Acoustic scene and events recognition:
how similar is it to speech recognition and music genre/instrument recognition ?

G. Richard
DCASE 2016

Thanks to my collaborators:
S. Essid, R. Serizel, V. Bisot

Content

- **Some tasks in audio signal processing:**
  - What is scene recognition and sound event recognition ?
  - What is speech recognition/speaker recognition/Music genre recognition,… ?
  - How similar are the different problems ?
  - Are the tasks difficult for humans ?

- **(Very) Brief historical overview of speech/audio processing**
- **Looking at recent trends for acoustic scenes recognition (DCASE2016)**
- **A recent and specific approach**
- **Discussion/Conclusion**

Acoustic scene and sound event

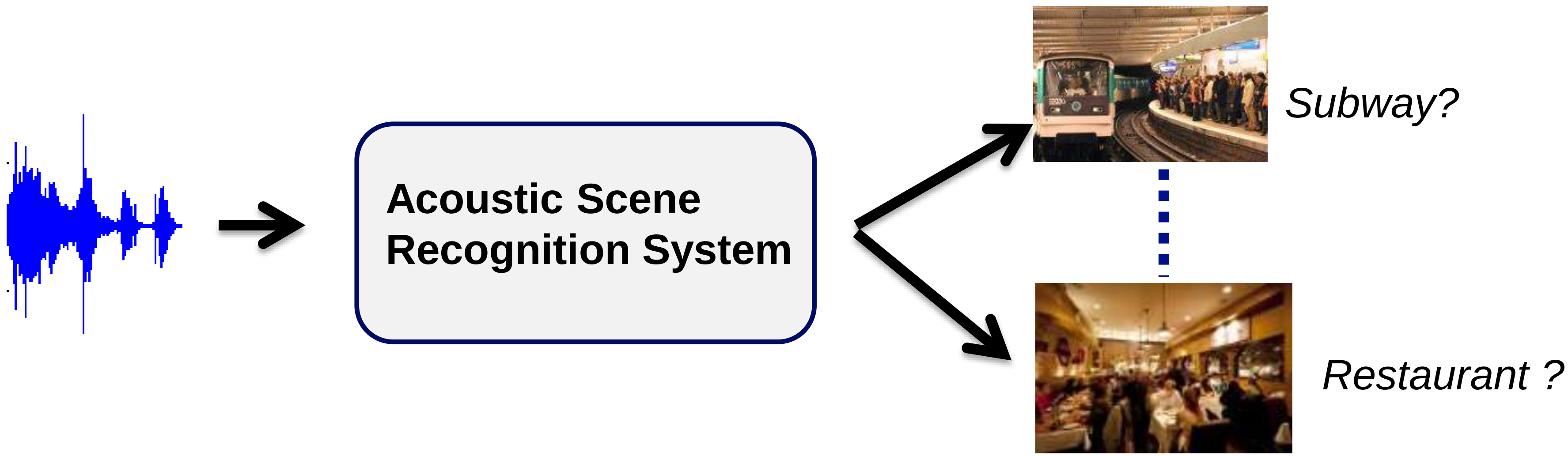
- **Some example of acoustic scenes**

- **Some example of sound events**

Acoustic scene and sound event recognition

Acoustic scene recognition:

- « associating a semantic label to an audio stream that identifies the environment in which it has been produced »

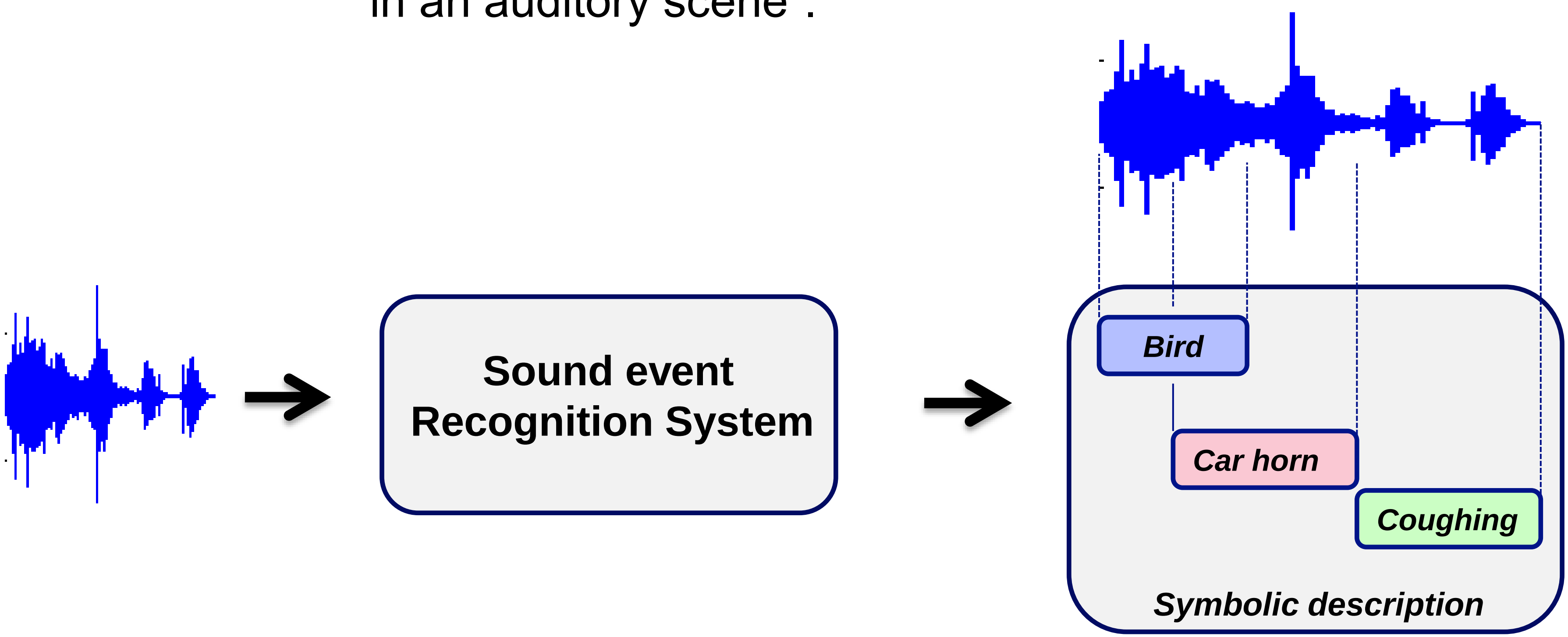


- Related to CASA (*Computational* Auditory Scene Recognition) and SoundScape cognition (*psychoacoustics*)

D. Barchiesi, D. Giannoulis, D. Stowell and M. Plumbley, « Acoustic Scene Classification », IEEE Signal Processing Magazine [16], May 2015

Acoustic scene and sound event recognition

- **Sound event recognition**
  - “aims at transcribing an audio signal into a symbolic description of the corresponding sound events present in an auditory scene”.



Applications of scene and events recognition

- **Smart hearing aids (Context recognition for adaptive hearing-aids, Robot audion,..)**
- **Security (*see for example the LASIE project*)**
- **indexing,**
- **sound retrieval,**
- **predictive maintenance,**
- **bioacoustics,**
- **environment robust speech reco,**
- **ederly assistance**
- **…..**

Use Case 3: The Missing Person: http://www.lasie-project.eu/use-cases/

- **Is « Acoustic Scene/Event Recognition » just the same as**
  - Speech recognition ?
  - Speaker recognition ?
  - Music genre recognition ?
  - Music instrument reccognition ?
  - …

What is speech recognition ?

- **From Speech to Text**

« I am very happy to be here…. »

Input is an audio signal
Output: sequence of words
Associates an « acoustic recognition » model and a « language model

Acoustic model:

- Classification of an audio stream in 35 classes (« phonemes ») … but many more if triphones are considered (even with *tied-states*)
- Class should be independant of the speaker and of pitch

What is speaker recognition ?

- **Recognizing who speaks**

« Tuomas Virtanen »

Input is an audio signal
Output: name of a person
No language model

Acoustic model:
- Classification of an audio stream in N classes (« speakers »)
- Class should be independant of the individual events (phonems) pronounced

What is Music genre recognition ?

- **From music to genre label**

« Modern Jazz »

Input is an audio signal
Output: Genre of the music
No language model, but hierarchical model possible

Acoustic model:

- Classification of an audio stream in N classes (« genre »)
- Class should be (more or less) independant of the individual events (instruments, pitch, harmony, …).

What is Music instrument recognition ?

- **From music to instrument labels**

« Tenor saxophone, Bass, piano »

Input is an audio signal
Output: name of the instrument playing concurrently
No language model, but hierarchical model possible

Acoustic model:

- Classification of an audio stream in N classes (« instruments »)
- Multiple classes active concurrently
- Class should be (rather) independant of pitch.

- **Is « Acoustic Scene/Event Recognition » as difficult for humans as**

  - Speech recognition ?
  - Speaker recognition ?
  - Music genre recognition ?
  - Music instrument recognition ?
  - …

Complexity of the tasks for humans ….

- **Speech recognition :**
  - 0.009% error rate for connected digits
  - 2 % error rate for non sense sentences (1000 words vocabulary)
  - Phoneme recognition (CVC or VCV) in noise: 25% error rate at -10db SNR

- **Speaker recognition**
  - About 1.3% of False Alarm and 3% Misses in a task « are the two speech signals from the same speaker ? »

R. Lippmann, *Speech recognition by machines and humans*, Speech Communication, Vol. 22, No 1, 1997
B. Meyer & al. "Phoneme confusions in human and automatic speech recognition", Interspeech 2007
W. Shen & al., "Assessing the speaker recognition performance of naive listeners using mechanical turk," in Proc. of ICASSP *2011*

Complexity of the tasks for humans ….

Music Genre recognition

- 55% accuracy (on average) for 19 musical genres including « Electronic&Dance”, “Hip-Hop », « Folk » but also « easylistening », « vocals »

Music instrument recognition

- 46% for isolated tones to 67 % accuracy for 10s phrases for 27 instruments

Sound scenes recognition

- 70% accuracy for 25 acoustic scenes

K. Seyerlehner, G. Widmer, P. Knees “Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems”, In Proc. of Workshop on Adaptive Multimedia Retreival (AMR-2010), 2010.
Martin. (1999). “Sound-Source Recognition: A Theory and Computational Model”. Ph.D. thesis, MIT
V. Pelton & al., “Recognition of everyday auditory scenes : Potentials, latencies and cues, in Proc. AES, 2001

- **A (very) brief historical overview of**
  - Speech Recognition
  - Music instrument/genre recognition
  - Acoustic scenes/Event recognition

An overview of speech recognition

1952: Analog Digit Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1956: Analog 10 syllable recognition
1 speaker
Features: Filterbank (10 filt.)

1962: Digital vowel Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1971: Isolated word Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U detection, Formant center frequencies, energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of speech recognition

1962: Digital vowel
Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1952: Analog Digit
Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

1956: Analog 10 syllable
recognition
1 speaker
Features: Filterbank (10 filt.)

1971: Isolated word
Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based
Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U
detection, Formant center frequencies,
energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of speech recognition

1952: Analog Digit Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1956: Analog 10 syllable recognition
1 speaker
Features: Filterbank (10 filt.)

1962: Digital vowel Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1971: Isolated word Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U detection, Formant center frequencies, energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of speech recognition

1952: Analog Digit Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1956: Analog 10 syllable recognition
1 speaker
Features: Filterbank (10 filt.)

1962: Digital vowel Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1971: Isolated word Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U detection, Formant center frequencies, energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : Genre recognition
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : Instrument recognition (polyphonic music)
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : Instrument recognition : exploiting source separation, dictionary learning
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition, *Baker, Jelinek, Rabiner , …*

1993 Computational ASA (Audio stream segregation)
Use of auditory periphery model
Blackboard model ('IA)
M. Cook & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2014 - :
DNN for acoustic event recognition
Gencoglu & al..

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner , …

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA
(Audio stream segregation)
Use of auditory periphery model
Blackboard model ('IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al..

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner ,…

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA
(Audio stream segregation)
Use of auditory periphery model
Blackboard model (‘IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al..

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner ,…

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA
(Audio stream segregation)
Use of auditory periphery model
Blackboard model ('IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al..

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner ,…

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA (Audio stream segregation)
Use of auditory periphery model
Blackboard model ('IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al, ...

- **And in 2016 ….**
  - The example of Acoustic Scene recognition (DCASE2106)

The (partial) figure in 2016
(from DCASE 2016 – Acoustic Scene Detection)

Accuracy (Eval)	Input	Features	Classifier
89.7 %	mono+binaural	MFCC+spectrograms	fusion
88.7 %	binaural	MFCC	I-vector
87.7 %	mono	spectrogram	NMF
87.2 %	mono	various	fusion
86.4 %	mono	various	fusion
86.4 %	binaural	MFCC	I-vector
86.2 %	mono	mel energy	CNN
85.9 %	mono	MFCC distribution	SVM
85.6 %	mono	MFCC	DNN-GMM
85.4 %	mono	unsupervised	CNN ensemble
84.6 %	mono	mel energy	CNN
84.1 %	mono	various	ensemble
84.1 %	mono	spectrogram	CNN-RNN
83.8 %	mono	MFCC+mel energy	fusion
83.6 %	mono	MFCC+mel energy	fusion
83.3 %	mono	CQT	CNN
83.3 %	mono	spectrogram	CNN
83.3 %	mono	label tree embedding	CNN
83.1 %	mono	MFCC+mel energy	fusion
83.1 %	mono	MFCC	kNN

The (partial) figure in 2016
(from DCASE 2016 – Acoustic Scene Detection)

- **Some observations:**
  - Few systems exploit spatial information
  - … even though it is one of the important ideas of CASA…
  - It seems that spatial information helps

  (as in speech recognition but has probably more potential here)

Input
mono+binaural
binaural
mono
mono
mono
binaural
mono
mono
mono
mono
mono
mono
mono
mono
mono
mono
mono
mono
mono
mono

The (partial) figure in 2016
(from DCASE 2016 – Acoustic Scene Detection)

- **Some observations:**
  - MFCC are still very popular which seems surprising since an audio scene is not a speech signal :
    - 11 of the top 20 systems use MFCC

Features
MFCC+spectrograms
MFCC
spectrogram
various
various
MFCC
mel energy
MFCC distribution
MFCC
unsupervised
mel energy
various
spectrogram
MFCC+mel energy
MFCC+mel energy
CQT
spectrogram
label tree embedding
MFCC+mel energy
MFCC

Are MFCC appropriate for acoustic scene/event recognition ?

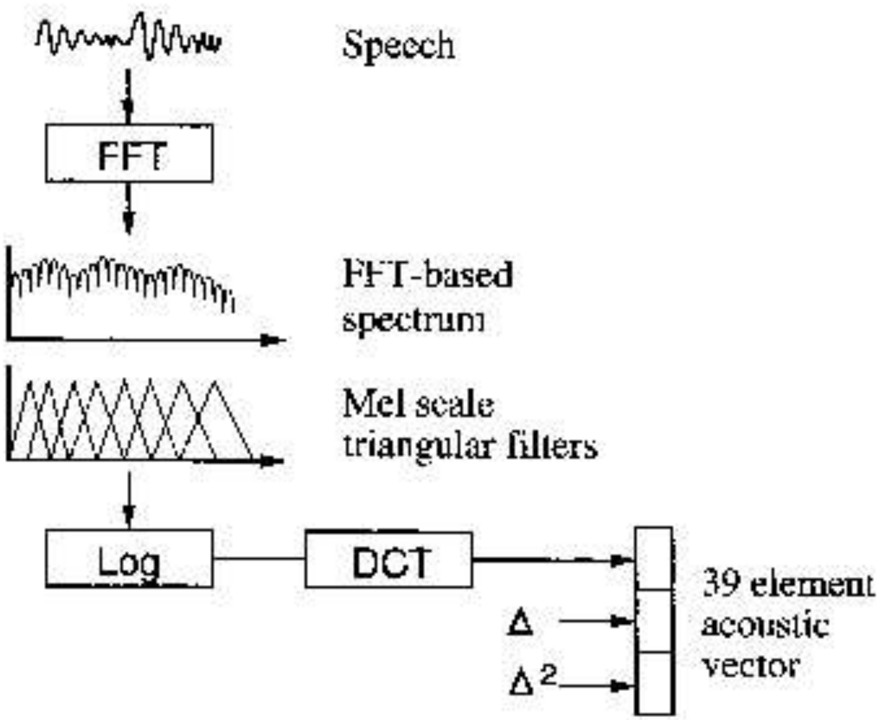
- Pitch range is much wider in audio signal than in speech
- For high pitches the deconvolution property of MFCCs does not hold anymore (e.g. MFCC become pitch dependent)
- Their global characterization prevents MFCCs to describe localised time-frequency information and in that sense they fail to model well-known masking properties of the ear.
- MFCC are not highly correlated with the perceptual dimensions of “polyphonic timbre” in music signals despite their widespread use as predictors of perceived similarity of timbre.
- Sometimes MFCC are used exactly as for 8kHz sample speech (e.g. 13 coefficients) …

➡ **Their use in general audio signal processing is therefore not well justified**

G. Richard, S. Sundaram, S. Narayanan "An overview on Perceptually Motivated Audio Indexing and Classification", Proceedings of the IEEE, 2013.
A. Mesaros and T. Virtanen, “Automatic recognition of lyrics in singing,” EURASIP Journal on Audio, Speech, and Music
B. Processing, vol. 2010, no. 1, p. 546047, 2010.
V. Alluri and P. Toiviainen, “Exploring perceptual and acoustical correlates of polyphonic timbre,” Music Perception, vol. 27, no. 3, pp. 223–241, 2010

What are MFCC ?
« Mel-Frequency Cepstral Coefficients »

- **The most widely spread speech features (before 2012…)**



What do the MFCC model ?

- **Interest**
  - Speech source-filter production model (*Fant 1960*)

$$s(t) = g(t) * h(t)$$

✓ The model in spectral domain

$$S(\omega) = G(\omega)H(\omega)$$

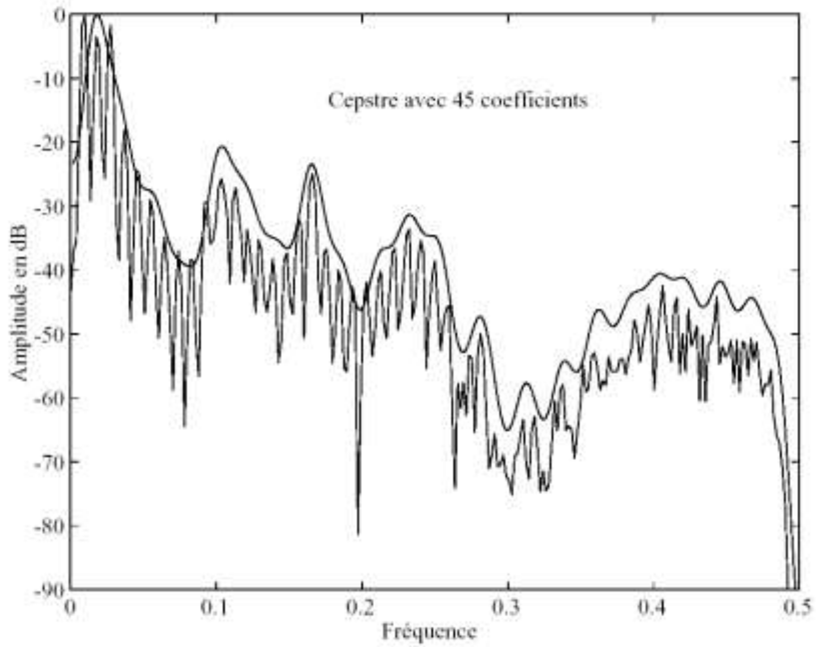
✓ Cepstre (real): a sum of two terms

$$c(\tau) = FFT^{-1} \log |S(\omega)| = FFT^{-1} \log |G(\omega)| + FFT^{-1} \log |H(\omega)|$$

✓ Source contribution is removed by selecting the first few cepstral coefficients

MFCC capture "global" spectral envelope

- **Fourier transform of the cepstrum (first 45 coefficients)**



- **It seems that MFCC's capacity to capture "global" spectral envelope properties is the main reason of their success in audio classification tasks.**

The (partial) figure in 2016 (from DCASE 2016…)

- **Some observations:**
  - All but 4 systems use Neural Networks
  - ….. But the best systems without fusion do not use Neural networks
  - Other recent ideas:
    - Use of i-vectors (from speaker recognition)
    - Exploit decomposition techniques (NMF)

Classifier
fusion
I-vector
NMF
fusion
fusion
I-vector
CNN
SVM
DNN-GMM
CNN ensemble
CNN
ensemble
CNN-RNN
fusion
fusion
CNN
CNN
CNN
fusion
kNN

- **A (very) recent system for Acoustic Scene recognition proposed in DCASE2016**
  - An alternative approach to DNN

V. Bisot, R. Serizel, S.Essid and G. Richard, “Supervised NMF for Acoustic Scene Classification, techn rep. DCASE2016 challenge, 2016.
V. Bisot, R. Serizel, S.Essid and G. Richard, Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification, submitted to special issue of IEEE Trans. On ASLP, 2016
Available at: https://hal.archives-ouvertes.fr/hal-01362864

Some hypotheses

- **Hypotheses**
  - An acoustic scene is characterised by the nature and occurrence of specific events
    - *A car horn is mostly present in streets*
  - Most of the events have specific time-frequency content
- **Objective :** to find a mean to capture event occurencies and time-frequency content for acoustic scene recognition

An Acoustic Scene recognition system

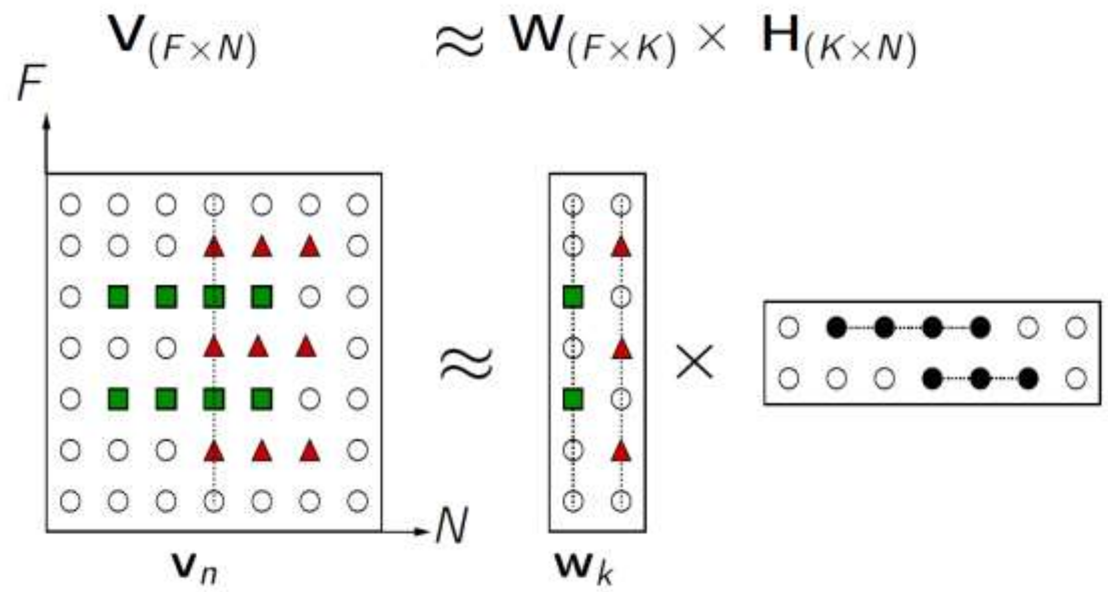
- Aim t**o decompose audio scene spectrograms in events using matrix factorization**
  - Learn a dictionary of audio event
  - Use as features the projections on the learned dictionary

- **Additional possibility:**
  - Jointly learn the dictionary and the classifier
  - Take into account the multi-class aspect of the problem

V. Bisot, R. Serizel, S.Essid and G. Richard, “Supervised NMF for Acoustic Scene Classification, techn rep. DCASE2016 challenge, 2016.

V. Bisot, R. Serizel, S.Essid and G. Richard, Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification, submitted to special issue of IEEE Trans. On ASLP, 2016

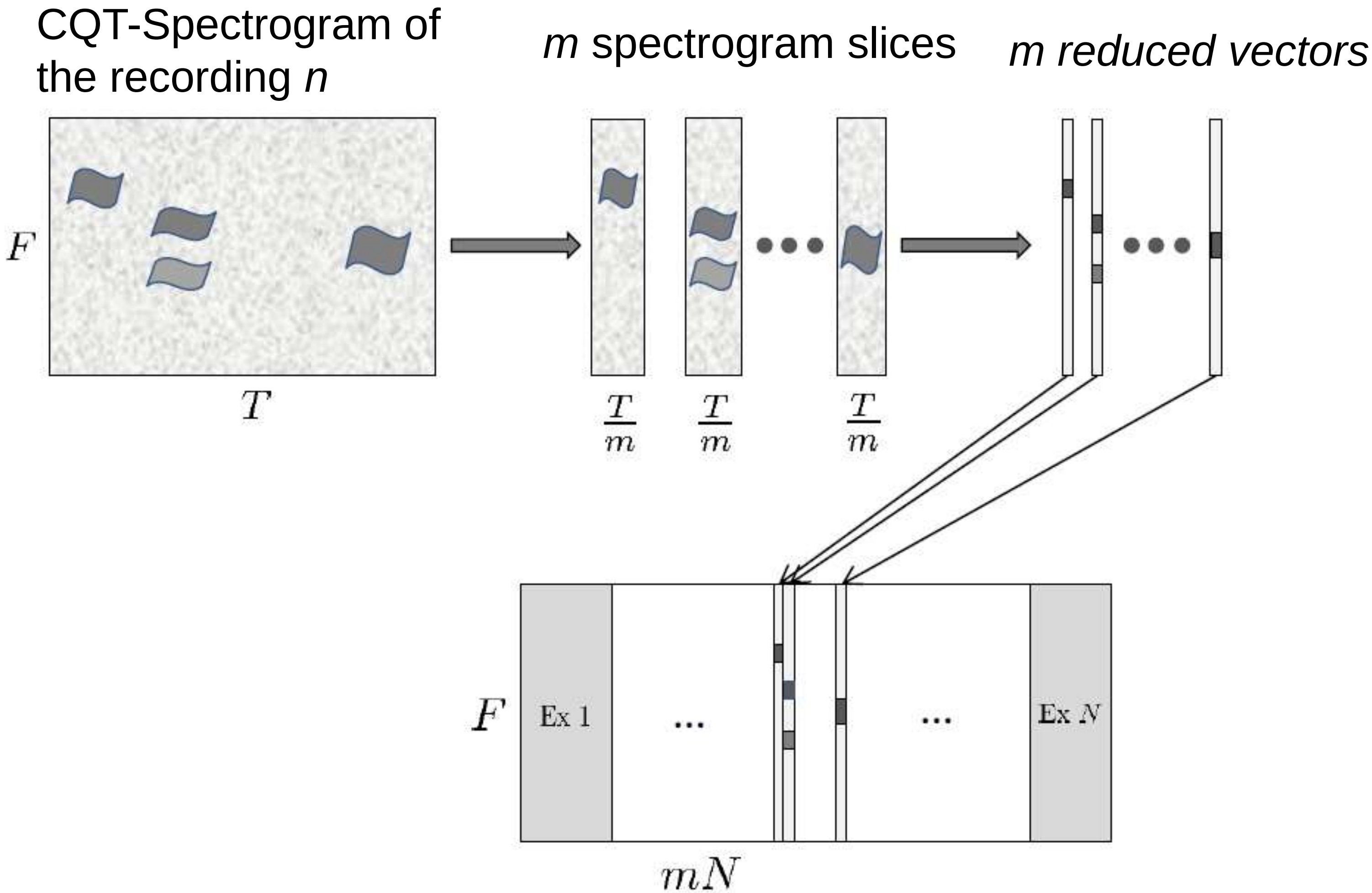
Matrix factorization for feature learning



- **V** is the data Matrix
- **W** is the learned « dictionary » Matrix
- **H** is the « activation » matrix and the learned features

D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, 1999.

Data matrix

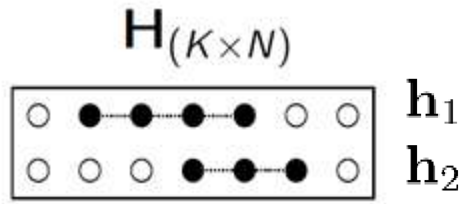


Data Matrix $\mathbf{V} \in \mathbb{R}^{F \times mN}$

Feature and Classifier

Input feature for each recording

- The average of each $\mathbf{h}_k$



$$\overline{\mathbf{h}}_i = \frac{1}{M} \sum_m h_i(m)$$

Classifier

- Multinomial Linear Logistic Regression

Multinomial Linear Logistic Regression

- **Classifier cost to be minimized:**

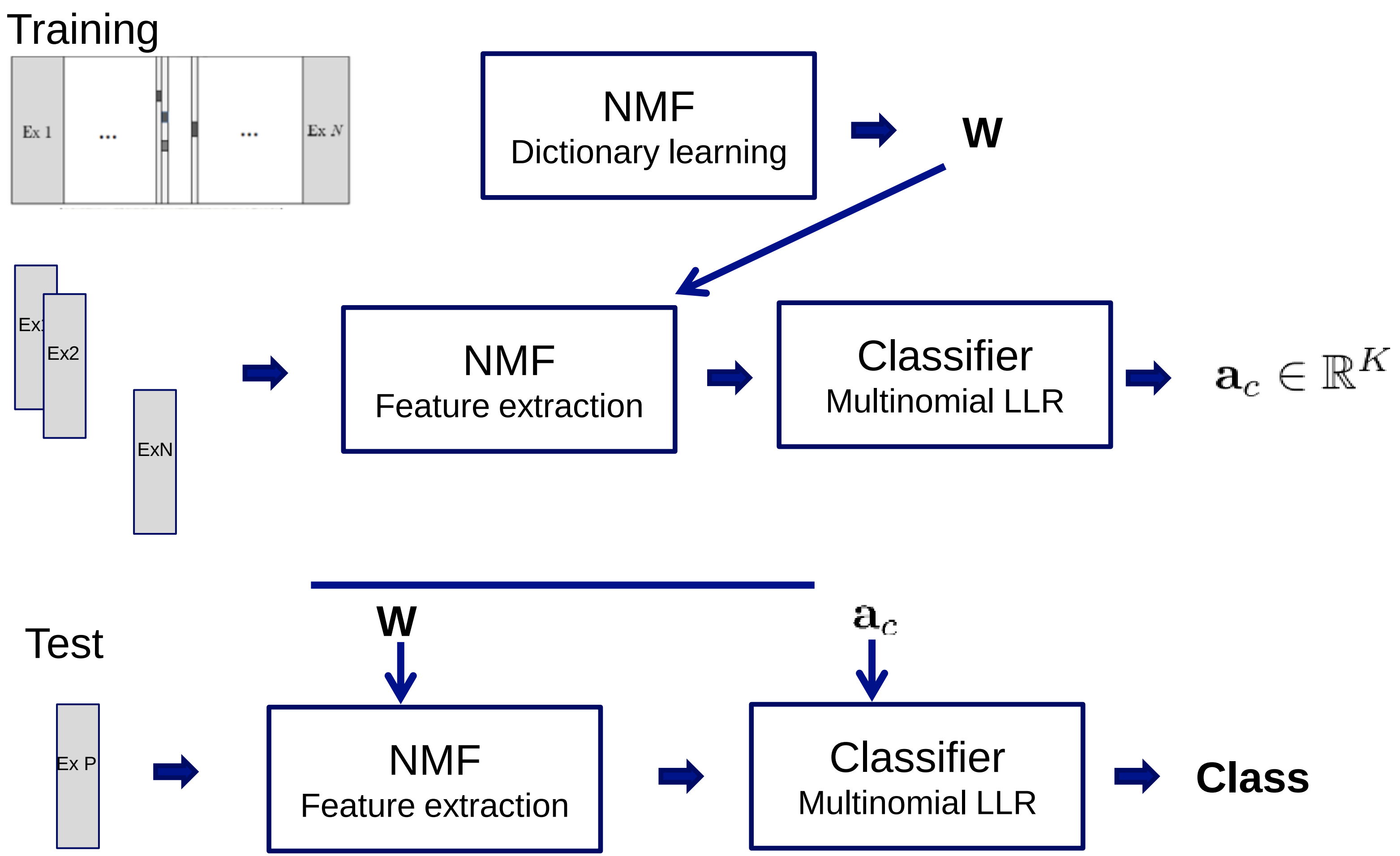
$$\ell_s(y, \mathbf{h}) = -\log(P(y = c|\mathbf{h}))$$

- **With**

$$\begin{cases} \mathrm{P}(y = c|\mathbf{h}) = \dfrac{e^{(b_c + \mathbf{a}_c^T \mathbf{h})}}{1 + \sum_{j=1}^{C-1} e^{(b_j + \mathbf{a}_j^T \mathbf{h})}} \ ; c = 1, ..., C-1 \\ \\ \mathrm{P}(y = C|\mathbf{h}) = \dfrac{1}{1 + \sum_{j=1}^{C-1} e^{(b_j + \mathbf{a}_j^T \mathbf{h})}} \end{cases}$$

  - $\mathbf{a}_c \in \mathbb{R}^K$ are the classifier weights
  - y is one of the possible label

In summary


Training
Ex 1
...
...
Ex N
NMF
Dictionary learning
W
Ex1
Ex2
ExN
NMF
Feature extraction
Classifier
Multinomial LLR
$\mathbf{a}_c \in \mathbb{R}^K$
W
$\mathbf{a}_c$
Test
Ex P
NMF
Feature extraction
Classifier
Multinomial LLR
Class

What can be improved ?

- **Exploit more sophisticated and task-adapted NMF**
  - Sparse NMF: towards more interpretable decomposition
  - Convolutive NMF: to exploit 2D dictionnary elements
  - …
- **Jointly learn the dictionnary for feature extraction and the classifier**
  - For example : *Task driven Dictionnary Learning*

J. Mairal, F. Bach, and J. Ponce, “Task-driven dictionary learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 791–804, 2012.

Task driven Dictionnary Learning (TDL)

- **Supervised dictionary learning**

- Aim of TDL: jointly learn a good dictionary and the classifier along with activation sparsity constraints

  ➡ Classify optimal projections on the dictionary

- Solving the following problem:

$$\mathbf{h}_i^\star(\mathbf{v}_i, \mathbf{W}) = \min_{\mathbf{h}\in\mathbb{R}^K} \ \frac{1}{2}||\mathbf{v}_i - \mathbf{W}\mathbf{h}||_2^2 + \lambda_1 \left|\left|\mathbf{h}\right|\right|_1 + \frac{\lambda_2}{2} \left|\left|\mathbf{h}\right|\right|_2^2$$

Adapted algorithm

- **Adaptation to our task**
  - Classifying averaged projections $\overline{\mathbf{h}}_i = \frac{1}{M}\sum_m h_i(m)$
  - Exploit a Multinomial Linear Logistic Regression classifier (as before)
  - Force non negativity for activations (e.g. projections)

V. Bisot, R. Serizel, S.Essid and G. Richard, “Supervised NMF for Acoustic Scene Classification, techn rep. DCASE2016 challenge, 2016.
V. Bisot, R. Serizel, S.Essid and G. Richard, Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification, submitted to special issue of IEEE Trans. On ASLP, 2016
Available at: https://hal.archives-ouvertes.fr/hal-01362864

Results

- **This approach is efficient for Acoustic scene classification**
  - Ranked 3rd in DCASE2016 challenge without exploiting DNN (but a little bit of fusion).
  - Is better than our DNN approach using the same datamatrix for the DCASE2016 development dataset
  - But less good (but not statistically significant) than DNN on LITIS dataset which is larger

Discussion / Wrap up

- **Acoustic Scene Recognition and Audio event recognition is a more recent field than speech recognition, speaker recognition, MIR, …**

- **The problems are « similar »**
  - The input signal is an audio signal
  - The problem is to classify the input signal in different classes

- **… but also different**
  - The classes are very different and always well defined
  - The audio signal is a complex mixtures of overlapping individual sounds which may be never observed in isolation or quiet environment
  - Cannot really use a « Language » model, but taxonomy is possible
  - The number of classes may differ very significantly…

Discussion / Wrap up

- **The influence of Speech domain is natural**
  - Due to the proximity of the different problems,
  - Due to the fact that the speech community is much larger and has a stronger past history
  - Due to the fact that speech models are trained on much larger and varied datasets
  - Speech recognition is a complex audio signal classification problem.

- **it is then natural to find in Acoustic Scene and Event Recognition the solutions proposed for speech/speaker**
  - MFCC, i-vectors, GMM, HMM, ….and now DNNs
  - And DNNs do work in scene/event recognition

Discussion / Wrap up

- **But the problem is also different and calls for task designed and adapted methods**
  - Adapted to the specificities of the problem
  - Adapted to the scarcity of training (*annotated*) data
  - Adapted to the fact that individual classes (especially events) may be only observed in mixtures
  - Potential of novel paths is shown in the DCASE2016 results

Conclusion

- Yes, we are right in looking what the speech processing community is doing
- … but we should adapt their findings to our problem
- … and It is worth looking other domains…
- … and it is worth developping new methods which are not a direct application of speech methods
- … There may be a life besides DNNs especially for Acoustic Scene and Event recognition