Detection of Acoustic Events in Meeting-Room Environment

Presented by
Andriy Temko

11/Dec/2008
Content

- Introduction
- State of the Art
- Acoustic Event Classification
- Acoustic Event Detection
- Participation in International Evaluations
- Speech Activity Detection
- UPC Smart-Room Activities
Introduction

- **Need of environments (Smart-rooms)** in which computers do not need much human attention and humans focus on interaction with humans (not with the computer)
  e.g. CHIL project
- **Multimodality**
- **Technologies** (Person ID, Person Localization & Tracking, Meeting Summarization)
- Need a large set of **unobtrusive Sensors**
Introduction

- In such environments the human activity is reflected in a rich variety of acoustic events (AEs)

- Need of perceptual components (activity detection, automatic speech recognition, speaker identification and verification, and speaker localization)

- Speech is the most informative AE however…

- Detection and classification of other AEs may:
  - Describe social activity that takes place in the room (e.g. laughter or clapping inside meeting, door slam when a meeting just started)
  - Increase the robustness of other (audio) technologies e.g. speech recognition / speech activity detection
State of the art

- Sound taxonomy
- Applications of audio recognition
- Features and classifiers
- Methods of detection of AEs
Sound Taxonomy

**Acoustic**
- Sound
  - Continuous tone
  - Continuous non-tone
  - Non-repetitive non-continuous
  - Regular repetitive non-continuous
  - Irregular repetitive non-continuous
  - Other

**Semantic**
- Sound
  - Speech noises
    - Repetitions
    - Corrections
    - False starts
    - Mispronunciation...
  - Human vocal tract non-speech noises
    - Breathes
    - Laughter
    - Throat/cough
    - Disagreement
    - Whisper
    - Mouth (generic)...
  - Non-speech environment noises
    - Clapping
    - Beep
    - Chair moving
    - Door slams
    - Footsteps
    - Key jingling
    - Music (radio, handy)...

---

A. Temko, Detection of Acoustic Events in Meeting-Room Environment
Applications

- Audio indexing and retrieval (speech, background noise, etc)
- Audio recognition for a **given environment** (kitchen, shops, hospitals, etc)
- Recognition of **generic sounds** (birds, animals, alarms, etc)
- Classification of **acoustic environments** (street, restaurants, library, airport, etc)
Features and Classifiers

Features
- Conventional ASR features
- Perceptual features (energy, silence ratio, spectral tilt, SBE, spectral flux, zero-crossing etc)
- Concatenation of the ASR and perceptual features
- Feature selection (feature space reduction): LDA, PCA, ICA, etc

Classifiers
- Distance based classifiers (Euclidean, Mahalanobis, kNN, etc)
- Generative classifiers (conventional ASR): GMM, HMM, etc
- Discriminative classifiers: ANN, Decision trees, SVM, etc
Methods of Detection

- Detection by classification

- Detection and classification
  (classification of the segment bounded by detection algorithm)
Acoustic Event Classification

- SVM-based Clustering Schemes (Multi-class problem)
- Comparison of Sequence Discriminant SVM (dynamic data problem)
- Fuzzy Integral Based Information Fusion for Classification of Highly Confusable Non-Speech Sounds (Fusion and Feature Selection)
SVM-based Clustering Schemes (I)

First attempt to deal with AEC comparing different feature sets and two distinct classifiers – SVM and GMM

Database of AEs:
16 classes (Cough, Laugh, Paper, Door slam, Steps, etc)

Defined acoustic feature sets:
9 Feature sets (different combination of perceptual features and ASR features)
SVM-based Clustering Schemes (II)

Results of AEC with binary tree multi-class SVM and GMM. Best feature set for SVM is 8 and for GMM is 9.

Best feature sets for “liquid”, “sneeze”, and “sniff” are different and none is the best overall (8).
SVM-based Clustering Schemes (III)

Outline of variable-feature-set clustering algorithm

• Split classes into two clusters with the least possible confusion based on confusion matrices (exhaustive search)
• The algorithm returns both the two clusters of classes and the feature set that provide the smallest number of confusions

88.29 % classification rate - 31.5% relative average error reduction
Acoustic Event Classification

- SVM-based Clustering Schemes (Multi-class problem)
- Comparison of Sequence Discriminant SVM (dynamic data problem)
- Fuzzy Integral Based Information Fusion for Classification of Highly Confusable Non-Speech Sounds (Fusion and Feature Selection)
Comparison of sequence discriminant SVM (I)

*A drawback of SVMs when dealing with audio data is their restriction to work with a fixed-length input that represents a time sequence of feature vectors*

**Solution 1:** normalize the size of the time sequence of feature vectors  
**Solution 2:** find a suitable kernel function that can deal with sequential data of variable length
Comparison of sequence discriminant SVM (II)

AEs with temporal structure “sneeze”, “music”

Comparison Results

- Fisher kernel obtained the best overall results
- DTW-based kernels work well for sounds that show a temporal structure
- The observed bias of the classifiers to specific types of classes is a good condition for a successful application of fusion techniques
Acoustic Event Classification

- SVM-based Clustering Schemes (Multi-class problem)
- Comparison of Sequence Discriminant SVM (dynamic data problem)
- Fuzzy Integral Based Information Fusion for Classification of Highly Confusable Non-Speech Sounds (Fusion and Feature Selection)
Fuzzy Integral Information Fusion (I)

Feature and decision level information fusion

Acoustic Features
1. Zero crossing rate (1)
2. Short-time energy (1)
3. Fundamental frequency (1)
4. Sub-band log energies (4)
5. Sub-band log energy distribution (4)
6. Sub-band log energy correlations (4)
7. Sub-band log energy time differences (4)
8. Spectral centroid (1)
9. Spectral roll-off (1)
10. Spectral bandwidth (1)
Fuzzy Integral and Fuzzy Measure

Weighted Arithmetical Mean

\[ M_{WAM} = \sum_{i \in Z} \mu(i)h_i \]

\[ \sum_{i \in Z} \mu(i) = 1 \]

\[ \mu(i) \geq 0 \text{ for all } i \in Z \]

\( h_i \) are the information sources

\[ h_1 \leq \ldots \leq h_z \]

Fuzzy Integral (Choquet)

\[ M_{FI}(\mu, h) = \sum_{i=1}^{z} [\mu(i, \ldots, z) - \mu(i+1, \ldots, z)] h_i \]

\[ S \subseteq T \Rightarrow \mu(S) \leq \mu(T) \]

\[ \mu(\emptyset) = 0, \mu(Z) = 1 \]

The lattice representation of fuzzy measure for \( n = 4 \).

FM can be
- learnt from data
- provided by an expert
- calculated from fuzzy densities
Fuzzy Integral Information Fusion (II)

Performance for the 10 SVM systems running of each feature type, the combination of the 10 features at the feature-level with SVM, and the fusion on the decision-level with WAM and FI operators. FM is learn from data.
**Fuzzy Integral Information Fusion (III)**

Importance and Interactions are extracted from Fuzzy Measure

The most important

4, 6, 7

Highly redundant

(light cell) \{4,5\}, \{1,8\}

**Feature selection:**

- With feature selection based on information extracted from FM an improvement over the baseline was obtained

**Fusion of different classifiers:**

- SVM with signal level features + HMM with frame-level features
Fuzzy Integral Information Fusion (IV)

- Fusion of several information sources with FI formalism showed significant improvements with respect to single information source
- Decision-level FI fusion showed comparable results with SVM feature level fusion
- Importance and the degree of interaction can be used for feature selection and give a valuable insight into the problem
- FI may be a good choice when feature-level fusion is not an option (different nature of the involved features, application/technique dependence)

Finalist of the Best Journal Paper Award of 2008 in Speech Technologies
Acoustic Event Detection

- UPC Acoustic Event Detection Systems 2006
- UPC Acoustic Event Detection Systems 2007
1st SVM for pre-segmentation based on silence/non-silence detection.
2nd SVM trained on 12 classes + speech + unknown for event detection.
UPC Acoustic Event Detection Systems 2007

1 second

200ms

SVM SVM SVM SVM

Post-processing

Final decision windows

System output AE assignment segments
Participation in International Evaluations

- CHIL Dry-Run Evaluations 2004
  (1 place – 2 participants)
- CHIL Evaluations 2005
  (1 place – 2 participants)
- CLEAR Evaluations 2006
  (1 place – 3 participants)
- CLEAR Evaluations 2007
  (2 place – 8 participants)
Speech Activity Detection

- **AED is event-based, SAD is frame-based**
- **Dataset reduction algorithm**
- **Metric**
  \[
  \text{NIST error rate} = \frac{\text{Duration of Incorrect Decisions}}{\text{Duration of all Speech Segments}}
  \]
- **Features**

![Diagram](image)
Enhanced SVM Training for SAD

Dataset reduction algorithm

**Step 1.** Divide all the data into chunks of 1000 samples per chunk.

**Step 2.** Train a PSVM on each chunk performing 5-fold cross-validation (CV)

**Step 3.** Apply an appropriate threshold to the highest/lowest CV accuracy

**Step 4.** Train a classical SVM on the amount of data selected in Step 3.

NIST SAD Evaluations 2006 – best SAD system out of 10 systems with 4.88 % error rate
UPC Smart-Room Activities

- UPC Smart-Room
- Database Recording
- Real-Time Implementation
- Demonstrations
UPC Smart-Room

Cam1
Cam2
Cam3
Cam4
Cam5
Cam6
Cam8

Hammer_xx
Hammer_xx
Hammer_xx
Hammer_13 to Hammer_20

5.245 m

3.966 m

MkIII_001
MkIII_064

PTZ Cam
Real-Time Implementation

- Implementation of AED component as a smart-flow client
Demonstrations (I)

- Participation in several demonstrations
  - Mockup demo:
    - Journalist scenario
    - Reliably detected AEs: door knock, door opening/closing, speech, applause, and key jingles.
Demonstrations (II)

- Participation in several demonstrations
  - Virtual UPC smart-room:
    - 3D person tracking, ASL, and AED
Demonstrations (III)

- Participation in several demonstrations
  - Demo of AED & ASL

Real-time video  Graphical representation

VIDEO >>