An investigation into Spoken Audio Topic Identification using the Fisher Corpus

Neil Kleynhans
Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
Email: ntkleynhans@csir.co.za

***Abstract*—There are many important informational aspects associated with the audio data, but one common component is the topic under discussion. Knowing the topic can help process the data in a variety of ways – cluster similar audio recordings based on the topic or improve ASR recognition outputs by using appropriate language models.**

In this work, the best spoken audio topic identification system achieved an error rate of 17.6%, using an ASR system that produced an average word error rate of 57% and supervised latent Dirichlet allocation topic modelling technique. The proposed language model topic modelling technique, produced the worst results, highlighting the sensitivity to high ASR word error rates. The support vector machine topic classifier, which made use of a simplified term-weighted feature vector, performed comparably to that of the term frequency inverse document frequency feature vectors.

I. Introduction

In the current ICT age, there are many media sources that generate and archive large volumes of audio data, that contains spoken audio. These sources are varied:

- Call centres, which generally archive incoming calls.
- Broadcasters, such as television stations (SABC, e-TV), radio stations (702, RSG) or Internet sites (YouTube, Vimeo, SoundCloud).
- Government institutes, such as Parliament or courts.
- Military and security organizations who comb audio for intelligence.
- Universities recording and distributing lectures.

Given the massive volume of audio data, it becomes impractical for humans to listen to and annotate the data. Thus, there is a definite need to find methods that can be used to automatically process and extract useful information from the data. The information that needs to be extracted, is defined by the particular application, but one general information component found throughout spoken audio is the topic under discussion. Identifying the topic, allows clustering of similar audio streams or archived recordings. Furthermore, an audio stream can be segmented by topic, which provides a means to search within the audio stream and can vastly reduce the time needed to search through the audio and find regions of interest. When performing automatic transcribing, an automatic speech recognition (ASR) system can use dynamic-topic-tracking information to adapt and specialize a language model, used during the decoding operation when searching for the most likely spoken text, and improve the recognition results.

II. Background

A *topic* is described as a probability distribution over a set of words or phrases [1]. Each document in a corpus, contains words that can be thought of, as being drawn from a mixture of topics [1], [2]. A *topic model* specifies the topics that occur in a corpus and the topical proportions found in the documents. Using the topic model, one can work back from the words found in a document and determine the distribution of topics or most likely topics used to generate the document. There are a number of text-based topic modelling approaches.

Latent semantic analysis (LSA), maps a high-dimensional word space to a lower dimensional representation, named the latent semantic space[3]. The data mapping function is determined by finding a representation that preserves the most relevant information for the given topics. An extension to LSA is probabilistic latent semantic analysis (PLSA), which introduces a latent topic space and estimates joint probability models, that model the relationship between the hidden latent topics and the documents as well as the relationship between the hidden latent topics and the words [3], [4]. One drawback of PLSA is that the model parameters increase as more documents are added to the corpus.

In unsupervised latent Dirichlet allocation (LDA) approach, a document is assumed to be drawn from a weighted mixture of topics. Similar to PLSA, hidden latent variables are used to model topical structures given a set of observed words [5] but the distributions, however, are drawn from Dirichlet distributions. When training with the unsupervised LDA approach, no topic labels are needed as the topics are "discovered" during the training process. A supervised version of LDA [2] is avaliable – supervised latent Dirichlet allocation (SLDA) – which makes use of the labels during the model training process. During estimation, a response variable is assigned to each document, and, the documents and response variables are jointly modelled to maximise the likelihood between response variables and labelled documents.

Topic modelling and identification is performed on text resources, which can extended to spoken audio streams by using an automatic speech recognition system (ASR). Current state-of-the-art large vocabulary speaker independent speech recognition systems are used to generate text transcriptions for the unannotated audio, but these typically produce word error rates (WER) greater than 10%. Hazen et. al. [6] reported that their ASR system (trained on 553 hours of spoken audio) would on average achieve a 40% WER but still yielded a topic identification error rate of 9.6%. Similarly, Wintrode and Kulp [7] achieved a topic identification error rate of 10.1% with an ASR system that delivered a WER of 34%. This highlights

that topic identification is possible even at high WERs which relaxes the needed for finely-tuned ASR systems.

A suitable classifier used for spoken audio topic identification is support vector machines (SVMs). Suitable SVM topic feature vectors are constructed using term frequency inverse document frequency (TF-IDF) values [6], [7] . TF-IDF are calculated by normalising a within-document term's frequency by the frequency of occurrence across all documents. This in effect, reduces the weight of terms that are common to all documents and increases the weight of terms that appear in a few documents but occur frequently within a document. Using a SVM topic classifier and TF-IDF features, Wintrode and Kulp [7] achieved a topic identification error rate of 10.1%. Other approaches can also be used to select topic specific words, such as χ^2 statistics, as investigated by Hazen *et. al.* [6], who managed to achieve topic identification error rates of 16.8% for words and 35.3% for 3-gram phones on call sides (single channel of a two-way telephone call) and 9.6% for words and 22.9% for 3-gram on the whole call. Both investigations performed the topic identification on the Fisher corpus – a corpus that contains conversational telephone audio where the participants discuss 40 topics.

The primary aim of this work, was to investigate spoken audio topic identification, which could be used by a system to find structure in audio data. In this domain, the Fisher corpus is commonly used to develop and evaluate topic identification systems. As the Fisher corpus contains topic labels, techniques such PLSA or unsupervised LDA are not needed. More appropriate topic identification approaches are ones that make use of SVMs or supervised LDA. Therefore, in this investigation, a spoken audio topic identification system was developed and evaluated on the Fisher corpus, using SVM and SLDA topic classifiers. In addition, language modelling techniques were also investigated, to determine their suitability in topic identification tasks.

III. Method

A. Fisher corpus

The Fisher corpus [8] contains two speaker telephone conversations, where the participants were instructed to discuss a certain topic for a duration of ten minutes. There are 40 topics, covered by the corpus, such as "Movies" and "Foreign Relations". For this investigation only the training part of the English phase 1 corpus was available.

To proceed with the development and evaluation of the spoken audio topic identification system, the available Fisher corpus data was divided into speaker independent training, development and evaluation sets. The splitting process created two gender-dependent sub-corpora that contained 100 hours of training audio data and 50 hours of audio data for the development and evaluation datasets. The more traditional 80%-10%-10% dataset split was not chosen as to reduce the training and decoding recognition times. The data selection process made use of the topic and speaker labels when partitioning the data. Each dataset had mutually exclusive speaker sets and uniform topic selection was also performed. Table I shows the modified English Fisher corpus used during this investigation.

The gender split audio and text data were used to train gender-dependent acoustic models, however, when training the SVM classifier and language models, the text data was combined. Topic identification was performed using the call sides – a call side contains a single speaker on one channel.

TABLE I. Training, development and evaluation set partitions of the English Fisher corpus.

Data Set	Male		Female	
	# Call Sides	# Utterances	# Call Sides	# Utterances
Train	1084	85617	1154	101412
Development	518	46646	693	67669
Evaluation	519	47384	672	66600

B. Topic identification system

Figure 1 shows a flow diagram that describes the topic identification system. First, the audio data was split by gender and recognised by a speech recognition system, that used gender-dependent acoustic models and a bi-gram language model. The bi-gram language model was trained on all the training text data. Next, the automatically generated transcriptions were processed by the various topic identification classifiers, which identified the most likely topic, given the recognised words. The topic identification approaches processed all the text generated from a call side, when estimating the topic under discussion.

C. Pronunciation dictionary

The CMUDict0.7a [9] pronunciation dictionary was sourced as a seed North American English pronunciation dictionary. It contains over 125k words and uses 39 stress-marked phonemes. For this investigation, the stress markings were removed. Phonetisaurus [10] was used to perform grapheme-to-phoneme (G2P) prediction for words not found in the seed pronunciation dictionary. Phonetisaurus implements a WFST-driven G2P framework that is used to rapidly develop high quality G2P or P2G (phoneme-to-grapheme) systems – it does this by learning statistical rules from a seed pronunciation dictionary. The final pronunciation dictionary contained 44385 words in total (all unique).

D. Speech recognition systems

Audio data was converted to Perceptual Linear Prediction (PLP) coefficients where each 52 dimensional feature vector was created by appending the first, second and third derivatives to the 13 static coefficients (including the 0'th component). The frame length was 25 ms and the frame shift was 10 ms. The only form of feature normalisation was corpus-wide mean and variance normalisation.

The acoustic models (AM) were developed in an iterative training scheme. Firstly, 32-mixture context-independent (CI) AMs were trained and used to produce state aligns 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, except the AMs, generated during the previous iteration, were used to produce state alignments up and till the mixture incrementing phase. The process was repeated twice for all experiments.

All hidden Markov models (HMM) employed a three state left-to-right structure and each CD HMM state contained 16

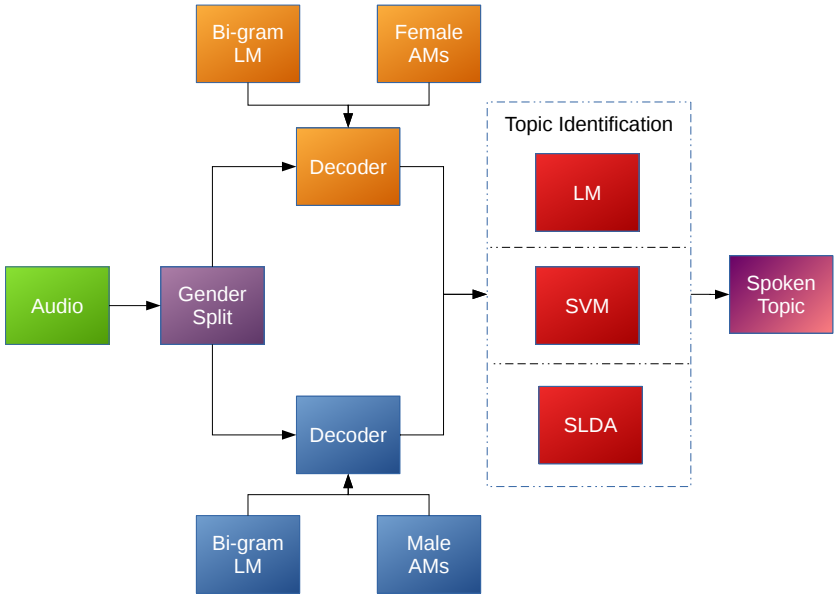

Fig. 1. A high-level flow diagram of the topic identification system components.

mixture diagonal covariance Gaussian models. A question-based tying scheme was followed to create a tied-state data sharing system [11], where any context-dependent triphone, that has the same central context, could be tied together.

Once the CD AM development was completed, heteroscedastic linear discriminant analysis (HLDA) was applied to reduce the 52-dimensional PLP feature vectors to a 39-dimensional vector. A global transform was used for the estimation – a single class for all triphones. After estimating the HLDA transform, the CD AMs' parameters were updated while applying the transform. Only the weight and mean parameters were updated and two iterations were performed.

Lastly, speaker adaptive training (SAT) was applied using constrained maximum likelihood linear regression (CMLLR) transformations. The same HLDA global transform was used for the CMLLR transform estimation, and the CD AMs were updated twice – again only weights and means.

The decoding task was a two-step process: the HLDA CD AMs were used to automatically generate transcriptions and a speaker-based CMLLR transform estimated. Then, the CMLLR was applied on the second decoding pass.

All ASR related-tasks were performed using HTK [12] and gender-dependent models were trained. A bi-gram back-off language model was trained on all the training text data, found in the customised English Fisher corpus detailed in section III-A.

E. Topic classifiers

1) Weighted term frequency: Support vector machines have been used previously to perform accurate topic identification, as reported in [6], [1], [7]. General topic identification SVM features are weighted word frequency values, such as the term frequency inverse document frequency (TF-IDF) features, which weights a term's frequency, found within a document, by a logarithm transformed cross-document term frequency. Another aspect of using the TF-IDF, is that a document containing N words is converted to a term (usually words) vector of length M, where M is the vocabulary size. Thus, arbitrary length documents are all normalised to a consistent length. Lastly, to improve the classification accuracy, a *stop-word* list is used to reject words with low topic discriminative information. These stop-words are usually articles, conjunctions and auxiliary verbs.

In this investigation, however, a simpler weighted term frequency vector was used. Firstly, a super vector of terms was created. To do this, all documents belonging to a specific topic were grouped into single topic-specific document and a single background document containing all topics was created. Next, for each word, the frequency of occurrence within a topic (f_i) and across all topics (f_{all}) was calculated. All words with a ratio between the within-topic term frequency and cross-topic term frequency less than 0.1, $\frac{f_i}{f_{all}} < 0.1$, were rejected. This is similar to the approach followed by Hazen [4], where they used an automatic process to create a stop-word list. Following this, for each topic, a topic-specific term vector was created by using only a limited number of the words – the words were ordered by their frequency of occurrence and the top number of words selected. Finally, a super vector was created by concatenating all the frequency values found across the topic term vectors.

To train the SVM and predict unseen documents, a feature vector for each document was created by calculating the within-document word frequencies for all words found in the super vector. For this investigation the top 5, 10 and 20 words per topic were chosen – this relates to a super vector of 200, 400 and 800, respectively. The SVM used radial-basis function kernels and a grid search was performed on the development

dataset, to find the optimal parameters. LibSVM toolkit was used to train and evaluate the SVMs [13].

2) Language models: The N-gram language model (LM) provides a method for estimating the probability of a word sequence, which is estimated on written text data. The perplexity measure gives an indication on how well a N-gram model predicts a text sample. If we assume that topics produce different word sequences, then it may be possible to perform topic identification using topic-specific N-gram models – select the N-gram model that produces the lowest perplexity. Given this assumption, the viability of topic-specific N-gram language modelling was investigated.

To produce topic-specific N-gram LMs, all topic-related documents were concatenated into a single document. Additionally, a "background" document, containing all training text data, was created and used to estimate a background N-gram LM. Then, for each topic, a N-gram LM was created by interpolating from the background LM, using the topic-specific documents. For this investigation, tri-gram back-off LMs with fixed Kneser-Ney smoothing were developed. The MIT-LM language modelling software was used [14] to develop the various LMs.

3) SLDA: SLDA makes use of provided topic labels to estimate the LDA model parameters in a supervised manner. The SLDA models were estimated using an implementation provided by Wang [15]. The model parameters were estimated on the combined training text data. A light preprocessing was performed similar to that detailed in section III-E1 where, words were rejected if the ratio between the within-topic word frequency and cross-document word frequency was less than 0.05. Out-of-vocabulary words were ignored. A linear search was performed to find the optimal model parameters using the development dataset.

IV. Experimental results

In this section, the ASR system performance and closed-set topic identification rates are reported.

A. Speech recognition word error rates

Table II shows the WER for gender-dependent recognitions obtained on the development and evaluation sets of the customised Fisher corpus. There is about a 4-5% absolute difference in the WERs, when comparing the female to male results for both the development and evaluations sets. This difference is most likely caused by the telephone channel bandwidth restrictions, which seems to affect female speech more than male speech.

TABLE II. WERs obtained on the development and evaluation sets of the customised Fisher corpus, for the gender-dependent ASR systems.

	WER		
	Female	Male	Average
Dev	54.91	59.72	57.32
Eval	55.44	59.77	57.61

It should be noted that the average WER achieved by the gender-dependent ASR systems was around 57%, which is larger than the WERs reported by Hazen *et. al.* [6] (around 40%) and Wintrode and Kulp [7] (around 30-50%). The difference in the WERs may be due to a few factors such as acoustic modelling techniques, the amount of data used to develop the acoustic models or the use of more robust language models.

B. Topic identification

Table III shows the closed-set topic identification error rates for LM, SVM, and SLDA approaches. The LM approached achieved an error rate of roughly 47%, for both the development and evaluations sets. In contrast, if the orthographic transcriptions were used instead of the recognised text, then the LM approach produced results of 11.14% and 12.96% for the development and evaluation sets, respectively. This highlights that the LM is extremely sensitive to the high WER delivered by the ASR systems.

TABLE III. Closed-set topic identification error rates obtained by the various topic classifiers.

Approach	Dev	Eval
LM (trans.)	11.14	12.96
LM	47.85	47.55
SVM Top 5	26.75	28.15
SVM Top 10	21.86	24.55
SVM Top 20	20.17	22.40
SLDA	17.1	17.6

The term-weighted SVM results show a consistent improvement, as the number of top words per topics were increased for each topic. A better performance may be achieved, if a greater number of top words per topic is chosen but this does introduce an increase in training and prediction times. The evaluation error rates are relatively close to the development set results, roughly 2% absolute, which implies the model parameters seem robust across differing datasets.

The SLDA approach produced the lowest error rate at 17.6%, which is significantly better than the LM approach. A significant improvement is also observed when compared to the SVM Top 20 approach, around 5% absolute.

V. Conclusion

The investigation into spoken audio topic identification has shown, that the system can make use of ASR recognisers, with high WER, and, still produce comparable topic identification error rates, using standard topic modelling and identification approaches, which is agreement with previously published work.

The language model topic classifier approach produced the worst results and is sensitive to the ASR recognition errors, which may be a result of poor topic-representative word sequences produced by the recognisers. The term-weighted SVM approach showed consistent improvements as the number of top words per topic were increased, but never out-performed the SLDA classifier, which gave the best results.

For the 0.1xRT experimental results reported in Wintrode and Kulp [7], the ASR system had a WER of 47% and produced a topic identification error rate of 19.2%. The SLDA results of 17.6%, at an ASR WER of 57%, are therefore comparable to that of the term-weighted SVM topic models. The best performing simplified term-weighted SVM (SVM

Top 20), used in this investigation, produced results that are also in the region of the topic identification error rates.

VI. Future work

The results presented by Wintrode and Kulp [7] show that there is a correlation between the WER and topic identification error rates. Therefore utilising better acoustic modelling techniques and more robust language models would help somewhat in reducing the topic identification error rates. Implementing ASR system adaptation, as used in Wintrode and Kulp [7], can also help to improve the performance.

During text processing of the conversations, a few interesting trends were noticed: the two callers were asked to discuss a certain topic for ten minutes, however, the following deviations were observed;

- Each call had an introduction and concluding phase not relevant to the topic at hand.
- The callers often drifted to different topics during the course of the conversation.

The topic modelling results show that the SLDA approach absorbed these artefacts quite well but employing text processing techniques, to the training and recognition texts, to partly isolate these regions may improve the results.

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] 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.

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

[5] 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.

[6] 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.

[7] 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.

[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] C. M. University, "The Cmu pronouncing Dictionary," 2014. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] J. Novak, D. Yang, N. Minematsu, and K. Hirose, "Initial and evaluations of an open source WFST-based phoneticizer," *The University of Tokyo, Tokyo Institute of Technology*.

[11] 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.

[12] 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//.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[14] 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.

[15] C. Wang, "Supervised latent Dirichlet allocation for classification," 2014. [Online]. Available: http://www.cs.cmu.edu/ chongw/slda/