Unsupervised Topic Modelling on South African Parliament Audio Data

Neil Kleynhans
Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
Email: ntkleynhans@gmail.com

***Abstract*—Using a speech recognition system to convert spoken audio to text can enable the structuring of large collections of spoken audio data. A convenient means to summarise or cluster spoken data is to identify the topic under discussion. There are many text-based topic modelling and identification techniques that become available once the audio to text conversion has occurred. These approaches allow the management and presentation of spoken audio data in a more structured way.**

In this work, an accurate spoken topic identification system was developed to identify a dominant topic discussed in a South African parliamentary session. This was achieved by using CMU Sphinx word recognisers to convert the conversations to word representations and latent Dirichlet allocation topic modelling techniques. The best topic identification accuracy of 92.3% was obtained on 40 topics, derived from speech recogniser transcriptions and compared to the Hansard transcriptions of National Assembly sessions of the South African Parliament.

I. Introduction

Large unstructured data is constantly being generated in the information age, with most captured in text and/or audio form. One way to make sense of the data is to automatically annotate the data so that some structure can be imposed. An automatic approach is preferred over manually annotating the data, as the latter is resource intensive and in some cases impractical. Annotations can be added to the data on creation but there are many systems that do not have this functionality.

Considering written text, one way to structure or cluster the data is to analyse the *topics* that occur in the documents. The topic defines a probability distribution over a set of words [1] which implies topics are correlated to certain word patterns and use. Therefore, documents that cover the same or similar topics should make use of similar words and patterns that can be exploited to cluster the documents. To discover the topic, *topic models* are used, in a probabilistic manner, to infer the underlying semantic structure of a collection of documents [2]. A few approaches to generate topic models are latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA).

LSA, projects word or term counts, extracted from a set of documents, from a high-dimensional term-space to a lower dimensional latent semantic space [3]. The mapping is achieved by applying singular value decomposition to a term matrix that finds the components with the most variability. Weaknesses of this approach are aspects around the theoretical framework [3] and generative properties [4].

To address the limitations of LSA, Hofmann [3] proposed an extension in the form of PLSA. In PLSA, each word in a document belongs to a topic and is generated by sampling a mixture model where the mixtures represent the topics. The documents are created by mixing the topics proportionally. PLSA makes use of an aspect model that adds latent variables to the joint probability model which represents the probability between the documents and the words. PLSA suffers from two problems [4]: (1) the number of model parameters grow with an increase in corpus documents and (2) the model learns topic mixtures from documents found in the training dataset which makes it difficult to assign a probability to an unseen document.

To overcome the shortcomings of the PLSA approach, LDA has been proposed [2], [4]. With LDA, a fully generative probabilistic model is defined, that assumes that the documents are represented by a mixture of latent topics and each topic is characterised by a word distribution. In the LDA model, the latent topic structure is hidden, and, is inferred from the words in the documents, using posterior probabilistic inference. The benefits of LDA over PLSA are (1) that the model size does not grow with training data increase, (2) it is less sensitive to over-fitting and (3) is able to generalise to unseen documents.

Topic modelling approaches can easily be extended into the spoken audio domain by making use of a speech recognition system that can convert the spoken audio into a text representation. Topic identification on spoken audio has been investigated by Hazen et. al. [5] and Wintrode and Kulp [6]. The focus is different to the topic discovery approaches highlighted in [2], [4], [3] as the topics, under discussion for each spoken audio utterance, was pre-defined. Hazen et. al. [5] reported a 9.6% topic identification error rate using an automatic speech recognition (ASR) system with a 40% word-error-rate (WER) and using a naive Bayes topic classifier. Similarly, Wintrode and Kulp [6] achieved a topic identification error rate of 10.1% using a Support Vector Machine (SVM) topic classifier and an ASR system that delivered a WER of 34%. Both these investigations highlight that topic identification is possible, even at high WERs, and it stands to reason that lowering the WER should result in improved topic modelling.

In terms of unsupervised topic modelling techniques applied to spoken audio, Hazen [7] made use of ASR and PLSA topic modelling to automatically summarise the Fisher corpus [8] – the Fisher corpus contains topic labels for all conversations (40 in total). The task focused on determining the best number of topics, ranking the importance of the topics and selecting appropriate keywords for topic summaries. The task was complicated by making use of conversational speech which results in less focused topical content. At an optimal latent topic number of 35, the system managed to

select summary keywords, found in the reference topic word set, 85% of the time.

The results presented by Hazen [7] have shown that latent topic estimation and summarisation from ASR word hypotheses are indeed possible and that reliable results can be achieved. The PLSA approach, however, has a few limitations that are addressed by LDA, which makes LDA an attractive replacement. Given this background, the aim of this work is to develop a spoken audio topic identification system that makes use of LDA to infer latent topics from a text corpus and given the topics, can estimate the most probable topic being discussed in an utterance. The envisioned use of the system is to be able to cluster audio recordings that contain similar topics under discussion.

The South Africa Parliament audio and Hansard text data were identified as a suitable corpus, which was used to develop and evaluate the spoken audio topic identification system. Many parliamentary sessions can be accessed, but for this investigation only the National Assembly debates were considered. Some of the Hansard text contains a number of topics that were purportedly discussed during the session, which makes it difficult to assign a single topic to an entire document. The LDA method, however, assumes that each document is created by sampling a mixture of topics, which is well supported by the parliamentary data. Therefore, the meta-topic information was ignored and the LDA latent topics were only considered for this investigation.

Section II describes the development of the system components and in section III experimental results are reported. Concluding remarks and future work can be found in section IV.

II. Method

Three main components were identified in the development of a spoken audio topic identification system for audio data from the South African Parliament. These components are:

- Audio Diariser
- Speech Recogniser
- Topic Modeller

Under these components there are various supportive subtechnologies that will be discussed in the subsections that follow.

The flow of the overall system is as follows:

- The audio is extracted from the video recordings of Parliament and sent through the *Audio Diariser*, which extracts spoken audio and marks the audio with meta-information such as gender and spoken language.
- The processed audio is then converted to a text representation using the *Speech Recogniser* – this requires acoustic and language models.
- The text representation of the audio is finally sent to the *Topic Modeller* which analyses the text and determines the most likely topic. As input to the topic modeller, previously trained topic-specific models are needed.

A. Audio data from the South African Parliament

The sourced parliamentary data was packaged as a DVD video for each session. The open-source software *transcode*[1] was used to extract the audio from the video. Following the extraction, the audio was converted PCM WAVE format and 16-bit signed integers used to represent a data sample. In addition, the audio was down-sampled to 16 kHz as speech recognition does not benefit from audio sampled at greater sampling rates. Lastly, it was observed that the audio clipped at some portions during the session so the volume of the audio was decreased by 3 dB. The open-source software package *Sox*[2] was used to perform all post audio extraction operations. Audio was extracted from 17 video National Assembly parliamentary sessions.

B. Audio diariser

The parliament audio is heterogeneous in nature. There is a mixture of spoken audio, babble noise, clapping and silence. Sending non-spoken audio through the speech recogniser will result in nonsensical outputs that will confuse the topic modelling and detection results. Therefore, an audio diariser was developed to automatically identify and organise the audio into homogeneous audio portions. Of interest are the spoken audio portions, while the remaining audio portions were marked but not used for topic detection.

The audio diarisation comprised the following operations (as applied to each parliamentary session):

- The entire audio file was volume normalised using automatic gain control (AGC).
- PRAAT application [9] was used to identify voiced and un-voiced portions using pitch features. Non-speech portions were automatically marked by identifying un-voiced segments with a duration longer then 250 ms.
- Voiced segments were combined together to form longer speech portions. An objective of an average of five seconds per speech segment was set. These segments were extracted from the larger audio session and saved to audio chunks.
- Mel-frequency cepstral coefficients (MFCC) were extracted from the audio chunks.
- Audio segments were clustered (using the MFCCs) by following a speaker identification process. The procedure was:
    - Three Gaussian mixture models (GMM) were trained for two adjacent segments – one for each segment and one for the combined segments.
    - If the average log-likelihoods for two separate segments was less than the combined segment then the segments were combined, otherwise the segments were left as is.
    - The process was then repeated by considering the next adjacent segment. The process stop after all segments were compared.

[1] http://www.transcoding.org/
[2] http://sox.sourceforge.net/

- The combined segments were further classified based on gender (male or female), spoken language (English or non-English) and signal-type. Possible signal-types were:
    - Applause
    - Applause and speech
    - Babble speech
    - Music
    - Music and speech
    - High-bandwidth speech (8 kHz)
    - Low-bandwidth speech (less than 4 kHz)

512-mixture GMMs were used for the signal-type, gender and language classifiers. The data for the signal-type classifiers were sourced from various downloadable audio datasets, TIMIT [10] and NTIMIT [11] corpora. The gender classifier was trained on all signal-type data that had gender tags (sourced from the TIMIT and NTIMIT labels). Finally, to train the language identification classifier the NCHLT corpora were used – all languages not English were grouped into one dataset (non-English). The initial language identification was performed using all eleven South African languages but the accuracy of the classifier was too low. Only the audio segments marked as English and high-bandwidth speech were passed on to the speech recognition system and the remaining portions were discarded. All GMM classifiers were developed using SPTK [12].

C. Hansard text data from Parliament

Each parliamentary session had an accompanying Hansard transcription that captured the session's proceedings. The text from the Hansard transcriptions was used to train topic and language models. To analyse the text in the document, the document had to be converted to a more easily-accessible format, as the documents were saved in MS-WORD DOC format. The text, however, was not formatted in a consistent manner, and text normalisation had to be applied. The broad text normalisation steps were:

1) Convert MS-WORD document to UTF-8 text.
2) Character map or remove non-UTF-8 characters.
3) Extract spoken text only and remove all other text – the document contained superfluous information.
4) Parse the text and mark entities. The defined entities were:
    - TITLE - The title of the talk
    - INFO - Extra information about the talk usually placed under the title
    - SPK - The speaker who is speaking
    - TEXT - The text representation of the audio spoken by the speaker
    - IGN - Ignore the text mark with this entity.
5) Apply first iteration of normalisation:
    - Whitespace normalisation
    - Remove transcriber inserted annotations such as [Applause.], [Interjections.]
    - Normalise spacing and punctuation around numbers
    - Remove punctuation.
6) Remove non-English text.
7) Mark acronyms and spelled words. These must be marked as the pronunciation of the words are different to general words.
8) Apply second iteration of text normalisation (see first iteration of normalisation 5).
9) Convert all numbers (numbers, monetary text, dates, etc.) to written form, e.g. $1234 \rightarrow$ one thousand two hundred and thirty four.
10) Extract text portions only, ignoring TITLE, SPK, and INFO.

759 Hansard transcriptions, downloaded from the South African Parliament website, plus the Hansard transcriptions that accompanied the 17 video sessions, were processed using the above text normalisation procedure. The 759 text transcriptions were used as a training set and the remaining 17 sessions were used as an evaluation set. The parliamentary sessions for these two datasets were mutually exclusive but the speakers were not. This is due to the limited number of parliamentary members and staff.

D. LDA topic modelling

The parliamentary Hansard transcriptions contain topic labels that were discussed by the speakers, but as stated previously, for the initial investigation these were ignored. The consistency of the topic labels had to be verified as different transcribers were used to generate these. The topic labels indicated that, for the majority of cases, there are many topic changes throughout the course of a parliamentary session, which fits well with the LDA topic modelling assumption: a document contains a sampling from many topics. As the multiple Hansard topic labels were ignored, the unsupervised LDA approach was used to infer the topics from the normalised Parliament text data. Before modelling the documents and inferring topics, further text processing was performed where NLTK [13] was used to categorise the words and discard any word that was not an adjective, noun or verb. LDA was used to infer topics from the training set (759 documents) for 10, 20, 40 and 80 topics in total. Table I shows the top 10 words for a selection of topics extracted from the parliamentary text data.

E. CMU Sphinx ASR

The the high-bandwidth read speech NCHLT English subcorpus [14] was used to train the acoustic models. Table II shows a few corpus statistics. The AGC'ed audio was converted to 39-dimensional MFCCs – 13 static, 13 delta and 13 delta-delta. Cepstral mean normalisation (CMN) was applied on a per utterance basis.

TABLE II. NCHLT ENGLISH CORPUS USED TO TRAIN THE ACOUSTIC MODELS.

# of utterances	77085
Duration in hours	50
# of speakers	210

Two speech recognition systems were developed using CMU Sphinx [15]: (1) Semi-continuous and (2) continuous. The semi-continuous system applies vector quantization to the

TABLE I. TOP 10 WORDS FOR VARIOUS TOPICS EXTRACTED FROM THE TEXT NORMALISED PARLIAMENTARY HANSARD DOCUMENTS, USING LDA.

Topic 001	Topic 002	Topic 003	Topic 004	Topic 005
sector	price	financial	money	prices
schools	teachers	educators	bodies	quality
bill	training	da	legislation	fet
tax	taxation	income	amendment	sars
skills	works	industry	growth	dont
madam	departments	food	private	reports
school	learners	teacher	parents	students
colleges	industry	skills	amendment	madam
bill	fund	funds	laws	revenue
departments	sector	programmes	madam	asgisa

speech features and creates a codebook of feature entries. The codebook contained 128 clusters and had the same number of mixtures in the final acoustic models. This style of system performs very quick recognitions and was chosen as part of the adaptation strategy to generate unsupervised transcriptions for the parliamentary audio data. Acoustic model adaptation can be applied to this style of system, but it is not as effective as applying the adaptation to the continuous systems. Therefore, a continuous acoustic model recognition system was trained on the same data that performed the same recognition task, at slightly elevated recognition times.

The continuous system was developed using standard HMM training techniques, with the following additional refinements:

- Linear discriminative analysis applied to the features to reduce 39 dimensional MFCCs to 29.
- Maximum likelihood linear transformation (MLLT) was applied. This technique estimates a linear transform, applied to the features, that increases the modelling likelihood of the different classes.
- 16 mixtures per class were used to model the distribution of the features belonging to the class.

Speech recognition systems perform poorly if there is a mismatch between the acoustic models and the audio data that has to be recognised. To reduce the performance loss, acoustic model adaptation can be implemented. Our speech system employed an unsupervised maximum likelihood linear regression (MLLR) adaptation. To apply the adaptation, the diarised parliamentary audio data was first split by gender using the diariser gender tags. Only the high-bandwidth English speech data was used. Next, the semi-continuous system was used to generate text transcriptions of both audio data sets. Following this, gender-specific global MLLR transformation matrices were estimated. Lastly, when recognising the parliamentary audio data, the MLLR matrices are used by the continuous speech recognition system to reduce the error rate for the final text transcriptions. The adaptation gains were not verified as there were no manually corrected transcriptions of the parliamentary data.

Figure 1 shows a flow diagram that details how the audio was processed and analysed by the topic identification system.

F. Language modelling

A statistical language model captures the probability of a sequence of words. With all languages, certain word sequence patterns are more likely to occur than others. A speech recognition system can use this information to refine the search and speed up the recognition process by discarding low probability paths. Training a language model on text data sourced from the domain is important as similar terms are more likely to occur and the out-of-vocabulary word count can be lowered. Therefore, a tri-gram back-off language model was developed on the text-normalised training data (759 documents). The language model was used by both the semi-continuous and continuous ASR systems during decoding. Table III shows the normalised parliamentary text data used to train the tri-gram language model. The MIT-LM [16] was used to develop the language models.

TABLE III. SOME STATISTICS EXTRACTED FROM THE TEXT-NORMALISED PARLIAMENTARY HANSARD DOCUMENTS.

# of sentences	63034
Total number of words	16390372
Total number of unique words	65442

III. EXPERIMENTAL RESULTS

The entire system is made up of three sub-systems: diariser, speech recogniser and topic modeller. In this section the related performances achieved by these sub-systems is reported.

A. Diarisation

The main task of the diariser was to annotate the segments created from the parliamentary audio stream – section (II-B) details the functionality of the diariser. Table IV shows the top occurring annotations per gender (for all the diariser processed parliamentary data) as well as the amount of audio classified as non-speech and total amount of audio processed. The diarisation classification tags have the following form: gender, signal-type, language. The results show that roughly a third of the audio data is not usable. There was a greater amount of male audio data compared to female but the majority of the audio data, per gender, is high-bandwidth English speech.

To investigate the accuracy of the audio diariser a portion of the automatically generated labels were manually verified. Only audio segments marked as English and high-bandwidth were considered and were selected at random as these segments are only sent through to the ASR systems. Table V shows the accuracy of the diariser at classifying English high-bandwidth audio segments of the parliamentary data. Surprisingly, the optimal performance of the diariser, was for audio segments with a duration of 5 to 10 seconds, where it was expected that longer segments should allow for a better performance.

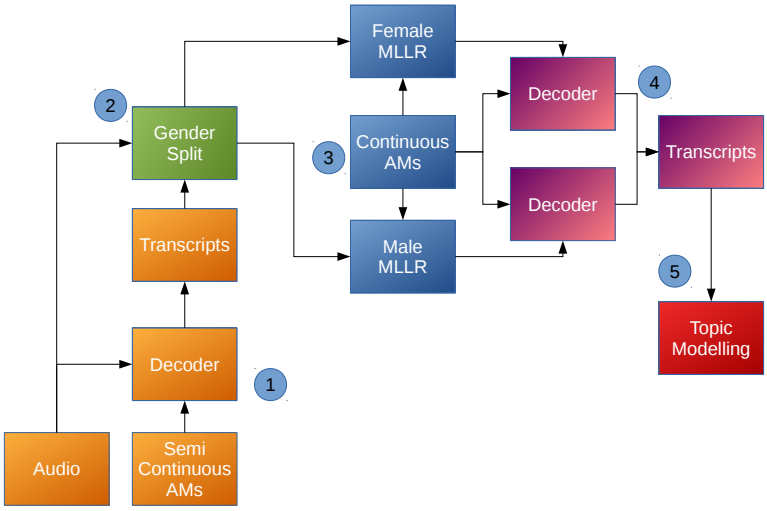

Fig. 1. A high-level flow diagram of audio processing and topic identification on the recognition text outputs.

TABLE IV. THE TOP OCCURRING DIARISER PRODUCED ANNOTATIONS (PER GENDER), THE TOTAL AMOUNT OF NON-SPEECH AND TOTAL AUDIO DATA AMOUNT.

Diarisation classification	# hours
Female,Speech(HB),English	9.21
Female,Speech(HB),Non-English	3.03
Female,Music+Speech,Non-English	0.39
Female,Music+Speech,English	0.13
Female,Applause+Speech,Non-English	0.12
Male,Speech(HB),English	14.75
Male,Speech(HB),Non-English	1.78
Male,Babble,English	0.36
Male,Babble,Non-English	0.15
Male,Applause+Speech,Non-English	0.11
Non-speech	16.43
Total	46.82

TABLE V. DIARISATION ACCURACY AT CLASSIFYING ENGLISH HIGH-BANDWIDTH AUDIO SEGMENTS.

Segment duration (s) #	correct	# incorrect	Total	Accuracy (%)
0 to 5	578	141	719	80.39
5 to 10	569	88	657	86.61
more than 10	182	44	226	80.53

B. Speech recognition

The NCHLT evaluation set was used to test the accuracy of the two speech recognition systems, which were developed on a NCHLT English sub-corpus. Table VI shows the number of speakers, utterance amount and the total amount of hours for the NCHLT evaluation set.

TABLE VI. THE NCHLT ENGLISH EVALUATION SET USED TO MEASURE THE PERFORMANCE OF THE ASR SYSTEMS.

# speakers	8
# utterances	3232
# hours	2.24

Table VII shows the word-error-rates (WERs) of the developed speech recognition systems. From the results, it can be seen that the semi-continuous systems perform better than the continuous version. The difference in WER may be due to the increased number of mixtures utilised by the semi-continuous system. The mixture count can be increased for the continuous system but this would increase recognition times.

TABLE VII. THE ASR WERS FOR THE CMU SEMI-CONTINUOUS AND CONTINUOUS SYSTEMS.

	Semi-continuous	Continuous
WER (%)	7.0	8.3

No ASR evaluations were performed on the parliamentary data. The Hansard transcriptions are not verbatim, as highlighted in [17], where the measured accuracy of the Hansard are in the range of 59 % to 68 %. An intensive process would have to be run to correct the contents and produce an accurate evaluation set.

C. Topic identification

To determine how accurately the system could identify topics from the audio, the following evaluation procedure was devised:

- The LDA topic modelling software was used to extract topics from the text normalised training document set (759 documents) – the number of topics extracted was 10, 20, 40 and 80.
- The Hansard transcriptions (referred to as the reference set) accompanying the parliamentary audio were text normalised.
- The LDA software was used to infer the topics from the evaluation document set. The inference produces a topic vector, where each element in the vector indicates the likelihood (weight) of the document belonging to a specific topic
- The CMU Sphinx speech recognisers were used to generate approximate transcriptions from the audio.

- The output text transcriptions (referred to as the evaluation set) were analysed by the LDA software and topics were inferred.
- To evaluate the accuracy, the topic vectors from the reference and evaluation sets were compared and only the most likely topic was considered. If the reference and evaluation set gave the same topic number, measuring the highest likelihood, then the detection was marked as correct, otherwise it was marked as incorrect.

Table VIII shows the topic identification accuracy in detecting the topics using the speech recognition text.

TABLE VIII. SPOKEN AUDIO TOPIC IDENTIFICATION ACCURACIES WHEN DETECTING VARIOUS NUMBER OF TOPICS USING AUDIO FROM PARLIAMENT.

# of Topics	Topic Identification Accuracy (%)
10	77.78
20	77.78
40	92.3
80	70.37

The results in table VIII show that, for both 10 and 20 topics, the system produced the same results 77.78% (verified to be the same). The best accuracy was produced at 40 topics at 92.3%, while at 80 topics the system produced the worst results of 70.35%.

IV. CONCLUSION AND FUTURE WORK

It has been demonstrated that an accurate spoken topic identification system can be developed to identify a dominant topic in the audio from South African Parliament. A follow-up process could use this information to cluster a corpus of parliamentary sessions. Using CMU Sphinx word recognisers and LDA topic modelling techniques, an accuracy of 92.3% was obtained on 40 topics, inferred from the text normalised Hansard and speech recogniser transcriptions.

The evaluation procedure detailed in section III may be improved and made more stringent by using the Kullback-Leibler divergence to measure the inferred topic distributions between the Hansard and speech recognition transcriptions. Furthermore, the evaluation set needs to be increased to cover all the discovered topics adequately. Unfortunately, 17 documents do not sufficiently cover all topics to reliably report a robust topic identification rate.

For future work the following optimisations are proposed:

- Improve the speech recognition system by increasing the recogniser's accuracy.
- Extend the system to other languages used during the parliamentary session, as all non-English was ignored for this investigation.
- Split the parliamentary documents and audio into smaller topic-based portions, as there are many topics per document. A supervised topic modelling approach can then be used, such as supervised LDA [18], for more refined topic clustering.

REFERENCES

[1] N. Pansare, C. Jermaine, P. J. Haas, and N. Rajput, "Topic models over spoken language," in *IEEE International Conference on Data Mining series (ICDM)*, Brussels, Belgium, December 2012, pp. 1062–1067.

[2] D. Blei and J. Lafferty, *"Topic Models," in A. Srivastava and M. Sahami, editors, Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. Berkeley, USA: ACM, August 1999, pp. 50–57.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[5] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *Automatic Speech Recognition and Understanding, 2007. ASRU'07. 2007 IEEE Workshop on*. Kyoto, Japan: IEEE, December 2007, pp. 659–664.

[6] J. Wintrode and S. Kulp, "Confidence-based techniques for rapid and robust topic identification of conversational telephone speech," in *Proceedings of INTERSPEECH*. Brighton, United Kingdom: ISCA, September 2009, pp. 1471–1474.

[7] T. J. Hazen, "Latent topic modeling for audio corpus summarization," in *Proceedings of Interspeech*. Florence, Italy: ISCA, August 2011, pp. 913–916.

[8] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, vol. 4, Lisbon, Portugal, May 2004, pp. 69–71.

[9] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program] version. 5.3.84," 2014. [Online]. Available: http://www.praat.org/

[10] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[11] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Albuquerque, New Mexico, USA: IEEE, April 1990, pp. 109–112.

[12] S. Imai and T. Kobayashi, "Speech signal processing toolkit (SPTK)," 2014. [Online]. Available: http://sp-tk.sourceforge.net/

[13] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, vol. 1. Association for Computational Linguistics, 2002, pp. 63–70. [Online]. Available: http://www.nltk.org/

[14] 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.

[15] C. M. University, "CMU sphinx," 2014. [Online]. Available: http://cmusphinx.sourceforge.net/

[16] B.-J. Hsu and J. Glass, "Iterative language model estimation: efficient data structure & algorithms," in *Proceedings of Interspeech*, vol. 8, Brisbane, Australia, September 2008, pp. 1–4.

[17] N. Kleynhans and F. de Wet, "Aligning audio samples from the south african parliament with hansard transcriptions," in *Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, November 2014.

[18] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Twenty-First Annual Conference on Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2008, pp. 121–128.