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 [8]. 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, 8, 9]. Typical datasets are the RAI dataset [9], Quaero dataset [10] 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 [8] 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% [9]). On a custom YouTube dataset [8, 9], 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. [8] 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 recur, 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. [8] 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 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 collections of discrete data. Since speech observations are continuous and thus the dataset is more diverse, the same baseline models performs much worse [13].

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 posteriors as features for classification. Discriminative classifiers such as SVMs have been used successfully for genre classification tasks before [8,20] including our previous work [13].

Kim et al. [8] 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] 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:
### 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 2. 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
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. 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 be 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 6 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.

### 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. 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. 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% accuracy 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). The audio and subtitle data used for these experiments was distributed as part of the MGB Challenge (mgb-challenge.org) through a licence with the BBC.
8. References

[1] M. Larson, X. Anguera, T. Reuter, G. Jones, B. Ionescu, M. Schedl, T. Patrak, 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] S. Kim, P. Georgiou, and S. Narayanan, “On-line genre classification of TV programs using audio content,” in Proc. of ICASSP, Vancouver, Canada, 2013.

[9] 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.

[10] “Quaero programme website,” 2011. [Online]. Available: http://www.quaero.org

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

[12] 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.

[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. Breitenieder, “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, “FoCa! toolkit for evaluation, fusion and calibration of statistical pattern recognisers,” 2010. [Online]. Available: https://sites.google.com/site/nikobrummer/focal