Speech Recognition Combining MFCCs and Image Features

S. Karlos from Department of Mathematics
N. Fazakis from Department of Electrical and Computer Engineering
K. Karanikola from Department of Mathematics
S. Kotsiantis from Department of Mathematics
K. Sgarbas from Department of Electrical and Computer Enginnering

University of Patras, Greece

Aim

- Combination of audio signal and image features
- Exploitation of larger frames for speech signals
- Increase of classification accuracy without using complex algorithms

Contents

- Speaker Identification problem
- Attributes of speech signals
- Examine Content Based Image Features (CBIR)
- Combination of MFCCs + CBIR
- Experiments
- Conclusion

Speaker Identification Problem

- Determines the speaker from a set of registered speakers
  - This is called a “closed” set identification
  - Result is the best speaker matched
- What if the speaker is not in the database?
  - This is called an “open” set identification
  - Result can be a speaker *or* a no-match result
- Our experiment is a closed set identification problem

Extraction of audio characteristics

- Different representations of speech signals:

1. **M**el-**F**requency **C**epstral **C**oefficients (**MFCC**)
2. **L**inear **P**redictive **C**odes (**LPCs**)
3. **P**erceptual **L**inear **P**rediction (**PLP**)
4. PLP-Relative Spectra (**PLP-RASTA**)

- Non-linear behavior of speech
- Need for adapting signal to human ear scale
- Most efficient solution: MFCCs features

Extraction of image characteristics

- Spectrogram: time-frequency representation of an audio signal
- Short-Term Fourier Transform (STFT)

- Different approaches of image processing :

1. Content-Based
2. Feature-Based
3. Appearance-Based

- Determine the similarity through distances of feature vectors

Related works

- Content Based Image Processing (CBIR) techniques have been widely used
- Exploitation of color content and texture information

- Most known approaches:

1. Local gradient features along with PCA + HMMs
2. Delta MFCCs
3. 2D Gabor Features + MLP
4. Feature-Finding Neural Network (FFNN)
5. Wavelet package transform + MKL
6. RANSAC algorithm

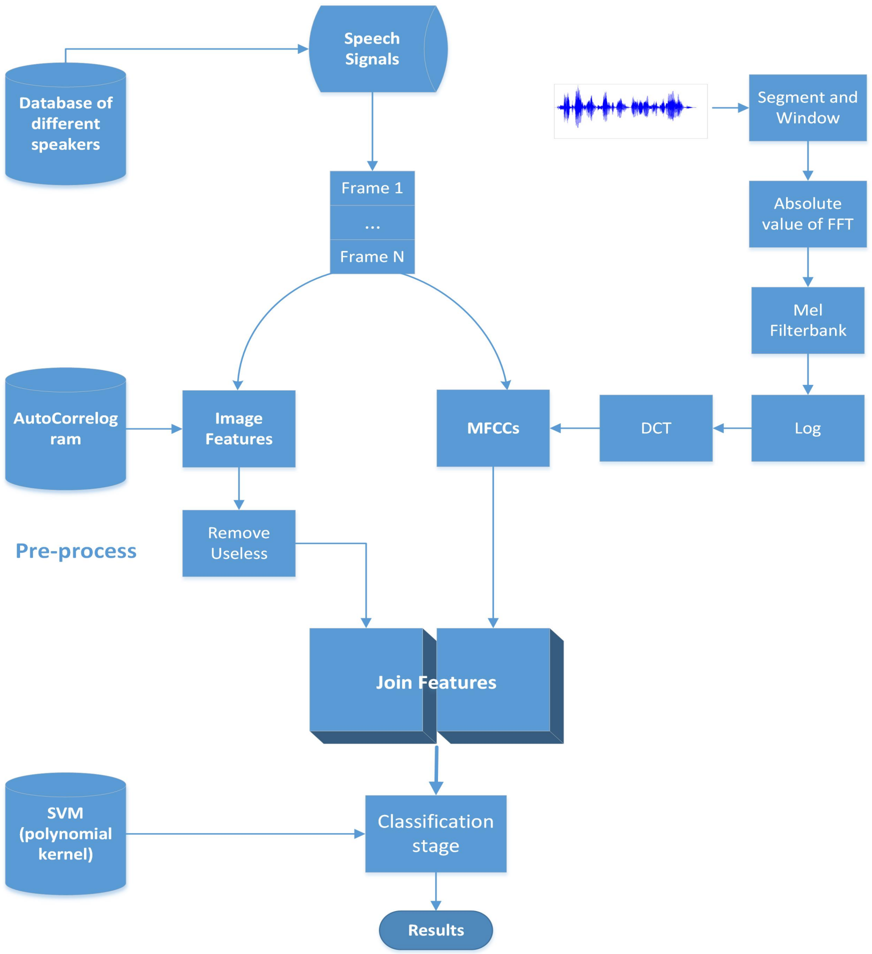
Proposed Technique – 1st view

- Acquire the first 25 coefficients of MFCCs (0th has been rejected)
- Hamming window has been preferred
- Time duration of each frame equals to 0.5 seconds
- Overlap factor equals to 50%
- Highest band edge of Mel filters equals to 4kHz
- Use of 40 warped spectral bands
- Logarithmical scale of magnitude spectrum
- **D**iscrete **C**osine **T**ransformation (**DCT**)

Proposed Technique – 2nd view

- Use of AutoColorCorrelogramFilter (**autocor**)
- $a_c^{(k)}(I) = \gamma_{c,c}^{(k)}(I), \quad \gamma_{c1,c2}^{(k)}(\mathrm{I}) = \ Pr_{p1\,\in\, Ic1,\, p2\,\in\, I}\left[p_2 \in \ I_{c_2} \mid \mathrm{dist}(p_1, p_2) = k\right]$
- Spatial correlation of colors from each image is distilled
- Not based on purely local properties
- Effective in recognizing large changes of shape
- Efficiently computed

MFCCs + autocor + SVM



Proposed Technique – Learning stage

- **S**upport **V**ector **M**achines (**SVMs**)
- Hyperplanes that separate two classes
- Maximizing the margin for reducing the generalization error
- Can deal with very high dimensional data
- Efficient implementation through LibSVM library
- Use of polynomial kernel (degree = 3)

Data

- CHAINS Corpus
- Selected mode: Solo speech
- 36 speakers (28 from Eastern Ireland – 8 from UK and USA)
- 19 different sentences out of the 33
- 3 scenarios: 8, 16 and 36 speakers
- Equal male and female speakers during each scenario

Experimental procedure

- Comparison with another 9 image filters
- Supervised classifiers:

1. **SVMs**
2. **M**ulti-**L**ayer **P**erceptron (**MLP**)
3. **Log**istic **Reg**ression (**LogReg**)

- 10-cross-validation technique
- WEKA tool was used along with libraries of **L**ucene **I**mage **Re**trieval (**LIRe**)
- Record computational time (Intel i3 – 64bit system - 8GB RAM)

Experimental procedure

CBIR Filters	Initial Number of features	Useful Number of features
autocor	1024	57
binpyr	756	131
clay	33	33
edhist	80	80
fcth	192	18
fuzzy	576	17
gabor	60	60
jpeg	192	192
phog	630	44
simplehist	64	11

Reduction of dimensionality: Remove useless attributes

Size of datasets on instances has been reduced dramatically:

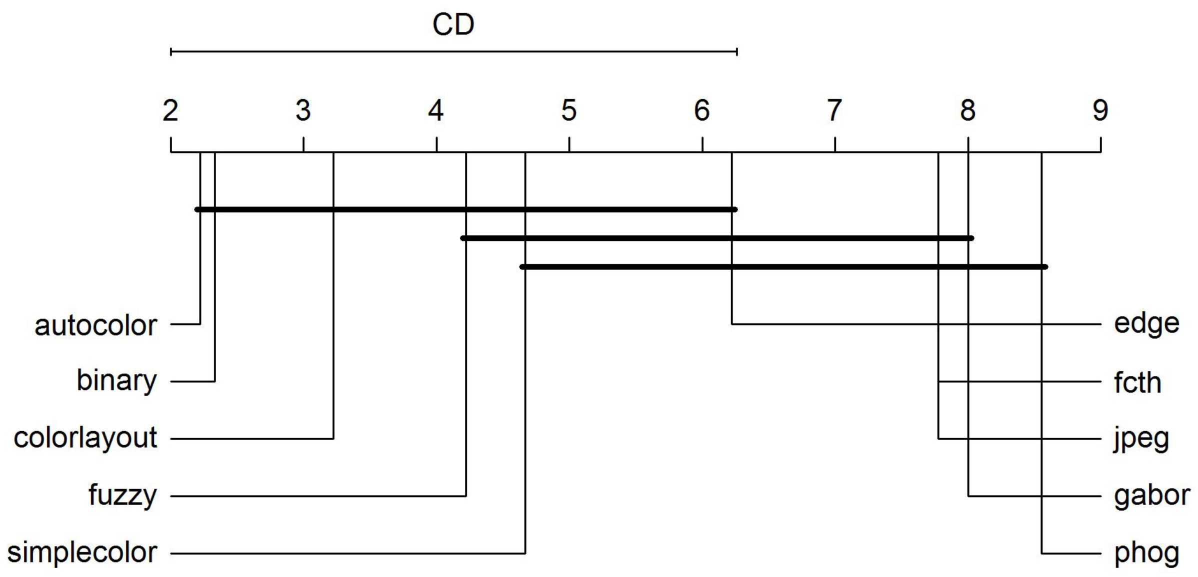
- 8speakers: about 32.000 -> 1.298
- 16speakers: about 65.000 -> 2.577
- 36speakers: about 146.000 -> 5.818

Results

	8 speakers		16 speakers		36 speakers	
Classifiers	MFCCs	MFCCs + autocor	MFCCs	MFCCs + autocor	MFCCs	MFCCs + autocor
SVM	79.89	87.44	75.90	83.70	66.74	76.64
Time(sec)	0.45	0.88	1.29	2.09	5.93	9.62
MLP	69.49	82.42	69.03	80.36	60.1581	66.33
Time(sec)	10.71	60.80	35.43	121.04	179.89	452.50
LogReg	66.41	76.96	73.38	79.74	60.89	67.13
Time(sec)	0.26	1.08	1.71	4.06	5.46	27.98

Statistical comparison

- Post-hoc test of Nemenyi
- CD's length depicts the needed distance for significant difference



Experiments

- A boost of accuracy was recorded for all the tested scenarios
- 11.5%, 7.8% and 9.9% improvement compared with standalone MFCCs
- Building of classification model demands a few seconds
- Fuzzy filtering techniques performed fluctuations
- *MFCCs+autocor* and *MFCCs+binpyr* achieved the best results
- The proposed technique requires much less computational resources

Conclusions

- Tackle with **A**utomatic **S**peech **R**ecognition (**ASR**) tasks
- Increase the feature vector of audio signals
- Reduce the training time
- Methods based on local features performed poor results
- Improved generalization behavior for the most SI filters

Promising points

- Extract more specialized features under MFCCs + SI features scheme
- Parallel implementation
- Apply multi-view Semi-supervised techniques
- Combination of magnitude with phase related features (Hartley Phase Spectrum)

References

- M. Lux and S. A. Chatzichristofis, “Lire: lucene image retrieval,” *Proceeding 16th ACM Int. Conf. Multimed. - MM ’08*, p. 1085, 2008.
- F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “The CHAINS Speech Corpus: CHAracterizing INdividual Speakers,” *Proc SPECOM*, pp. 1–6, 2006
- J. Dennis, H. D. Tran, and H. Li, “Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions,” *IEEE Signal Process. Lett.*, vol. 18, no. 2, pp. 130–133, Feb. 2011
- M. Mayo, “ImageFilter WEKA filter that uses LIRE to extract image features,” 2015. [Online]. Available: https://github.com/mmayo888/ImageFilter
- I. Paraskevas and M. Rangoussi, “The hartley phase spectrum as an assistive feature for classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5933 LNAI, pp. 51-59, 2010