Model-based Event Labeling in the Transcription of Percussive Audio Signals

Jouni Paulus, Anssi Klapuri

Tampere University of Technology / Institute of Signal Processing

paulus@cs.tut.fi

September 9th 2003

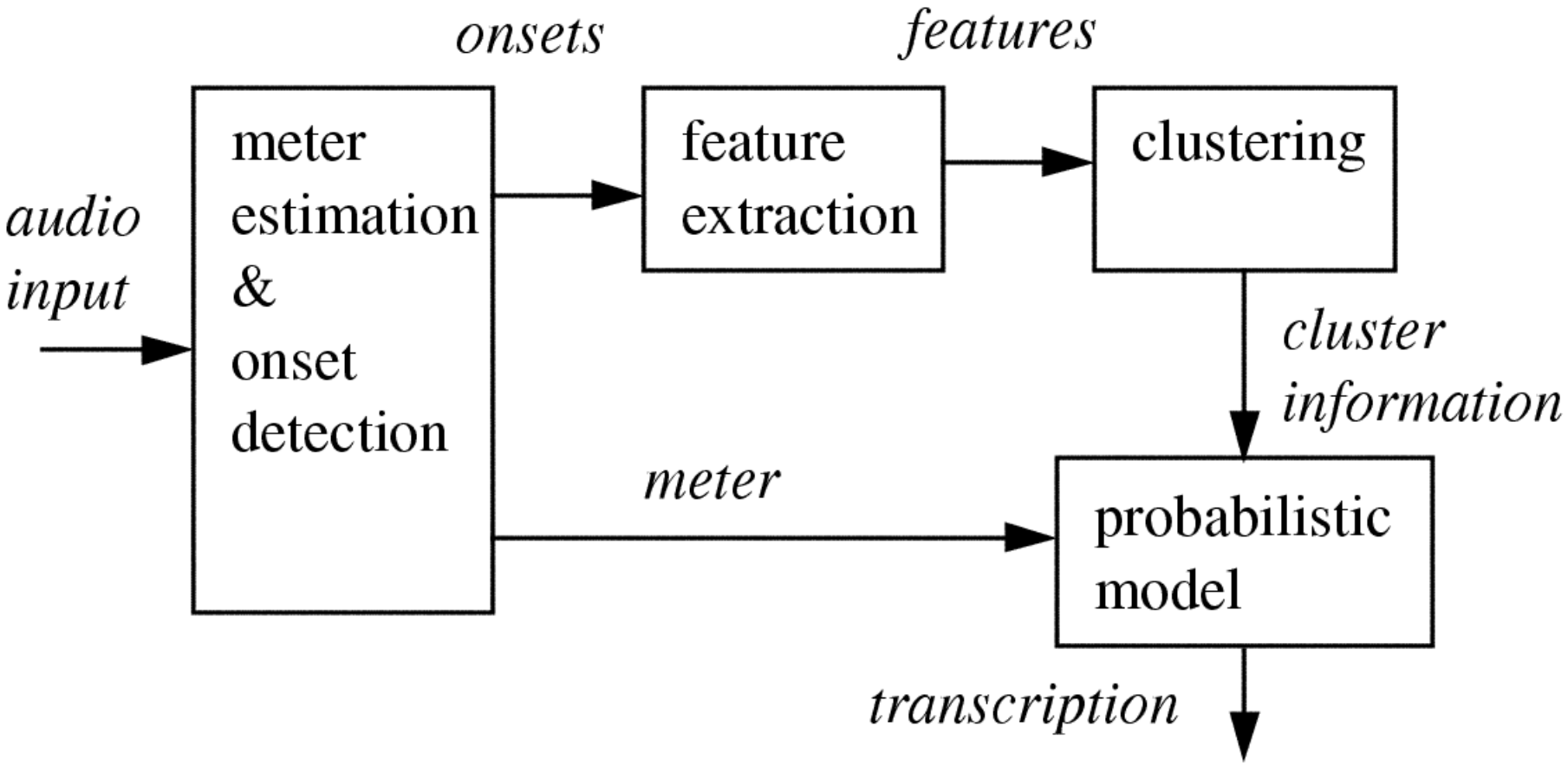
Contents

- Introduction
- System overview
- Rhythmic roles
- Probabilistic model
- Post-labeling cluster changing
- System evaluation
- Results
- Demos
- Conclusions

Introduction

- The task is to transcribe percussive audio signals to a symbolic representation, e.g. drum tracks in MIDI.
- Input signal can be created using arbitrary sounds, e.g. tapping with fingers, pencil clicking, scat singing.
  - Acoustic models can not be constructed and used.
- Limit the task to three different rhythmic labels.
  - Use them as abstractions for the real sounds.
- Utilize a statistical model based on the metrical (temporal) positions of the labels in transcription.

System overview



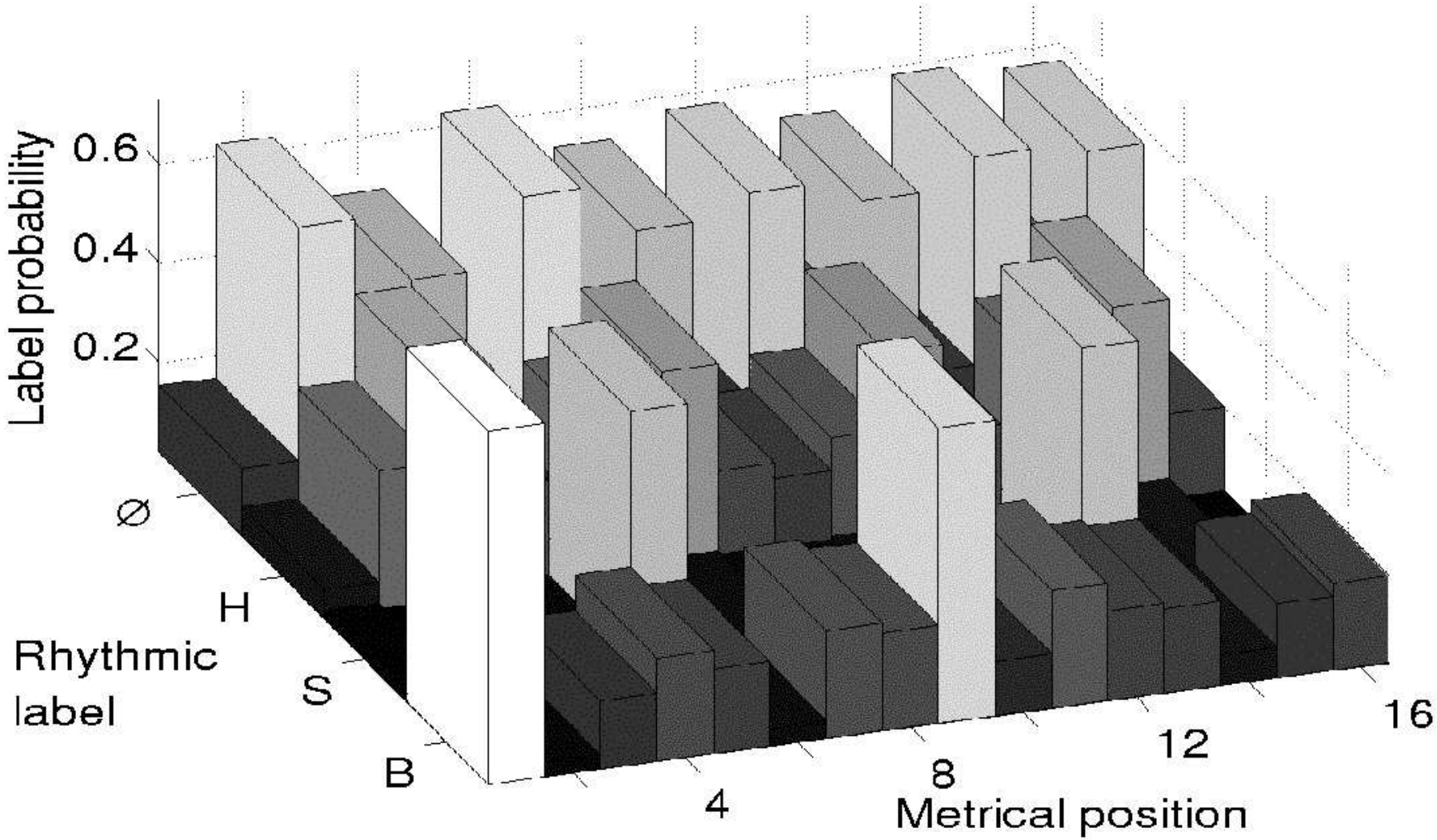
Rhythmic roles and labels

- Rhythmic percept of a percussive signal is simulated with three abstract rhythmic roles: **B**ass drum, **S**nare drum and **H**i-hat.
  - Role names are based on the typical instruments used.
  - Possible to represent more complex rhythms than with just **B** and **S**.
- Sound events in input signal are assigned with labels.
- How to determine roles/labels for the sounds?

Probabilistic model

- Observation: probability of a certain label to occure depends on the metrical position.
  - E.g. **B** is present often at the beginning of a measure etc.
  - (Metrical position = position within a musical *measure*, quantized to steps of one *tatum*.)
- Use this observation in constructing a simple statistical model for the labels.
  - Independent models for different time signatures.
- *P(q|(n,m))*, the probability of label *q* to be present at the tatum *n*, when the measure length is *m* tatums.

Probabilistic model, an example



Model usage

- After clustering and time grid quantization: a sequence of cluster numbers

$$c_i \in \{0, 1, 2, \ldots, K\}$$

- Find mapping

$$L:\{0, 1, 2, \ldots, K\} \rightarrow \{\varnothing, B, H, S\}$$

from cluster numbers to rhythmic role labels

$$q \in \{\varnothing, B, H, S\}$$

by maximising

$$P(L) = \prod_i P(q_i | (n_i, m))$$

over the whole signal.

Post-labeling cluster changes

- Likely that not all sound events are clustered correctly.
- After cluster number to label mapping has been set, allow to change labeling of individual events.
- Which events to change?
  - If fuzzy K-means clustering used, cluster membership value can be used.
  - Or try changing all...
- If the total probability P(L) over signal increases after the change, retain it, otherwise change back to original.

System evaluation

- Synthesized monophonic 30 sec clips using different sound sets
  - from 13 genres/categories, 2/genre randomly
  - tests repeated 10 times, results averaged
- System evaluated in four steps
  - clustering & onset detection performance from input signal
  - rhythmic role labeling
  - whole system without post-labeling fixes
  - whole system with post-labeling fixes

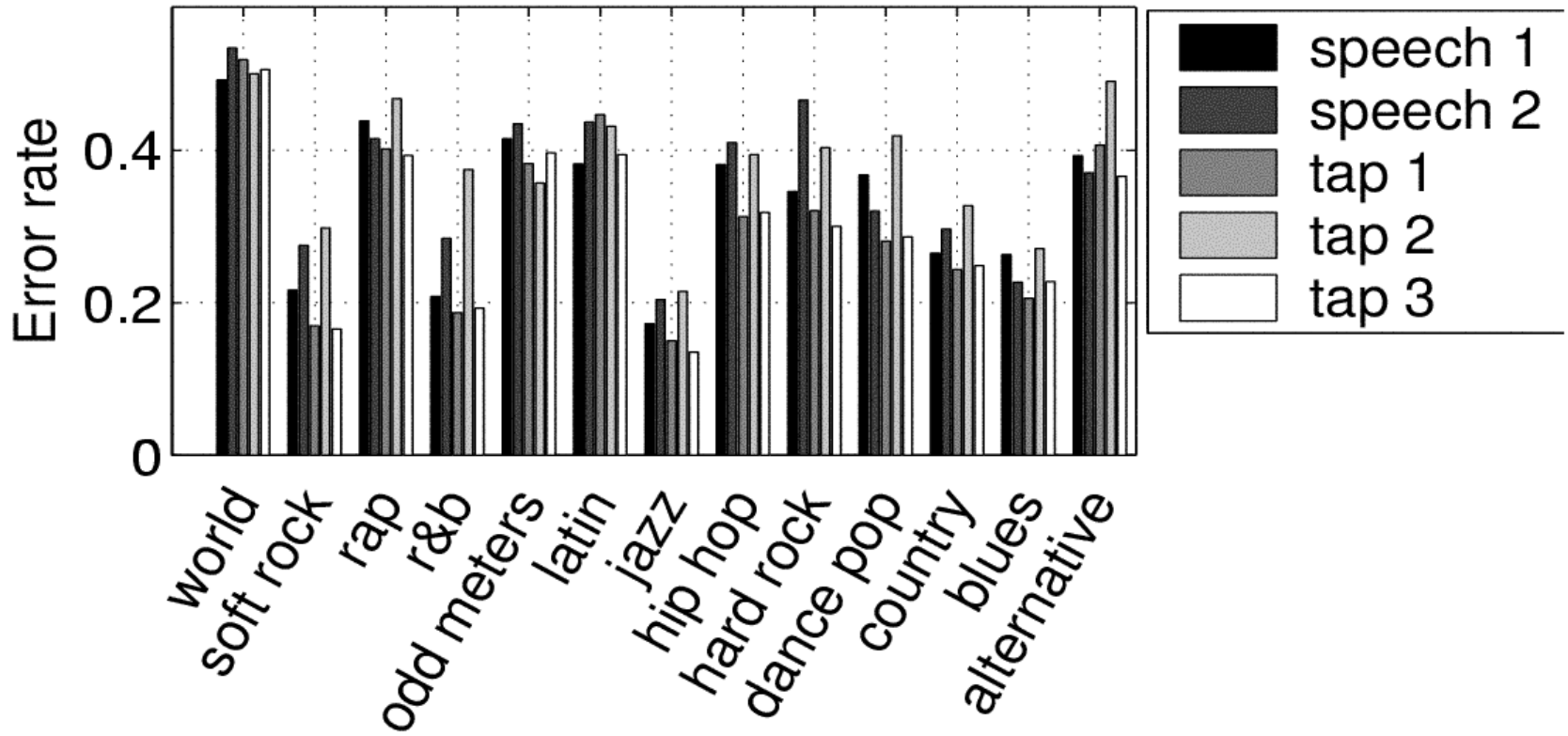
Results, subsystems

- Clustering and onset detection performance relatively good.
- Role label assignment generates errors.
- Post-labeling fixes have very small effect on total performance, number of happened changes usually low.
- Sound set related performance difference also present.

test step	speech ER	tapping ER	total avg ER
1. clustering	15.28%	13.12%	13.98%
2. role labeling	27.91%	27.91%	27.91%
3. whole w/o post fix	34.91%	33.15%	33.85%
4. whole w/ post fix	35.68%	33.01%	33.67%

Results, by genre

- Some genre related differences in performance are present.



Demos

original	speech	transcription	tapping	transcription
blues				
dance/pop				
hard rock				
rap				
soft rock				

Conclusions

- Transcribing percussive tracks created with arbitrary sounds need a mapping from sound events to notes.
- Rhythmic role abstraction allows handling complex patterns in more general way.
- Metrical position model can be used to some extent in event labeling.
  - Alone prone to errors.
  - Perhaps as an top-down addition to acoustical models.
- Post-labeling cluster changes effect very little in total.

Questions

?