Aligning Audio Samples from the South African Parliament with Hansard Transcriptions

Neil Kleynhans and Febe de Wet
Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
Email: {ntkleynhans,fdwet}@csir.co.za

***Abstract*—Most of the developing world can still be classified as under-resourced in terms of the resources that are available in their languages. Harvesting suitable and relatively easily accessible spoken resources can drastically improve the situation. One such resource is parliamentary sessions, which in general are publicly available and are most often manually transcribed.**

In this investigation we present an automatic harvesting procedure which makes use of the "islands of certainty" principle to segment long utterances into more manageable shorter chunks and a garbage model to improve alignment by absorbing superfluous speech. The final harvesting approach was used to harvest 50 hours of South African Parliament audio data from a total 105 hours of raw audio data, at a goodness-of-pronunciation (GOP) score of 1.9. The word alignment accuracy, performed on two parliamentary sessions, showed that over 96% of the words are within 1.0 seconds of their true position in the audio stream.

I. Introduction

The majority of the official languages of South Africa can still be classified as being under-resourced. The situation has been alleviated to some extent by the successful completion of various resource development projects sponsored by the Department of Arts and Culture of the South African government. One such project was the National Centre for Human Language Technology (NCHLT) Speech project which produced broadband speech corpora for all official languages [1]. Each corpus contains between 50 and 60 hours of transcribed speech in all eleven official South African languages.

The current project aims to extend the resources created during the NCHLT Speech project by aligning audio data from National Parliament (NP) with corresponding Hansard transcriptions, thereby creating a multilingual corpus of transcribed speech data. Although the primary aim of the project is to extend existing speech resources by aligning parliamentary speech data with Hansard transcriptions, attention will also be paid to the possibility of enhancing the transcription process with automatic speech recognition ASR technology.

II. Background

The TC-STAR [2] project has a primary goal of improving all core technologies used in speech-to-speech translation systems – automatic speech recognition (ASR), spoken language translation (SLT) and text-to-speech (TTS). The domain of development is unconstrained speech, selected from speeches and broadcast news and focusses on three languages: European English, European Spanish, and Mandarin. The European Parliament Plenary Sessions (EPPS) were identified as an initial reference task for speech translation. The aim was to perform English to Spanish translation and vice versa. In the TC-STAR ASR evaluation performed by Lamel *et. al.* [3] over 90 hours of audio recordings from the parliamentary sessions were used to train the ASR systems. The word-error-rates (WERs) were 8.2% for English and 10.7% for Spanish on the evaluation set. It is unclear which text representations of the sessions were used during the acoustic model development, since for each session a final text edition (FTE) is generated which strives to achieve high readability and does not serve as a direct verbatim transcription. In addition, members are allowed to make edits to the transcriptions.

Kawahara [4] describes the use of automatic transcription generation in the Japanese Parliament to assist transcribers. To achieve this, an ASR system is used which has to produce an accuracy greater than 90%. As highlighted, this goal is difficult to achieve in committee meeting contexts as the speech is spontaneous, interactive and emotive. There is an abundance of data to train the acoustic models – 1200 hours per year – but the accompanying transcriptions are approximate due to the parliamentary editing processes. To overcome this problem, a statistical machine translation (SMT) approach is used to convert the approximate transcriptions into faithful transcriptions. Then, a lightly-supervised acoustic model training scheme is used to develop in-domain acoustic models. A 2011 evaluation showed that the system achieved an error rate of 10%.

The following parliaments make use of the ASR technology to varying degrees: Australian NP [5], Australia: Western Australia [6], Australia: New South Wales [7], Denmark [8] and Isle of Man [9]. The main aim of the using ASR technology is to reduce the transcription effort.

In a previous study, an aligned corpus of South African English was created using South African English radio broadcast audio data and associated transcriptions harvested from the internet [10]. Davel *et. al.* [10] showed that lightly-supervised automatic harvesting for ASR resource creation in a resource-scarce environment does not require well-trained language models. In their approach, a phone-based ASR system was used to automatically generate transcriptions, for roughly 100 hours of South African English radio broadcast audio data, using a flat-phone task grammar. The seed models were initially developed on US English data and gradually replaced by the in-domain SAE dialect. Data filtering was achieved by using a garbage model which absorbed badly-aligned audio portions. The data is in a different domain, but a similar

approach can be used to align the speech data from Parliament with the Hansard transcriptions. Use of the garbage model that was introduced to absorb superfluous speech and replacing the seed acoustic models with ones trained or adapted on in-domain data were especially useful.

Practically, aligning long audio segments is challenging but this can be overcome by using the iterative procedure proposed by Moreno *et. al.* [11]. Effectively, the task is converted to a recursive speech recognition problem. The first step is to decode the entire utterance using a large-vocabulary ASR system that makes use of a biased language model trained on the utterance transcriptions. A dynamic programming algorithm is then used to align the transcriptions and recognised text. Anchor points are identified by selecting text regions that show agreement between the transcription and recognised text. The utterance and transcription are then chunked based on the anchor points and the process repeated. At each chunking stage the language model is retrained on the text that occurs within the segment. The results showed that after running the automatic alignment procedure, 98.5% of the words were within 0.5 seconds of their true alignments.

From this brief review it is clear that parliamentary sessions can be a rich source of both audio and text language resources. Although the domain has some challenges, such as non-verbatim transcripts and unconstrained spontaneous speech, these can be overcome through a variety of approaches. If the approximate transcriptions are not too far removed from what was actually said, then the machine translation approach can be omitted. Furthermore, given enough raw data, a sufficient amount of data can be harvested. Aspects of the harvesting approaches proposed by Davel *et. al.* [10] and Moreno [11] can be combined to create a suitable approach that can be used to harvest long unconstrained spontaneous speech found in parliamentary sessions.

The audio and text data that was obtained from NP as well as the NCHLT data that was used for system development are described in the next section of the paper. ASR tool development and the procedure that was followed to perform the alignment are described in section IV. The performance of the alignment strategy is reported on in section V followed by a discussion of the results and comments on the alignment procedure in section VI.

III. Resources

This section gives an overview of the speech and text data that were used to develop the alignment tools and to perform the alignment itself. Two sets of speech data were used in this study: 1) English data from the NCHLT Speech corpus that was used to develop seed acoustic models and 2) speech data obtained from Parliament that needed to be aligned with Hansard transcriptions. A number of Hansard texts were provided by NP. Additional text data was downloaded from the Parliament website.

A. Speech data

1) NCHLT speech data: Seed acoustic models, needed to start the harvesting procedure [10], were trained on NCHLT speech data [1]. A modified "English" NCHLT corpus was created by supplementing the English sub-corpus with English prompts sourced from the other 10 languages. The reason for adding accented English was to improve the acoustic coverage of the acoustic models, as the parliamentary data contains speech from many speakers who have accents. Table I shows the number of utterances, hours and speakers found in the modified "English" NCHLT corpus.

TABLE I
Enhanced English NCHLT speech corpus.

# of utterances	87557
# of hours	64.54
# of speakers	1673

2) Speech data from National Parliament: Negotiations to obtain audio recordings of the speech data generated in Parliament were initiated in 2011. Initial investigations revealed that it was not possible to obtain audio only recordings and video material was therefore acquired. The audio was extracted from the video using *transcode*[1], converted to PCM WAVE format, with 16-bit signed integers used to represent the data samples. The audio data was also down-sampled to 16 kHz and the volume was decreased by 3 dB due to audio clipping observed in some instances. All post audio extraction operations were performed using *Sox*[2].

The transcription unit at NP publishes an index of each session that is transcribed. One of the fields in the index file indicates which language was used by the speaker. These overview documents were used to select the most diverse debates from a set that was made available by NP. The aim was to select debates that contained examples of as many different languages as possible. 32 debates from the National Assembly were identified in this manner and the corresponding video material was obtained from NP. Although an attempt was made to select debates in which languages other than English were spoken, an initial analysis of the data revealed that almost all the speech was South African English - the other 10 languages are only used incidentally.

B. Hansard text data

In addition to the 32 Hansard documents corresponding to the selected debates, a further 759 Hansard transcriptions of the National Assembly were downloaded from the Parliament website[3]. The documents were all in MS-WORD DOC format and had to undergo extensive pre-processing and normalisation before the text could be used for language modelling and alignment purposes.

[1] http://www.transcoding.org/
[2] http://sox.sourceforge.net/
[3] http://www.parliament.gov.za/live/content.php?Category_ID=119

C. Ground truth transcriptions

A set of manually corrected reference alignments were needed to verify the quality of the generated alignments. At this preliminary stage, only two parliamentary sessions, namely 1_March_2012 and 11_May_2012, were corrected. To start the manual correction process, both sessions were automatically aligned using the procedure detailed in section IV-B. The lengthy alignments were further segmented into 30 seconds chunks to ease the manual correction effort. The 1_March_2012 session contained 145 alignment chunks, while the 11_May_2012 session contained 142 chunks, which were stored in Master Label File (MLF) HTK-format. Each MLF was converted to a PRAAT [4] TextGrid as the transcribers made use of PRAAT to correct the transcriptions. Transcribers were guided by a thorough protocol that dealt with uncertainties that arose during the correction process.

The manually verified TextGrids were subsequently converted back to MLFs for evaluation. A word list was generated from the MLFs in each of the parliamentary sessions and from the verified MLFs. Table II summarises the results obtained by comparing the manual and reference MLFs.

TABLE II
COMPARISON BETWEEN THE HANSARD AND MANUALLY CORRECTED TRANSCRIPTIONS.

Session	Corr. %	Acc. %
1_March_2012	77.87	59.13
11_May_2012	86.07	68.03

The low accuracy values that are observed in table II show that there is significant mismatch between the transcription sets. This result indicates that many words were omitted during the parliamentary transcription process.

IV. METHOD

Two factors complicated the automatic alignment task, namely (1) the huge discrepancy between the Hansard transcriptions and the content of the audio, and, (2) the duration of the parliamentary sessions. This section describes the harvesting procedure and decisions taken through the course of the investigation to compensate for these complicating factors.

A. NCHLT seed acoustic models

The speech recognition system development followed a similar structure to that described in Kim *et. al.* [12] and made use of HTK [13]. All audio data was converted to perceptual linear prediction (PLP) coefficients. A 52 dimensional PLP feature vector was created by appending the 13 static, 13 delta, 13 delta-delta and 13 delta-delta-delta coefficients. Global mean and variance normalisation was applied to all the features.

Acoustic models (AMs) were developed by following an iterative training scheme where previous models were used to perform alignments during training. Firstly, 32-mixture context-independent (CI) AMs were trained and used to produce state alignments for the CI AMs trained in the initial development of cross-word triphone context-dependent (CD) AMs. Once the CD AMs were trained, the process was repeated and the previous AMs were used to produce all state alignments before increasing the mixture number.

All hidden Markov models (HMMs) were three state left-to-right in structure. Each CD HMM's state contained eight mixture diagonal covariance Gaussian models. A question-based tying scheme was followed to create a tied-state data sharing system [14] – where any context-dependent triphone having the same central context could be tied together.

A last step in the AM development, was to estimate a heteroscedastic linear discriminant analysis (HLDA) transformation, which was applied to the 52-dimensional PLP feature vectors in order to reduce the dimension to 39. After estimating the HLDA transform, the CD AMs' parameters were updated using two iterations. As a large percentage of model variances were floored during re-estimation only the weights and means were updated.

Davel *et. al.* [10] made use of a garbage model in their harvesting procedure, therefore a garbage model was trained and inserted into the acoustic model set. The garbage model was a 64-mixture general model trained on data where the phone labels were all converted to the same value. The garbage model was inserted into the specialised HTK "sp" model and modifications made to the model structure to accommodate the garbage model.

B. Harvesting procedure

ASR tools can be used to generate alignments between spoken audio and text, but practically the audio durations are usually many seconds to a few minutes long. A large portion of the parliamentary sessions are well over an hour long, with some approaching five hours in duration. With current ASR technology, the tools fail to achieve good alignment or any alignment for such audio segments. Thus, the problem had to be broken up – using a divide-and-conquer approach. To this end, work presented by [11] was used as a starting point for the automatic alignment approach, as these authors undertook similar work.

The algorithm breaks up a long audio file into smaller chunks by finding "islands of certainty" between the transcriptions and automatic text outputs generated by an ASR system recognition. At each iteration, the audio is segmented into smaller chunks and the pronunciation dictionary and language model used during the recognition phase are restricted for each chunk. Our algorithm follows a similar approach but due to some of the parliamentary session's audio approaching five hours in duration, such an iterative process would be too time consuming. Thus, for our implementation only a single iteration was used.

Figure 1 shows a high-level flow diagram of the harvesting procedure. The numbers in the figure correspond to the numbers of the processing steps in the procedure:

1) For each session, the audio track was extracted from the parliamentary video, Automatic Gain Control (AGC) was applied to the audio and converted to PLP feature

[4] http://www.praat.org/

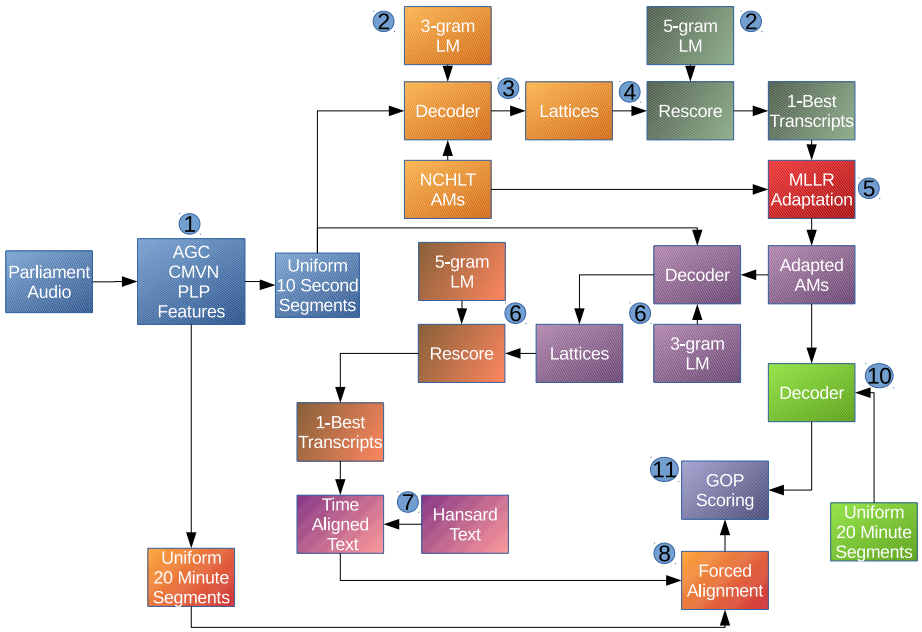

Fig. 1. High-level flow diagram of the automatic harvesting procedure. Tasks are grouped by colour and the order of execution are shown by numeric values.

vectors. To improve feature robustness, utterance-based mean and variance normalisation was applied.

2) Tri-gram and quin-gram language models were trained on the normalised parliamentary text produced by the document parsing and normalisation task.
3) The PLP feature file was uniformly segmented into non-overlapping ten second chunks and each chunk was decoded using the tri-gram language model and the seed NCHLT acoustic models. The output from the decoder was a lattice capturing a subset of possible word paths.
4) Each lattice was re-scored using the quin-gram language model.
5) Acoustic model mean and variance parameter adaptation was performed using cascaded Maximum Likelihood Linear Regression (MLLR). The global MLLR transforms were estimated using the transcriptions generated by the quin-gram lattice re-scoring.
6) Processes 3 and 4 were repeated using the adapted acoustic models.
7) The transcriptions generated from the lattice re-scoring, using the adapted models, were used to find approximate alignments between the recognition outputs and the normalised Hansard transcriptions. To perform the alignment, four to seven word sequences were used to find anchor points – preference was given to longer word sequences. This alignment process gave rough ten second alignment granularity.
8) Using the approximate alignments, the feature file was segmented into 20 minute chunks. A transcription file was created for each 20 minute chunk by extracting the words from the corresponding normalised Hansard transcription. The adapted acoustic models were used to force align the feature chunks and the corresponding transcriptions.
9) Once aligned, each 20 minute alignment output was post-processed by adding silence boundaries. Silence portions greater than 150 ms were marked as boundaries.
10) A recognition process was run for each 20 minute chunk as well. A task grammar was created using the 20 minute transcriptions generated in step 8 and modified such that words could be skipped.
11) The outputs of the forced alignment (step 8) and recognition (step 10) were used to derive confidence scores. Goodness-Of-Pronunciation (GOP) scores were used to gauge the match between the audio and the text [15]. The scores were not calculated for the entire 20 minute chunk, but for shorter segments that had a corresponding audio duration greater than five seconds. These segments were selected between silence boundaries as marked in the 20 minute forced alignment outputs.
12) Lastly, the GOP outputs were re-formatted to allow easier prompt selection. The output file format contained a possible prompt per line and each line had the following information: GOP score, start time, end time and transcription word sequence. This output file was used in the corpus packaging step to create the final NCHLT II Speech Resource Development ASR corpus.

C. Alignment evaluation procedure

An ASR corpus requires accurate audio and transcription pairs. The harvesting approach, described in section IV-B, aligns the normalised parliamentary session audio with the

corresponding Hansard transcriptions and outputs GOP scores used to select prompts. This measure, however, does not indicate for each word, how accurately the individual words are aligned. Therefore, a verification step was performed to evaluate how accurately the words are aligned between the Hansard transcriptions and the audio.

The evaluation procedure described in Moreno *et. al.* [11] was used to perform the alignment verification. Their verification process compares alignments between manually corrected transcriptions and automatically generated ones. For each word, the start and end times are compared between the reference and automatic sets and the difference recorded over the entire set of words.

The verification process requires both manually corrected and automatically generated transcriptions. The process followed in creating the manually edited transcriptions is described in section III-C and the automatic transcriptions were generated using the alignment procedure highlighted in section IV-B. PRAAT TextGrids were used to format the manual transcriptions, while the automatic transcriptions are stored in HTK MLF format.

The first step for the verification process was to create alignments between the two transcriptions sets. The alignments were generated using *Sclite* application from the "Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0" [16], which require the conversion of the PRAAT TextGrids and MLF format files to "CTM" format. The sclite alignment procedure segments the utterances based on common silence boundaries that results in multiple alignment chunks per utterance. Non-English alignment chunks were removed by using a 124k English word lookup table.

Given the alignments, the difference in word start and end times, for each word, could be calculated. The final output from the verification analysis was (1) a table capturing the total number of comparisons, substitution errors, insertion errors and deletion errors, (2) a file containing the difference in word start times, and, (3) a file containing the difference in word end times. To date, the verification process was performed on parliamentary sessions 11_May_2012 and 1_March_2012.

V. Results

This section presents results of prompt selection and the alignment accuracy between the audio and text.

A. GOP prompt selection

The last phase of the harvesting procedure was to create an ASR corpus. This was done by selecting text portions from the aligned text and assessing the corresponding GOP scores. Figure 2 shows the amount of harvested audio (in hours) packaged in the final ASR corpus as a function of the GOP score. At a high GOP score the total number of audio hours is around 78 hours, meaning that 27 hours (around 25%) of audio is unsuitable for harvesting. A GOP score of between 0 and 4 sees the greatest change in the amount of harvested data.

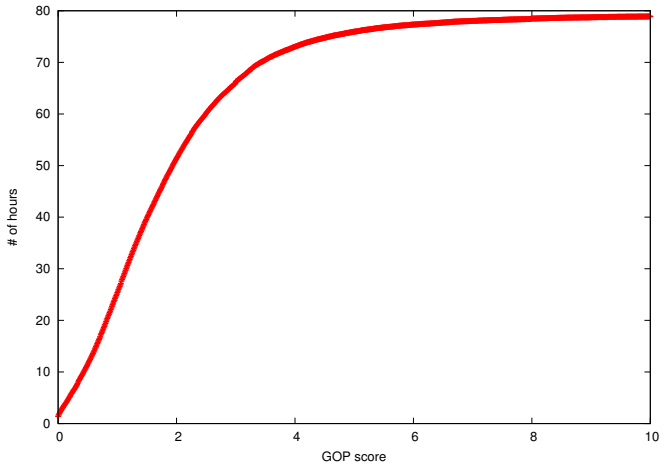

Fig. 2. The amount of audio data harvested as a function of the GOP score.

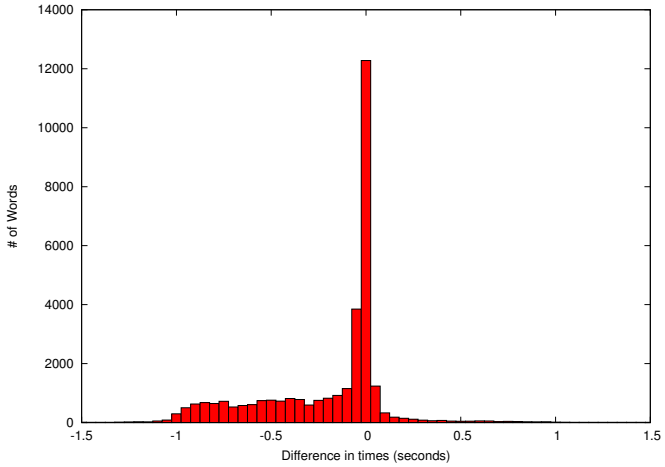

Fig. 3. Difference in starting and ending word times between the ground truth transcriptions and automatically harvested transcriptions.

Lastly, with the GOP threshold set to 1.9, 50 hours of aligned data were selected from the 105 hours of processed parliamentary audio. This almost doubles the NCHLT English data set and increases the accent variability of the audio data.

B. Alignment evaluation

The ground truth data described in section III-C was used to evaluate the accuracy of the time boundaries assigned by the automatic alignment system. Figure 3 shows a histogram of the time differences between the word start and end times. The histogram shows that 96.36% of the words are within 1.0 second of the true alignments. A large portion of the word mismatches are less than 0 which indicates that automatic word alignments occur after the actual word occurrences in the audio. The artefact that causes this has still to be identified.

Table III shows values for the total number of word comparisons, substitution, deletion and insertion errors. For both the 1_March_2012 and 11_May_2012 parliamentary sessions, over 80% of the words were used for the alignment accuracy

comparison while around 20% of the words were rejected due to alignment errors.

TABLE III
Total comparisons made, substitution, deletion and insertion errors made between the ground truth transcriptions and automatically harvested transcriptions.

Session	Total Comparisons	Subs.	Dels.	Ins.
1_March_2012	12327	1157	1109	597
11_May_2012	9249	550	1030	588

VI. Discussion

The primary aim of this investigation was to extend existing speech resources by developing an automatic harvesting approach that could be used to create an ASR corpus from parliamentary audio and Hansard transcriptions. Unfortunately, only the English corpus could be extended as the parliamentary data contained less than 5% of other language data. The benefit, however, is that the final corpus contains accented English data, which should enable the development of acoustic models suited for the South African context. Furthermore, it should be relatively easy to extended our approach to other similar datasets, which may contain a greater diversification of South African languages.

The final harvesting approach was a modified version of the algorithm proposed by Moreno *et. al.* [11] and managed to harvest 50 hours of audio data from a total 105 hours of raw data at a GOP score of 1.9. The alignment accuracy evaluation, on two parliamentary sessions, showed that over 96% of the words are within 1.0 seconds of the true position in the audio stream. These results show the promise of pursuing data harvesting of parliamentary data, which is a challenging environment.

Main conclusions that can be drawn from the investigation are:

- From Section III-C we can see that there is a large discrepancy between the audio recorded in Parliament and the Hansard transcriptions. This complicates the alignment task, as finding aligned portions is more difficult and reduces the quality of the full transcription alignments.
- The immediate goal of the Hansard transcriptions is to succinctly capture the proceedings of a parliamentary session. In order to create an accurate ASR corpus, however, more data is needed as verbatim portions are rather scattered throughout the transcriptions.
- The above points are not unique to South Africa as this has been seen in the Japanese Parliament [4]. Their approach to the problem was different as machine translation was used to "normalise" the transcriptions.
- Despite the big discrepancies some segments could still be aligned accurately and included in an accented South African English ASR corpus.

The immediate goal of future work is to improve the harvesting approach to produce more accurate alignments. This will allow the harvesting of more usable data from the raw parliamentary sessions and increase the final ASR corpus audio-text accuracy.

Acknowledgment

This research was conducted within a project supported by the Department of Arts and Culture of the South African government. The authors would like to acknowledge the contributions of the other members of the NCHLT II team.

References

[1] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proccedings of the 4^{th} Workshop on Spoken Language Technologies for Under-resourced Languages*, St Peterburg, Russia, May 2014, pp. 194–200.

[2] "Technology and corpora for speech to speech translation (TC-STAR)," September 2014, http://tcstar.org/.

[3] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, "The LIMSI 2006 Tc-Star transcription systems," in *Proc. TC-STAR Workshop*, 2006, pp. 123–128.

[4] T. Kawahara, "Transcription system using automatic speech recognition for the Japanese parliament (Diet)," in *Proceedings of the Twenty-Fourth Innovative Applications of Artificial Intelligence Conference*, Toronto, Ontario, Canada, July 2012, pp. 2224–2228.

[5] "Austrlia: National Parliament," September 2014, http://www.aph.gov.au/about_parliament/senate/about_the_senate.

[6] "Australia: Western Australia," September 2014, http://www.parliament.wa.gov.au/webcms/webcms.nsf/content/hansard.

[7] "Australia: New South Whales," September 2014, http://www.parliament.nsw.gov.au/Prod/parlment/publications.nsf/key/ParliamentHouse,HansardintheParliamentofNSW/$File/History+Bulletin+7.pdf.

[8] "Denmark," September 2014, http://www.sail-labs.com/news-events/press-releases.html?tx_clpresse_pi1%5BshowUid%5D=121.

[9] "Isle of man," September 2014, http://www.tynwald.org.im/business/hansard/Documents/voice-recognition.pdf.

[10] M. H. Davel, C. van Heerden, N. T. Kleynhans, and E. Barnard, "Efficient harvesting of Internet audio for resource-scarce ASR," in *Proceedings of INTERSPEECH*. Florence, Italy: ISCA, August 2011, pp. 3153–3156.

[11] P. J. Moreno, C. Joerg, J. M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proceedings of INTERSPEECH*. Sydney, Australia: ISCA, November 1998.

[12] D. Kim, G. Evermann, T. Hain, D. Mrva, S. Tranter, L. Wang, and P. Woodland, "Recent advances in broadcast news transcription," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. St. Thomas, U.S. Virgin Island: IEEE, November 2003, pp. 105–110.

[13] 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. revised for HTK version 3.4," March 2009, http://htk.eng.cam.ac.uk//.

[14] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[15] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, pp. 95–108, 2000.

[16] Multimodal Information Group (NIST), "NIST Speech Evaluation Tools," 2014. [Online]. Available: http://www.itl.nist.gov/iad/mig//tools/