# Language detection in audio content analysis

**Conference Paper** *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · March 2008

DOI: 10.1109/ICASSP.2008.4518058 · Source: DBLP



**3 authors**, including:

Vikramjit Mitra
Apple Inc.
**124** PUBLICATIONS **2,328** CITATIONS
SEE PROFILE

Carol Espy-Wilson
University of Maryland, College Park
**224** PUBLICATIONS **4,934** CITATIONS
SEE PROFILE

# LANGUAGE DETECTION IN AUDIO CONTENT ANALYSIS

*Vikramjit Mitra*[1], *Daniel Garcia-Romero*[2], *Carol Y. Espy-Wilson*[3]

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD
[1]vmitra@umd.edu, [2]dgromero@umd.edu, [3]espy@umd.edu

## ABSTRACT

Experiments have shown that Language Identification systems for telephonic speech using shifted delta cepstra as the feature set and Gaussian mixture models as the backend, offers superior performance than other competing techniques. This paper aims to address the task of Language Identification for audio signals. The abundance of digital music from the Internet calls for a reliable real-time system for analyzing and properly categorizing them. Previous research has mainly focused on categorizing audio files into appropriate genres; however genre types vary with language. This paper proposes a systematic audio content analysis strategy by initially detecting whether an audio file has any vocals present in it and, if present, then detecting the language of the song. Given the language of the song, genre detection becomes a closed set classification problem.

***Index Terms***— Language Identification, Audio Content Analysis, Gaussian Mixture Model, GMM-supervector.

## 1. INTRODUCTION

Extensive use of audio/visual files over the internet has resulted in customization of search engines to incorporate audio/video file searching. ISO standardization [1] for multimedia file content may not be coherent since users can alter the fields; hence audio file search engines that rely upon such information may not be reliable. Automated genre classification addresses this issue to reliably categorize audio files into respective genres by analyzing the audio signal content. However the set of possible genres vary by language. For example, the genre 'Rabindra-sangeet' and 'Adhunik' are unique to Bengali audio files. Likewise, there are other genres that are unique to a particular language or a set of languages, such as 'HipHop' is common to English and Spanish, but does not exist in Bengali. This necessitates the need to detect the language of an audio sample prior to its genre detection.

Automatic Language Identification (LID) systems have been primarily used for detecting the language of telephonic speech. It has been shown [2] that phonotactic content based LID systems offers superior results; however they suffer from computational complexity. Torres-Carrasquillo [2] et al. has reported that Shifted Delta Cepstral (SDC) [6] coefficients can be used as the feature set for constructing Gaussian Mixture Model (GMM) based LID systems. Manual audio content analysis requires huge effort [3]. Microsoft® employed 30 musicologists for one year to perform such a task, which claims the necessity of an automated procedure. Pachet et al [4] have shown that a taxonomical description of audio genres is a difficult task. However most of the initial results have been devoted on categorizing audio samples belonging to a specific language.

To address a systematic description of an audio file, the first thing that needs to be known is whether the file has any vocals in it or is purely instrumental. If vocals are present then the knowledge about the language of an audio file is of paramount importance. Not only does it determine the set of possible genres for that file, but it also gives some insight about the origin of the file. The present work reports the performance of a Mel Frequency Cepstral Coefficient (MFCC) based GMM classifier that automatically detects whether a given audio file has any vocals in it or is purely instrumental. Given that an audio file has vocals, the LID system detects the language of the song. Different feature sets are considered as input to the GMM based LID system. A GMM-supervector feature set [11, 12] is also constructed, which is used as the input to a Linear kernel Support Vector Machine (SVM) and Least Square Support Vector Machine (LS-SVM). A fused GMM-SVM LID system is also developed, which offers the best performance. The results claim that despite the presence of background music, the proposed approach offers a high degree of accuracy in detecting the language of an audio file.

The organization of the paper is as follows: Section 2 presents the corpora used, Section 3 presents the Vocals-Instrumental Detector (VID). Section 4 present the GMM based LID system. Section 5 presents the supervector-SVM based LID system and the SVM-GMM fused model, followed by Section 6, which presents conclusion and future work.

## 2. AUDIO CORPORA

There are no standard corpora available, which can be used to perform LID experiments on audio files. The corpora used in this paper was obtained manually, where 1358 audio segments of approximate duration 10 seconds were created,

out of which 186 segments were purely instrumental. The remaining 1172 segments belonged to 6 different languages: English (ENG), Bengali (BEN), Hindi (HIN), Spanish (ESP), Russian (RUS) and Chinese (CHN). Their distribution is given in Table 1. Approximately 67% of the files were used for training and the rest for testing. The sampling frequency of the files varied from 44.1 KHz to 11.025 KHz. However the sampling frequency is reduced to 8 KHz, as it was observed that the vocals dominated over the background music within 0 to 4 KHz. The major difference between LID systems in conversational speech and audio files is the presence of background music acting as colored noise in the later case, which increases the difficulty of the problem. This difficulty can be circumvented by considering the bandwidth where vocals dominate, as well as by ensuring that vocal files actually have vocals (ignoring nonsense words or babbles) for more than 50% of the entire duration. Each 10 sec audio segment was manually segmented from a larger audio file. From each audio file 3 to 8 audio segments were generated. The number of genres considered in each language is shown in Table. 1. The data corpus has been intentionally biased towards vocals and English audio files, to observe how the VID and the LID systems perform inspite of the bias.

Table 1. Number of segments and genres vs Language

| Language | ENG | BEN | HIN | ESP | RUS | CHN |
|---|---|---|---|---|---|---|
| # Segments | 465 | 234 | 210 | 103 | 85 | 75 |
| # Genres | 8 | 3 | 3 | 2 | 2 | 2 |

## 3. VOCALS-INSTRUMENTAL DETECTION (VID)

GMM has been widely used in speech and speaker recognition systems [5]. The GMM-VID system consists of a feature extraction preprocessor, two GMM models for Vocals and Instrumentals and a Universal Background Model (UBM), as shown in Fig. 1

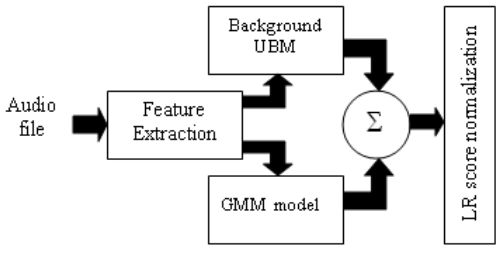


Fig. 1 Block Diagram of the GMM-UBM based VID system

Two different types of feature extractions are considered: MFCCs and SDCs. In the case of the MFCCs, the first 23 coefficients are considered ignoring the $0^{th}$ coefficient and their corresponding deltas were obtained using a delta spread of d = 2, which is represented as MFCC23,23.

The SDCs are stacked delta cepstra coefficients, usually parameterized as *N-d-P-k* [7, 8], where *N* is the number of cepstral coefficients computed at each frame, *d* is the spread for delta computation, *P* is the time shift between consecutive blocks of delta coefficients and *k* denotes the number of blocks of delta coefficients considered. Hence the dimensionality of such a vector is *Nk*. For the VID system, three different SDC parameterization has been considered: SDC7,1,3,7, SDC10,1,3,7 and SDC7,1,3,10, where SDC 7,1,3,7 has been proposed [8] as the ideal set for telephone speech based LID system. Two 2048 mixture GMM models were trained for Vocals and Instrumentals. A single GMM-UBM of 2048 mixtures was trained using the entire training set. The UBM is trained using the Expectation Maximization (EM) Algorithm and the GMM based Vocal and Instrumental models were trained using Bayesian adaptation of the UBM parameters [5, 9]. The performance of different feature sets are compared initially by keeping the number of mixtures equal to 2048 and the best 3 Average Error Rate (AER) obtained is shown in Table 2 where the AER is measured as:

$$AER = \sum_{i=1}^{M} f_i e_i \qquad (1)$$

where $M$ is the number of categories, $f_i$ is the frequency and $e_i$ is the percentage error rate of the $i^{th}$ category.

Table 2. AER for each parameter

| Parameters | MFCC23,23 (60ms, 46dim) | SDC7,1,3,7 (220ms, 49dim) | SDC7,1,3,10 (310ms, 70dim) |
|---|---|---|---|
| AER (%) | **3.318** | 7.964 | 8.405 |

The AER is represented as a percentage and gives a metric of the average false detection. From Table 2, it can be observed that MFCC23,23 offered the best AER, however the AER can be further reduced by altering the number of mixtures of the GMM. Fig. 2 shows the plot of AER versus the number of Gaussian mixtures for MFCC23,23, using two different adaptation schemes. As evident from Fig. 2, 512 mixture GMM, with mean adaptation provided the best AER (=1.76%). The confusion matrix is given in Table 3.

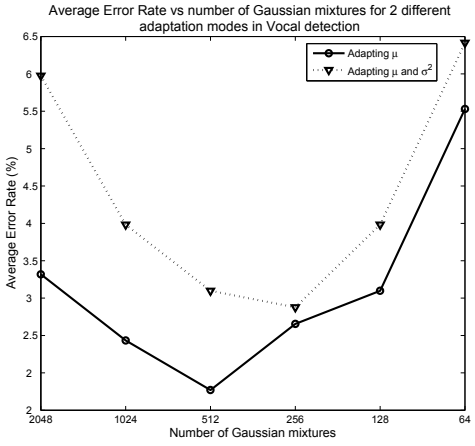


Fig. 2. Plot of AER versus number of Gaussian mixtures for MFCC23,23.

Table 3. Confusion matrix for 512-order GMM with μ-adaptation

| | Vocals | Instrumental |
|---|---|---|
| Vocals | **100** | 0 |
| Instrumental | 12.9 | **87.1** |

## 4. THE GMM BASED LID SYSTEM

Different parameterizations of MFCC, SDC, MSDC (mel-frequency cepstral coefficients + SDC) and FSDC (Fused SDC) were considered as possible features for the GMM based LID system. MSDCs are parameterized similar to SDC, [*N*,*d*,*P*,*k*] except that they have *N* mel-frequency cepstral coefficients stacked before the *Nk* SDC coefficients. FSDCs, proposed in this paper are fused multiple SDCs, parameterized as $[i,\{N,d,P,k\}_i]$, where *i* denotes the number of SDCs concatenated and $\{N,d,P,k\}_i$ denotes the parameter of the $i^{th}$ SDC. Given an SDC parameter set, the number of analysis frames, $n_f$, associated is given by:

$$n_f = \{(k-1)p+1\} + 2d \qquad (2)$$

SDCs are identified as pseudo-prosodic feature vectors [10] that give a quick approximation to the true prosodic modeling. The temporal resolution of the prosodic modeling is governed by the parameter *d* and $n_f$. Standard SDCs offer a single temporal resolution, whereas FSDCs offer multiple resolutions based upon *i*. However FSDCs usually suffer from the increased dimensionality of the feature space.

For each feature set a single UBM model was trained. GMM Language models of 2048 mixtures were constructed from the training sample, and means were adapted from the UBM using Bayesian adaptation. The obtained AER for the different parameter sets are given in Table 4.

Table 4. AER for each parameterization in LID system

| | MFCC 19,19 (60ms, 38dim) | SDC 19,2,4,2 (100ms, 38dim) | MSDC 19,1,2,2 (60ms, 57dim) | FSDC 2,{19,1,2,2}{17,2,4,2} (100ms, 72dim) |
|---|---|---|---|---|
| AER(%) | **17.30** | 20.73 | 18.32 | 20.16 |

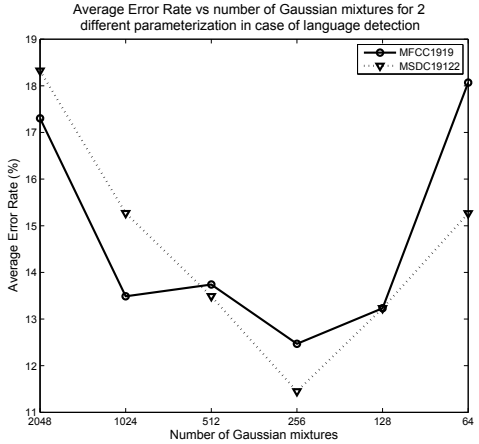


Fig. 3. Plot of AER versus number of Gaussian mixtures for the best two parameters.

Twelve different parameterizations of SDC were considered, among them SDC19,2,4,2 gave the best AER. Comparing the results obtained from SDC and FSDC in Table 4, it can be seen that varying the time resolution helped to reduce the AER. It should be noted that MFCC23,23 feature set provided better AER than SDCs, which is contrary to the LID results for conversational telephonic speech [2]. The MSDC feature set performed comparatively well offering an AER of 18.32%. The order of GMM mixture significantly contributes to the obtained AER, which is demonstrated in Fig. 3. With 256 order GMM for feature set MSDC19,1,2,2, the least AER is obtained (=11.45%). Table 5 shows the confusion matrix for 256 order GMM-LID using MSDC19,1,2,2.

Table 5 Confusion matrix for MSDC19,1,2,2

| | MSDC19,1,2,2 | | | | | |
|---|---|---|---|---|---|---|
| | CHN | RUS | BEN | HIN | ESP | ENG |
| CHN | **84.00** | 0.00 | 0.00 | 0.00 | 0.00 | 16.00 |
| RUS | 0.00 | **88.24** | 0.00 | 0.00 | 0.00 | 11.76 |
| BEN | 0.00 | 0.00 | **84.62** | 3.85 | 0.00 | 11.54 |
| HIN | 0.00 | 0.00 | 2.86 | **81.43** | 0.00 | 15.71 |
| ESP | 0.00 | 0.00 | 0.00 | 3.57 | **67.86** | 28.57 |
| ENG | 0.00 | 0.00 | 1.29 | 0.00 | 0.65 | **98.06** |

Each parameterization uses '*x* msec' of the input audio segment and represents it by an '*n*' dimensional vector, which are shown in Table 2 and 4, as (x msec, n dim)

## 6. THE SUPERVECTOR-SVM LID SYSTEM

The Supervector-SVM LID system is based upon the model described in [11, 12]. Given a training audio file, a GMM model for that file can be obtained by adapting the means from the UBM. The resultant mean vector, $\mu$, of the model forms the GMM-supervector. The GMM-supervector is then used to train a Linear kernel SVM, where the kernel function is given as [12]:

$$K(as_a, as_b) = \sum_{i=1}^{N} \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a\right)^T \left(\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b\right) \qquad (3)$$

where $as_a$ and $as_b$ are two audio segments, $w_i$ and $\Sigma_i$ are the mixture weights and covariance of the Gaussians. Two different GMM-supervectors were considered, obtained from MFC19,19 and MSDC19,1,2,2 with 256 mixture GMMs and they are represented as SV-MFCC19,19 and SV-MSDC19,1,2,2. The dimension of GMM-supervectors, $d_{SV}$, is given by: $d_{SV} = d_f \times n_{GMM}$ (4)

where, $d_f$ is the dimension of the features used as the input to the GMMs and $n_{GMM}$ is the number of GMM mixtures. Thus, dimension of SV-MFCC19,19 is 9728 and that of SV-MSDC19,1,2,2 is 14592. The two GMM-supervectors are used to train linear kernel SVMs (using SVMTorch [13]) and LS-SVMs [14]. The supervector-SVM LID system offered better AER than the supervector-LS-SVM LID

system, where the latter offered a minimum AER of 12.72%. The AER obtained from supervector-SVM LID and 256-order GMM-LID systems are given in Table 6. The AER is found to decrease further by fusing the MSDC19,1,2,2 supervector-SVM LID system with the MSDC19,1,2,2 based 256 order GMM-LID system, where the fusion was performed by optimizing the expression:

$$\arg\min_{\alpha\in(0,1)}\left[ AER_{FUSED} \mid S_{FUSED} = \left(\alpha \times S_{SVM} + (1-\alpha)\times S_{GMM}\right)\right] \quad (5)$$

where $S_{SVM}$ and $S_{GMM}$ are the scores obtained from the MSDC19,1,2,2 supervector-SVM and GMM LID system. The minimum AER (=8.39%) of the fused model is found to be at $\alpha$ = 0.425. The confusion matrix obtained from the GMM-SVM fused LID system is given in Table. 7

Table 6. AER for supervector-SVM and GMM LID

| | SVM | | GMM | |
|---|---|---|---|---|
| | SV-MFCC 19,19 | SV-MSDC 19,1,2,2 | MFCC 19,19 | MSDC 19,1,2,2 |
| AER (%) | 10.43 | **8.65** | 12.47 | **11.45** |

Table. 7 Confusion matrix for GMM-SVM fused LID

| | MFCC1919 | | | | | |
|---|---|---|---|---|---|---|
| | CHN | RUS | BEN | HIN | ESP | ENG |
| CHN | **80.00** | 0.00 | 4.00 | 0.00 | 4.00 | 12.00 |
| RUS | 0.00 | **94.12** | 0.00 | 2.94 | 0.00 | 2.94 |
| BEN | 0.00 | 0.00 | **89.74** | 1.28 | 0.00 | 8.97 |
| HIN | 0.00 | 0.00 | 4.29 | **87.14** | 0.00 | 8.57 |
| ESP | 0.00 | 0.00 | 0.00 | 3.57 | **75.00** | 21.43 |
| ENG | 0.00 | 0.00 | 1.29 | 0.00 | 0.00 | **98.71** |

## 6. CONCLUSION

We have demonstrated a systematic technique to analyze audio files prior to performing genre classification. The Vocals-Instrumental detection efficiently distinguishes the audio segments that contain vocals in them with an AER of 1.77%. Given that we know an audio segment has vocals in it, we can detect the language of the song with an AER of 8.39%, which means an accuracy of 91.61%. For faster implementation GMM based LID systems can be considered using MSDC19,1,2,2, which offers an AER of 11.54%. The supervector-SVM LID system reduces the AER further, but at the cost of more computation. The best AER is obtained from the SVM-GMM fused LID system. For the SDC based GMM LID system, it was observed that using higher values of $N$ ( $\geq 17$) significantly improve the performance, which is unlike the SDC parameters used for telephonic speech, where N is typically $\leq 10$. Future direction should address the problem of high dimensionality of the supervectors and implement the SVM nuisance attribute projection (NAP) [12] method to observe if the AER can be improved further. Given that the presence of vocals in an audio file is known, this paper proposes a technique that detects the language of the vocals. Future research should consider analyzing the vocals to detect the gender of the vocalist. Given the language and the gender, the set of possible vocalists reduces to a smaller subset, making the problem of vocalist identity easier. Given the fact that the language of the audio file is known, genre classification becomes a closed set problem, since for a given language, the number of possible genres is well defined.

## 7. REFERENCES


[1] "MPEG-7: ISO/IEC 15938-5 Final Committee Draft-Information Technology-Multimedia Content Description Interface-Part 5 Multimedia Description Schemes", ISO/IEC JTC1 /SC29/WG11 MPEG00/N3966, Singapore, Wesley, May2001.

[2] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, J.R. Deller Jr, 'Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features', Proc. International Conference on Spoken Language Processing, ISCA, pp. 89-92, September 2002.

[3] R. Dannenberg, J. Foote, G. Tzanetakis, C. Weare, "Panel: new directions in music information retrieval", Proc. International Computer Music Conference. Havana, Cuba, September 2001.

[4] F. Pachet, D. Cazaly, "A taxonomy of musical genres", Proc. Content-based Multimedia Information Access, France, 2000.

[5] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 2000, Vol.10, pp. 19-41.

[6] B. Bielefeld, "Language identification using shifted delta cepstrum", Proc. 14th Annual Speech Research Symposium, 1994.

[7] M.A. Kohler, M. Kennedy, "Language identification using shifted delta cepstra", Proc. 45th Midwest Symposium on Circuits and Systems, MWSCAS'02, Vol.3, pp. 69-72, 2002.

[8] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, D.A. Reynolds, "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Recognition", Proc. Eurospeech, ISCA'03, pp. 1345-1348, September 2003.

[9] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification". Proc. EU Conf. on Speech Comm. & Tech., pp.963–966, Sep 1997.

[10] J. Lareau, "Application of shifted delta cepstral features for GMM language identification", MS Thesis, RIT, CS. Dept, 2006.

[11] W.M. Campbell, D.E. Sturim, D.A. Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification", IEEE Signal Processing Letters, Vol.13, No.5, pp. 308-311, May 2006.

[12] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A. Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector Kernel and NAP variability compensation", Proc. IEEE International conf. on Acoustics, Speech and Signal Processing, ICASSP'06, Vol.1, pp. 97-100, May 2006.

[13] R. Collobert, S. Bengio, "SVMTorch: Support Vector Machines for Large-Scale Regression Problems", Journal of Machine Learning Research, Vol.1, pp. 143-160, 2001.

[14] J. A. K Suykens, J. Vandewalle, "Least squares support vector machine classifiers"', Neural Processing Letters, Vol.9, No.3, pp. 293-300, Jun. 1999.



**ACKNOWLEDGEMENT**

The authors would like to acknowledge Dr. D.A. Reynolds of MIT Lincoln Laboratory for his valuable comments and advice regarding the GMM-LID system.