A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations

Juhan Nam
CCRMA, Music
Stanford University

Jiquan Ngiam
CS Department
Stanford University

Honglak Lee
EE & CS Department
University of Michigan

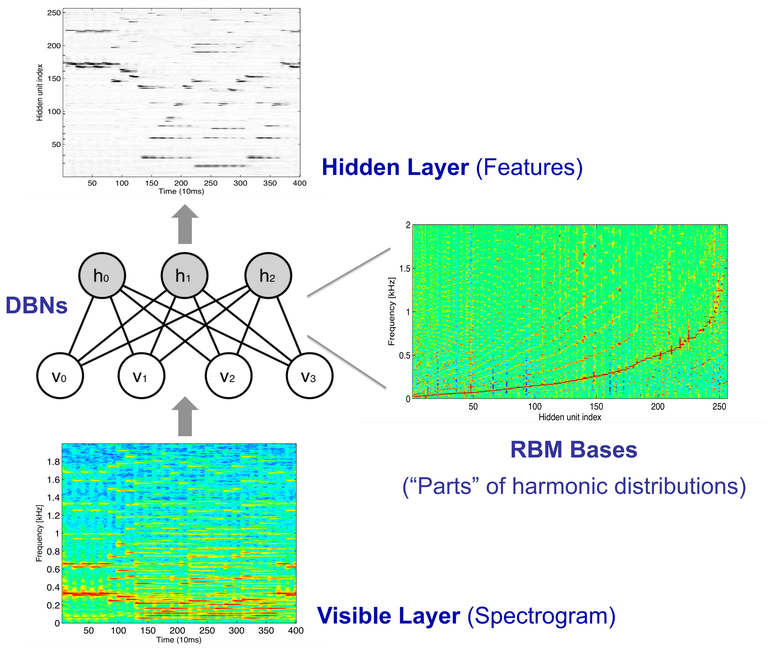
Malcolm Slaney
CCRMA, Stanford
Yahoo! Research

Summary

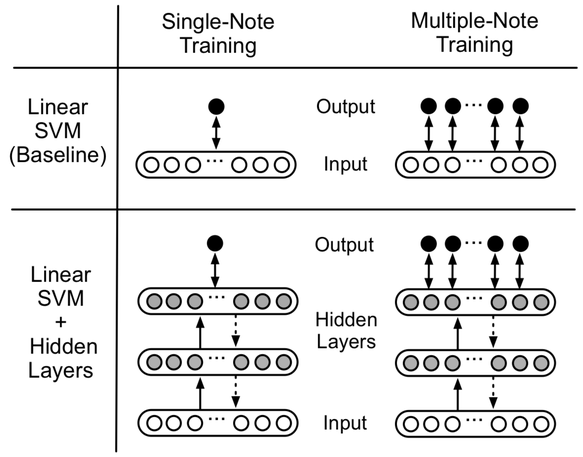
- Piano Transcription: audio recording to music score
- Classification approach
  - 88 binary classifiers: each detects the presence of one note
- What's new
  - Feature learning by deep belief networks (DBNs)
  - Multiple-note training: multi-task learning
- Compared features
  - Normalized spectrogram (baseline feature)
  - DBN-based representation
  - Improvements using the learned features (frame-level accuracy)
    - Poliners and Ellis: up to +4.4%
    - MAPS: up to +6.6%

Feature Learning

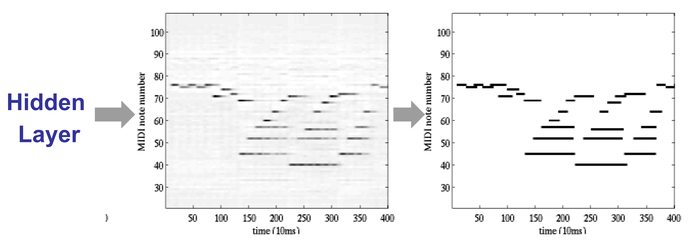
- Sparse Restricted Boltzmann Machines (RBMs)
  - RBM specifies the probability of possible assignments of visible and hidden layers (parameters are trained by ML)
  - **Sparse RBM** can control the activation of the hidden layer
- Deep Belief Networks (DBNs)
  - Trained by greedy layer-wise stacking of RBMs
  - Feed-forward "pre-training" of deep neural networks
  - Fine-tuning by back-propagation (supervised training)



Classification-based Transcription



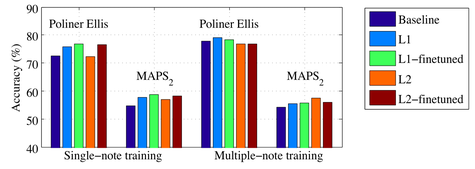
- Single-Note Training: Poliner and Ellis 07
  - 88 **linear SVMs** per note
  - Training data is separately sampled for each note
  - Fine-tuning 88 deep networks (slow)
- Multiple-Note Training
  - 88 concatenated **linear SVM**
  - Training data is **shared**
  - **Fine-tuning a single deep networks (fast)**
    - Multi-task learning or Multi-label classification
- Compare different levels of features
  - Baseline: normalized spectrogram
  - DBN-based features: hidden layer 1 and 2 (L1 and L2)
- HMM post-processing
  - Temporal smoothing
  - Independently for each note

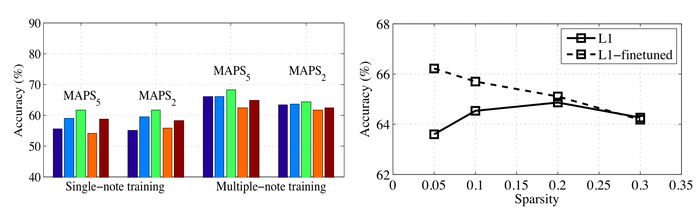


Evaluation

- Datasets
  - Poliner and Ellis: 124 MIDI and 29 piano recordings
  - MAPS: 9 sets of 30 songs with different pianos
  - Marolt: 3 synthetic and 3 piano recordings
- Evaluation Metrics
  - Frame-Level accuracy = TP / (TP+FN+FP)
  - F-measure: Precision and Recall
- Training
  - Scenario 1 (S1): trained on Poliner and Ellis set, validated on Poliner and Ellis and a subset of MAPS
  - Scenario 2 (S2): trained and validated on subsets of MAPS

Validation Results





Test set: Poliner and Ellis / Marolt

Algorithms	Poliner and Ellis	Marolt
Poliner and Ellis	67.7%	44.6%
Proposed (S1-L1)	71.5%	**47.2%**
Proposed (S1-L1-finetuned)	**72.5%**	46.5%
Marolt	39.6%	46.4%
Ryyananen and Klapuri	46.3%	50.4%
Proposed (S2-L1)	**63.8%**	**52.0%**
Proposed (S2-L1-finetuned)	62.5%	51.4%

Test set: MAPS

Algorithms	Precision	Recall	F-measure
Marolt	74.5%	57.6%	63.6%
Vincent et al.	71.6%	65.5%	67.0%
Proposed (S2-L1)	**80.6%**	67.8%	73.6%
Proposed (S2-L1-finetuning)	79.6%	**69.9%**	**74.4%**

Contact: juhan@ccrma.stanford.edu