AUTOMATIC LANGUAGE IDENTIFICATION IN MUSIC VIDEOS WITH LOW LEVEL AUDIO AND VISUAL FEATURES

Vijay Chandrasekhar, Mehmet Emre Sargin, David A. Ross

Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043

ABSTRACT

Automatic Language Identification (LID) in music has received significantly less attention than LID in speech. Here, we study the problem of LID in music videos uploaded on YouTube. We use a “bag-of-words” approach based on state-of-the-art content based audio-visual features and linear SVM classifiers for automatic LID. Our system obtains 48% accuracy for a corpus of 25000 music videos and 25 different languages.

Index Terms— automatic language identification, LID in music, audio-visual features

1. INTRODUCTION

Automatic Language Identification (LID) in spoken language processing has received a lot of attention. However, there is little work in the field of LID in songs and music videos. LID in music is important as it enables better categorization of audio and video collections. Often, the language of the audio or video title in the collection is not the language for the video, e.g., a song sung in Mandarin might have an English title. In this case, analysing the contents of the audio or video can be useful for better categorization. In this work, we focus on automatic LID of music videos.

There is a lot of work in LID for speech - see [1, 2, 3, 4, 5] for review and discussion of techniques used in this field. The classical approach for LID involves tokenization combined with phonotactic analysis [1]. Torres-Carrasquillo et al. [6] use Gaussian Mixture Models (GMM) with Shifted Delta Cepstral (SDC) coefficients for good performance. Another approach introduced by Campbell et. al [4] involves Support Vector Machines (SVM) with SDC features. Campbell et al. combine SVMs with GMM based techniques to improve performance beyond GMM-only based approaches [4]. Current state-of-the-art involves using GMM supervectors and SVM [5]. The basic idea here is to adapt a universal background model GMM on a per utterance basis and then use the resulting shift in means. The stacked adapted means form a GMM supervector, which is used for classification with SVMs. The GMM supervector technique was initially used for speaker recognition, but subsequently applied to LID too for good performance [5]. Campbell improves the performance of the GMM supervector technique in [7].

Intuitively, techniques used for LID in speech can be applied to music too. However, singing differs from speech in many ways [8]. Some key differences include extensive interference from background music, multiple noise sources and atypical music lyrics. Further, there are several phonological modifications in music made by singers, compared to conventional speech. Tsai and Wang [8] tackle some of these challenges in their early work on distinguishing between English and Mandarin songs. Tsai and Wang segment music into vocal/non-vocal segments, train vocabularies on spectrum-based features and use a sequence of phonological units for classification. The authors show that manual vocal/non-vocal segmentation improves performance compared to performing no segmentation at all. The authors also propose automatic techniques for vocal/non-vocal segmentation, which, however, do not provide the improvement in classification accuracy that manual segmentation does. A peak accuracy of 70% is obtained with automatic segmentation for 2 language classes: English and Mandarin.

Schwenninger et al. [9] propose using Mel Frequency Cepstrum Coefficients (MFCC) for classifying English and German songs, and obtain 64% accuracy. Schwnninger et al. experiment with several pre-processing techniques for obtaining the speech portions of the music, e.g., detecting energy in high frequency regions [10], distortion reduction of music from drums and bass guitars [11], and azimuth discrimination [12]. However, contrary to the results presented by Tsai et al. [8], none of these pre-processing techniques improve performance.

Jacob and Cox [13] focus on LID in videos using visual-only features. The authors segment the video and use lip-shape features, appearance and motion for classification. Such an approach would work for highly structured videos like news-clips, but not work for music videos.

Vijay Chandrasekhar is pursuing a Ph.D student in the Electrical Engineering Department at Stanford University. The work was done while at Google, Inc.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

In our work, we use a "bag-of-words" approach for classification with several state-of-the-art audio and visual features. Automatic segmentation of the music video into vocal/non-vocal is non-trivial, while manual segmentation is not feasible for large data sets. Further, based on prior work [9, 8], it is not clear whether vocal/non-vocal segmentation helps improve classification accuracy. As a result, we do not perform any segmentation in this work. Our contributions in this work are as follows:

- The authors in prior work [8, 9] consider only 2 languages for classification. In this work, we consider a large-scale data set with 25000 music videos and 25 languages.
- Prior work deals only with songs. Here, we consider songs along with their videos for classification.
- We combine both audio and visual features for classification, and show that it improves classification accuracy.

In Section 2, we describe our approach, and in Section 3, we discuss the experimental setup, training phase and test results. Finally, we discuss directions for future work to improve LID in music videos.

2. APPROACH

In this section, we describe the audio and visual features extracted on each video. Specifically, we consider user-generated playlists with titles of the format "Language songs" e.g. "English songs", "Arabic songs", etc. These are used to obtain a corpus of 25000 videos for 25 different languages, 1000 videos per language. The languages represented are: Arabic, Bangla, Chinese, English, French, German, Greek, Hindi, Irish, Italian, Japanese, Khmer, Korean, Malay, Malayalam, Nepali, Pashto, Punjabi, Russian, Sinhala, Spanish, Tagalog, Tamil, Telugu and Thai.

Song videos typically vary in length. For each video, we wish to create a feature of fixed size regardless of its length. One effective technique [14] is to generate a descriptor for each frame and map the set of descriptors to a histogram. We generate codebooks offline with training data using k means vector quantization, and each descriptor is quantized to the nearest codeword. Each histogram is normalized so that the sum of values in all bins is 1. The final feature vector is obtained by concatenating the histograms of each feature descriptor. Next, we discuss the audio and visual descriptors computed for each video. The size of codebooks for different features is listed in Tab. 1.

Audio Spectrogram and Volume: For audio features, the frame rate is set to 100 frames per second. For each audio frame, we compute a 32-bin audio spectrogram. In addition, we compute the volume of the audio stream, which is represented as a single floating point value.

Feature	Size of codebook
Audio Volume	64
Spectrogram	1024
MFCC	2000
SAI	28×256
Global Visual	1858
Motion cuboids - Pixel PCA	512
Motion cuboids - HoG	647

Table 1. Size of codebook for each feature.

Mel-Frequency Cepstral Coefficients (MFCC): For each audio frame, we generate the standard set of MFCC coefficients.

Stabilized Auditory Images (SAI): The auditory features that we use are based on models of the mammalian auditory system. Specifically, we use a cochlear-model filterbank followed by a correlation process that makes a stabilized auditory image (SAI) [15]. Computing the SAI starts with a set of band-pass filters, followed by an autocorrelation of each channel. This data is then vector-quantized at different scales to create a histogram. The histogram implicitly characterizes several aspects of music and speech of the audio track. For a detailed description of the features, please refer to the work by Lyon et. al. [15] which uses these features for ranking and retrieval of sound files. Since the SAI is very high dimensional, we divide the feature into 28 smaller blocks (see Tab. 1) and perform vector quantization on each one.

Global Visual Features: We list the set of global visual features computed on videos. An important visual feature is an 8x8 hue–saturation histogram. This captures how colors vary over the duration of the video, and act as a relatively strong contextual prior for the classifiers when combined with other local visual features. Another feature we compute is the output of a face detector. We compute several statistics based on it: the ratio of the largest face to the area of the image, number of faces, and various statistics based on the skin pixels. In addition, we compute textons for each video frame. For more details of global visual features, readers are referred to [16].

Motion Cuboids: For characterising motion in the video, we first compute spatio-temporal interest points using the detector proposed by Dollar et al. [17]. Next, we extract raw pixel 13×13×19 "cuboids" around spatial-temporal interest points. We compute two descriptors around these interest points: (1) We take the raw cuboid pixels and reduce the dimensionality using Principal Component Analysis (PCA) to 256. (2) At every image pixel in the cuboid, we extract an 1800-dimensional descriptor made up of 100 overlapping Histograms of Oriented Gradients (HOG) [18]. The descriptors are quantized into a bag-of-words representation using fast randomized decision trees [19].

Feature	Accuracy (%)
Baseline (random)	4.0
Audio spectrogram	19.6
MFCC	26.1
SAI	37.7
All audio	44.7
All video	14.3
All audio + video	47.8

Table 2. Classification accuracy for different features. We obtain a peak accuracy of 47.8% with all audio and visual features. Adding visual features improves performance by 3.1% compared to using just audio features.

Final feature vector: The final feature vector is obtained by concatenating all the histograms of audio-visual features described in this section.

3. EXPERIMENTAL RESULTS

We divide the corpus of 25000 music videos into training and test sets of 18750 and 6250 respectively. We train a set of "one-vs-all" linear SVMs for each language category. Each classifier learns to separate videos that belong to a certain language category from those that don't. The classifier with the highest score is treated as the output category for a video. The classification accuracy stated in Tab. 2 is obtained by considering the sum of all diagonal elements of the confusion matrix, divided by the sum of all elements.

First, we list the accuracy for different feature sets in Tab. 2 averaged across all languages. We note that we can achieve close to 50% classification accuracy for the 25 different language categories—a number much higher than chance. Using audio-only features, we achieve 44.7% accuracy. The SAI feature provides the highest accuracy amongst the different audio features, if considered individually. There's a significant gap between SAI, and conventional audio features like MFCCs for music LID. Adding visual features improves the accuracy by $\sim$3% averaged across all languages, resulting in a net accuracy of 47.8%. Thus, visual features help in LID as hypothesized.

Next, we provide a breakdown of classification accuracy for different languages in Tab. 3, using all audio and visual features. The languages are sorted in the increasing order of classification accuracy, from lowest to highest. The Pashto language performs the best, with the peak classification accuracy of 79%.

Finally, in Tab. 4, we present the improvement in classification accuracy using visual features over using audio-only features for different languages. The highest improvement is observed for Thai, with an increase of 10% accuracy using visual features.

CA (%)	Languages
<30	English, French, German, Tamil
30-40	Spanish, Hindi, Italian, Russian
40-50	Chinese, Bangla, Tagalog, Greek
50-60	Telugu, Sinhala, Punjabi, Korean Malay, Irish, Thai, Japanese
60-80	Nepali, Malayalam, Arabic, Khmer, Pashto

Table 3. Classification Accuracy (CA) for different languages. The languages are sorted in the order of increasing classification accuracy.

Increase in CA (%)	Languages
<1	Spanish, Malay, German, Arabic Khmer, Pashto, Telugu, Tamil
1-2	Irish, Punjabi
2-4	Russian, Sinhala, Italian, Korean
4-5	Bangla, Japanese, Hindi, Tagalog
>5	English, Malayalam, Greek, Nepali French, Chinese, Thai

Table 4. Increase in Classification Accuracy (CA) for different languages using visual features.

The results presented here are promising, as a peak accuracy of close to 50% is achieved for 25 different language categories. However, these results are still preliminary and a lot more can be done to improve performance. In future work, we plan to

- Explore pre-processing techniques like segmentation of music videos into vocal/non-vocal segments and distortion reduction of background music.
- After vocal/non-vocal segmentation and distortion reduction, apply state-of-the-art GMM techniques from speech-LID and compare its performance to current "bag-of-words" approach.
- Combine techniques from state-of-the-art speech-LID and current "bag-of-words" approach.
- Build higher level grammatical models from low level features to improve performance of music-LID.
- Learn language models in a hierarchical fashion, i.e., divide languages into high-level groups like Romanic, Slavic, etc., and then build models for each sub-group.

4. ACKNOWLEDGEMENT

We would like to acknowledge Dennis Strelow, Weilong Yang, George Toderici, Hrishikesh Aradhye, Chris Alberti, Thomas C. Walters and Richard Lyon for their input.

5. CONCLUSION

We present preliminary results for automatic LID for a large set of music videos uploaded on Youtube. We use a "bag-of-words" approach based on state-of-the-art content based audio-visual features and linear SVM classifiers for automatic music-LID. Our system obtains 48% accuracy for a large corpus of 25000 music videos and 25 different languages (compared to 4% for chance). We observe that SAI features and visual features provide a significant improvement in performance over using conventional audio spectrograms and MFCCs. Future work will focus on combining techniques from speech-LID and music-LID for improving performance.

6. REFERENCES

[1] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31, Jan. 1996.

[2] J. Navratil, "Spoken language recognition - a step toward multilinguality in speech processing," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 678–685, Sep. 2001.

[3] Pavel Matejka, "Review of automatic language identification," .

[4] W. Campbell, E. Singer, P.A.Torres-Carrasquillo, and D.A.Reynolds, "Language recognition with support vector machines," in *Proc. of Odyssey: The Speaker and Language Recognition Workshop, ISCA*, Toledo, Spain, June 2004.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Acoustic language identification using fast discriminative training," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.

[6] P.A.T.-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. of International Conference on Spoken Language Processing*, Denver, USA, September 2002.

[7] W. M. Campbell, "A covariance kernel for svm language recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, USA, March 2008.

[8] W. Tsai and H. Wang, "Towards automatic identification of singing language in popular music recordings," in *Proc. of the 5th International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 568–576.

[9] J. Schwenninger, R. Brueckner, D. Willett, and M. Hennecke, "Language identification in vocal music," in *Proc. of the 7th International Symposium on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006.

[10] T. L. Nwe and Y. Wang, "Automatic detection of vocal segments in popular songs," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 138–145.

[11] K. West and S. Cox, "Finding an optimal segmentation for audio genre classification," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[12] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *Proc. of International Conference on Digital Audio Effects*, Naples, Italy, October 2004.

[13] J. L. Newman and S. J. Cox, "Speaker independent visual-only language identification," in *Proc. of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, Texas, 2010.

[14] M. Pasca L. Sbaiz J. Yagnik G. Toderici, H. Aradhye, "Finding meaning on youtube: Tag recommendation and category discovery," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, SFO, USA, 2010.

[15] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik, "Sound retrieval and ranking using sparse auditory representations," in *Neural Computation*, 2010, vol. 22, pp. 2390–2416.

[16] H. A. Rowley, Y. Jing, and S. Baluja, "Large scale imagebased adult-content filtering," in *Proc. of International Conference on Computer Vision Theory and Applications (VISAPP)*, February 2006.

[17] P. Dollr, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of VS-PETS*, 2005, pp. 65–72.

[18] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, June 2005.

[19] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008.