CONVENTIONAL AND PERIODIC *N*-GRAMS IN THE TRANSCRIPTION OF DRUM SEQUENCES

Jouni K. Paulus, Anssi P. Klapuri
Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland
{paulus,klap}@cs.tut.fi

Introduction

- Scope: to transcribe polyphonic drum sequences from audio to a symbolic representation.
- Purpose: to improve acoustic recognition accuracy with *N*-gram based language model.

Notation

- Assign all possible drum sounds to seven categories: bass drum, snare, hi-hat, cymbal, tom tom, ride cymbal and percussions (B, S, H, C, T, R, P).
- Categories form an alphabet Σ of symbols, s_n, alphabet size is 7.
- Unordered subsets of Σ are words in vocabulary, e.g. $w_i = \{s_1, s_2, s_3\}$ or $w_i = \{\ \}$ which is silence.
- Time discretization to a grid of equidistant tatum pulses, one word in each grid point.
- Symbolic representation as a sequence of words, $w_1 w_2 \ldots w_K$ or w_1^K.

N-grams

- Create a language model for drum sequences with *N*-grams, based on (*N*-1)th order Markov assumption.
- Use *N*-1 previous words to predict the next word, see Figure 1.

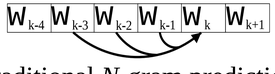

Fig 1. Traditional *N*-gram prediction with *N*=4.

Problem

- Size of the needed training corpus for estimating the probabilities increases exponentially with the vocabulary size.
- Most words which consist of more than two symbols occur relatively rarely, see Figure 2.

Solution

- Instead of word *N*-grams, use symbol *N*-grams and combine them to estimate word probabilities, e.g. the probability of word "BH" to appear after the word "H" is
  $P(\mathrm{BH}|\mathrm{H}) = P_B(1|0)P_H(1|1)P_C(0|0)\ldots$
- Since vocabulary size decreases, the size of needed training corpus decreases similarly.

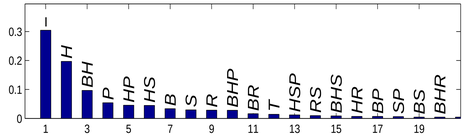

Fig 2. Occurrence frequencies of the 20 most probable words in the used song database.

Observation

- Conventional *N*-grams can predict quite successfully melodies following locally predictable patterns.
- Accompaniment and rhythm section tend to follow a pattern that is regular with a period L.

Utilisation of this observation

- Instead of using previous words $w_{k-(N-1)}^{k-1}$ to predict the word w_k, use the words $w_{k-(N-1)L}^{k-L}$, i.e. every Lth word, see Figure 3.

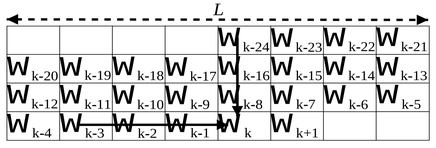

Fig 3. Traditional vs. periodic *N*-gram prediction with *N*=4 and period L=8.

Acoustic model

- Gaussian mixture model classifier for non-empty drum sounds.
- Support vector machine for "silence" detection.
- Features: MFCCs, ΔMFCCs, ZCR, crest factor, spectral kurtosis and skewness, RMS and temporal centroid
- For whole system overview, see Figure 4.

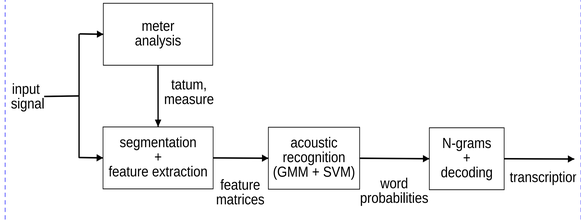

Fig 4. Overview of the whole transcription system.

Validation Experiments

- Evaluation with 65 synthesized test songs with the average length of 140 s, results are in Table 1.
- Word unigrams improve the accuracy of acoustic models by 35% and added *N*-grams even more.
- Periodic word *N*-grams outperform conventional word *N*-grams.
- Larger *N* value implies better prediction power.
- Symbol *N*-grams enable utilising relatively large *N*.
- Coarse error analysis in Figure 5.

Table 1: Performance of different tested *N*-grams, their symbol error rates (ER) and improvement from the baseline system (Imp).

method	ER (%)	Imp (%)
acoustic model	76.1	
baseline (word unigram)	49.5	
1. conv. word bigram	47.2	4.7
2. per. word bigram	46.6	5.8
3. conv.+per. word bigram	47.1	4.9
4. conv. word trigram	46.9	5.2
5. per. word trigram	46.8	5.5
6. conv.+per. word trigram	46.5	6.0
7. conv. sym. quintagram	46.8	5.5
8. per. sym. quintagram	45.9	7.3
9. conv. sym. decagram	45.7	7.6
10. per. sym. decagram	46.0	7.1

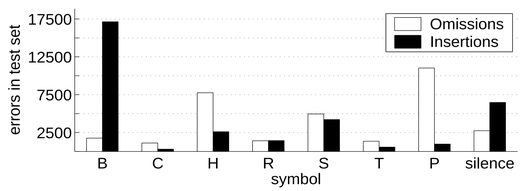

Fig 5. Symbol omission and insertion errors.