Project no. FP6-507752

MUSCLE

Network of Excellence
Multimedia Understanding through Semantics, Computation and Learning

DN4.2:
Report on Benchmark-Based Evaluations

Due date of deliverable: 29.02.2008
Actual submission date: 27.02.2008

Start date of project: 1 March 2004 Duration: 48 months

Deliverable Type: P
Number: DN 4.2
Nature: Report
Task: WP 4

Name of responsible:
Andreas Rauber, TU Vienna-IFS, Austria

Revision 1.1

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	×
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contents

I Audio **5**

1 MIREX 2007: Combining Audio and Symbolic Descriptors for Music Classification from Audio (TU Vienna-IFS **6**
- 1.1 Introduction . . . 7
- 1.2 System Description . . . 8
  - 1.2.1 Audio Feature Extraction . . . 8
  - 1.2.2 Symbolic Feature Extraction . . . 9
  - 1.2.3 Classification . . . 11
- 1.3 Evaluation . . . 12

2 IRIT System Description for NIST 2007 Language Recognition Evaluation **16**
- 2.1 Introduction . . . 16
- 2.2 Training and Development Data . . . 17
- 2.3 Primary System . . . 17
- 2.4 Contrastive System . . . 18
- 2.5 Evaluation . . . 18
- 2.6 Conclusions . . . 19

II Text **21**

3 Benchmarking activities at CLEF'2007 (CEA LIST) **22**
- 3.1 Introduction . . . 22
- 3.2 CLEF 2007 . . . 22
  - 3.2.1 CLEF 2007 Tracks . . . 22
  - 3.2.2 CLEF 2007 Participation . . . 24
- 3.3 INFILE - A new Track proposed by CEA LIST . . . 25
  - 3.3.1 Main characteristics of the campaign . . . 26
  - 3.3.2 Brief description of the protocol . . . 26
  - 3.3.3 The corpus . . . 27

3.3.4 Metrics and evaluation 27

Introduction

Thomas Lidy, Andreas Rauber
Vienna University of Technology (TU Vienna-IFS)

MUSCLE, the Network of Excellence on Multimedia Understanding through Semantics, Computation and Learning consolidates research groups that are exploring methods for knowledge extraction from multimedia data. Research within work-package 4 focusses on processing of text and audio data (including speech, sound and music). The aim of this work-package is to extract semantic concepts from these modalities which are then utilized for knowledge extraction, for instance in topic or genre recognition tasks. The extracted numerical data from the different modalities can be used individually, or in a combined manner, which is of particular value to the Cross-Modal Integration research work of work-package 5.

Comparability of results is a key issue in many disciplines and has proven to be extremely helpful for both collaboration and competition among research groups. Benchmarking corpora are clearly essential in order to devise well-defined tasks on which researchers can test their algorithms. In this context, we can define a corpus as an annotated collection of files, documents, or digital objects, where the annotations represent the criteria the algorithms will be evaluated against.

The evaluation of the effectiveness of MUSCLE techniques for knowledge extraction is an important factor for the assessment of the techniques devised. The plethora of tasks and approaches make clear the necessity for a suitable means of comparing results amongst researchers across institutional backgrounds and domains. While data sets and evaluation settings are provided within the work-package 2 on Evaluation, Integration and Standards, the scientific research communities provide well-defined and well-known test corpora for mutual comparison of algorithms. It is very important for MUSCLE to participate in these international scientific benchmarking forums and

to compare and collaborate with the research community outside MUSCLE. Besides evaluation of the methods on common standard benchmark corpora, these benchmark-based evaluations moreover allow a comparison of the novel methods to international state-of-the-art concepts.

This report on benchmark-based evaluations within MUSCLE WP4 describes the results that have been achieved by MUSCLE teams in scientific benchmarking campaigns.

The report is organized in two parts: Part I includes evaluations of the methods for the different modalities of audio and Part II describes the participation in benchmarking campaigns that use textual information as input.

In Part I, Chapter 1 reports the achievements of TU Vienna-IFS on various different music similarity and classification tasks of the Music Information Retrieval Evaluation eXchange (MIREX) 2007 with a novel music classification approach. Chapter 2 describes the system that IRIT participated with in the Language Recognition Evaluation of NIST 2007 (US National Institute of Standards and Technology).

Chapter 3 in Part II reports about CLEF (the Cross Language Evaluation Forum) 2007 and a new track proposed by CEA LIST.

Part I

Audio

Chapter 1

MIREX 2007: Combining Audio and Symbolic Descriptors for Music Classification from Audio (TU Vienna-IFS

Thomas Lidy, Andreas Rauber
Department of Software Technology and Interactive Systems
Vienna University of Technology, Austria

Antonio Pertusa, José Manuel Iñesta
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain

Abstract

Recent research in music genre classification hints at a glass ceiling being reached using timbral audio features. To overcome this, the combination of multiple different feature sets bearing diverse characteristics is needed. We propose a new approach to extend the scope of the features: We transcribe audio data into a symbolic form using a transcription system, extract symbolic descriptors from that representation and combine them with audio features. With this method, we are able to surpass the glass ceiling and to further improve music genre classification. In this work, the methodology of

the system presented in [3] is described and evaluated.

1.1 Introduction

Audio genre classification is an important task for retrieval and organization of music databases. Traditionally the research domain of genre classification is divided into the audio and symbolic music analysis and retrieval domains. The goal of this work is to combine approaches from both directions that have proved their reliability in their respective domains. To assign a genre to a song, audio classifiers use features extracted from digital audio signals, and symbolic classifiers use features extracted from scores. These features are complementary; a score can provide very valuable information, but audio features (e.g., the timbral information) are also very important for genre classification.

To extract symbolic descriptors from an audio signal it is necessary to first employ a transcription system in order to detect the notes stored in the signal. Transcription systems have been investigated previously but a well-performing solution for polyphonic music and a multitude of genres has not yet been found. Though these systems might not be in a final state for solving the transcription problem, our hypothesis is that they are able to augment the performance of an audio genre classifier. In this work, a new transcription system is used to get a symbolic representation from an audio signal.

The overall scheme of our proposed genre classification system is shown in Figure 1.1. It processes an audio file in two ways to predict its genre. While in the first branch, the audio feature extraction methods described in Section 1.2.1 are applied directly to the audio signal data, there is an intermediate step in the second branch. A polyphonic transcription system, described in Section 1.2.2, converts the audio information into a form of symbolic notation. Then, the symbolic feature extractor is applied on the resulting representation, providing a set of symbolic descriptors as output. The audio and symbolic features extracted from the music serve as combined input to a classifier.

This study is described in [3], and it's an extension of [2], as our goal is to improve previous music genre classification results by extension of the feature space through the novel approach of including features extracted from symbolic transcription.

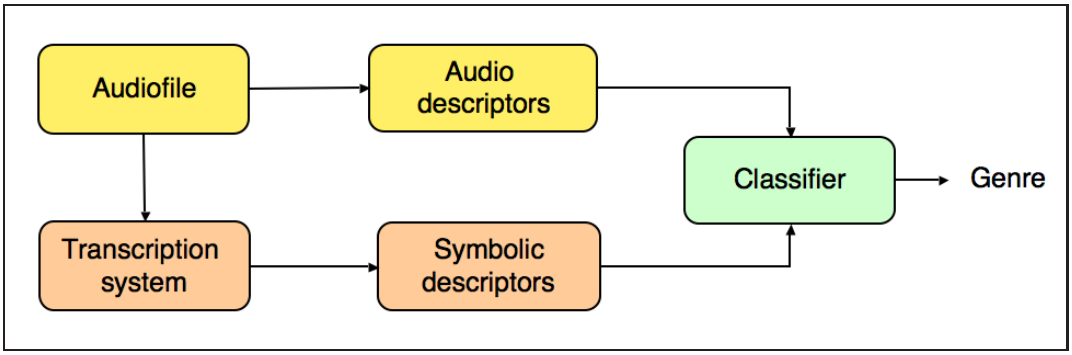

Figure 1.1: General framework of the system

1.2 System Description

1.2.1 Audio Feature Extraction

Rhythm Patterns

The feature extraction process for a Rhythm Pattern [6, 2] is composed of two stages. First, the specific loudness sensation on 24 critical frequency bands is computed, by using a Short Time FFT, grouping the resulting frequency bands to the Bark scale, applying spreading functions to account for masking effects and successive transformation into the Decibel, Phon and Sone scales. This results in a psycho-acoustically modified Sonogram representation that reflects human loudness sensation. In the second step, a discrete Fourier transform is applied to this Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. After additional weighting and smoothing steps, a Rhythm Pattern exhibits magnitude of modulation for 60 modulation frequencies (between 0.17 and 10 Hz) on 24 bands, and has thus 1440 dimensions.

Rhythm Histograms

A Rhythm Histogram (RH) aggregates the modulation amplitude values of the individual critical bands computed in a Rhythm Pattern and is thus a lower-dimensional descriptor for general rhythmic characteristics in a piece of audio [2]. A modulation amplitude spectrum for critical bands according to the Bark scale is calculated, as for Rhythm Patterns. Subsequently, the magnitudes of each modulation frequency bin of all critical bands are summed up to a histogram, exhibiting the magnitude of modulation for 60 modulation frequencies between 0.17 and 10 Hz.

Statistical Spectrum Descriptors

In the first part of the algorithm for computation of a Statistical Spectrum Descriptor (SSD) the specific loudness sensation is computed on 24 Bark-scale bands, equally as for a Rhythm Pattern. Subsequently, the mean, median, variance, skewness, kurtosis, min- and max-value are calculated for each individual critical band. These features computed for the 24 bands constitute a Statistical Spectrum Descriptor. SSDs are able to capture additional timbral information compared to Rhythm Patterns, yet at a much lower dimension of the feature space (168 dim.), as shown in the evaluation in [2].

Onset Features

An onset detection algorithm described in [4] has been used to complement audio features. The onset detector analyzes each audio frame labeling it as an onset frame or as a not-onset frame. As a result of the onset detection, 5 onset interval features have been extracted: minimum, maximum, mean, median and standard deviation of the distance in frames between two consecutive onsets. The relative number of onsets are also obtained, dividing the number of onset frames by the total number of frames of a song. As this onset detector is based on energy variations, the strength of the onset, which corresponds with the value of the onset detection function $o(t)$, can provide information about the timbre; usually, an $o(t)$ value is high when the attack is shorter or more percussive (e.g., a piano), and low values are usually produced by softer attacks (e.g., a violin). The minimum, maximum, mean, median and standard deviation of the $o(t)$ values of the detected onsets were also added to the onset feature set, which finally consists of 11 features.

1.2.2 Symbolic Feature Extraction

Transcription System

To complement the audio features with symbolic features we developed a new polyphonic transcription system to extract the notes. This system converts the audio signal into a MIDI file that will later be analyzed to extract the symbolic descriptors. It does not consider rhythm, only pitches and note durations are extracted. Therefore, the transcription system converts a mono audio file sampled at 22 kHz into a sequence of notes. First, performs a Short Time Fourier Transform (STFT) using a Hanning window with 2048 samples and 50% overlap. With these parameters, the temporal resolution is 46 ms. Zero padding has been used, multiplying the original size of the window by 8 and adding zeroes to complete it before the STFT is computed. This

technique does not increase resolution, but the estimated amplitudes and frequencies of the new spectral bins are usually more accurate than applying interpolation.

Then, the onset detection stage described in [4] is performed, classifying each time frame t_i as onset or not-onset. The system searches for notes between two consecutive onsets, analyzing only one frame between two onsets to detect each chord. To minimize the note attack problems in fundamental frequency (f_0) estimation, the frame chosen to detect the active notes is t_o+1, being t_o the frame where an onset was detected. Therefore, the spectral peak amplitudes 46 ms after an onset provide the information to detect the actual chord.

For each frame, we use a peak detection and estimation technique proposed by Rodet called Sinusoidal Likeness Measure (SLM) [8]. This technique can be used to extract spectral peaks corresponding to sinusoidal partials, and this way residual components can be removed. SLM needs two parameters: the bandwith W, that has been set as $W = 50$ Hz and a threshold $\mu = 0.1$. If the SLM value $v_\Omega < \mu$, the peak will be removed. After this process, an array of sinusoidal peaks for each chord is obtained.

Given these spectral peaks, we have to estimate the pitches of the notes. First, the f_0 candidates are chosen depending on their amplitudes and their frequencies. If a spectral peak amplitude is lower than a given threshold (experimentally, 0.05 reported good results), the peak is discarded as f_0 candidate, because in most instruments usually the first harmonic has a high amplitude. There are two more restrictions for a peak to be a f_0 candidate: only f_0 candidates within the range [50Hz-1200Hz] are considered, and the absolute difference in Hz between the candidate and the pitch of its closest note in the well-tempered scale must be less than f_d Hz. Experimentally, setting this value to $f_d = 3$ Hz yielded good results. This is a fixed value independent of f_0 because this way many high frequency peaks that generate false positives are removed.

Once a subset of f_0 candidates is obtained, a fixed spectral pattern is applied to determine whether the candidate is a note or not. The spectral pattern used in this work is a vector in which each position represents a harmonic value relative to the f_0 value. Therefore, the first position of the vector represents f_0 amplitude and will always be 1, the second position contains the relative amplitude of the second partial respect to the first, one and so on. The spectral pattern sp used in this work contains the amplitude values of the first 8 harmonics, and has been set to $sp = [1, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01]$, which is similar to the one proposed by Klapuri in [1]. As different instruments have different spectra, this general pattern is more adequate for some instruments, such as a piano, and less realistic for others, like a violin. This

pattern was selected from many combinations tested.

An algorithm is applied over all the f_0 candidates to determine whether a candidate is a note or not. First, the harmonics h that are a multiple of each f_0 candidate are searched. A harmonic h belonging to f_0 is found when the closest spectral peak to $f_0 h$ is within the range $[-f_h, f_h]$, being f_h:

$$f_h = h f_0 \sqrt{1 + \beta(h^2 - 1)} \tag{1.1}$$

with $\beta = 0.0004$. There is a restriction for a candidate to be a note; a minimum number of its harmonics must be found. This number was empirically set to half of the number of harmonics in the spectral pattern. If a candidate is considered as a note, then the values of the harmonic amplitudes in the spectral pattern (relative to the f_0 amplitude) are subtracted from the corresponding spectral peak amplitudes. If the result of a peak subtraction is lower than zero, then the peak is removed completely from the spectral peaks. The loudness l_n of a note is the sum of its expected harmonic amplitudes.

After this stage, a vector of note candidates is obtained at each time frame. Notes with a low absolute or relative loudness are removed. Firstly, the notes with a loudness $l_n < \gamma$ are eliminated. Experimentally, a value $\gamma = 5$ reported good results. Secondly, the maximum note loudness $L_n = \max l_n$ at the target frame is computed, and the notes with $l_n < \eta L_n$ are also discarded. After experiments, $\eta = 0.1$ was chosen. Finally, the frequency and loudness of the notes are converted to MIDI notes.

Symbolic Features

A set of 37 symbolic descriptors was extracted from the transcribed notes. This set is based on the features described in [5], that yielded good results for monophonic classical/jazz classification, and on the symbolic features described in [7], used for melody track selection in MIDI files. The number of notes, number of significant silences, and the number of non-significant silences were computed. Note pitches, durations, Inter Onset Intervals (IOI) and non-diatonic notes were also analyzed, reporting for each one their highest and lowest values, their average, relative average, standard deviation, and normality. The total number of IOI was also taken into account, as the number of distinct pitch intervals, the count of the most repeated pitch interval, and the sum of all note durations, completing the symbolic feature set.

1.2.3 Classification

There are several alternatives of how to design a music classification system. The option we chose is to concatenate different feature sets and provide the

combined set to a standard classifier that receives an extended set of feature attributes on which it bases its classification decision (c.f. Figure 1.1). For our experiments we chose linear Support Vector Machines. We used the SMO implementation of the Weka machine learning software [9] with pairwise classification and the default Weka parameters (complexity parameter $C = 1.0$).

1.3 Evaluation

A first evaluation using three different datasets was presented in [3]. Despite the system was originally developed for genre classification, it's suitable to be applied to other similar music classification tasks, so it was presented for MIREX evaluation in different contests. The results showed that the system yielded a high success rate for genre classification (see tab. 1).

Participant	Hier.	Raw
IMIRSEL (svm)	76.56%	68.29%
Lidy, Rauber, Pertusa & Iñesta	**75.57%**	**66.71%**
Mandel & Ellis	75.03%	66.60%
Mandel & Ellis (spec)	73.57%	65.50%
G. Tzanetakis	74.15%	65.34%
Guaus & Herrera	71.87%	62.89%
IMIRSEL (knn)	64.83%	54.87%

Table 1.1: Genre classification results. The second column shows to the average hierarchical classification accuracy, and the third to the average raw classification accuracy.

For the audio music similarity contest, two systems were submitted; (1) is the system described in this work, and (2) a previous system presented in MIREX'06, containing audio (SSD + RH) features only, and presented this year to compare both. As shown in the table 1.3, the whole set of features extracted (1) yielded better results than the audio features only (2).

In the case of audio artist identification and classical composer identification, the system yielded encouraging results, and for mood classification the results were satisfactory.

These hopeful results open a new research line by combining audio and symbolic features, and wide future work includes, for example, the use of classifier ensembles.

Participant	F-score
Pohle & Schnitzer	0.568
G. Tzanetakis	0.554
Barrington, Turnball, Torres & Lanskriet	0.541
C. Bastuck (1)	0.539
Lidy, Rauber Pertusa & Iñesta (1)	0.519
Mandel & Ellis	0.512
Lidy, Rauber Pertusa & Iñesta (2)	0.491
C. Bastuck (2)	0.446
C. Bastuck (3)	0.439
Bosteels & Kerre (1)	0.412
Paradzinets & Chen	0.377
Bosteels & Kerre (2)	0.178

Table 1.2: Audio similarity results. The second column shows the sum of fine-grained human similarity decisions (0-10).

Participant	Avg. Raw Acc.
IMIRSEL (svm)	48.14%
Mandel & Ellis (spec)	47.16%
Mandel & Ellis	40.46%
Lidy, Rauber, Pertusa & Iñesta	**38.76%**
G. Tzanetakis	36.70%
IMIRSEL (knn)	35.29%
K. Lee	9.71%

Table 1.3: Audio artist identification results. The second column corresponds to the raw classification accuracy.

Participant	Avg. Raw Acc.
IMIRSEL (svm)	53.72%
Mandel & Ellis (spec)	52.02%
IMIRSEL (knn)	48.38%
Mandel & Ellis	47.84%
Lidy, Rauber, Pertusa & Iñesta	**47.26%**
G. Tzanetakis	44.59%
K. Lee	19.70%

Table 1.4: Audio classical composer identification results. The second column corresponds to the raw classification accuracy.

Participant	Avg. Raw Acc.
G. Tzanetakis	61.50%
C. Laurier	60.50%
Lidy, Rauber, Pertusa & Iñesta	**59.67%**
Mandel & Ellis	57.83%
Mandel & Ellis (spec)	55.83%
IMIRSEL (svm)	55.83%
L. Lee (1)	49.83%
IMIRSEL (knn)	47.17%
K. Lee (2)	25.67%

Table 1.5: Audio mood classification. The second column corresponds to the raw classification accuracy.

Bibliography

[1] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. ISMIR*, pages 216–221, Victoria, Canada, 2006.

[2] T. Lidy and A. Rauber. Evaluation of feature extractors and psychoacoustic transformations for music genre classification. In *Proc. ISMIR*, pages 34–41, London, UK, September 11-15 2005.

[3] T. Lidy, A. Rauber, A. Pertusa, and J.M. Iñesta. Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In *Proc. ISMIR*, Vienna, Austria, 2007.

[4] A. Pertusa, A. Klapuri, and J.M. Iñesta. Recognition of note onsets in digital music using semitone bands. In *Proc. 10th Iberoamerican Congress on Pattern Recognition (CIARP)*, LNCS, pages 869–879, 2005.

[5] P. J. Ponce de León and J. M. Iñesta. A pattern recognition approach for music style identification using shallow statistical descriptors. *IEEE Trans. on Systems Man and Cybernetics C*, 37(2):248–257, 2007.

[6] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003.

[7] D. Rizo, P.J. Ponce de León, C. Pérez-Sancho, A. Pertusa, and J.M. Iñesta. A pattern recognition approach for melody track selection in midi files. In *Proc. ISMIR*, pages 61–66, Victoria, Canada, 2006.

[8] X. Rodet. Musical sound signals analysis/synthesis: Sinusoidal+residual and elementary waveform models. *Applied Signal Processing*, 4:131–141, 1997.

[9] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Chapter 2

IRIT System Description for NIST 2007 Language Recognition Evaluation

Eduardo Sánchez-Soto, Jérôme Farinas
IRIT-CNRS, Toulouse, France

Abstract

The score provided as results of IRIT primary system for LRE07 is based on a classical acoustic UBM/GMM modeling. The contrastive system uses a statistical speech segmentation approach and a vowels detection to obtain two different classes which are also modeled using a UBM/GMM. Score combination, issue of each class, is performed using a basic weighted addition. All models have been discriminatively trained on the data described in the evaluation plan Section 4 [LRE 2007].

2.1 Introduction

The LRE 07 evaluation are for us part of a project founded by the "Agence National de la Recherche" in France called MISTRAL (http://mistral.univ-avignon.fr). The goal of this project is to provide at the end with an open source platform for Language Recognition and Audio-Visual Biometric Authentication. MISTRAL is based on ALIZE and particularly the system submitted to this evaluations uses the ALIZE/LIA_RAL open source software for text independent speaker recognition as a guide and base to develop

our segmental approach. ALIZE is developed in C++ following an object oriented UML method, which include already UBM/GMM with unsupervised adaptation for example.

2.2 Training and Development Data

In order to train the different languages involved in the evaluation, we have used a subset of the data provided by LCD and NIST. The included data is the CALLFRIEND, the LRE05 OHSU and LRE07 training databases. To calibrate the scores the LRE05 database was used.

2.3 Primary System

The first system implemented for language recognition is closely inspired on the speaker verification technology based on GMMs and is the first attempt to use LIA_RAL in a Language Recognition task. In both systems, primary and contrastive, we use a common frontend which consisted of standard SDC (shift delta cepstra) with the next parameters (6, 1, 3, 3). A voice activity detector based on the energy is then used to discard the silence parts of the signal. Then a "UBM" model for each language is trained with all the available data. The skin of the system is presented in the next figure.

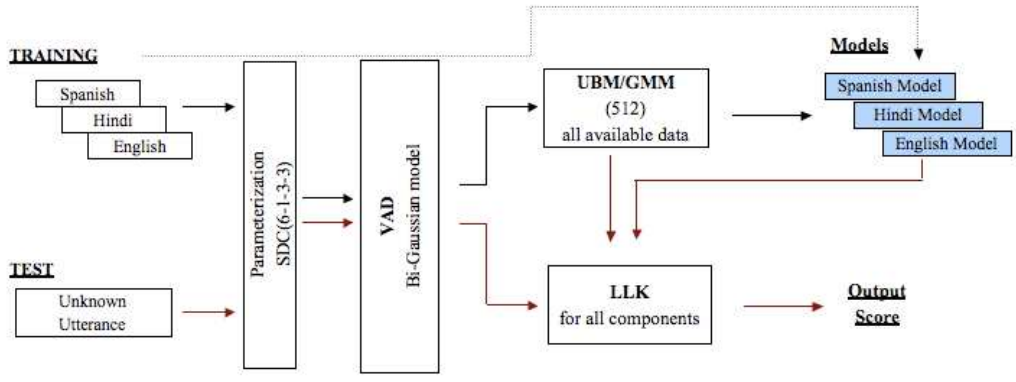

Figure 2.1: *Primary system.*

In the test phase, each input segment is compared with every language model for which a score is obtained. This score is the Log Likelihood value. All segments with a final score, after a normalization step, grater than 0 are considered to belong to the tested language, returning a True label.

2.4 Contrastive System

This system is based on a segmentation approach. The basic system is very close to the primary system given that both are based on UBM/GMM modeling. The basic difference is the segments used to create the model. In the primary system all the available data was employed. In the contrastive system two different classes are modeled by a UBM/GMM model. From the segmentation phase two classes are obtained : "vowels" and not "vowels". Each one of this classes is modeled in the same manner obtaining at the end an score, which are then combined to generate the final score by a weighted addition. This segmentation method is just the beginning of the integration of Language Recognition capabilities into MISTRAL.

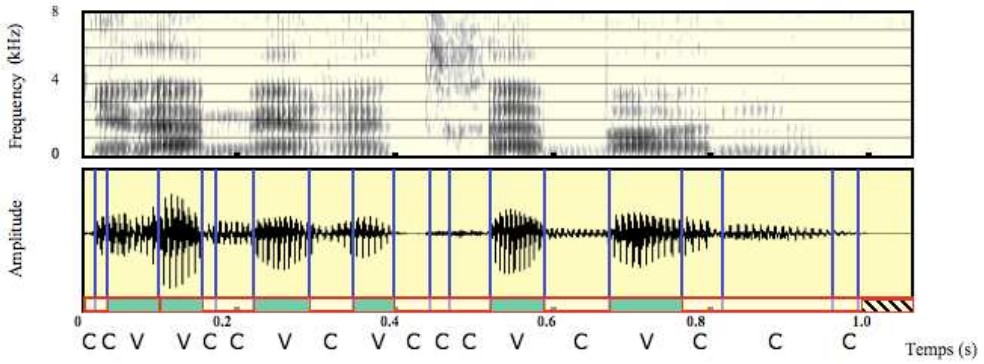

Figure 2.2: *Contrastive system.*

The fist step in the segmentation phase is based on a statistical speech segmentation [Obretch 1988]. This segmentation results into short segments (bursts, transient part of voiced sounds) and longer segments (steady parts of sound). This phase is followed by a segmental activity detector which is used to discard pauses not related to rhythm. Finally, vowels detection is performed via a spectral analysis of the signal [Pellegrino 2000].

2.5 Evaluation

The system has been submitted to the *general_lr* condition, in the closed set condition. For 30s speech files, a $Cavg = 0.4973$ was achieved and we finished 18th over 18th. After the evaluation, the system has been better tuned and achieved a Cavg of 0.37 under the same conditions. That represents 16th rank with the results of the evaluation.

2.6 Conclusions

We have presented the first step of the evolution of ALIZE/SpkDet into MISTRAL and its first step to use the platform in a Language Recognition task.
The challenge is to obtain a platform open source for Audio-Visual Biometric Authentication based on ALIZE.

Bibliography

[ALIZE] Bonastre J. F., Scehher N., Matrouf D., Fredouille C., Larcher A., Preti A., Pouchoulin G., Evans N., Fauve B. and Mason J. ALIZE/SpkDet: a state of the art open source software for speaker recognition. Speaker Odyssey, 2008.
[MISTRAL] http://mistral.univ-avignon.fr
[Obretch 1988] André-Obrecht R. A new statistical spproach for automatic segmentation of continuous speech. Trans. IEEE on Acoustics, Speech and Signal Processing. Vol. 36, January 1988.
[Pellegrino 2000] Pellegrino F. and André-Obrecht R. Authomatic Language identification: an alternative approach to phonetic modelling. Signal Processing, Elsevier Science, Nort Holland, Vol 80, July 2000.
[LRE 2007] The 2007 NIST Language Recognition Evaluation Plan (LRE07), NIST, version 8b, July 20, 2007.

Part II

Text

Chapter 3

Benchmarking activities at CLEF'2007 (CEA LIST)

Gregory Grefenstette, CEA LIST, France

Abstract

CLEF (the Cross Language Evaluation Forum) is yearly benchmarking conference based around shared tasks that participants perform and report on, following the evaluation model developed by the National Institute of Standards and Technology (NIST) in the USA. CLEF differs from the yearly, US-based TREC (Text Retrieval Conferences) conferences in its emphasis on field of multilingual system development.

3.1 Introduction

The CLEF 2007 was the eighth in the series of evaluation campaigns. A conference discussing the results was held in September 19-21, 2007 in Budapest, Hungary. MUSCLE representatives from the CEA LIST, Christian Fluhr and Halima Dahmani, attended the conference and proposed a new form of benchmarking. Here, we present an overview of CLEF 2007 and a description of this new campaign called INFILE which had been funded by the French ANR (Association Nationale pour la Recherche).

3.2 CLEF 2007

3.2.1 CLEF 2007 Tracks

CLEF 2007 offered seven tracks designed to evaluate the performance of systems for:

- monolingual, bilingual, and multilingual text retrieval on newspaper collections (called the "Ad Hoc" track after TREC's terminology)

This year the monolingual emphasis was on central European languages (Bulgarian, Czech and Hungarian). For cross language retrieval (finding English documents using non-English, queries were made available in Amharic, Chinese, Oromo, Indonesian, Hindi, Bengali, Tamil, Telugu and Marathi). These new queries add to the growing number of languages for which CLEF benchmarks exist. A benchmark consists of list of English documents from a large collection which are pertinent for each query. The CLEF multilingual comparable corpus of more than 3 million news documents in 13 languages was used in this and other tracks.

- monolingual and cross-language information on structured scientific data (the "Domain-Specific" track)

This year "social sciences" was the special domain, and search was performed over structured data (e.g. bibliographic data, keywords, and abstracts) from scientific reference databases: GIRT-4 for German/English, INION for Russian and Cambridge Sociological Abstracts for English.

- question answering in multiple language (QA@CLEF)

Question answering was performed this year over both newspaper texts and Wikipedia entries. In question answering, a natural language query is answered by a short text segment (for example, 50 characters, or a single phrase) from a document. CLEF provides human judges that decide response accuracy. Sub tasks were monolingual – where the questions and the target collections searched for answers are in the same language - and bilingual – where source and target languages are different. Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were offered as target languages

- cross-language image retrieval (ImageCLEF)

Image retrieval techniques can use only image captions found under images (which are provided by CLEF) or using image processing techniques, or both. Four search tasks were perfromed: (i) multilingual caption retrieval (collection with mixed English/German/Spanish annotations, queries in more languages), (ii) medical image retrieval (medical notes in English/ French/German; visual, mixed, semantic queries in same languages), (iii) hierarchical automatic image annotation for medical images (fully categorized

in English and German, a purely image processing task), (iv) another image processing task: image annotation via object detection (using the same collection as (i)). ImageCLEF is one of the principal workshops in object detection.

- cross-language speech retrieval (CL-SR)

searching spontaneous speech from oral history interviews from the Survivors of the Shoah Visual History Foundation (VHF) English (750 hours) and Czech (approx 500 hours) speech transcribed by automatic speech recognition was provided to search, along with additional manually and automatically assigned controlled vocabulary descriptors for concepts, dates and locations, manually assigned person names, and hand-written segment summaries. Text queries were available in Czech, Dutch, English, French, German and Spanish.

- multilingual Web document retrieval (WebCLEF)

At WebCLEF 2007, undirected informational search goals were tested in a web setting: "I want to learn anything/everything about my topic." EuroGOV, a multilingual collection from European governmental sites of about 3.5M webpages, was used in this track.

- Cross-Language Geographical Retrieval (GeoCLEF):

Cross-language geographic information retrieval (GIR) is meant to identify queries with q geographical component. For the GeoCLEF 2007 search task, twenty-five search topics were defined in English, German and Spanish for search over English, German, Portuguese and Spanish document collections. This may seem like a trivial task but place names are difficult because of a wide variety of spelling variations (within and between languages) and because of confusion between ordinary terms and geographical terms.

Figure 3.1 shows a graphic providing an overview of participation in various CLEF tracks over the years.

3.2.2 CLEF 2007 Participation

81 groups submitted runs in CLEF 2007. 51 from Europe, 14 from North America, 14 from Asia, 1 each from South America and Australia.

The breakdown per track is as follows: Ad Hoc 22; Domain-Specific 5; QAatCLEF 28; ImageCLEF 35; CL-Speech Retrieval; WebCLEF 4; GeoCLEF 13. The apparent lack of interest in WebCLEF is surprising to CLEF organizers, who expected a larger participation, given the importance of Internet and web search engines. There is growing interest in the ImageCLEF track, which is nonetheless the least multilingual of the CLEF tracks as much of the work is done in a language-independent context, which may indicate

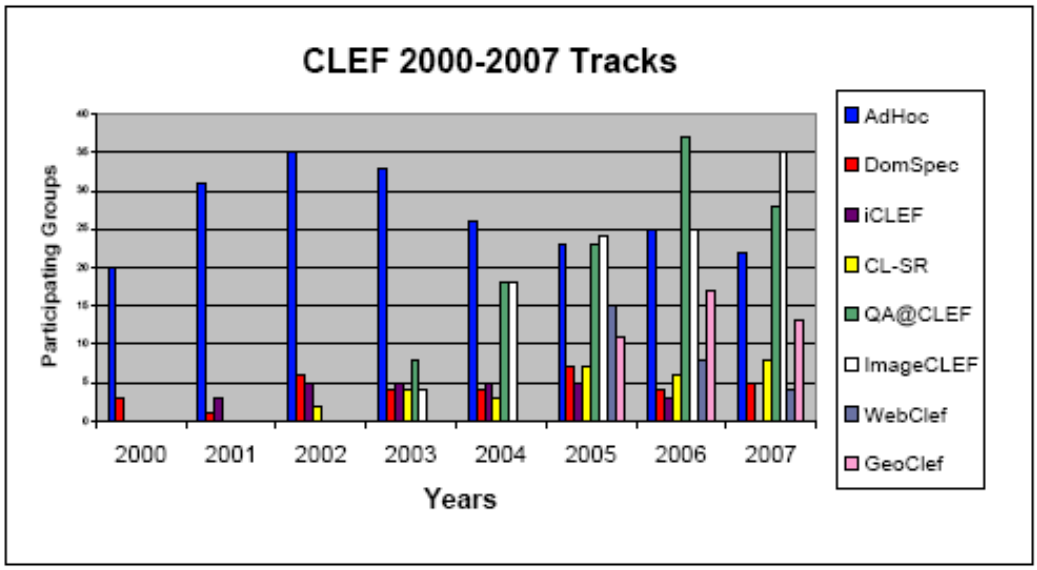

Figure 3.1: CLEF 2000 - 2007: Participation per Track in Tracks

that more workshops like ImagEval, sponsored by MUSCLE, might find a large following.

3.3 INFILE - A new Track proposed by CEA LIST

The InFile project (INformation, Filtering, Evaluation) is a cross-language adaptive filtering evaluation campaign. It will be a pilot track of the next CLEF 2008 campaigns, and it is financed by the French National Research Agency, and organized by the CEA LIST (MUSCLE), ELDA and the University of Lille3-GERiiCO.

The InFile evaluation campaign will measure the ability of filtering systems to successfully separate relevant and non-relevant documents in an incoming stream of textual information. An information filtering system is a system designed to manage unstructured or semistructured data. Information filtering systems deal primarily with textual information, involve large amounts of data incoming through permanent streams such as newswire services. Filtering is based on individual or group information profiles which assume to represent consistent and long-term information needs. From the

user point of view, the filtering process is usually meant to extract relevant data from the data streams, according to the user profiles.

Information filtering systems may be used in different business contexts of use : for example, text routing which involves sending relevant incoming data to individuals or specific groups, categorization process which aims at attaching one or more predefined categories to incoming documents, or anti-spamming which tries to remove "junk" e-mails from the incoming e-mails.

For the InFile project, the CEA LIST retained the context of competitive intelligence in which the information filtering is a very specific subtask of the information management process. In this approach, the information filtering task is very similar to Selective Dissemination of Information (SDI), one of the original and usual functions assumed by documentalists and, more recently, by other information intermediaries such as technological watchers or business intelligence professionals.

3.3.1 Main characteristics of the campaign

The InFile evaluation campaign is:

- crosslingual: English, French and Arabic are concerned by the process but participants may be evaluated on mono or bilingual runs
- the corpus will be composed of approximatively 100,000 newswires selected from the Agence France Presse (AFP) stream.There will be two groups of topics, one concerning general news and events, and a second on scientific and technological subjects
- the evaluation task will be performed using an automatic interrogation of participating systems with a simulated user feedback. System will be allowed to use the feedback at any time to increase performance
- systems will provide a Boolean decision for each document according to each profile.

3.3.2 Brief description of the protocol

The campaign will consist in a dry run following by the evaluation run. The dry run will be organized to control that the automated submission process runs correctly for each system.
Before the evaluation run, some general information about the two domains of interest will be given to the participants in order to adapt the systems, if necessary. Approximatively 15 days afterwards, profiles will be given to participants. Profiles will be composed of a list of keywords (simple and complex noun phrases) and up to 3 documents illustrating the information interest.

The data will be transmitted by the organizer to an automated interface of each participating system. The interface will return a Boolean response for each newswire according to each profile. After reception of this response, the organizer will send a feedback consisting of the expected assignment for each document submitted.

Participants will be allowed to adapt their system at any time using this feedback.

3.3.3 The corpus

In order to ensure that none of the participants already worked on the test corpus in previous evaluation campaigns, the InFile project will create a very new test corpus. This corpus will be composed of :

- a collection of 100,000 recent newswires of general and scientific interest from Agence France Presse. This collection is composed of three sets :
- a set of relevant documents provided by information professionals (assessors)
- a set of non relevant but close-to-profile documents will be specially chosen to ensure some confusion with the profiles
- and the rest of the collection which will be a large set of irrelevant documents in which the two previous subsets will be hidden. The size of this subset will be large enough to prevent any manual examination of the corpus. In order to ensure that this corpus does not contain any relevant document, assessors will use state-of-the-art information retrieval systems on the full data to detect and eliminate such documents.
- in addition, information professionals will create a set of 30 to 50 profiles.

3.3.4 Metrics and evaluation

To measure adaptive filtering performance, after every N documents (for example N=20 000) precision, recall and average measures will be computed to plot an effectiveness evolution curve. At the end of the campaign a mean effectiveness will be computed for each system.
We will also measure the number of documents that each system uniquely and correctly filters (originality measure). Other measures will be eventually included and are presently discussed.

To be as close as possible to the ground truth and because of the corpus size and the fact that feedback must be sent immediately after answer reception, a pooling methodology is not available. Evaluation will be based on the set of relevant documents provided by the human experts but, at the

end of the run, a limited control (via limited pooling) will be performed on documents considered as relevant by at least 2 systems for a specific profile.

This control will allow the organizers to eventually adjust the set of relevant documents, to improve the reliability of the feedback given to the participants during the run and to check the performance measures. If few modifications are needed, the way each system uses the feedback to increase performance can be considered as representative. If this limited pooling control detects that many modifications are needed, the results of using feedback will be less reliable to judge.

InFile will be a pilot track of the next CLEF 2008 campaigns.