

# **Where am I? Scene Recognition for Mobile Robot using Audio Features**

Selina Chu, Shrikanth Narayanan, C.-C. Jay Kuo, and  
Maja Matarić

**Signal Analysis and Interpretation Laboratory**

Department of Computer Science  
Department of Electrical Engineering  
University of Southern California  
Los Angeles, CA 90089

Email: [selinach@sipi.usc.edu](mailto:selinach@sipi.usc.edu)

# Outline

- Using audio for robots?
- Audio features and classification
- Framework of system
- Data collection and experimental setup
- Feature selection
- Results
- Conclusion and Future work

# Robots with Vision

- Vision-based robot has limitations
  - Requires much world knowledge
    - *Model-based approaches*: typically built for highly constrained environment; Reliable landmarks are required for model matching [Dickmanns and Mysliwetz]
    - *Mapless approaches (view-based)*: incoming images are matched against learned ones. Effective for small area, but generalize poorly (new scenes) [Matsumoto et al.]
  - Image processing and segmentation algorithms are computationally expensive
  - Lighting problems (or lack of), and angle of the camera

# Robots with Ears

- Use audio information to assist robot navigation
- Audio data can be obtained at anytime, if available on robot, neglecting external condition
- However, little research has been done on using audio to recognize the environment
- Environmental sounds are unstructured, similar to noise; Speech and music are structured data;

# Overview

*Building a scene recognition system using audio features:*

- Collect real-world audio data with a robot and extract relevant features
- Build classifier based on audio features
- Discriminate between five types of environment
- Investigate on features for discrimination

# Type of Environmental Sounds

Street



Elevator



Café



Hallway

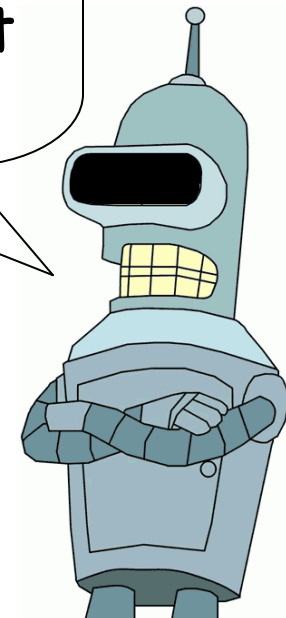


Lobby



I don't need vision...

What kind of sounds  
are we talking about  
here?



# Experimental Setup

- Hardware
  - Pioneer robot and Player for control
  - Edirol USB audio interface,
  - Sennheiser microphone mounted on the chassis of robot
- Record audio data with robot for training and testing
- 5 different scenes / environments (within and around a research facility building):
  - *Café* area, crowds of people
  - *Hallways* of research labs
  - *Elevator*, around and inside of
  - *Lobby* area, people walking thru with talking
  - *Street* area, next to building with pedestrian and vehicle traffics.



# Data Collection

- Collected about 2-3 hours of data from the five environments all together.
- Multiple clippings were taken at each location.
- Each clipping was 10-15 minutes long
  - Taken at multiple days and at various times.
    - Incorporate larger varieties of sound
    - Prevent biases of recording the same situation.
- Robot was deliberately being driven around with its sonar sensors turned on (and sometimes off)
  - Resembling a more realistic situation, including motor- and sonar-sounds
- Laser and camera were not used because they produce little, if any, noticeable sound.

# Dataset and Feature Extraction

**Dataset:** audio clips of environmental sounds

- Manually labeled and separated into 4-second segments
- Audio streams are sampled at 16 bits, mono-channel and 44 KHz
- 200 instances for each class (~13 minutes)
- Data was normalized to zero mean and unit variance

**5 Classes:** café, elevator, hallway, lobby, street

**Feature Extraction:** features are analyzed and extracted for every 20ms rectangular window frame, with 10 ms overlap. All spectra were computed with a 512-point FFT

# Typical Features for Audio Classification in Literature

Features	Eronen, 2006	Radhakrishnan, 2005	Eronen, 2005	Malkin, 2005	Rajapakse, 2005	Cano, 2004	Essid, 2004	Ahrendt, 2004	Herrera, 2002	De Santo, 2001	Peltonen, 2001	Moreno, 2000	Lu, 2000	El-Maleh, 2000	Nakajima, 1999	Srinivasan, 1999	Zhang, 1999	Carey, 1999	Liu, 1997
MFCC	X	X	X	X	X	X	X	X				X						X	
MFCC 1 <sup>ST</sup> Derivative	X		X				X					X							
Energy	X							X	X	X	X				X	X	X		
Zero-crossing	X							X			X		X	X		X	X	X	
Sub-band energy	X								X	X	X								X
Fundamental frequency															X		X		
Spectral centroid	X			X			X	X	X		X			X		X			X
Spectral bandwidth	X						X	X	X		X				X				X
Spectral roll-off	X										X								
Spectral flux	X										X								
Spectral flatness							X	X	X										X
Linear predictive coding	X							X			X								

# Typical Features for Audio Classification in Literature

Features	Eronen, 2006	Radhakrishnan, 2005	Eronen, 2005	Malkin, 2005	Rajapakse, 2005	Cano, 2004	Essid, 2004	Ahrendt, 2004	Herrera, 2002	De Santo, 2001	Peltonen, 2001	Moreno, 2000	Lu, 2000	El-Maleh, 2000	Nakajima, 1999	Srinivasan, 1999	Zhang, 1999	Carey, 1999	Liu, 1997
MFCC	X	X	X	X	X	X	X	X				X						X	
MFCC 1 <sup>ST</sup> Derivative	X		X				X					X							
Energy	X							X	X	X	X				X	X	X		
Zero-crossing	X							X			X		X	X		X	X	X	
Sub-band energy	X								X	X	X								X
Fundamental frequency															X		X		
Spectral centroid	X			X			X	X	X		X			X		X			X
Spectral bandwidth	X						X	X	X		X				X				X
Spectral roll-off	X										X								
Spectral flux	X										X								
Spectral flatness							X	X	X										X
Linear predictive coding	X							X			X								

Features used in this work...

# Types of Features

## Time-domain features:

- Zero-crossing
- Standard deviation of Zero-crossing
- Energy range
- Standard deviation of energy

## Frequency-domain features:

(Used fast Fourier transform (FFT) to convert signal)

- 1st – 12th MFCCs
- Standard deviation of 1st – 12th MFCCs
- Spectral centroid
- Spectral bandwidth
- Spectral asymmetry
- Spectral flatness
- Frequency roll-off
- Standard deviation of roll-off

# Classification Results of Related Work

	Number of classes	Classifiers	Overall Accuracy (%)	
<b>Unstructured</b>	<b>Environmental</b>	18	HMM	61
		13	KNN, GMM, HMM	56
		11	GMM	77
		4	GMM	81
<b>Structured</b>	<b>Music</b>	10	SVM, GMM	75
		9	KNN, decision tree	86
		8	Knowledge-based	77
		6	Knowledge based	85
		2	KNN, decision tree	99
<b>Structured</b>	<b>Speech, music, non-speech, silent</b>	5	KNN	88
		4	HMM	78
		4	Rule-based	82
		3	Thresholding,	90
		3	Thresholding	86
		2	Bayes, KNN, GMM	98
		2	GMM	89

# Experimental Setup & Result

## Classifiers used for comparison:

- Nearest Neighbor Classifier (1-NN)
  - Euclidean distance
- Support Vector Machine (SVM)
  - Kernel: 2-degree polynomial with  $C=10$  and  $\epsilon=1e-7$ , where  $C$  is the regularization parameter and  $\epsilon$  controls the fitting the training data, affecting number of support vectors
  - SVM is a 2-class classifier,
  - For 5-class, we use One-against-the-rest algorithm
- Gaussian Mixture Model (GMM)
  - Number of mixtures: 5

## Leave-One-Out cross-validation

# Results

	<b>KNN</b>	<b>GMM</b>	<b>SVM</b>
<b>Recognition Accuracy</b>	89.5%	89.5%	95.1%
<b>Time (sec)</b>	1.1	148.9	1681.8

- Highest recognition accuracy for discriminating 4-6 classes in audio: ~88% [Cano 2004, Eronen 2003, Herrera 2000]
- Results we obtained are competitive, but could be better

# Problem with Too Many Features

- Many irrelevant features might reduce the quality of classification
- Increasing number of features
  - Increase number of dimensions in search space
  - Data points become more sparse

We can alleviate problem by choosing smaller set

# Boosting with Feature Selection

- Select a subset from all possible features, yielding most 'effective' subset
- Optimal solution is an exhaustive search of all the features:  $\frac{m!}{d!(m-d)!}$ ,  $m = \#$  of features
- Resulting in  $2^{34} \approx 10^{10}$  combinations

# Forward Feature Selection

- Initialize selected set  $S = \text{empty set}$
- Initialize unselected set  $F = \{1, \dots, M\}$
- Repeat:
  - Evaluate performance with  $S \cup f_i$  for each  $f_i \in F$
  - $S := S \cup f_m$  and  $F := F \setminus f_m$ , where  $f_m$  give maximum performance
  - Stop when no significant improvement in classification

# More Results

Recognition Accuracy (percentage)

	KNN	GMM	SVM
Full Feature set (34)	89.5	89.5	95.1
Selected* Features (16~25)	94.3	93.4	96.6

\* Classifier dependent – 16 for KNN and SVM, 25 for GMM

## Selected 16 most effective features:

MFCC 1-3, 5-9, 12, Std dev of MFCC 1,4,5,7, spectral flatness, energy range, and frequency roll-off

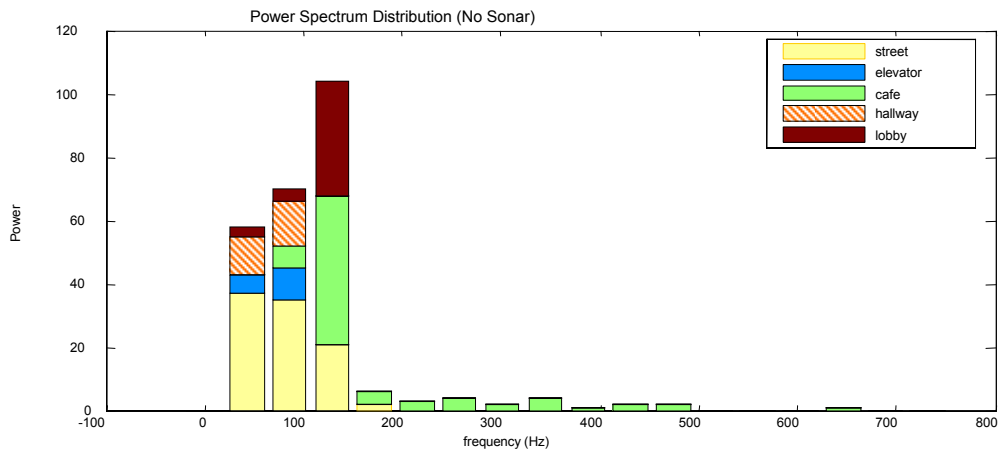
# Improving on Feature Selection

- Check for over-fitting and confirm validity of the selected features
- Repeat for 100 times
  - Randomly pick half of the dataset
  - Repeat the forward feature selection algorithm on the subset
  - Record the features selected
- Tally the selected features and picked the ones used more than half of the times
- Further refine the search using backward feature selection search

## Results:

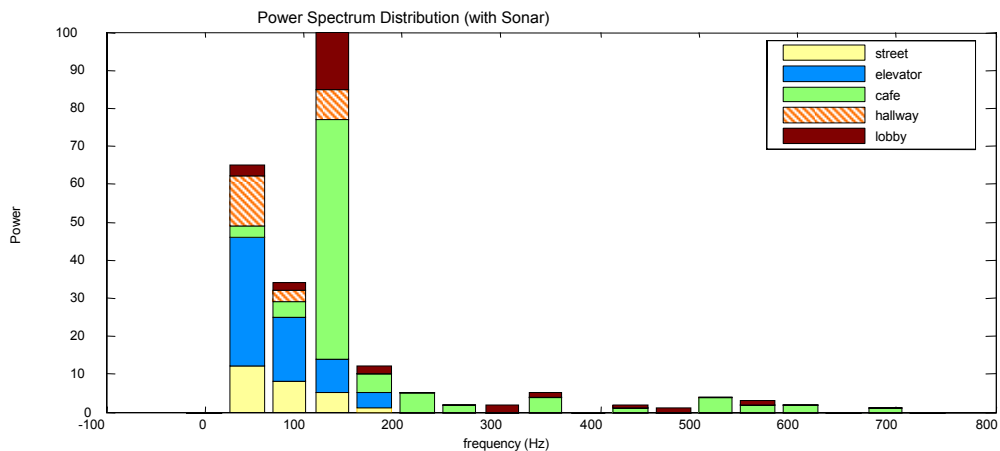
- Recognition accuracy: 94.2% using 9 features  
(compared with 89.5% for 34 features and 94.3% for 16 features)
- Selected 9 most relevant features for discriminations of environments:  
*MFCC1-3 and 7-9, zero-crossing, energy range, and the roll-off frequency*<sub>20</sub>

# A signal analysis perspective...



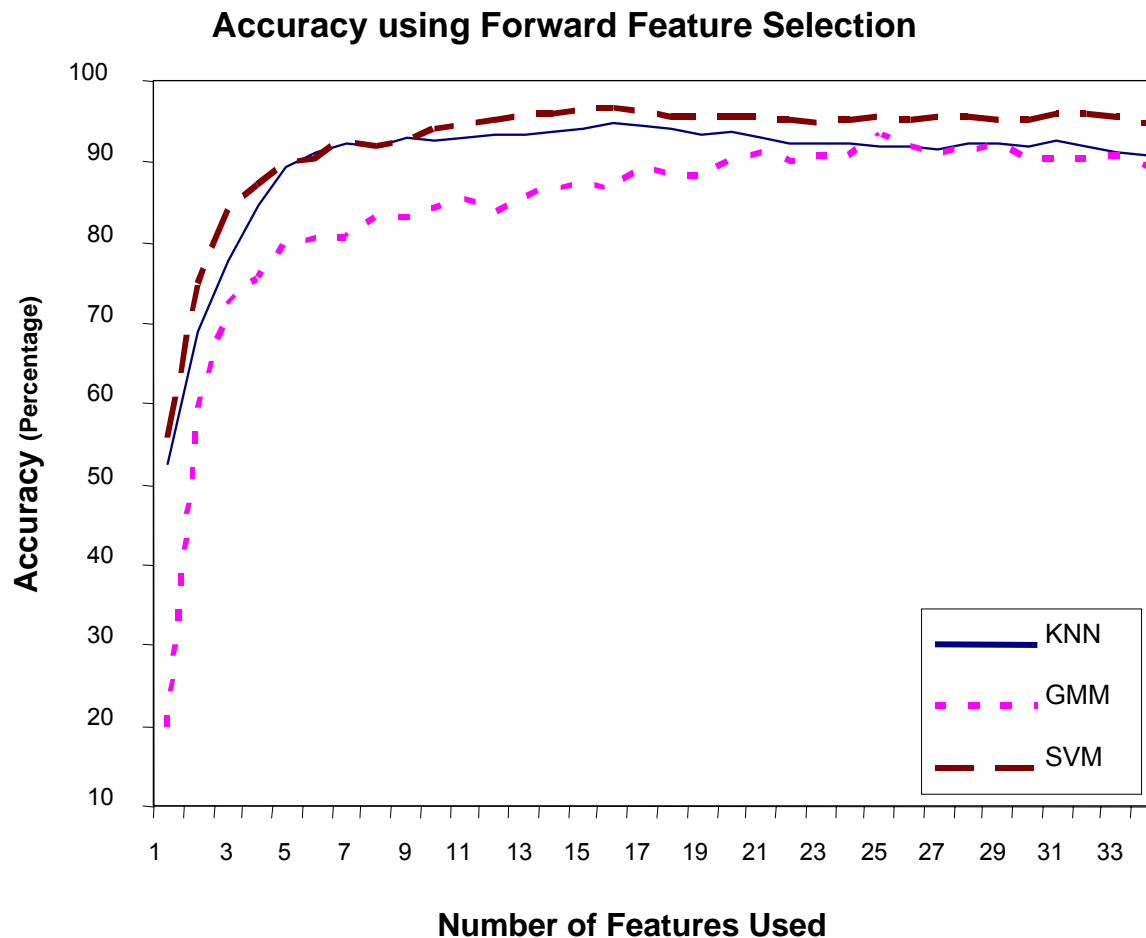
Class	Frequency Range (Hz)	Mean (Hz)	Mode (Hz)	% of data w/ mode value
<b>Street</b>	0-172	74.5	86	26.7
<b>Elevator</b>	0-172	72.1	43	43.3
<b>Café</b>	0-603	178.6	129	66.7
<b>Hallway</b>	0-172	66.0	58	64.4
<b>Lobby</b>	0-560	164.3	129	56.7

Café and Lobby have a wider distribution of frequencies, with both centering around 129 Hz



Class	Energy Range	Zero-Crossing	Frequency Roll-off
<b>Street</b>	81.7	14.2	3207
<b>Elevator</b>	87.5	18.6	2191
<b>Café</b>	75.3	30	2818
<b>Hallway</b>	91.3	11.9	2357
<b>Lobby</b>	84.3	29.1	2903

# A Closer Look ...



## Observation:

- Small number of features is needed to achieve high accuracy rate - ex. 6 features yields 91% for KNN and SVM
- KNN and SVM is relatively similar, but KNN is 1000x faster

# Conclusion

- Identified features relevant for discriminating the environments using audio
- Demonstrated feasibility of using audio for scene recognition and with competitive results
- Less is better -- demonstrated effectiveness of feature selection

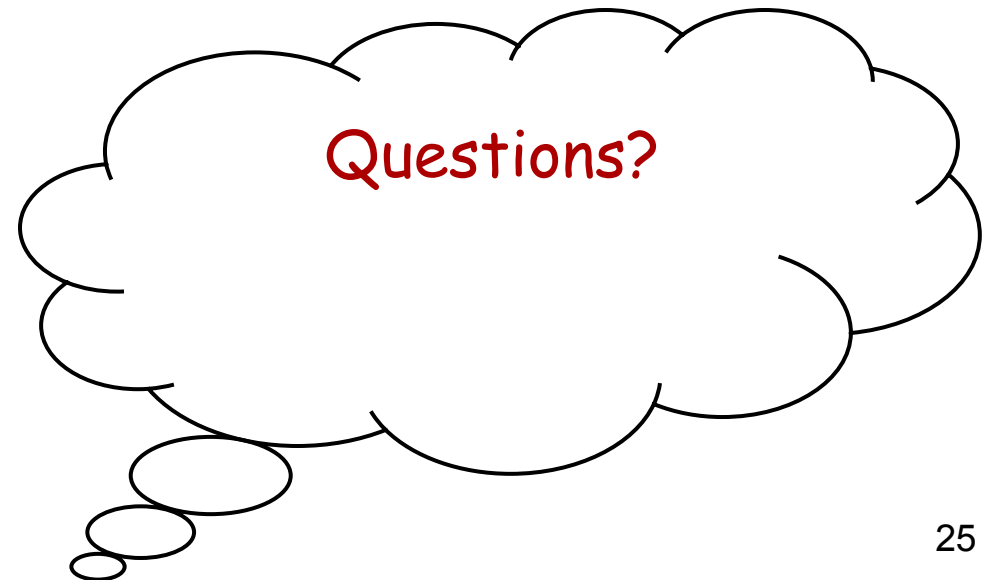
# Direction for Future Work

- Increase the types of environment for classification
  - Currently working on analysis of 20+ types of environmental sounds
- Investigate on mapping between audio and visual information
- Incorporate audio into a vision-based robot

# Thank you

[selinach@sipi.usc.edu](mailto:selinach@sipi.usc.edu)

<http://sail.usc.edu>



# References

- [1] DeSouza, G.N. and Kak, A.C. "Vision for mobile robot navigation: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2, pp. 237-267, 2002.
- [2] Matsumoto, Y., Inaba, M. and Inoue, H. "View-based approach to robot navigation," Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems, pp. 1702-1708, 2000.
- [3] Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. "Towards robotic assistants in nursing homes: challenges and results," Robotics and Autonomous Systems, Volume 42, Issues 3-4, pages 271-281, 2003.
- [4] Thrun, S., Bennewitz, M., Burgard W., Cremers, A.B., Dellaert, F., Fox, D., Haehnel, D. Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. "MINERVA: A second generation mobile tour-guide robot," Proc. of IEEE International Conference on Robotics and Automation, 1999.
- [5] Yanco, H.A. "Wheelesley, A Robotic Wheelchair System: Indoor Navigation and User Interface," Lecture Notes in Artificial Intelligence: Assistive Technology and Artificial Intelligence, Springer-Verlag, 1998
- [6] Fod, A., Howard, A., and Mataric, M. J., "Laser-Based People Tracking", Proc. of IEEE Int. Conf. Robotics and Automation, pages 3024-3029, 2002
- [7] <http://playerstage.sourceforge.net/>
- [8] Mitchell, T. M. "Machine Learning," Mc Graw-Hill, 1997
- [9] Moore, A. "Statistical Data Mining Tutorial on Gaussian Mixture Models," [www.cs.cmu.edu/~awm/tutorials](http://www.cs.cmu.edu/~awm/tutorials), CMU, 2004.
- [10] Scholkopf, B. and Smola, A.J. "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, 2002.
- [11] Burges, C.J. "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, 1998.
- [12] Silvia Allegro, Stefan Launer, Michael Büchler, Automatic Sound Classification Inspired by Auditory Scene Analysis, " Eurospeech, Aalborg, Denmark, 2001
- [13] Tong Zhang, C.-C. Jay Kuo, Audio content analysis for online audiovisual data segmentation and classification, in: IEEE Transactions on Speech and Audio Processing, May, Vol: 9, Issue: 4, pp: 441-457, 2001.
- [14] Antti Eronen, Juha Tuomi, Anssi Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, Jyri Huopaniemi, Audio-based context awareness – acoustic modeling and perceptual evaluation. In Proc. Intern. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, 2003
- [15] Robert Malkin, Alex Waibel, Classifying User Environment for Mobile Applications using Linear Autoencoding of Ambient Audio, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2005
- [16] Antti Eronen, Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Seppo Fagerlund, Timo Sorsa, Gaëtan Lorho, Jyri Huopaniemi, J. "Audio-based context recognition". IEEE Transactions on Speech and Audio Processing, 2006
- [17] Pedro Cano, Markus Koppenberger, Sylvain Le Groux, Julien Ricard, Nicolas Wack, and Perfecto Herrera, 'Nearest-neighbor generic sound classification with a wordnet-based taxonomy', in Proceedings of AES 116th Convention; Berlin, Germany, 2004.

# References

- [18] Massimo De Santo, Gennaro Percannella, Carlo Sansone, Mario Vento, "Classifying Audio Streams of Movies by a Multi-Expert System". 11th International Conference on Image Analysis and Processing, IEEE Computer Society Press, Palermo, Italy, pp. 386 - 391, 26 - 28 September , 2001.
- [19] Zhu Liu, Jincheng Huang, Yao Wang, and Tshuan Chen, Audio feature extraction & analysis for scene classification, in IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing, 1997.
- [20] Slim Essid, Gaël Richard, Bertrand David, Efficient musical instrument recognition on solo performance music using basic features," in AES 25th International Conference, London, UK, June 2004.
- [21] Perfecto Herrera, Alexandre Yeterian, Fabien Gouyon, Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques, Proceedings of Second International Conference on Music and Artificial Intelligence; Edinburgh, Scotland, 2002.
- [22] Pedro J. Moreno, Ryan Rifkin, Using The Fisher Kernel Method For Web Audio Classification, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2000.
- [23] Guojun Lu and Templar Hankinson, An investigation of automatic audio classification and segmentation, in Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on Signal Processing.
- [24] Khaled El-Maleh, Mark Klein, Grace Petrucci, and Peter Kabal, Speech/Music Discrimination for Multimedia Applications, in Proc. IEEE Intl. conf. on Acoustics, Speech, Signal Processing, pp. 2445-2446, 2000.
- [25] Yasuyuki Nakajima, Yang Lu, Masaru Sugano, Akio Yoneyama, Hiromasa Yanagihara, and Akira Kurematsu. A fast audio classification from MPEG coded data. In International Conference on Acoustics, Speech and Signal Processing, volume VI, pages 3005--3008. IEEE, 1999.
- [26] Savitha Srinivasan, Dragutin Petkovic, Dulce Ponceleon, Towards robust features for classifying audio in the CueVideo system, Proceedings of the seventh ACM international conference on Multimedia, 1999.
- [27] Vesa Peltonen, Computational Auditory Scene Recognition, M.S. thesis Department of Information Technology, Tampere University of Technology, Tampere Finland, 2001.
- [28] Michael J. Carey, Eluned S. Parris, and Harvey Lloyd-Thomas, "A comparison of features for speech, music discrimination," in Proceedings of ICASSP 99, 1999, pp. 149-152, Phoenix, Arizona
- [29] Menaka Rajapakse and Lonce Wyse, Generic Audio Classification Using a Hybrid Model Based on GMMs and HMMs, Proceedings of the 11th International Multimedia Modelling Conference (MMM'05), 2005.
- [30] Peter Ahrendt, Anders Meng and Jan Larsen, Decision time horizon for music genre classification using short time features, in proc. of EUSIPCO 2004, 6-10 Sep 2004, Vienna.
- [31] Regunathan Radhakrishnan, Ajay Divakaran, Paris Smaragdis, Audio analysis for surveillance applications," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.

Quick comments..

slides are over all good, and flow well.

minor comments

slide 9: may be play one example max (or none). cartoon is nice!

slide 10: difficult to read legend and, importantly, what the figures tell us

slides 11,12: make table a little bigger so it is readable

slide 14: "Works" --> "Work"

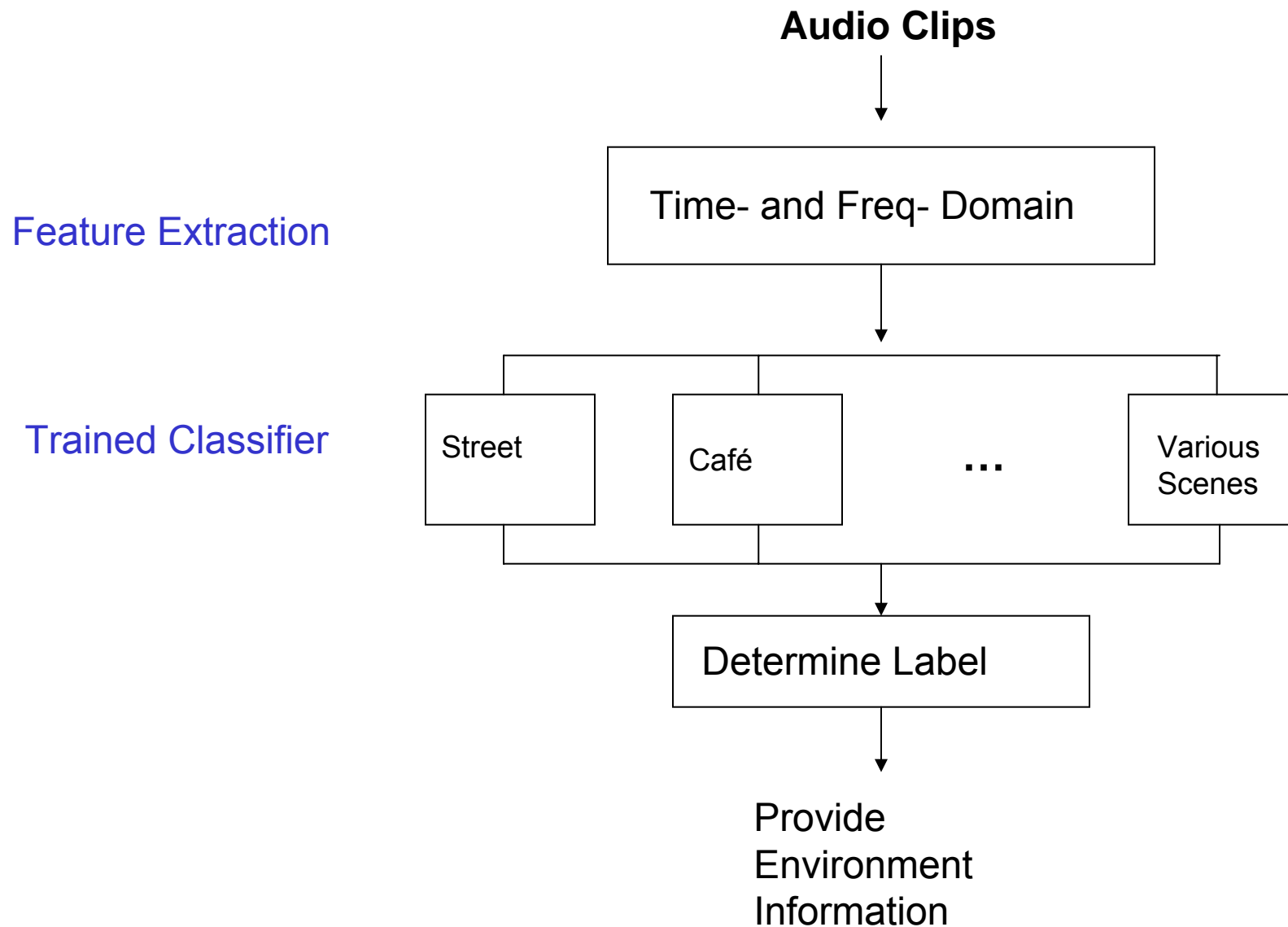
don't need to show slides 25 & 26; move 27 right after conclusions.

Practice! And good luck.

# Confusion Matrix

	Street	Elevator	Café	Hallway	Lobby
Street	94.4	0	0	0	5.6
Elevator	0	90.0	1.1	7.8	1.1
Café	0	0	95.6	0	4.4
Hallway	0	0	0	100	0
Lobby	2.2	0	3.3	0	94.4

# Framework of Classifier



# Types of Features

## Time-domain features:

- Zero-crossing
- Standard deviation of Zero-crossing
- Energy range
- Standard deviation of energy

$$Z(n) = \frac{1}{2} \sum \text{sgn}[S(m)] - \text{sgn}[S(m-1)]w(n-m)$$

$$\text{where } \text{sgn}[S(n)] = \begin{cases} 1 & S(n) \geq 0 \\ -1 & S(n) < 0 \end{cases}$$

and  $w(n)$  is the window frame.

## Frequency-domain features:

(Used fast Fourier transform (FFT) to convert signal)

- 1st – 12th MFCCs
- Standard deviation of 1st – 12th MFCCs
- Spectral centroid
- Spectral bandwidth
- Spectral asymmetry
- Spectral flatness
- Frequency roll-off
- Standard deviation of roll-off

$$S_c = u_1$$

$$S_w = \sqrt{u_2 - u_1^2}$$

$$S_a = \frac{2u_1^3 - 3u_1u_2 + u_3}{S_w^3} - 3$$

$$S_f = \frac{-3u_1^4 + 6u_1u_2 - 4u_1u_3 + u_4}{S_w^4}$$

$$u_i = \frac{\sum_{k=0}^{N-1} f_k^i A_k}{\sum_{k=0}^{N-1} A_k}$$

# Observation of Each Scene

- *Hallway*: mostly quiet, with occasional doors opening/closing, distant sound from the elevators, and individuals quietly talking, and some footsteps
- *Café*: many people talking, ringing of the cash registers, moving of chairs
- *Lobby*: footsteps with certain echo, but different from hallways, due to the type of flooring, people talking, sounds of rolling dollies from deliveries being made
- *Elevators*: bells and alerts from the elevator, footsteps, rolling of dollies on steel surface of the elevator
- *Outside*: footsteps on concrete, traffic from buses and cars, bicycles, and occasional planes and helicopters

# Robots with Vision

- **Vision-based system has limitations**
  - Requires much world knowledge
    - *Model-based approaches*: typically built for highly constrained environment; Reliable landmarks are required for model matching [Dickmanns and Mysliwetz]
    - *Mapless approaches (view-based)*: incoming images are matched against learned ones. Effective for small area, but generalize poorly (new scenes) [Matsumoto et al.]
  - Image processing and segmentation algorithms are computationally expensive
  - Lighting problems (or lack of)
- **Determine the modality of control**
  - Mobile robots designed for indoor/outdoor navigation usually needs to switch between indoor and outdoor modes of control [Yanco, Kosaka and Kak.]