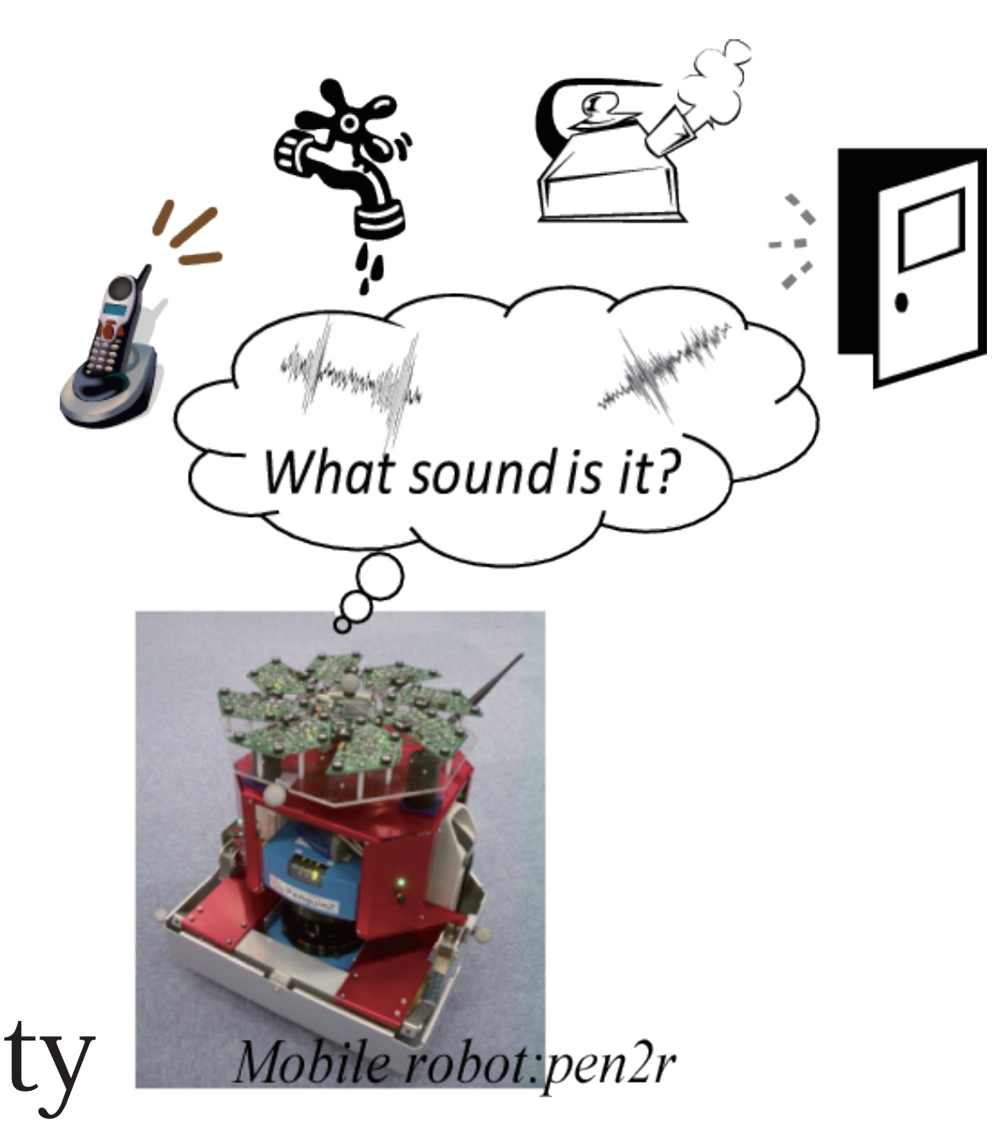
A Nested Infinite Gaussian Mixture Model for Identifying Known and Unknown Audio Events

Yoko Sasaki Kazuyoshi Yoshii Satoshi Kagami (AIST)

National Institute of Advanced Industrial Science and Technology
AIST
Digital Human Research Center

Summary

Goal:
Identify audio events occurring in the real environment



Problems:
1) All kinds of audio event classes cannot be defined in advance
2) Each class has unique acoustic characteristics with varying complexity

Approach:
Formulate a nonparametric Bayesian model called a nested infinite GMM (Gaussian mixture model) consisting of
- infinitely many GMMs (infinitely many classes considered)
- infinitely many Gaussians in each GMM (flexible acoustics)

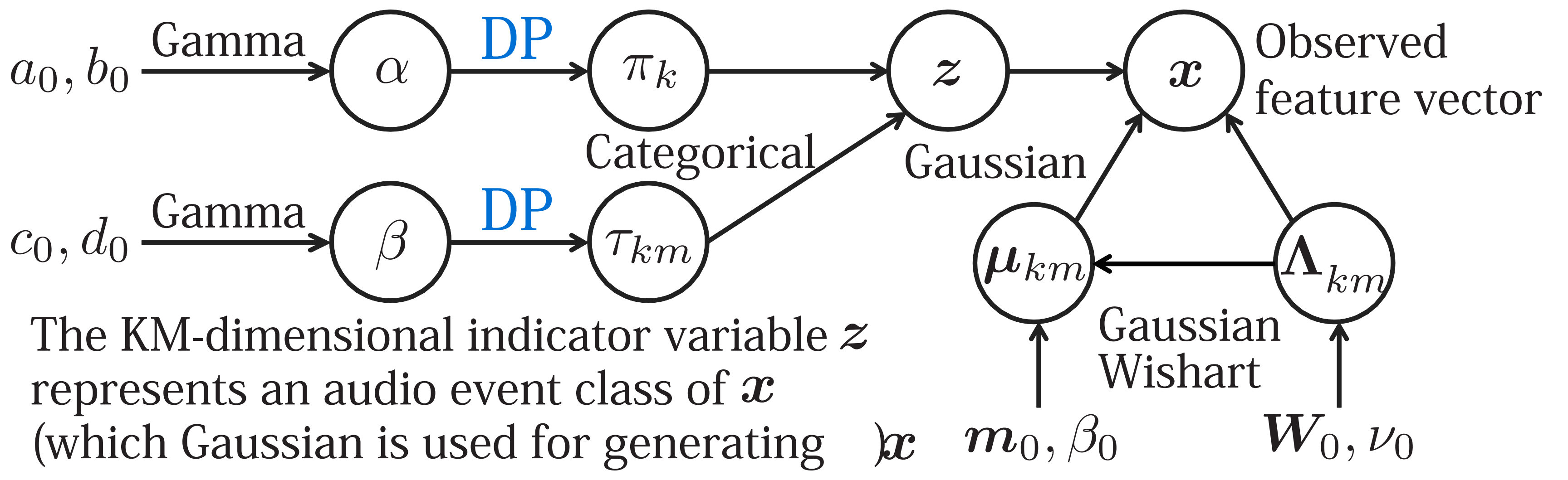
Given a finite amount of observed data, only a finite number of GMMs and a finite number of Gaussians are activated (effective model complexity is automatically estimated)

The Proposed Model

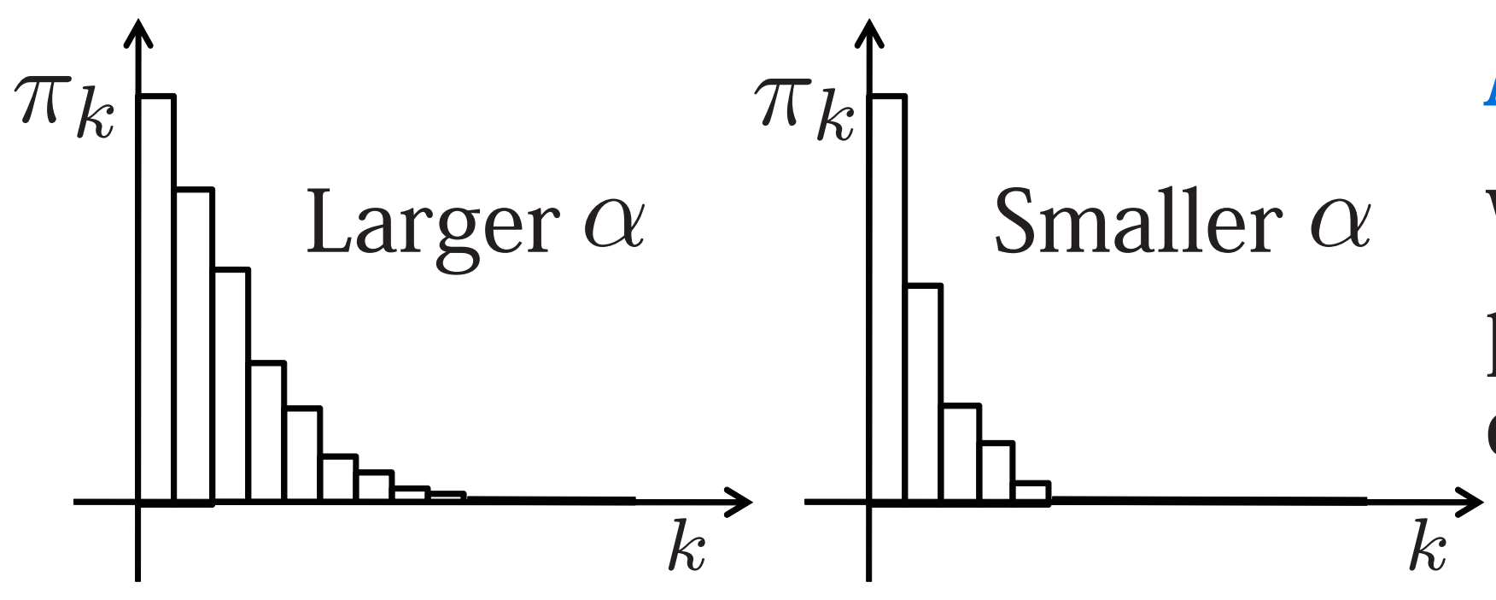
Nested infinite GMM — Infinite GMM for each class k

$$\mathcal{M}(\boldsymbol{x}) = \sum_{k=1}^{\infty} \pi_k \mathcal{M}_k(\boldsymbol{x}) \quad \xleftarrow{\text{Mix}} \quad \mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{\infty} \tau_{km} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{km}, \boldsymbol{\Lambda}_{km}^{-1})$$

How to make K and M go to infinity? → Dirichlet processes!



The KM-dimensional indicator variable $\boldsymbol{z}$ represents an audio event class of $\boldsymbol{x}$ (which Gaussian is used for generating)$\boldsymbol{x}$

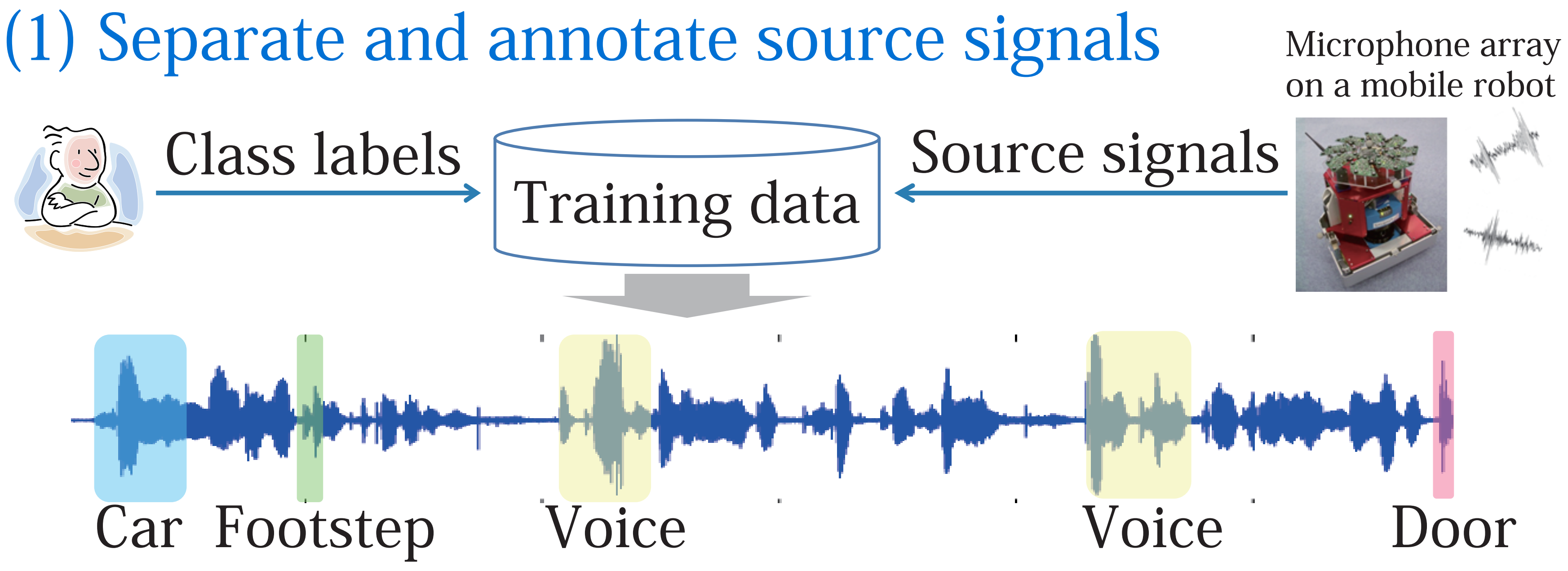


Automatic complexity control

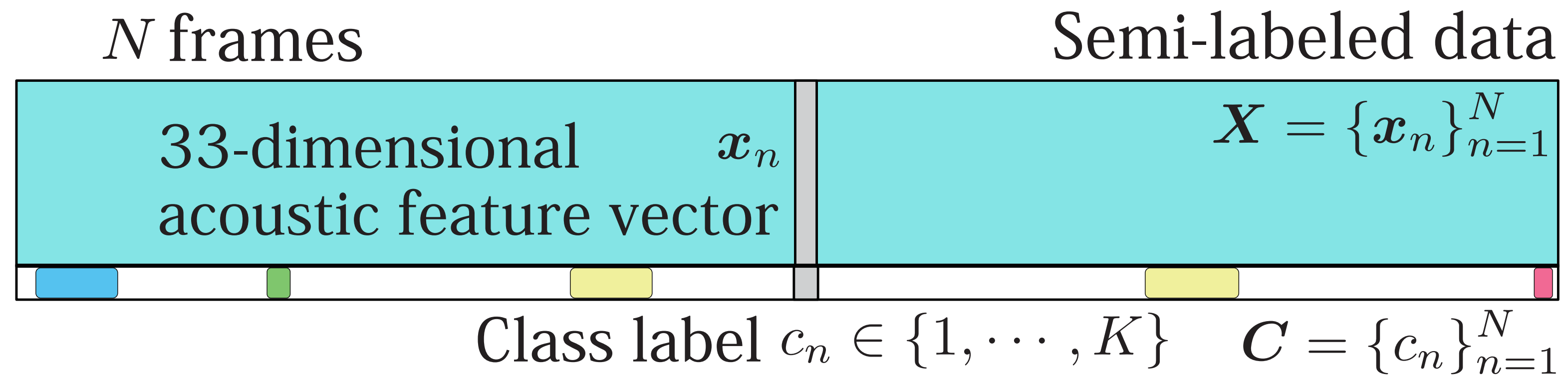
We place a non-informative Gamma prior on the concentration parameter of each DP (hierarchical Bayes)

System Overview

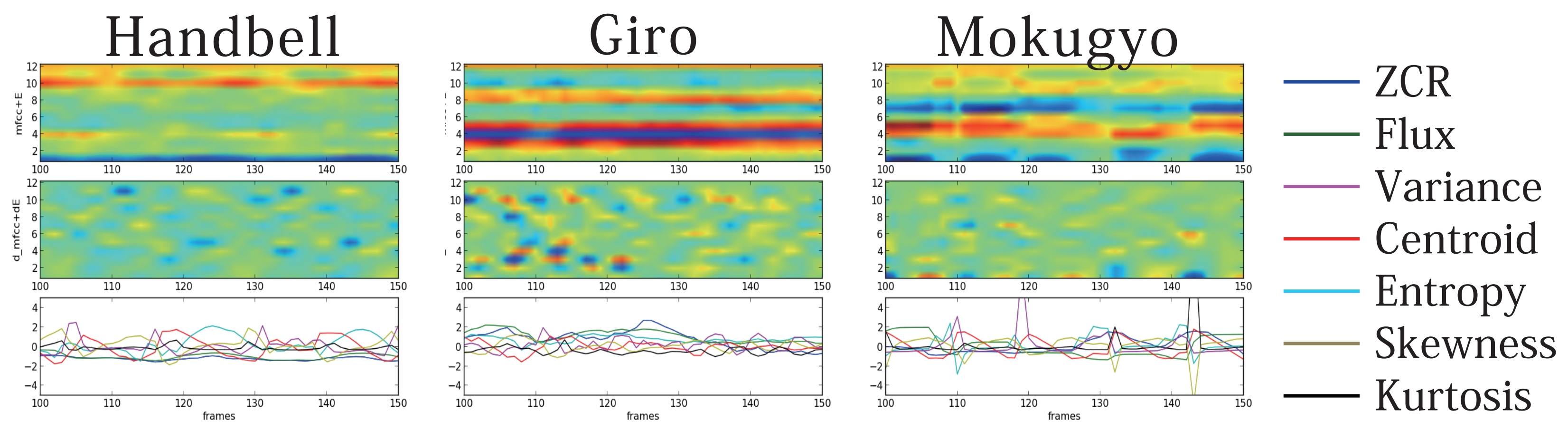
(1) Separate and annotate source signals



(2) Extract acoustic features

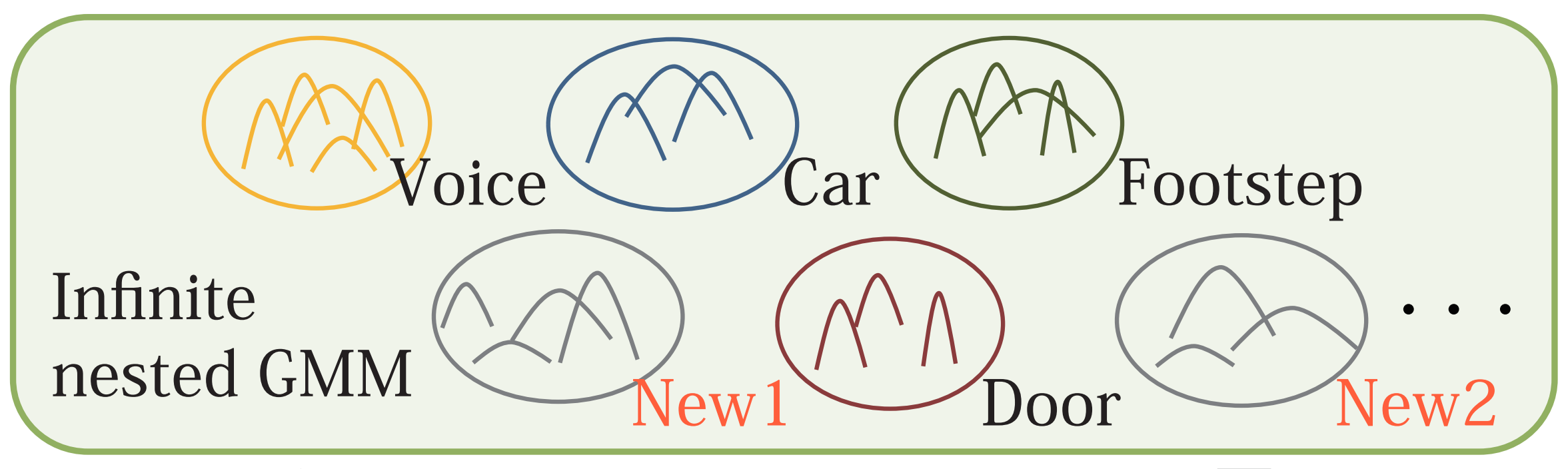

N frames
Semi-labeled data
33-dimensional acoustic feature vector $\boldsymbol{x}_n$
$\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$
Class label $c_n \in \{1, \cdots, K\}$
$\boldsymbol{C} = \{c_n\}_{n=1}^{N}$

33-dimensional acoustic features:
- 12-dimensional MFCCs, energy, and their delta values
- Zero-crossing rate (ZCR)
- 6-dimensional spectral features (flux, centroid, entropy, variance, skewness, and kurtosis)



(3) Train and use an infinite nested GMM

The effective numer of classes (K) and the effective number of Gaussians (M) in each class are inferred

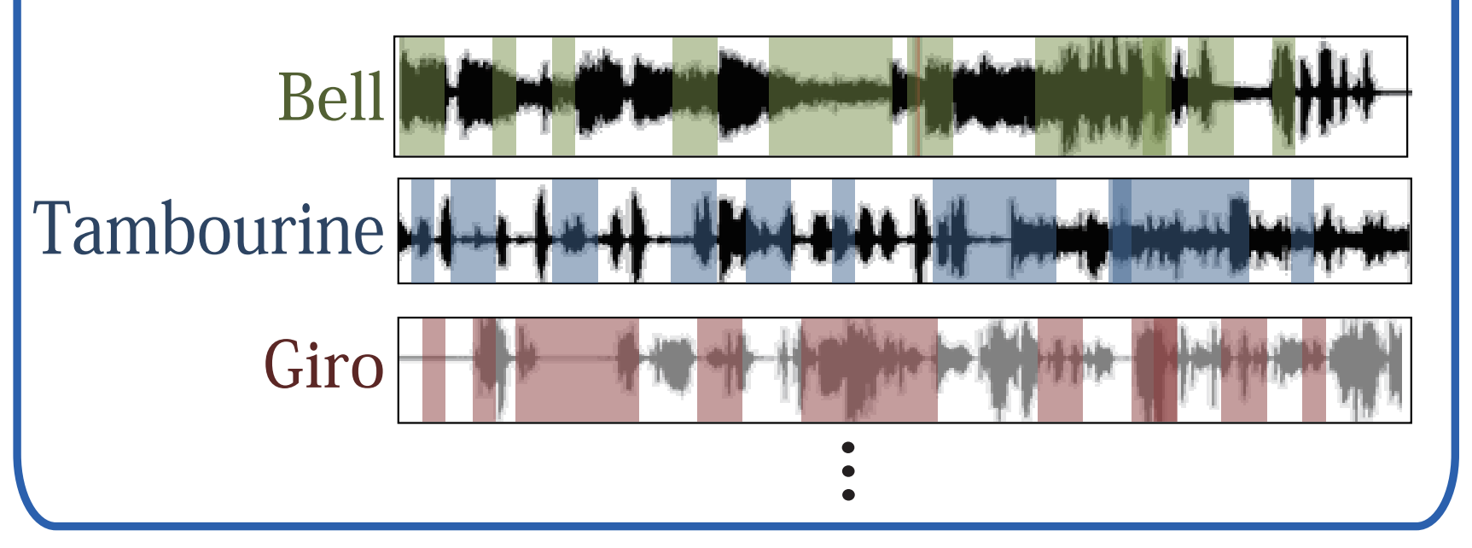


Training phase
New classes can be instantiated if necessary

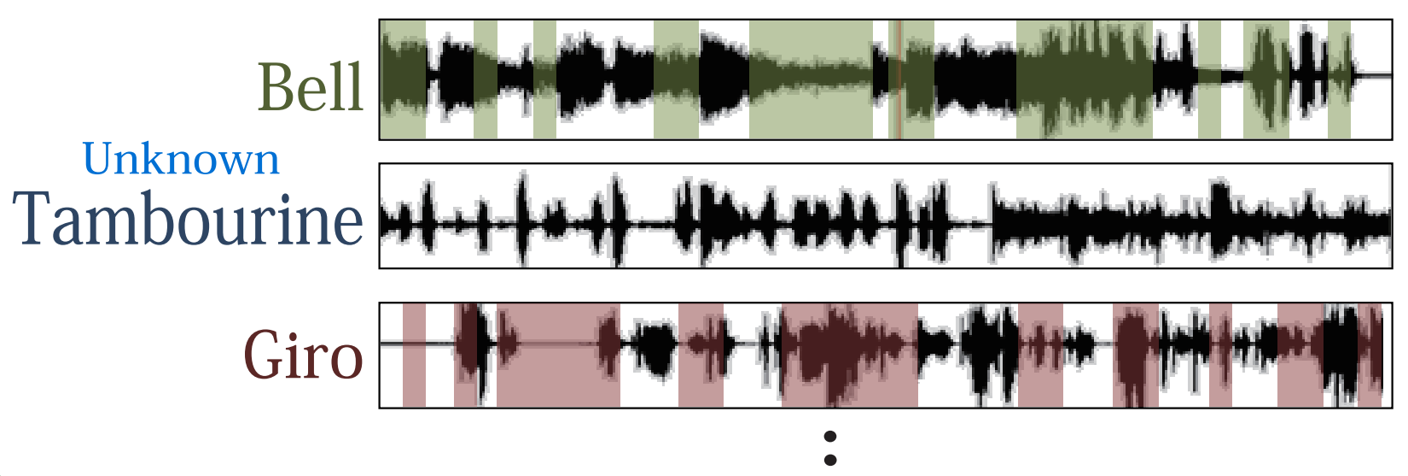
Identification phase
Given feature vector $\boldsymbol{x}$ — Posterior probability of each class

Experimental Results

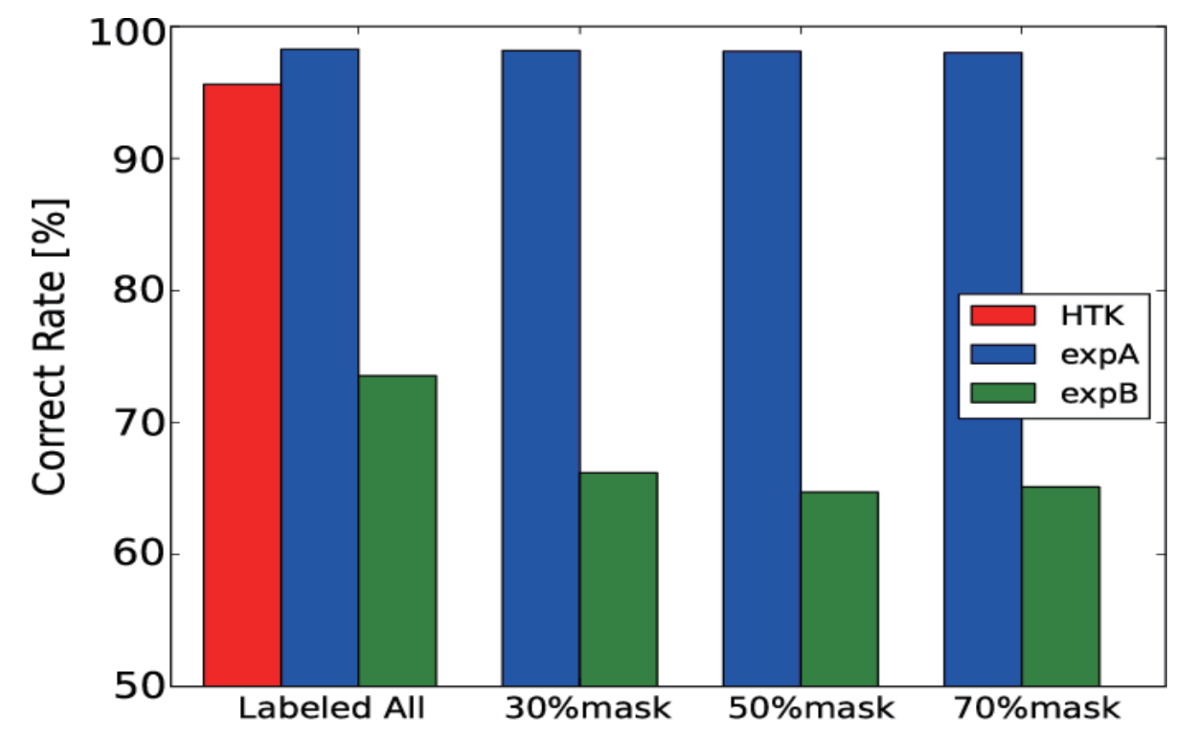
[Experiment A]
All kinds of class labels appeared in training data



[Experiment B]
Some kinds of class labels did not appear in the data



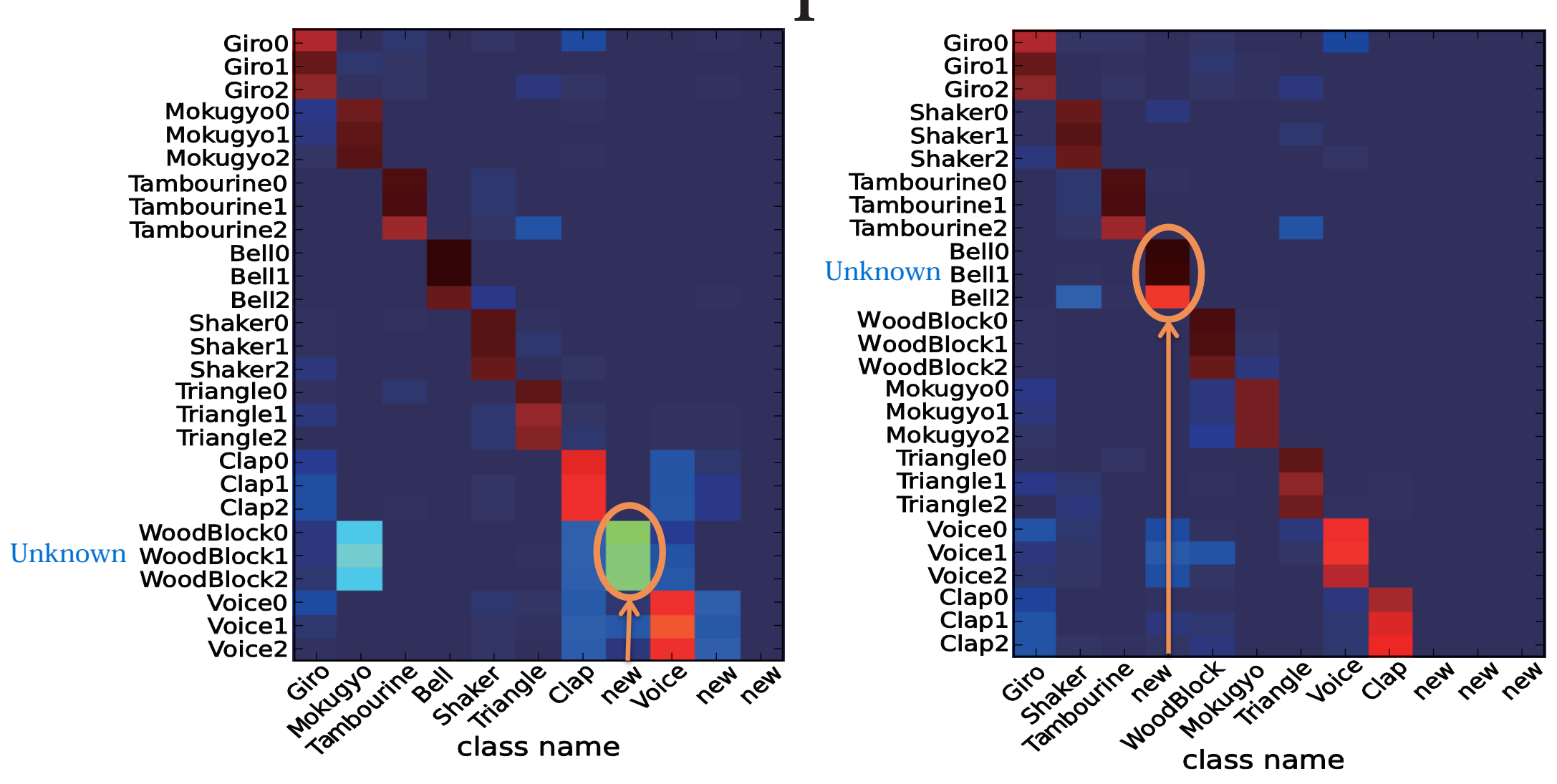
9 classes (7 percussions, voice, hand clap)
13-min audio signal for each class
(10 min for training and 3 min for evaluation)



The proposed model outperformed a classical finite model (K=9, M=12) in Exp. A (no new classes)

(HTK was used for training the finite model)

Posterior class probabilities in Exp.B

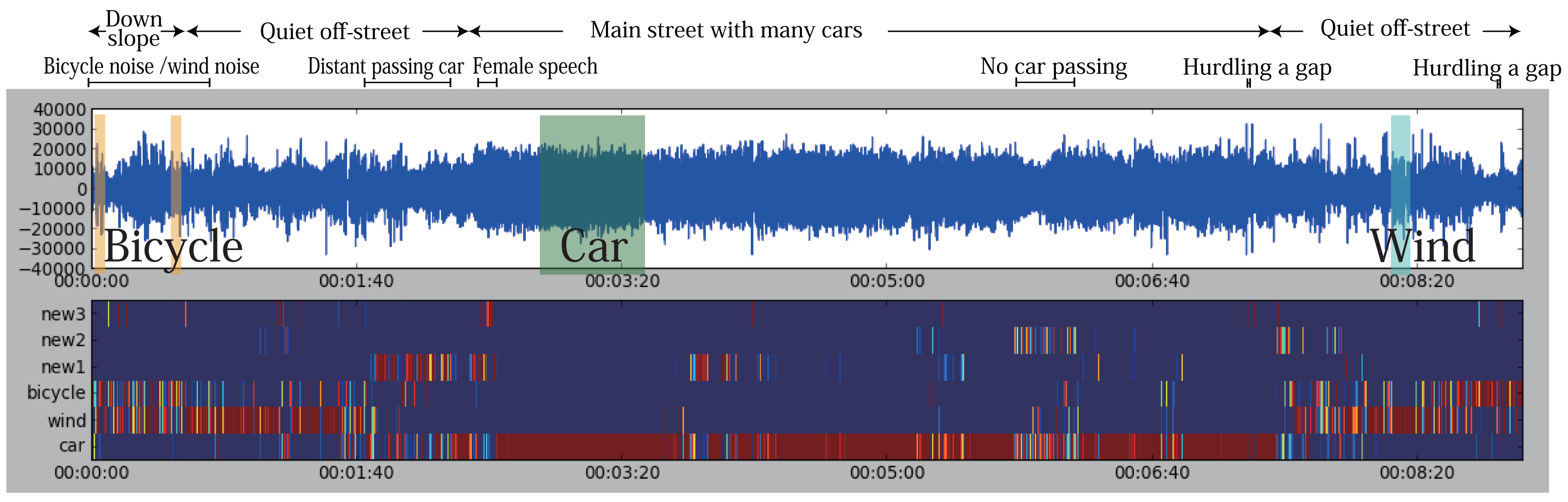


- Known classes were correctly identified as in Exp. A
- Some new classes were added in a data-driven manner
- It is difficult to identify unknown classes having similar acoustic features to known classes

Environmental Sound Modeling

Input: 9-min audio signal (27490 frames)
Label: 3 classes (bicycle, car, and wind)
4314 frames (15.7%) were annotated

6 classes were discovered!
- New class 1: distant car noise
- New class 2: bicycle road noise
- New class 3: female voice, etc.
The 3 known classes were identified correctly


Down slope
Quiet off-street
Main street with many cars
Quiet off-street
Bicycle noise /wind noise
Distant passing car
Female speech
No car passing
Hurdling a gap
Hurdling a gap
Bicycle
Car
Wind
00:00:00
00:01:40
00:03:20
00:05:00
00:06:40
00:08:20
new3
new2
new1
bicycle
wind
car