Interspeech 2018
2-6 September 2018, Hyderabad

Improved Acoustic Modelling For Automatic Literacy Assessment Of Children

Mauro Nicolao[1], *Michiel Sanders*[2], *and Thomas Hain*[1]

[1]Speech and Hearing Research Group, The University of Sheffield, UK
[2]ITSLanguage BV, The Netherlands

m.nicolao@sheffield.ac.uk, michiel.sanders@itslanguage.nl, t.hain@sheffield.ac.uk

Abstract

Automatic literacy assessment of children is a complex task that normally requires carefully annotated data. This paper focuses on a system for the assessment of reading skills, aiming to detection of a range of fluency and pronunciation errors. Naturally, reading is a prompted task, and thereby the acquisition of training data for acoustic modelling should be straightforward. However, given the prominence of errors in the training set and the importance of labelling them in the transcription, a lightly supervised approach to acoustic modelling has better chances of success. A method based on weighted finite state transducers is proposed, to model specific prompt corrections, such as repetitions, substitutions, and deletions, as observed in real recordings. Iterative cycles of lightly-supervised training are performed in which decoding improves the transcriptions and the derived models. Improvements are due to increasing accuracy in phone-to-sound alignment and in the training data selection. The effectiveness of the proposed methods for relabelling and acoustic modelling is assessed through experiemnts on the CHOREC corpus, in terms of sequence error rate and alignment accuracy. Improvements over the baseline of up to 60% and 23.3% respectively are observed.

1. Introduction

Speech technology advances in recent years have allowed automatic assessment tools to permeate education methodologies. Interactive computer assisted language learning (CALL) tools incorporate a variety of approaches [1], such as spoken word assessment [2], pronunciation assessment [3], and literacy assessment [4]. In particular, literacy assessment may involve a wide range of language-related skills, such as decoding words, fluently reading sentences aloud, reading comprehension, and writing [4].

The use of speech technology in reading assessment has been extensively investigated for almost three decades. Most of the studies have focused on children who read in their native language [5, 6, 7], or on adults that learn a second language [8]. Assessing reading skills is a particularly challenging task, especially with young children (6 to 12 years of age). Automatic reading assessment tools are crucial in primary education because they can compensate for different learning rates, and can provide personalised exercises and auxiliary support, when necessary.

This paper focuses on developing a system to assess children reading skills by detecting a range of typical fluency and pronunciation errors.

Automatic reading assessment for children is a complex task that relies on speech recognition methodologies. Thus it requires carefully transcribed data for training of children specific acoustic models. Accurate error-labelled annotation is also essential for developing and testing the error classifiers and prediction models that are required for this task. It is often very difficult to gather such high-quality material to train in-domain models due to the high labelling cost. Many corpora of children read speech provide only the prompted text and an overall speaker assessment score. The PF_STAR corpus [9] and the TBALL corpus [4] are examples of word level annotated data sets in English. Two ad-hoc corpora are available for the Dutch language: JASMIN-CGN [10] and CHOREC [11]. Both provide careful manual annotation of words and reading errors. Even though these corpora are very useful for research, the diversity of the speech material is often limited. Limitations are for example minimal vocabulary or use of specific microphones or recording conditions. Hence models derived cannot be easily transferred to different conditions. If the training material needs to be extended, the effort in providing the required level of accuracy is often overwhelming. Manually transcribed data such as children's read speech recorded in real environment is a very expensive and requires great deal of time and expertise. An alternative method to manual annotation is to automatically enhance approximate transcriptions of unseen data, allowing for iterative expansion of training sets.

Reading assessment is a somewhat unusual task as the spoken words should be identical to the original text prompted to the learner. However the realisiation can be regarded as a specifically constrained variation of the original text.

The correction of audio transcription is common problem in training statistical acoustic models with real audio recordings, and a *lightly supervised* approach to acoustic modelling as outlined in [12] vis often adopted. This training method is based on the opportunity of automatically improving the accuracy of speech transcriptions using available prior knowledge. The recovered transcript can originate from an inaccurate annotation, an extended summary of the speech content, or a prompted script.

Compensating for inaccurate annotations has been extensively researched in the domain of broadcast news to correct recognition errors of automatic speech transcriptions [13, 14, 15]. Weighted finite state transducers (WFST) are the most often adopted models to detect the variations from a given script. A WFST-based approach to improve the automatic alignment is for example proposed in [16]. In order to detect reading and pronunciation errors, transducers are used in [17].

In this paper, a flexible WFST-based language model is adopted to improve not only the recognition results in presence of a pre-trained model, but also the model training itself by providing a more accurate word-level alignment and segmentation.

2. Lightly-supervised training

The lightly-supervised training regime is designed to compensate for the mismatch between the spoken words and the pro-

10.21437/Interspeech.2018-2118

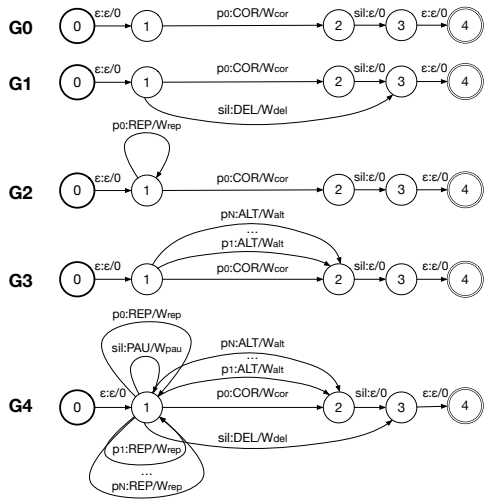

Figure 1: *WFST topologies modelling typical reading events: deletions (G1), repetitions (G2), and substitutions (G3). The no-error (G0) and the combined (G4) grammars are also displayed.*

vided transcriptions. Since for reading assessment the speaker is requested to only say the words on a prompted script, the assumption is that the discrepancies between the expected word sequence and the speech outcome cannot be extensive. The possible variations are therefore quite limited, and the most frequent differences can be modelled by small controlled changes of the word sequence. A WFST is introduced to describe these variations. This transducer implements the language model that drives the automatic speech recognition (ASR) decoding and provides the transcription in lightly-supervised training. Iterative cycles of the training regime are performed. At each iteration, the WFST-based decoding improves the transcriptions and by expectaction that also improves the derived acoustic models.

2.1. Typical reading error modelling

A WFST is a flexible structure which models word sequences as transitions from a series of nodes. Each transition is triggered by an input symbols, is associated with a cost, and may generate output symbols. The recognition WFST is a composition of the following four elements:

$$\mathcal{D} = \mathcal{H} \circ \mathcal{C} \circ \mathcal{L} \circ \mathcal{G} \tag{1}$$

where $\mathcal{H}$ represents the statistical description of context-dependent phoneme features, $\mathcal{C}$ is a transducer mapping context-dependent phonemes to monophones, $\mathcal{L}$ links monophones and words (lexicon), and $\mathcal{G}$ (or grammar) models the sequence of words. $\mathcal{H}$ and $\mathcal{C}$ mainly derive from the acoustic model training, $\mathcal{L}$ is defined by the pronunciation dictionary. The grammar $\mathcal{G}$ is the component that is crafted to model the common reading behaviours, such as deletion, repetition, and mispronunciations. Figure 1 illustrates the grammar topologies representing the typical reading events at word level. The G0 transducer models a word as it appears in the prompted text. G1 introduces word deletions superimposing a silence transition. G2 implements repetitions with word-level loops. G3 allows for multiple parallel transitions that model alternative word realisations, such as mispronunciations, false start, and word-spelling. G4 combines the above grammars to allow for recovering the greatest possible amount of mismatching annotation. The input symbols sil and p_i, $i \in 0, \ldots, N$, on the arcs accept the recognition engine output. The output symbols consist of the labels (COR, DEL, REP, PAU, ALT) which correspond to the recognised events (correct word, deletion, repetition/insertion, pause/silence, and substitution respectively) combined with the identifiers of the linked word. The likelihood of these transitions is defined by the costs w_i, and their values are normally learned from data (see § 3.1).

The WFST of a complete reading task can be automatically derived from the prompted text by selecting one transducer of Figure 1 for each word in the text, and concatenating them. This modular structure allows for several layers of error-modelling complexity. For example, single-word restarts are implicitly represented by G4 as a repetition/false start followed by a correct/deletion. The efficacy of these grammars in correcting the original prompted text is investigated in § 4.

2.2. Iterative acoustic model training

Figure 2 depicts the iterative process adopted to improve the lightly supervised acoustic model training. White blocks represent the steps required by iterative training with both supervised and lightly-supervised transcripts. These consist of two parts: the bootstrap and the optimisation loops. The audio as segmented with the original transcriptions is the input to the maximum-likelihood (ML) training of a generative model (a hidden Markov model with Gaussian mixtures, HMM-GMM). At each iteration, the new model is used to produce new transcriptions. The segmentation step also includes data filtering. The audio fragments that obtain likelihoods lower than the overall corpus average are discarded.

The green blocks in Figure 2 are related to the WFST-based *ASR decoding*. Depending on the input prompt and the degree of allowed variation, a dedicated grammar for each type of prompt and selected error category can be created by the *WFST grammar generator*.

The proposed iterative training is tested with two types of features: perceptual linear prediction (PLP) features and feed-forward deep neural network (DNN) bottle-neck (BN) features. The PLP-based model training (PLP-HMM) uses the prompt text and an out-of-domain (OOD) acoustic model at the bootstrap stage (red block in Figure 2) to generate the first transcriptions. Due to the sensitivity of DNN training to inaccurate segmentation, BN-based bootstrapping (blue blocks) takes advantage of the segmentation derived from previously-trained in-domain PLP-HMM models.

If accurate transcriptions (AT) are available for the corpus, i.e. when all acoustic events (words and errors) are labelled, an *oracle* acoustic model can be trained. These transcriptions along with their time information provide both the most accurate audio segmentation and the most effective filtering of the too-distorted speech segments. The resulting acoustic models (PLP-HMM+AT and BN-HMM+AT) can be addressed as the best possible ones that can be trained on such data. Their performance hence represents the upper limits towards which the proposed iterative regimes should converge.

2.3. The scoring system

The quality of the reading error recovering and of the acoustic modelling is assessed by computing a sequence error rate measure and an alignment accuracy measure.

The sequence error rate is the word error rate (WER) of the ASR output against an accurate manual transcription with

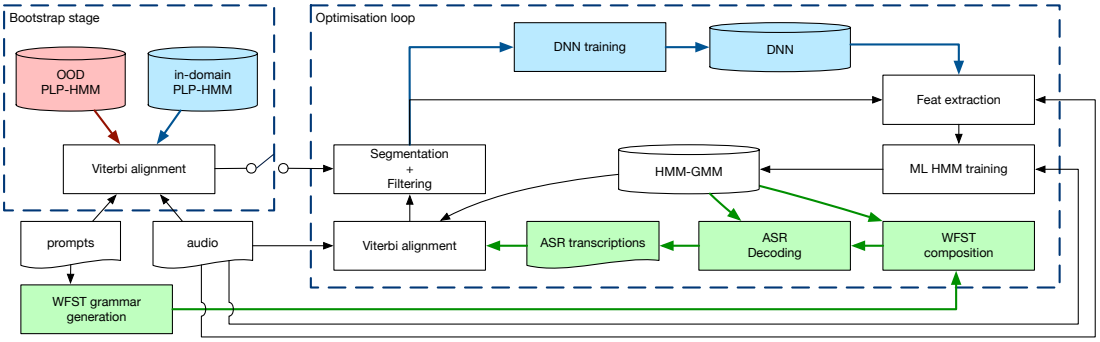

Figure 2: *The complete iterative lightly-supervised training process. Different colours identify the paths and blocks belonging to different sections of the process. The PLP-HMM and BN-HMM trainings are red and blue respectively. Green identifies the WFST creation and the ASR decoding. The bootstrap stage is also represented.*

all the reading errors. This score measures the quality of the grammar at predicting the reading errors in the audio.

The alignment accuracy score of the ASR output is performed with the method used in [18]. A precision/recall measure is calculated with respect to a manual transcription with accurate timing. A word is considered to be a match if both start and end times fall within a 100ms window of the associated reference word. The fragments that are filtered out during the segmentation stage of the iterative training are excluded from scoring.

3. The CHOREC corpus

The training and recognition process is evaluated on the CHOREC (CHildren's Oral REading Corpus) [11, 19], a database of recorded, transcribed and manually annotated children's oral readings. The corpus consists of recordings from 400 Dutch speaking children if 6 to 12 years of age. The children were asked to complete several reading tasks. 130 hours of audio are carefully annotated at several levels of descriptive details, among which the most interesting are: 1. the orthographic transcription tier (PMT) with the text prompted to the reader; 2. the accurate transcription tier (AT) with the automatically aligned complete description of what is in the audio.

Three reading tasks providing the largest sets of recordings are considered here: isolated words, LG ($\sim$ 28h), nonsense pseudo-words, LGP ($\sim$ 37h), and long paragraphs, AVI ($\sim$ 36h). The available material is split into training and test sets. A speaker does not appear in both sets and a fair distribution ($1/3$ and $2/3$) of sentences without/with errors in the test set is ensured. Table 1 shows the principal statistics for these.

Table 1: *Characteristics of training and test set.*

Name	Purpose	Files	Segments	Included tasks
chotrain.1	training	2445	$\sim$ 60000	AVI, LGP, LG
chotest.1	test	415	$\sim$ 15000	AVI, LGP, LG

The PMT transcriptions are used at the training bootstrap stage and as a input for WFST creation. The text and the timing information in the AT transcription constitute the reference against which the WFST and the acoustic training regime are scored. AT is also used as transcription to train an oracle system. As the timings of the manual transcription was obtained from forced alignment, the CHOREC word-level time information may not entirely accurate. Though, manual inspection confirmed that the alignment mismatches are minimal and probably can be ignored.

Text normalisation is conducted on the original text to transform it into a scoring-compatible format. The not-prompt related labels, such as background noise or external speaker speech, are discarded. When possible, error labels are linked to the prompt words that are related to them by explicitly duplicating the word labels in the final reference (ATR).

3.1. The CHOREC error label distribution

The best method to derive the weights which define the WFST transducer of § 2.1 is to directly learn them from children real behaviour. For this reason, the labels in the AT transcription are scrutinised and the overall distribution of the error labels is displayed in Figure 3. The LG, LGP, and AVI reading tasks are plotted in separate bar charts. For simplicity, the original

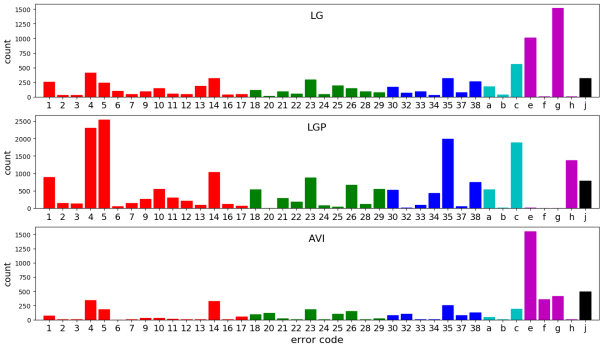

Figure 3: *Error type distribution in the CHOREC corpus annotation. Colours are used to group errors of similar nature.*

47 error codes, described in the CHOREC annotation protocol manual [11], are grouped into 6 main categories: 1. substitution errors (in red colour) which label phone-level error; 2. deletion errors (in green), which identify words or phones are missing in the audio; 3. insertion errors (in blue), which show words with extra phone insertion; 4. decoding errors (in cyan), such as letter-by-letter or syllable-by-syllable spelling; 5. word substitution errors (in purple); 6. unidentified events (in black). A dependency linking different types of read material and error categories can be easily extrapolated from Figure 3. In the LGP task, for example, phone substitutions are the most common er-

rors. On the other hand, these errors are generally less frequent in the AVI task, because, in such task, prior knowledge helps the reader predicting word sequences and their realisation.

The event occurrences for each of these categories are used to compute the log prior values that are used as weights $w_{\hat{E}}$ in the WFST, according to function expressed in:

$$w_{\hat{E}} = -\log\left(\frac{|\hat{E}|}{\sum_{E\in\mathcal{C}}|E|}\right) \quad (2)$$

where $\hat{E}$ is a set of specific events (correct, deletion, repetition, substitution, etc.) observed in the transcription, $|.|$ is the cardinality operator, and $\mathcal{C}$ is set of annotation from the entire corpus.

The log-prior values extracted from the corpus are reported in Table 2. Phone substitution and decoding, along with unknown events are not modelled in the transducers as these are very unlikely in the most realistic reading tasks (LG and AVI).

Table 2: *Log-prior of the error categories in CHOREC. Links between CHOREC events and WFST weights are highlighted.*

CHOREC event	WFST weight	LG	LGP	AVI
correct	w_{COR}	0.0900	0.3836	0.1940
phone substitution	–	5.6511	2.0449	6.5725
deletions	w_{DEL}	5.3453	3.1964	5.7884
insertions	w_{REP}	6.8397	3.1307	6.5260
decoding	w_{ALT}	4.0403	2.8448	4.1364
word substitution	–	2.8324	3.3991	1.8699
unknown	–	8.1834	4.3394	8.0076

4. Experiments

The experimental implementation of the system described in § 2.2 uses the HTK toolkit [20] to segment the audio and extract spectral audio features, the Juicer recogniser [21] to perform the WFST-based decoding, and the OpenFST library to automatically compose the WFST for the typical reading error modelling. Initial bootstrap out-of-domain models are trained on the children speech data from the JASMIN-CGN corpus.

The experiments conducted on the CHOREC corpus test the different error modelling configuration (G0, . . ., G4) of Figure 1. Different acoustic models are computed, and the derived automatic transcriptions are scored against the ATR reference, according the measures of § 2.3. The PLP-HMM+G0 and BN-HMM+G0 systems provide the baseline results against which all the other trainings are compared. The OOD PLP-HMM+PMT bootstrap results are also reported in the Figures to emphasise the mismatch between OOD and in-domain acoustic model performance. The sequence error rate results are reported as the WER of training and test sets against the ATR reference in Figure 4. This score measures the effectiveness

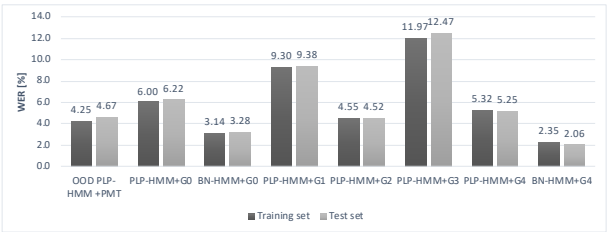

Figure 4: *WER on training and test sets w.r.t different error models.*

of the error prediction models in compensating for missing or repeated words. As expected, the combined grammar G4 is the WFST configuration that provides the lower WER. PLP-HMM+G4 and BN-HMM+G4 respectively obtain 11.3% and 60.8% relative WER reduction w.r.t the PLP baseline on the training set. It appears that repetition prediction grammar, G2, alone is responsible of most of the performance improvement in the error prediction (24.2% relative). The filtering step associated to the segmentation stage is also influencing the quality of the acoustic modelling. The training set size considerably varies depending on the quality of the segmentation. E.g., the OOD PLP-HMM+PMT system, even though it has a low WER, discards large portions of usable speech (size $\sim$ 13h) whilst BN-HMM+G4 manages to recover up to 20h of data. The alignment accuracy scores for each grammar are plotted in Figure 5. Precision and recall measures are combined in a *F-measure* value for clarity. These scores assess the accuracy of the model in positioning the recovered speech events (correct words and errors) in the audio. Along with the baseline results, the oracle system scores (PLP-HMM+AT and BN-HMM+AT) are computed. These define the upper boundary for lightly supervised training in which all transcription errors are completely recovered. The system using the G4 grammar also produces the best alignment accuracy scores. PLP-HMM+G4 and BN-HMM+G4 achieve 14.3% and 23.3% relative improvement respectively w.r.t. the PLP baseline on the training set. It is worth to notice that these scores are only few percentage points lower than the oracle results, 10.0% and 16.2% relative reduction, respectively.

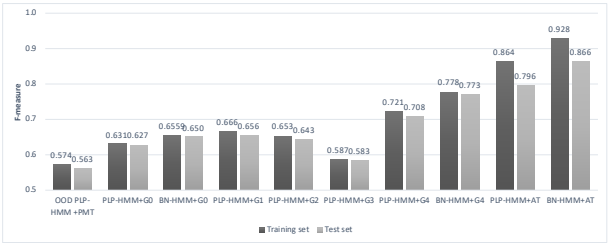

Figure 5: *Alignment score on training and test sets w.r.t. different error models.*

5. Conclusions

An iterative lightly supervised training regime was proposed to obtain acoustic models for children automatic reading assessment. A WFST was employed to model the typical reading errors observed in the CHOREC children recordings. A constrained recognition stage can provide transcriptions that recover most discrepancies with the original prompted text. Experiments conducted on the CHOREC corpus show that this training regime successfully improves the quality of the segmentation and labels, and hence of the derived acoustic models. Repetition error recovery is most important. Best results are obtained with a model that takes repetitions, substitutions, and deletions into account. Compared to a PLP HMM-GMM baseline WER is reduced by 11.3%, and by 60.1% with BN-HMM models. The process also improves alignment accuracy score by 14.3% and 23.3%, respectively.

6. Acknowledgements

This research has been funded by ITSLanguage BV `http://www.itslanguage.nl`.

7. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, Oct. 2009.

[2] R. C. van Dalen, K. M. Knill, and M. J. Gales, "Automatically Grading Learners' English Using a Gaussian Process ," in *SLATE 2015*, ISCA, Ed. Leipzig: ISCA, Sep. 2015.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000.

[4] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment," *IEEE Trans. Speech, Audio & Language Processing*, vol. 19, no. 4, pp. 1015–1028, Mar. 2011.

[5] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A Prototype Reading Coach that Listens," in *AAAI-94*, Aug. 1994, pp. 785–792.

[6] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*. IEEE, 2003, pp. 186–191.

[7] P. Cosi, R. Delmonte, S. Biscetti, R. A. Cole, B. L. Pellom, and S. van Vuren, "Italian literacy tutor-tools and technologies for individuals with cognitive disabilities," in *InSTILICALL Symposium*, ISCA, Ed., 2004.

[8] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology." *JASA*, vol. 107, no. 2, pp. 989–999, Feb. 2000.

[9] A. Batliner, M. Blomberg, S. D. o. Speech, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *EUROSPEECH 2005*, Lisbon. PT, 2005.

[10] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus," in *LREC 2008*, Mar. 2008, pp. 1–6.

[11] Katholieke Universiteit Leuven, Universiteit Gent, "Children's Oral Reading Corpus (CHOREC)," TST-Centrale.

[12] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *INTERSPEECH 2004*. IEEE, 2004, pp. I–737–40 vol.1.

[13] W. K. Seong, J. H. Park, and H. K. Kim, "Dysarthric Speech Recognition Error Correction Using Weighted Finite State Transducers Based on Context–Dependent Pronunciation Variation," in *Computers Helping People with Special Needs*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, Jul. 2012, pp. 475–482.

[14] Y. Long, M. J. F. Gales, P. Lanchantin, X. Liu, M. Seigel, and P. C. Woodland, "Improving Lightly Supervised Training for Broadcast Transcription," in *INTERSPEECH 2013*, 2013.

[15] J. Olcoz, O. Saz, and T. Hain, "Error Correction in Lightly Supervised Alignment of Broadcast Subtitles," in *INTERSPEECH 2016*. ISCA, Sep. 2016, pp. 2110–2114.

[16] P. Bell and S. Renals, "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *ASRU 2015*. IEEE, 2015, pp. 675–680.

[17] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquière, W. Verhelst, and H. V. hamme, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, Oct. 2009.

[18] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *ASRU 2015*, IEEE, Ed. Scottsdale, AZ: IEEE, 2015, pp. 687–693.

[19] L. Cleuren, J. Duchateau, P. Ghesquière, and H. van Hamme, "Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement," in *LREC 2008*, Mar. 2008, pp. 1–8.

[20] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *ICASSP 1994*. IEEE, 1994, pp. II/125–II/128 vol.2.

[21] D. Moore, J. Dines, M. M. Doss, J. Vepa, O. Cheng, and T. Hain, "Juicer: A weighted finite-state transducer speech decoder," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland. Springer, Berlin, Heidelberg, Dec. 2006, pp. 285–296.