.g. [10, 11]). Due to the irregular relationship between spelling and pronunciations in English, these G2P pronunciations tend to vary in quality, with proper nouns being particularly difficult to predict. Such terms are common as learners often refer to places and people they are familiar with.

Alternatively a word can be split into its constituent graphemes and the graphemes used as the subword units. Application of graphemic systems to English, where the graphemes are usually the letters of the alphabet, has typically shown the ASR error rate to degrade compared to phonetic systems due to the mismatch between spelling and sounds [12, 13]. [14] did show that graphemic English systems are complementary to phonetic systems and a lower WER can be achieved by system combination. By contrast, for low resource languages with limited training data, researchers have found graphemic systems to consistently match or outperform phonetic systems (e.g. [15, 16, 17, 18]). This paper, therefore, investigates whether a graphemic lexicon can yield better recognition accuracy for non-native spontaneous English speech and what effect this has on automatic assessment.

The graphemic lexicon and use in ASR and automatic assessment is presented in Section 2. Sections 3 and 4 present the experimental setup and results, respectively. Conclusions are given in Section 5.

2. Graphemic Lexicon

A graphemic lexicon replaces phones with graphemes. For English, this is straightforward with the alphabet letters /a-z/ forming the base grapheme set. Since the systems are designed to handle spontaneous effects hesitations and fillers have to be modelled. Following [18], two additional root graphemes are added, /G00,G01/, to model all events marked in the transcriptions as hesitations. Two attributes are also defined:

A APOSTROPHE P PARTIAL WORD

where the attribute P is used for partial words arising from the spontaneous speech transcriptions. All the graphemes are mapped into the set of 28 root graphemes plus the attributes. The graphemic lexicon also contains word boundary information [19] which is used as default in the Cambridge phonetic system [18]. For example,

Phonetic Lexicon

ABE'S	eyˆI bˆM zˆF
ABLE	eyˆI bˆM axˆM lˆF
ABOUT_%partial%	aeˆI bˆM awˆM tˆF
ABOUT_%partial%	axˆI bˆM awˆM tˆF

Graphemic Lexicon

ABE'S	aˆI bˆM eˆM;A sˆF
ABLE	aˆI bˆM lˆM eˆF
ABOUT_%partial%	aˆI bˆM oˆM uˆM tˆF;P

The graphemic attributes for the root grapheme set are restricted to assigning the graphemes /a,e,i,o,u/ to the vowel class, and /vowel,y/ to the vowely class.

2.1. Use of Graphemic Lexicon in ASR

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. To recognise the data for each section, a stacked hybrid DNN-HMM speaker independent (SI) MPE trained ASR system is used, Figure 2.

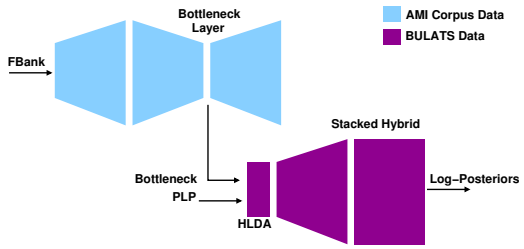

Figure 2: *Stacked hybrid ASR architecture.*

Bottleneck (BN) features for the BULATS data are generated using a bottleneck DNN trained on the AMI meeting corpus [21]. This corpus mostly consists of non-native English speakers. The BN DNN is sequence trained which is sensitive to transcription quality. It was found that using the professionally transcribed AMI corpus yields more robust BN features than training on the BULATS data, for which only crowd-sourced transcriptions exist. The BN DNN used phonetic state outputs, and was not retrained for these experiments.

The stacked hybrid DNN-HMM is trained on BN and PLP input features from the BULATS data. Global state-position context-dependent (CD) output targets for the DNN are taken from a PLP HMM graphemic model. Decision tree clustering [22] is used in the HMM training to tie the CD states to the desired number. Questions are asked in the decision trees relating to the grapheme identity, the presence of an attribute, the word boundary position and the graphemic attributes, of the grapheme in the centre, and directly to the left and right.

2.2. Use of Graphemic Lexicon in Assessment

Automatic assessment is done using the Gaussian Process (GP) grader proposed in [6], following the architecture in Figure 1. Each section of the BULATS test is graded between 0 and 6; giving an overall grade between 0 and 30. A GP is trained to map the input features to scores, and then used to predict a distribution over the grade, in the form of a mean and variance.

As shown in Figure 1, some grader input features are derived from the ASR hypothesis time aligned to the audio. The use of a graphemic lexicon will affect all of these features, however, since the grader is retrained for each system to ensure robustness to recognition errors the effect will be small in most cases. In the standard grader feature set (Section 3.2) a few features are directly related to phone level measurements, such as mean and median duration of phones. For the graphemic lexicon case these are mapped directly from phones to graphemes. Speaking rate is approximated by vowel frequency. For the grapheme lexicon case the vowels are considered to be the graphemes /a,e,i,o,u/.

As a candidate progresses up the CEFR levels, their pronunciation becomes more native, with commensurate reduction in strain to the listener caused by L1 effects [23]. Explicit features to represent pronunciation in the grader should therefore help assessment, however, there are a number of difficulties associated with extracting pronunciation features from 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, an approach based on distances between phones is presented in this paper. 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. For the graphemic system, the phones are replaced by graphemes.

Distances between acoustic models should be more robust to speaker variability than the models themselves. In [24] 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 cover-

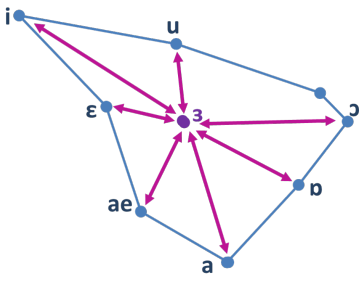

Figure 3: *Illustration of the phone distance concept*

ing all 47 phones (26 graphemes) in English and are applied to both read and spontaneous speech. This yields 1081 distances in total (326 graphemic distances). A set of statistical models is trained to represent the manner of pronunciation of each phone 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 [25] instead of Bhattacharyya distance in [24]. Suppose the statistical models for phones $\boldsymbol{\phi}_i$ and $\boldsymbol{\phi}_j$ are $p(\boldsymbol{\phi}_i)$ and $p(\boldsymbol{\phi}_j)$, respectively, the K-L divergence between the two phones is defined as

$$\mathrm{D}_{\mathrm{KL}}(p_i||p_j) = \int p(\boldsymbol{\phi}_i) \log\left(\frac{p(\boldsymbol{\phi}_i)}{p(\boldsymbol{\phi}_j)}\right) d\boldsymbol{\phi}_i. \qquad (1)$$

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 [26]) is used, which can be written as

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

Each phone is modelled 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 $\mathrm{D}_{\mathrm{JS}}(p_i||p_j)$ is calculated as

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

where $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ are the operators for the trace and determinant of the matrix, respectively. Strong negative correlations with grade are observed and a high K-L divergence seen to correlate with lower grades/scores.

3. Experimental Setup

3.1. ASR

The ASR system is trained on a 108 hour (1075 speaker) Gujarati L1 BULATS data set with merged crowd-sourced transcriptions [7], using the HTK toolkit [27, 28]. The BN DNN's structure is 720x1000^{4}x39x1000x6000. Its input consists of 9 consecutive frames of 40-D filterbank features with delta appended to each frame feature. This yields a input vector size of 720. After training on the AMI corpus [21], the 39-D BN features are extracted for BULATS data and transformed using a global semi-tied covariance matrix [29]. The transformed BN features are appended to HLDA [29] 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^{5}x6000. 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 [30].

Training of both DNNs is performed as follows. The DNN is initialised using layer-by-layer discriminative pre-training with context-independent states as targets. Fine tuning is done using the cross entropy criterion against global state-position context-dependent [31] output targets. Depending on the lexicon, the hybrid DNN targets will be phonetic or graphemic states. Finally MPE-based sequence training [32] is applied.

To assess ASR the grader evaluation sets (described in the next section) have only merged crowd-sourced transcriptions available for the spontaneous sections C-E (Section 2.1). To validate that these are sufficient to score against for ASR a test was run on a held-out development set which has both professional manual and crowd-sourced transcriptions. The difference between the two transcriptions was fairly small (1.5-2.7%) and the relative performance of the phonetic and graphemic systems unchanged illustrating that it is valid to assess against crowd-sourced transcriptions. There is a 1.5% increase in WER scoring over sections C-E vs all 5 grading sections.

3.2. Grader

Each section of the BULATS test is graded between 0 and 6; giving an overall grade between 0 and 30. In this work, the audio from all sections is used to predict the overall score. The GP grader [6] is trained on independent training data. Two BULATS grader train and evaluation set pairs are considered: matched (to ASR training) Gujarati L1, *Gujarati*; and mis-matched 6 L1 (Arabic, Dutch, French, Polish, Thai, Vietnamese), *Mixed*. Both consist of around 1000 training speakers and 225 evaluation speakers. The sets are designed so that the number of speakers per grade (and L1 for *Mixed*) is roughly equal, with the grades C1-C2 merged due to lack of data. Scores given by the original human examiner are used in training the GP grader. For assessment, the predicted scores are scored against scores provided by expert graders from Cambridge English. The Pearson correlation coefficient (PCC) is computed between the grader scores and the expert scores. These experts correlate at the 0.95-0.97 level. The examiner graders correlate with these experts at the 0.875 level. Crowd-sourced transcriptions for the spontaneous sections C-E (Section 2.1) are used to assess ASR WER.

The standard grader feature set [6] is based on a speaker's audio, fluency and basic text characteristics, similar to e.g. [33, 34, 2, 35]. A few features such as the mean energy are derived directly from the audio. Other features are derived from the ASR hypothesis time aligned to the audio. As noted in Section 2.2, for the graphemic lexicon systems, features based on phone measurements are replaced with grapheme measurements. The graphemes /a,e,i,o,u/ are used for estimating vowel frequency (speaking rate).

4. Experimental Results

From Table 1 it can be seen that the graphemic system consistently achieves a lower WER than the phonetic system of 1-1.4% absolute. The mismatch between the Mixed test set and the Gujarati L1 ASR training data can be seen in the large increase in WER from the Gujarati test set. The graphemic system produces a slightly bigger improvement (0.2-0.4%) in this mismatched case. The improvements observed are probably due to a combination of: variation from the native English pronunciations provided in the phonetic lexicon; the presence of a reasonably high number of proper nouns for which poor G2P pronunciations are produced; and odd pronunciations in the G2P relating to crowd-sourced mis-spellings which may be better modelled with graphemic lexicons. As in [14], system combination yields a further reduction in WER showing that the phonetic and graphemic systems are complementary. They could, therefore, be combined in a joint decoding system [36] to produce a lower WER without an increase in decoding time.

Test Set	Ph/Gr	tg	tg+CN
Gujarati	Ph	34.3	33.7
	Gr	33.3	32.5
	Ph $\oplus$ Gr	-	31.6
Mixed	Ph	48.6	47.5
	Gr	47.2	46.1
	Ph $\oplus$ Gr	-	44.2

Table 1: *Phonetic (Ph) and graphemic (Gr) trigram %WER.*

The pronunciation distance features depend on the accuracy of the phone or graphemic ASR output. To assess this, the ASR word hypotheses were converted into phone and grapheme sequences using the corresponding lexicons and then scored against the reference sequences. The results are shown in Table 2. It can be seen that the grapheme error rate (GER) is lower than the phone error rate (PER). Unexpectedly phone sequences derived from the lower WER graphemic word hypotheses saw an increase in PER.

Decoder (word)	Gujarati		Mixed	
	%PER	%GER	%PER	%GER
Ph	25.8	—	33.9	—
Gr	29.0	23.7	36.6	30.8

Table 2: *Phone/grapheme error rates (%PER/GER) for phonetic (Ph) and graphemic (Gr) lexicon systems.*

Using the phonetic systems for decoding and grader feature extraction (Ph-Ph), the pronunciation distance feature (Pron) grader is seen to perform almost as well as the standard grader for the Gujarati test in Table 3. Combining the two feature sets gives a further gain in PCC. Conversely for the Mixed L1 test set, the pronunciation distance features perform less well. This is because they are being averaged across a range of L1s and the pronunciation variations observed at different proficiency levels are both speaker and L1 dependent.

Performance using the graphemic lexicon system for decoding and grader feature extraction (Gr-Gr) shows similar performance to the Ph-Ph grader for the standard features (Table 3). A degradation, however, is observed for the Pron grader on both test sets. Combining with the standard features has little effect on the standard system performance. The pronunciation variation modelling which the ASR system can implicitly model with the graphemic set, is more complicated to handle in the grader as pronunciation distance features.

Test Set	ASR	Grd	Grader PCC		
			Std	Pron	Std+Pron
Gujarati	Ph	Ph	0.843	0.838	0.872
	Gr	Gr	0.832	0.771	0.849
	Gr	Ph	0.841	0.804	0.857
Mixed	Ph	Ph	0.852	0.806	0.852
	Gr	Gr	0.859	0.734	0.853
	Gr	Ph	0.863	0.804	0.870

Table 3: *Comparison of BULATS grader performance for phonetic and graphemic systems for standard (Std), pronunciation distance (Pron) and combined grader input features (Std+Pron).*

Using the graphemic lexicon system for decoding with phonetic feature extraction (Gr-Ph) the standard grader performance is similar to the Ph-Ph grader. For Gujarati a drop in the Pron grader PCC is seen, although less than for the Gr-Gr system. This is probably due to the increase in PER. The combined system performs close to the Ph-Ph system. For the Mixed test set the Gr-Ph Pron grader performs as well as the Ph-Ph grader and the overall combined performance is slightly higher.

5. Conclusions

This paper has presented the use of graphemic lexicons for automatic assessment of English spoken by non-native learners. The work is motivated by the limitations on tasks resulting from the current level of automatic speech recognition (ASR) accuracy for these speakers. Although reasonable assessment can be made of a speaker's proficiency, automatic systems are currently unable to provide anything other than basic feedback on performance in realistic communication settings with spontaneous speech. Non-native English speech shares many of the characteristics of limited resource languages - i.e. limited training data to cover a wide variety of speech due to speech variations caused by e.g. L1 and proficiency level, and by recording conditions - where graphemic lexicons have proven consistently better for ASR.

Unlike previous native English experiments, the graphemic lexicon was observed to improve the accuracy of the non-native English ASR systems on both matched and mis-matched L1 test sets. This comes with the advantage that no G2P system is required. For a standard grader feature set the graphemic lexicon system works as well as the phonetic lexicon system. Pronunciation distance features were applied to the grading of spontaneous speech for the first time. These features may be more important for feedback. Features extracted directly from a graphemic system were unable to discriminate as well as phonetic features, but deriving phonetic features from the graphemic decode reproduced the phonetic system grader performance. Since English is one of the hardest languages for graphemic lexicons, this suggests that this lower resource approach would also be of use for assessment of other languages. Expansion of the grader feature sets to take advantage of better WER should be investigated.

6. 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.

7. References

[1] B. Council, "The English Effect," Aug 2013, Research Report.

[2] 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.

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

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

[5] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Communication*, vol. 73, pp. 14–27, 2015.

[6] 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.

[7] 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.

[8] "CMUdict." [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[9] K. Richmond, R. A. J. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proceedings of INTERSPEECH*, 2010, pp. 1974–1977.

[10] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, pp. 434–451, May 2008.

[11] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[12] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 845–848.

[13] M. Killer, S. Stuker, and T. Schultz, "Grapheme Based Speech Recognition," in *Proceedings of INTERSPEECH*, 2003, pp. 3141–3144.

[14] C. Breslin, M. Stuttle, and K. Knill, "Compression Techniques Applied to Multiple Speech Recognition Systems," in *Proceedings of INTERSPEECH*, 2009, pp. 1407–1410.

[15] S. Stüker and T. Schultz, "A Grapheme Based Speech Recognition System for Russian," in *Proceedings of the International Conference on Speech and Computer (SPECOM)*, 2004.

[16] W. Basson and M. Davel, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans," in *Proceedings of Pattern Recognition Association of South Africa (PRASA)*, 2012.

[17] S.Sakti and S.Nakamura, "Recent progress in developing grapheme-based speech recognition for Indonesian ethnic languages: Javanese, Sundanese, Balinese and Bataks," in *Proceedings of International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014.

[18] M. Gales, K. Knill, and A. Ragni, "Unicode-based Graphemic Systems for Limited Resources Languages," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[19] F. Diehl, M. J. F. Gales, X. Liu, M. Tomalin, and P. C. Woodland, "Word Boundary Modelling and Full Covariance Gaussians for Arabic Speech-to-Text Systems," in *Proceedings of INTERSPEECH*, 2011, pp. 777–780.

[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] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 28–39.

[22] S. J. Young and P. C. Woodland, "State clustering in HMM-based continuous speech recognition," *Computer Speech and Language*, vol. 8, no. 4, pp. 369–394, 1994.

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

[24] 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.

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

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

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The* HTK *book (for* HTK *version 3.4.1)*. University of Cambridge, 2009. [Online]. Available: http://htk.eng.cam.ac.uk

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

[29] 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.

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

[31] K.Knill, M.Gales, A.Ragni, and S.Rath, "Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting," in *Proceedings of INTERSPEECH*, 2014.

[32] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative Training of Deep Neural Networks," in *Proceedings of INTERSPEECH*, 2013, pp. 2345–2349.

[33] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1997, pp. 622–629.

[34] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI Eduspeak system: Recognition and pronunciation scoring for language learning," in *Proceedings of InSTILL 2000*, 2000, pp. 123–128.

[35] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[36] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Proceedings of INTERSPEECH*, 2015, pp. 3660–3664.