This is a repository copy of *Automatic Genre and Show Identification of Broadcast Media*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/109228/

Version: Accepted Version

Proceedings Paper:
Doulaty, M., Saz, O., Ng, R.W.M. et al. (1 more author) (2016) Automatic Genre and Show Identification of Broadcast Media. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech). Interspeech 2016, 08-12 Sep 2016, San Francisco. ISCA .

https://doi.org/10.21437/Interspeech.2016

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Automatic Genre and Show Identification of Broadcast Media

Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain

Speech and Hearing Group (SpandH), Department of Computer Science, University of Sheffield

{mortaza.doulaty, o.saztorralba, wm.ng, t.hain}@sheffield.ac.uk

Abstract

Huge amounts of digital videos are being produced and broadcast every day, leading to giant media archives. Effective techniques are needed to make such data accessible further. Automatic meta-data labelling of broadcast media is an essential task for multimedia indexing, where it is standard to use multi-modal input for such purposes. This paper describes a novel method for automatic detection of media genre and show identities using acoustic features, textual features or a combination thereof. Furthermore the inclusion of available meta-data, such as time of broadcast, is shown to lead to very high performance. Latent Dirichlet Allocation is used to model both acoustics and text, yielding fixed dimensional representations of media recordings that can then be used in Support Vector Machines based classification. Experiments are conducted on more than 1200 hours of TV broadcasts from the British Broadcasting Corporation (BBC), where the task is to categorise the broadcasts into 8 genres or 133 show identities. On a 200-hour test set, accuracies of 98.6% and 85.7% were achieved for genre and show identification respectively, using a combination of acoustic and textual features with meta-data.

Index Terms: genre identification, show identification, broadcast media automatic labelling, latent Dirichlet allocation

1. Introduction

With the ever increasing amounts of digital media and requirements to process media archives, automatic labelling and classification of media recordings becomes more and more important. Multimedia data can be grouped by genre such as sports, news and comedy, which are categories that also imply other than purely semantic information. As such classification is easier to understand by viewers, is required for downstream processes such as indexing. Research in this field is pushed forward by initiatives such as the "MediaEval Benchmarking for Multimedia Evaluation" [1], or the "Robust, as Accurate as Human Genre Classification for Video" challenges within the Multimedia Grand Challenges of ACM Multimedia Conference [2]. Genre identification, and identification of shows can be considered as a core task in multimedia processing and is studied in this paper.

In a typical genre ID setting supervised methods learn from audio and/or video features extracted from the media streams. For audio-based classification mostly short-term features are used [3], such as Mel-Frequency Cepstral Coefficients (MFCC) [4]. The use of other features such as average speech rate, signal energy, zero crossing rate, duration of silence, noise and speech have also been studied [5]. Typical features extracted from video include colour statistics, camera motion and cut detection [5, 6, 7]. In the literature, audio based features usually have very similar performance compared to the video-based features [10]. Textual features such as subtitles and meta-data (e.g. title, tag, video description) contain semantic information and are believed to give promising results in genre ID [5].

This paper proposes new methods for automatic detection of media genre based on audio and explores what information sources are required to obtain very high levels of performance on a very large dataset of more than 1,200 hours of data. Also for the first time, to the best of our knowledge, the show identification task on very large datasets is studied in this paper.

This paper is organised as follows: Section 2 reviews the related work for genre identification. Section 3 describes the proposed method for genre and show identification, followed by the experimental setup in Section 4, results in Section 5 and a conclusion of this work in Section 6.

2. Related Work

Research on genre ID tasks typically report accuracies of over 90% [5, 6, 10, 11]. Typical datasets are the RAI dataset [11], Quaero dataset [12] and some custom YouTube videos. Both RAI and Quaero datasets are around 70 hours each and most of other datasets have similar or smaller sizes.

Genre labelling is difficult even for humans, mostly because of its subjectiveness. Labels of genres differ among datasets and this makes interpretations of results difficult. Also, the chosen labels do not always fully reflect multi-genre TV; for instance the RAI dataset has 7 genres labels. These 7 genres are cartoon, commercial, football, music show, news, talk show and weather forecast, which seem to be in some cases very specific, e.g. football which can be considered as a subset of a broader sport genre.

The proposed method in [10] uses acoustic features and using the RAI dataset, they reported accuracy of 94.3%. Using video, 99.2% was reported in [5] for the same dataset. For other similar datasets such as the Quaero dataset, similar classification accuracies are reported (e.g. 94.5% [5]). On a custom YouTube dataset [5], 87.3% was reported which was further improved by the use of meta-data to 89.7%

Genre ID can be addressed by using generative models. Kim et al. [10] reported an accuracy of 93.6% on a 11.5h test set with the RAI dataset using Gaussian Mixture Models (GMM) trained with the MFCC features. These features represent short-term characteristics of speech, such as the spectral properties of phonemes and speakers. In smaller and more homogeneous datasets where the same shows and speakers might often reoccur, the classification performance with those features are usually much better than the accuracies obtained on larger and more heterogeneous datasets [13].

The probabilistic approach using GMMs can be further extended using latent semantic indexing techniques. [10] had the accuracy improved by 0.7% absolute over their GMM baseline of 93.6% on the RAI dataset using acoustic topic models. They used vector quantisation to represent frames by discrete sym-

bols and trained Latent Dirichlet Allocation (LDA) models [14] followed by Support Vector Machine (SVM) classifiers. However when the amount of data is more and thus the dataset is more diverse, the same baseline models performs much worse [13].

Sageder et al. [15] tried to pool various types of features and then group and select a subset using canonical correlation analysis in order to identify low-correlated and complementary features. These features were then used to train different classifiers such as K-Nearest Neighbour, Random Forest and SVM. They reported very good classification performance on different datasets including some custom RAI and BBC shows, however the amount of data is tiny (less than 55h in total and in case of BBC, 4.5h with just 3 classes) and thus hard to directly compare with other approaches.

Other approaches try to identify certain audio-visual events, with the objective to model the semantics of the broadcast shows or YouTube videos [16, 17]. However, due to the complexity of the shows and videos, the performance of these techniques are not usually competitive with the previously mentioned methods.

3. Acoustic Latent Dirichlet Allocation

As shown in our previous work [18], acoustic LDA domain posteriors have a unique distribution across genres and shows. In this work we make use of acoustic LDA domain posterior features to classify broadcast media and investigate the use of other data sources such as subtitles, automatic speech recognition (ASR) output as well as meta-data.

LDA is an unsupervised probabilistic generative model for collections of discrete data. Since speech observations are continuous data, first it needs to be represented by some discrete symbols, here called acoustic words. A GMM with N mixture components is employed for this purpose. The index of Gaussian component with the highest posterior probability is then used to represent each frame with a discrete symbol. Frames of every acoustic document of length T, $\mathbf{d}_i = \{\mathbf{u}_1, ..., \mathbf{u}_t, ..., \mathbf{u}_T\}$ are represented as:

$$v_t = \arg\max_n P(G_n|\mathbf{u}_t), \quad 1 \leq n \leq N \tag{1}$$

Where G_n is a Gaussian component from a mixture of N components. With this new representation, document $\mathbf{d}_i$ is represented as $\tilde{\mathbf{d}}_i = \{v_1, ..., v_t, ..., v_T\}$. For each acoustic word v_t in each acoustic document $\tilde{\mathbf{d}}_i$, term frequency-inverse document frequency (tf-idf) can be computed as:

$$w_t = \text{tfidf}(v_t, \tilde{\mathbf{d}}_i, \tilde{\mathbf{D}}) = \text{tf}(v_t, \tilde{\mathbf{d}}_i)\ \text{idf}(v_t, \tilde{\mathbf{D}}) \tag{2}$$

Where $\tilde{\mathbf{D}}$ is the set of all acoustic documents represented with acoustic words. With each document now represented with tf-idf scores as $\bar{\mathbf{d}}_i = \{w_1, ..., w_t, ..., w_T\}$, the LDA models can be trained.

A graphical representation of the LDA model is shown at Figure 1, as a three-level hierarchical Bayesian model. In this model, the only observed variables are w_t's. α and β are dataset level parameters, $\theta_{\tilde{\mathbf{d}}_i}$ is a document level variable and z_t is a latent variable indicating the domain from which w_t was drawn. The following joint distribution is the result of the generative process of LDA:

$$p(\theta, \mathbf{z}, \bar{\mathbf{d}}|\alpha, \beta) = p(\theta|\alpha) \prod_{t=1}^{T} p(z_t|\theta) p(w_t|z_t, \beta) \tag{3}$$

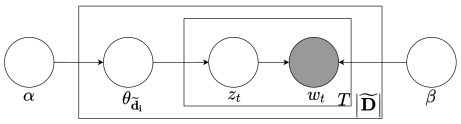

Figure 1: Graphical model representation of LDA

The posterior distribution of the latent variables given the acoustic document and α and β parameters is:

$$p(\theta, \mathbf{z}|\bar{\mathbf{d}}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \bar{\mathbf{d}}|\alpha, \beta)}{p(\bar{\mathbf{d}}|\alpha, \beta)} \tag{4}$$

Computing $p(\bar{\mathbf{d}}|\alpha, \beta)$ requires some intractable integrals. A reasonable approximate can be acquired using variational approximation, which is shown to work reasonably well in various applications [19]. The approximated posterior distribution is:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{t=1}^{T} q(z_t|\phi_t) \tag{5}$$

where γ is the Dirichlet parameter that determines θ and ϕ is the parameter for the multinomial that generates the latent variables.

Training minimises the Kullback-Leiber Divergence between the real and the approximated joint probabilities (equations 4 and 5) [19]:

$$\arg\min_{\gamma,\phi}\ \text{KLD}\left(q(\theta, \mathbf{z}|\gamma, \phi)\ ||\ p(\theta, \mathbf{z}|\bar{\mathbf{d}}, \alpha, \beta)\right) \tag{6}$$

The posterior Dirichlet parameter $\gamma(\bar{\mathbf{d}})$ can be used as feature for classification. Discriminative classifiers such as SVMs have been used successfully for genre classification tasks before [10, 20] including our previous work [13].

Kim et al. [10] used the whole shows to train the LDA models and used the domain posteriors as features for an SVM classifier. In this work we followed our previous setup [18, 21] where only speech segments are used to train the LDA model. For each show, the domain posteriors of its segments were accumulated and length normalised and used as features for the discriminative classifier in the later stage:

$$\mathbf{x}_i = \frac{1}{\sum_{s \in segs} \text{len}(s)} \sum_{i \in segs} \text{len}(i)\ \gamma(\bar{\mathbf{d}}_i) \tag{7}$$

4. Experimental Setup

4.1. Data

TV broadcasts provided by the British Broadcasting Corporation (BBC) were used for all experiments. The data is identical to the one defined and provided for the 2015 Multi-Genre Broadcast (MGB) Challenge [22] with a different training/testing set definitions. The shows were chosen to cover the full range of broadcast show types and categorised in 8 genres: advice, children's, comedy, competition, documentary, drama, events and news. All shows were broadcast by the BBC during 6 weeks in April and May 2008. There were more then 2,000 shows in the original MGB challenge data, from which 1,789 shows were selected for the experiments, 1,501 shows for the training set and 288 shows for test set, with 133 unique shows in total. The distribution of shows (time and count) across genres for the training and test data is shown in Table 1. Figure 2

Table 1: Amount of training and testing data per genre

Genres	Train Set		Test Set	
	# Shows	Dur (h)	# Shows	Dur (h)
Advice	189	135.3	35	24.4
Children's	301	112.7	60	25.0
Comedy	90	44.1	22	10.8
Competition	224	153.3	45	29.8
Documentary	90	57.4	29	19.3
Drama	102	69.0	21	14.6
Events	98	161.0	21	36.3
News	407	293.0	55	40.2
Total	**1501**	**1025.6**	**288**	**200.4**

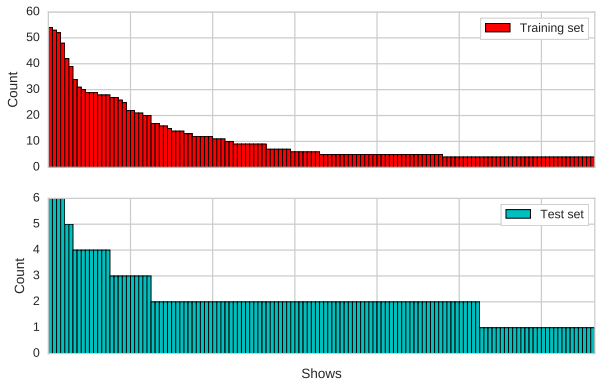

Figure 2: Distribution of 133 unique shows in training and test set

shows the distribution of the 133 unique shows for both training set and test set, where the horizontal axis represents unique shows and the vertical axis represents the number of episodes in that show. Order of the bars are identical in both plots and e.g. the first bar of both plots represents the same show.

It is important to note that this dataset is by orders of magnitude larger than most of the datasets used in the literature for the genre ID task [2, 4, 5, 6, 10, 15].

4.2. Baseline

As a baseline, GMM classifiers were used for both genre and show identification tasks. For the data as described above, genre ID task has 8 target classes and show ID task has 133 target classes. 13 dimensional PLP [23] features plus their first and second derivatives were used to train the genre-based and show-based GMMs using Expectation Maximisation algorithm and mix-up procedure to reach 512 mixtures. The optimal number of mixtures for a similar task was found to be 512 in our previous experiments [13]. Table 2 shows the classification accuracy for both tasks. Since there are fewer target classes, genre ID should be an easier classification task compared to show ID. However, GMMs are found to perform better in classifying shows than genres (70.1% compared to 61.5%), one reason for this could be the diversity of data as discussed in the introduction and the fact that PLP features are good for representing speaker specific characteristics [13] and for the show ID task the GMMs are learning speakers in re-occurring episodes. However they provide poor generalisation for the genre ID task. If show to genre mapping is assumed to be *a priori* knowledge, then the show ID GMMs can be used for the genre ID task. The accuracy for genre ID in such a setting would be 79.2%.

Table 2: Genre/show classification accuracy with GMMs

Model	Genre ID	Show ID
GMM	61.5 (79.2)	70.1

Table 3: Genre/show classification accuracy using whole show and segment based acoustic LDA models

#Domains	Whole Show		Segment Based	
	Genre ID	Show ID	Genre ID	Show ID
16	73.6	45.1	76.7	46.7
32	71.9	53.8	81.5	57.8
64	78.1	56.6	81.2	63.4
128	77.8	56.9	83.3	66.6
256	76.4	58.0	86.4	67.3
512	80.2	61.8	85.0	66.7
1024	77.1	65.3	85.7	63.8
2048	80.6	65.3	84.7	63.1

5. Results

5.1. Whole Show and Segment Based Acoustic LDA

Whole shows were used to train the LDA models with varying number of latent domains with the same procedure outlined in the previous section. The performance of these models is to be compared with the proposed segment based LDA models. The classification accuracy for the genre ID and show ID tasks are presented in Table 3. For the segment level models the posterior estimates on short segments can be noisy. Picking the domain with the highest posterior probability and representing the posterior vector as one-hot-vector may reduce the posterior estimate noise and it was found to slightly outperform the base case and was used in the experiments.

As the performance of segment level models was better than the whole show models, they were used in the rest of experiments. Segment based models also had higher accuracy with fewer latent domains. E.g. the highest accuracy with the segment based models for genre ID was 86.4% obtained with an LDA model with 256 latent domains. However, the best performance for the whole show models was 80.6%, with 2048 latent domains. A similar pattern was found for the show ID task as well.

5.2. Text Based LDA

Transcripts of the shows have valuable information for discrimination of genres and shows. In this section the classification is studied based on solely textual features. BBC TVs provide subtitles of the TV soundtrack, mostly for helping deaf and hard-of-hearing viewers. The quality of these subtitles varies considerably by genres. For example subtitles of live events and news are mostly re-spoken live ASR output and have higher errors, however for other genres which does not have the live nature, the quality is higher. For a detailed analysis of the subtitles quality refer to [22] and [24]. Subtitles were used as-is, without any preprocessing, to train the classifiers for both tasks. Although subtitles can be of varied quality, their correctness is still high. In a second experiment, ASR output is used instead of subtitles. The ASR systems used here were trained for participation in the MGB Challenge. For more details about these ASR systems, refer to [24] and [25]. The classification task here is similar to a document classification task, where each show's transcript is a document and the classes are either genres or shows. To have a

Table 4: Genre/show classification accuracy using text based LDA models

#Domains	Subtitles		ASR Output	
	Genre ID	Show ID	Genre ID	Show ID
16	77.4	41.3	70.1	29.2
32	81.3	50.7	71.9	34.0
64	85.4	62.1	81.6	45.8
128	89.2	68.8	87.5	55.2
256	91.0	77.1	88.2	65.6
512	91.0	76.7	87.9	63.9
1024	94.8	81.3	88.5	64.9
2048	96.2	79.9	89.9	64.9
4096	93.1	78.1	89.6	64.2

fair comparison with the acoustic LDA experiments, text based LDA models were trained and the domain posteriors were used as features in the SVM classifiers. A simpler approach would be SVMs with tf-idf features directly. However here the LDA model reduces the dimensionality of the tf-idf features to the number of latent domains, which is known to work better than tf-idf only features for document classification [19]. Table 4 summarises the results. LDA models trained with the subtitles performed substantially better than models trained on the ASR output. Note that the ASR models used here have around 30% WER on the official development set of the MGB challenge. The performance gap is even wider in case of the show ID task, 22.6% vs. 13.5% absolute difference. This could caused by some specific names that were present in the subtitles, but not in the ASR output. Such words may have considerable discriminability information.

The overall performance of text based classification with subtitles is generally better than with direct audio based classification (96.2% vs. 84.4% for the genre ID task and 81.3% vs. 67.3% for the show ID task) but when considering the ASR output only, the audio based classification is better for the show ID task.

5.3. Using Meta-Data

The data used in the experiments also includes some meta-data, such as the BBC broadcast channel number, the date and time of broadcast, and other unstructured information. Using some of the structured meta-data is studied next to learn how the classification accuracy can be improved further. Since these programmes were broadcast during 6 weeks in April and May 2008, using the date was not likely to be helpful which we verified in the experiments. Instead, the time of broadcast, splitting 24 hours into 8 chunks, and channel number, in this setup 1–4 corresponding to BBC1, BBC2, BBC3 and BBC4, were appended as one-hot-vectors to the inputs of the SVM classifiers and their effect is studied. Table 5 summarises the results of using the meta-data together with acoustic LDA features. Adding these meta-data helps for both tasks. When comparing channel and time, in both tasks appending time helps more and the difference is bigger in case of the show ID task (72.8% vs. 77.7%). Combining channel information and time of broadcast also helps further improve the classification accuracy in both tasks and overall with meta-data there is 5.9% and 15.3% absolute improvement in accuracies of genre ID and show ID tasks. The first row in Table 5 shows the accuracy when only meta-data is used (without any acoustic or textual features) which shows how much information with the meta-data is provided.

Table 5: Genre/show classification accuracy using meta-data

Meta-Data	Genre ID	Show ID
Only Channel & Time	46.7	22.0
Baseline (acoustic 256)	86.4	67.3
+ Channel	89.6	72.8
+ Time	89.9	77.7
+ Channel & Time	92.3	82.6

Table 6: Genre/show classification accuracy with system fusion

Method	Genre ID	Show ID
Baseline (acoustic 256)	86.4	67.3
Baseline (text 2048)	96.2	79.9
Acoustic & Text	97.2	85.0
Acoustic + Meta-data & Text	98.6	85.7

5.4. System Fusion

With the two systems based on acoustic and textual features, one can use a combination of both, assuming that they will make different classification errors and their outputs are complimentary. To combine the scores of the systems, logistic regression is used to find a linear combination of individual system scores to maximize the probability of correct classification [26]. Table 6 shows the classification accuracy with the system fusion. The combination of acoustic and text based systems improves the classification accuracy for both tasks, 97.2% and 85.0% accuracy for genre ID and show ID respectively, which shows the complementarity of the individual systems. Moreover, including meta-data further improves the accuracy to 98.6% and 85.7% which is near perfect for the genre ID task.

6. Conclusions

In this paper new methods for the genre classification of broadcast media based on audio were proposed. Furthermore, required information sources to obtain very high levels of performance was explored. Also for the first time, show classification task on very large datasets was studied. For the experiments more than 1,200 hours of data with more than 1,500 TV shows from the BBC which was broadcast in 2008 was used. These data was a part of the MGB 2015 challenge [22]. For the genre ID task there were 8 classes and for the show ID task there were 133 classes. Acoustic and textual LDA models were trained with the audio and subtitles to infer the posterior Dirichlet parameters which were then used in SVM classifiers to classify the genres and shows. On a 200h test set, combination of both acoustic and text based classifiers had accuracy of 97.2% and 85.0% for genre ID and show ID tasks respectively. Use of meta-data such as time of broadcast further improved the accuracies to 98.6% and 85.7%.

Future work can be exploiting more information from the unstructured meta-data and trying to deal with cases where some meta-data is missing.

7. Acknowledgements

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). All the data related to the MGB challenge, including audio files and subtitle text is available via special license with the BBC on `www.mgb-challenge.org`. Dataset definitions and other related files are available with DOI `10.15131/shef.data.3457541`

8. References

[1] M. Larson, X. Anguera, T. Reuter, G. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, and M. Soleymani, "Indexing multimedia documents with acoustic concept recognition lattices." in *Proc. of MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain, 2013.

[2] "Multimedia Grand Challenge (2009, 2010)." [Online]. Available: http://comminfo.rutgers.edu/conferences/mmchallenge

[3] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model," in *Proc. of Multimedia Signal Processing Workshop*, 1998, pp. 27–32.

[4] M. Roach and J. S. Mason, "Classification of video genre using audio," in *Proc. of Interspeech*, Aalborg, Denmark, 2001.

[5] H. K. Ekenel and T. Semela, "Multimodal genre classification of TV programs and YouTube videos," *Multimedia tools and applications*, vol. 63, no. 2, pp. 547–567, 2013.

[6] M. Montagnuolo and A. Messina, "Parallel neural networks for multimodal video genre classification," *Multimedia Tools and Applications*, vol. 41, no. 1, pp. 125–159, 2009.

[7] I. Mironica, B. Ionescu, P. Knees, and P. Lambert, "An in-depth evaluation of multimodal video genre categorization," in *Proc. of Content-Based Multimedia Indexing (CBMI) Workshop*, Veszprem, Hungary, 2013.

[8] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[9] R. W. M. Ng, M. Doulaty, R. Doddipatla, O. Saz, M. Hasan, T. Hain, W. Aziz, K. Shaf, and L. Specia, "The USFD spoken language translation system for IWSLT 2014," in *Proc. of IWSLT*, Lake Tahoe NV, USA, 2014.

[10] S. Kim, P. Georgiou, and S. Narayanan, "On-line genre classification of TV programs using audio content," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[11] M. Montagnuolo and A. Messina, "TV genre classification using multimodal information and multilayer perceptrons," in *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*. Springer, 2007, pp. 730–741.

[12] "Quaero programme website," 2011. [Online]. Available: http://www.quaero.org

[13] O. Saz, M. Doulaty, and T. Hain, "Background-tracking acoustic features for genre identification of broadcast shows," in *Proc. of SLT*, Lake Tahoe NV, USA, 2014.

[14] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proc. of WASPAA*, New Paltz NY, USA, 2009.

[15] G. Sageder, M. Zaharieva, and C. Breiteneder, "Group feature selection for audio-based video genre classification," in *MultiMedia Modeling*. Springer, 2016, pp. 29–41.

[16] K. Lee and D. P. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1406–1416, 2010.

[17] D. Castan and M. Akbacak, "Indexing multimedia documents with acoustic concept recognition lattices." in *Proc. of Interspeech*, Lyon, France, 2013.

[18] M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, "Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation," in *Proc. of ASRU*, Arizona, USA, 2015.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[20] M. Rouvier, D. Matrouf, and G. Linares, "Factor analysis for audio-based video genre classification." in *Proc. of Interspeech*, Brighton, UK, 2009.

[21] M. Doulaty, O. Saz, and T. Hain, "Unsupervised domain discovery using latent dirichlet allocation for acoustic modelling in speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[22] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Webster, and P. Woodland, "The MGB Challenge: Evaluating multi-genre broadcast media recognition," in *Proc. of ASRU*, Arizona, USA, 2015.

[23] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[24] O. Saz, M. Doulaty, S. Deena, R. Milner, R. W. M. Ng, M. Hasan, Y. Liu, and T. Hain, "The 2015 Sheffield system for transcription of multi-genre broadcast media," in *Proc. of ASRU*, Arizona, USA, 2015.

[25] R. Milner, O. Saz, S. Deena, M. Doulaty, R. Ng, and T. Hain, "The 2015 Sheffield system for longitudinal diarisation of broadcast media," in *Proc. of ASRU*, Arizona, USA, 2015.

[26] N. Brummer, "FoCal toolkit for evaluation, fusion and calibration of statistical pattern recognisers," 2010. [Online]. Available: https://sites.google.com/site/nikobrummer/focal