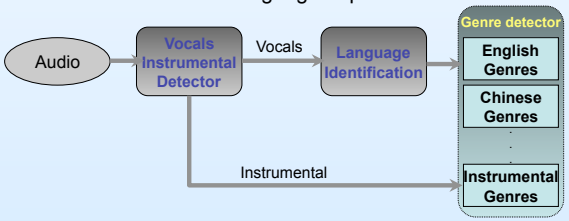
Language Detection for Music Information Retrieval

Vikramjit Mitra, Daniel Garcia-Romero and **Carol Y. Espy-Wilson**
Speech Communication Lab
SCL

Introduction

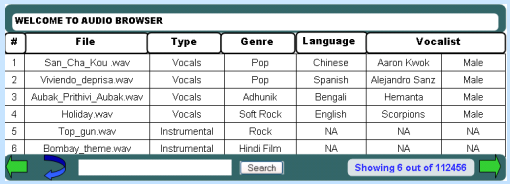
This research aims to introduce a systematic approach towards audio information retrieval using a Vocal-Instrumental detector (VID) and Language Identification (LID).

Cascade architecture for language dependent Genre detection



- **Since Genre types vary with language, knowledge of the language used in the vocals can improve Genre detection accuracy.**
- **Detecting language provides content related information that allows us to work with multi-lingual databases**

Example of a possible Information Retrieval based Audio-browser



Database

- **Since no standard multi-language audio corpora is available, our corpus was created in-house and manually tagged**
- **1358 audio segments of approximate duration 10 secs were created from larger audio files and downsampled to 8 kHz (region where speech is dominant)**

Distribution of files by Language

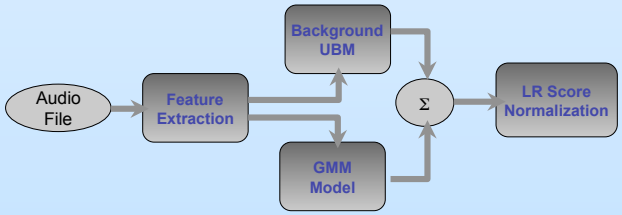
	Vocals						Instrumental
	English	Bengali	Hindi	Spanish	Chinese	Russian	
#Segments	465	234	210	103	85	75	186
#Genre	7	3	3	2	2	2	2

- **Since we are unable to separate the vocals from the background music, we segmented the audio segments that contain vocals in such a way that the vocals are dominant (i.e., the duration of the vocals is greater than 50%)**
- **Additionally, we made sure that the vocals are not dominated by nonsense words or humming.**
- **67% used for training, rest for testing**

Vocals – Instrumental Detector (VID)

- **Used Gaussian Mixture Model as the backend**
- **Two different feature sets considered**
  - **Mel-frequency Cepstral Coefficients (MFCC)**
  - **Shifted Delta Coefficients [1]**
- **UBM was trained using the entire training set**

The GMM-UBM based VID system [2]



Confusion matrix for 512 order GMM-VID

	Vocals	Instrumental
Vocals	**100**	0
Instrumental	12.9	**87.1**

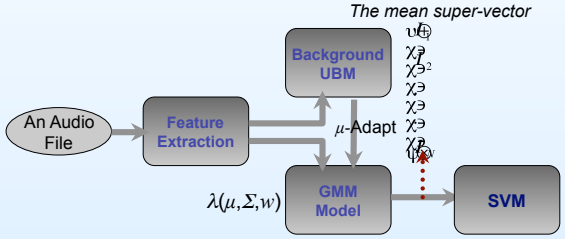
The GMM Language Identifier (LID)

- Different feature sets considered:
  - MFCC, SDC, MSDC, FSDC
  - FSDC provides multi-resolution, but dimension ↑

Avg. Error for GMM-LID using 2048 mixture GMM-UBM

	MFCC	SDC	MSDC	FSDC
Error (%)	**17.3**	20.7	**18.3**	20.2

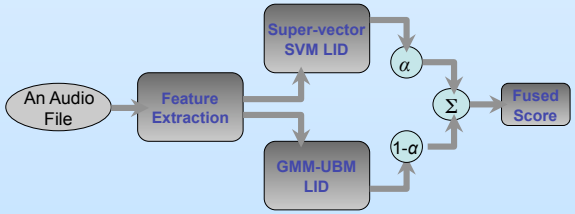
Super-vector SVM – LID System



2 GMM-super-vectors considered

- SV-MFCC: MFCC with 256 mix. GMM (Error 10.4%)
- SV-MSDC: MSDC with 256 mix. GMM (Error 8.65%)
- Error can be further ↓ by fusing super-vector LID with GMM LID (Error 8.39%)

The Fusion model



Confusion matrix for the Fused LID system

	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**

Conclusion

- Vocals can be detected from Instrumentals with high acc.
- Language can be detected with an average accuracy 91.6%.

References

[1] B. Bielefeld, "Language identification using shifted delta cepstrum", Proc. 14th Annual Speech Research Symposium, 1994
[2] 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.