Deep Learning for Classroom Activity Detection from Audio

Robin Cosbey[1], Allison Wusterbarth[3], Brian Hutchinson[1,2]

[1]Western Washington University [2]Pacific Northwest National Laboratory [3]Conversica

Overview

Motivation: Post-secondary instructors are increasingly incorporating innovative teaching practices into their classrooms to improve student learning outcomes, but manually quantifying the adoption of these techniques is costly and scales poorly.

Goal: Produce an automatic system to help university instructors rapidly understand how much time is spent on different types of activity in the classroom.

Approach: We introduce a set of deep learning models for activity annotation, evaluating them on a collection of university classroom recordings.

Background

- Many studies have shown that student-centered active learning strategies can improve the effectiveness of instruction; example activities include:
  - *Think-pair-share*: students reflect on a question, discuss in groups and share with the class.
  - *Polling*: students vote via polling device, often followed by a discussion of the results.

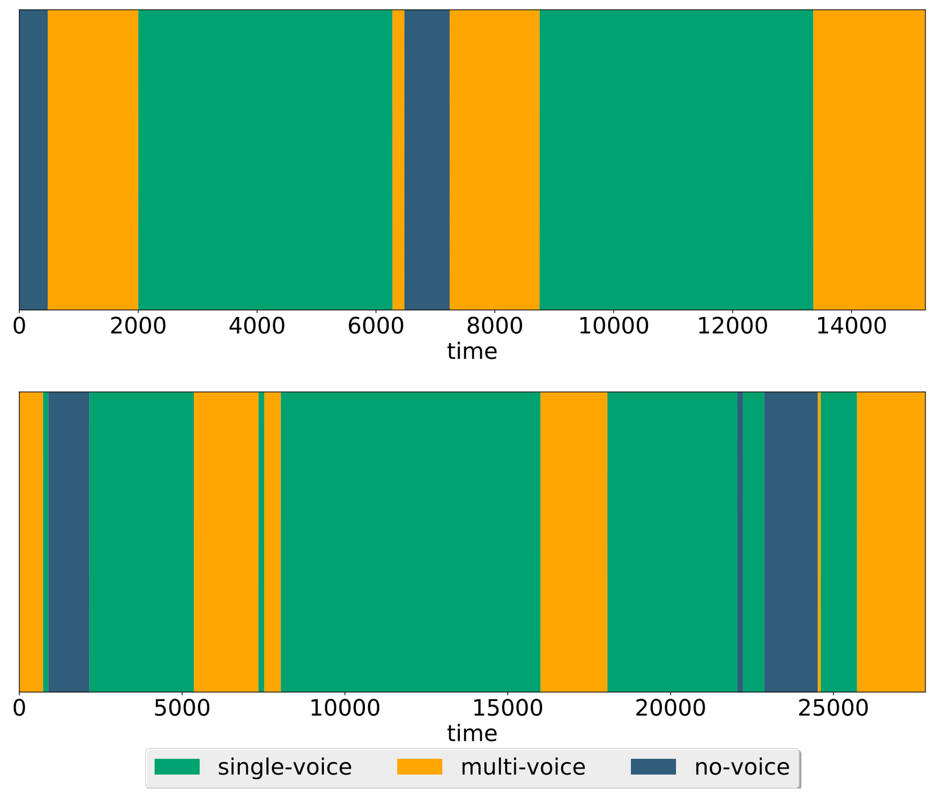

Figure: Illustration of activity in sample class sessions. The x-axis denotes time within the class.

DART Tool

- Researchers at San Francisco State University (SFSU) introduced the "Decibel Analysis for Research in Teaching" (DART) tool.
- A simple decision tree with features as energy statistics over a local window (15s).

DART Corpus

- The SFSU researchers collected a corpus of classroom recordings, using labels "single-voice," "multi-voice," "no-voice," and "other."
- The audio was collected with Sony ICD-PX333 handheld audio recorders placed at the front of the classroom and stored in a compressed (mp3) format.
- 85 hours of audio, 54 class sessions, seven instructors.

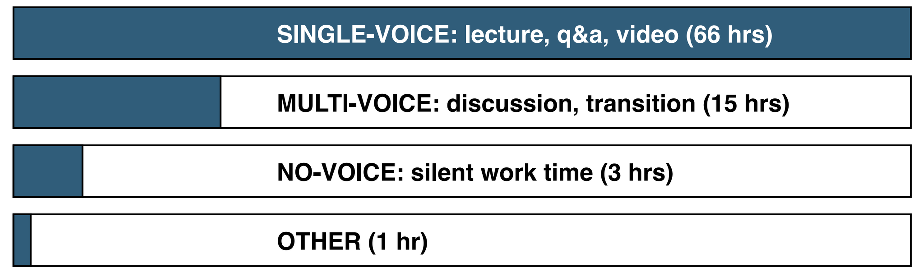

Figure: Breakdown of the DART corpus labels.

Methods

Deep Neural Networks (DNNs)

- Frame-level predictions, softmax output activation.
- Input $\boldsymbol{x}$: windowed acoustic features ($\boldsymbol{k}$ frames).
- Output $\boldsymbol{y} \in \mathbb{R}^4$: posterior probabilities over the four classes.

Recurrent Neural Networks (RNNs) & Gated Recurrent Units (GRUs)

- Standard (Elman) RNNs are fed a single frame's feature vector at each timestep and produce activity predictions for each frame given context from previous frames.
- GRU networks use gating mechanisms which, like LSTMs, more effectively propagate information across longer timespans.

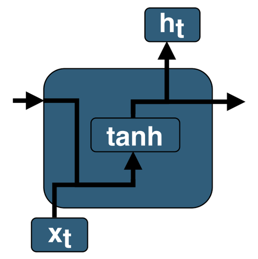

Figure: An RNN cell.

Baselines

- DART, previous state-of-the-art on the task.
- Logistic regression classifier, to assess the effect of model depth.
- Majority class (which predicts all frames as single-voice).

Experimental Setup

- We extract 40 log mel-filterbank features plus energy using HTK.
- The data is split into four sets to include train, development and two test sets.
- Train, development and test1 are an 80%-10%-10% split of the first five instructors and test2 contains the class sessions for the remaining two instructors.

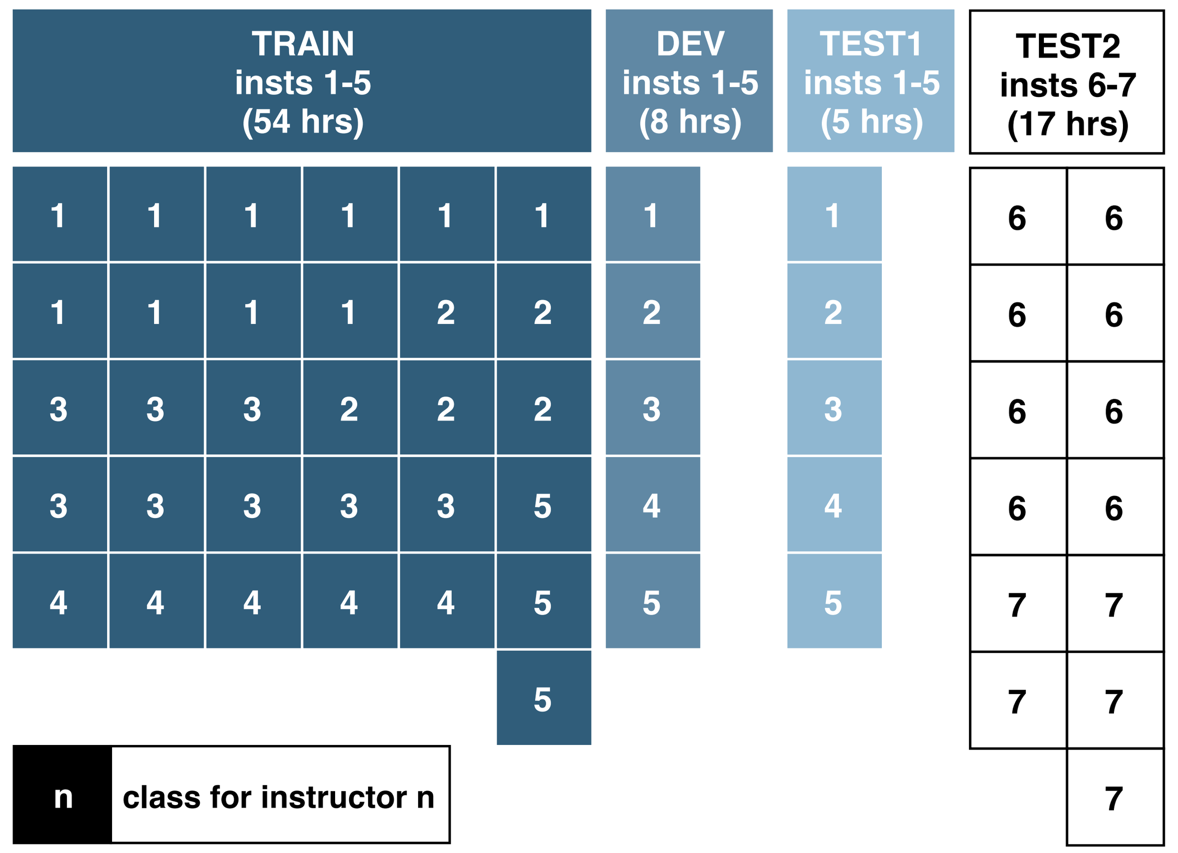

Figure: The number of classes per instructor in each set.

- We compare two frame sizes:
  - 0.5s frames with 0.25s offsets
  - 1s frames with 0.5s offsets
- We window the frames passed to the DNN and logistic regression models with total window sizes up to 31 to provide greater temporal context.
- No post-processing of frame-level predictions was done.

Results

Effect of Window Size

- We compare the window sizes on logistic regression and DNN models for frame sizes of 1s with 0.5s offsets, reporting frame error rate and weighted-F measure.

		test1		*test2*	
Model	*W Sz*	*Err*	*F*	*Err*	*F*
LR	*1*	0.158	0.826	0.235	0.711
LR	*3*	0.131	0.720	**0.225**	0.728
LR	*11*	0.105	0.892	0.227	0.742
LR	*17*	0.095	0.901	**0.225**	0.745
LR	*31*	**0.090**	**0.907**	0.227	**0.751**
DNN	*1*	0.120	0.876	0.215	0.777
DNN	*3*	0.093	0.903	0.171	0.819
DNN	*11*	0.080	0.916	0.155	0.832
DNN	*17*	0.076	0.921	**0.142**	**0.846**
DNN	*31*	**0.072**	**0.926**	0.177	0.821

Table: Effect of window size on logistic regression (LR) and DNN, measured with frame error rate and weighted F-measure 1s frame sizes with 0.5s offsets. Best models are bolded; best overall shaded blue.

- Larger window sizes show improved performance.
- The deeper DNN outperforms the shallow logistic regression classifier.
- Previously unseen instructors (test2) are more challenging overall.

Model and Frame Size Comparison

- Two frame sizes are compared across all baselines and models.
- DNN and logistic regression models use a window size of 31; others use 1.

		test1		*test2*	
Frame Size	*Method*	*Err*	*F*	*Err*	*F*
	MC	0.200	—	0.222	—
	DART	0.104	0.883	0.184	0.773
0.5s/0.25s	*LR*	0.097	0.899	0.225	0.742
	DNN	0.077	0.919	0.155	0.836
	RNN	0.076	0.918	0.140	0.850
	GRU	**0.071**	**0.927**	**0.101**	**0.891**
1s/0.5s	*LR*	0.090	0.907	0.227	0.751
	DNN	**0.072**	**0.926**	0.177	0.821
	RNN	0.077	0.919	0.154	0.838
	GRU	0.083	0.914	**0.108**	**0.883**

Table: Results on the test sets contrasting frame size and method: majority class (MC), DART, logistic regression (LR), DNN, RNN and GRU. Best models are bolded; best overall is shaded blue.

- The GRU gives strong performance in almost all cases.
- With larger frame sizes (and windowing), the DNN also performs well.
- The gap between test1 and test2 is not too large.

Analysis

Activity Time Correlation

- We compare the overall fraction of time on each activity predicted by DART and GRU with the true time spent on each activity, per class session.

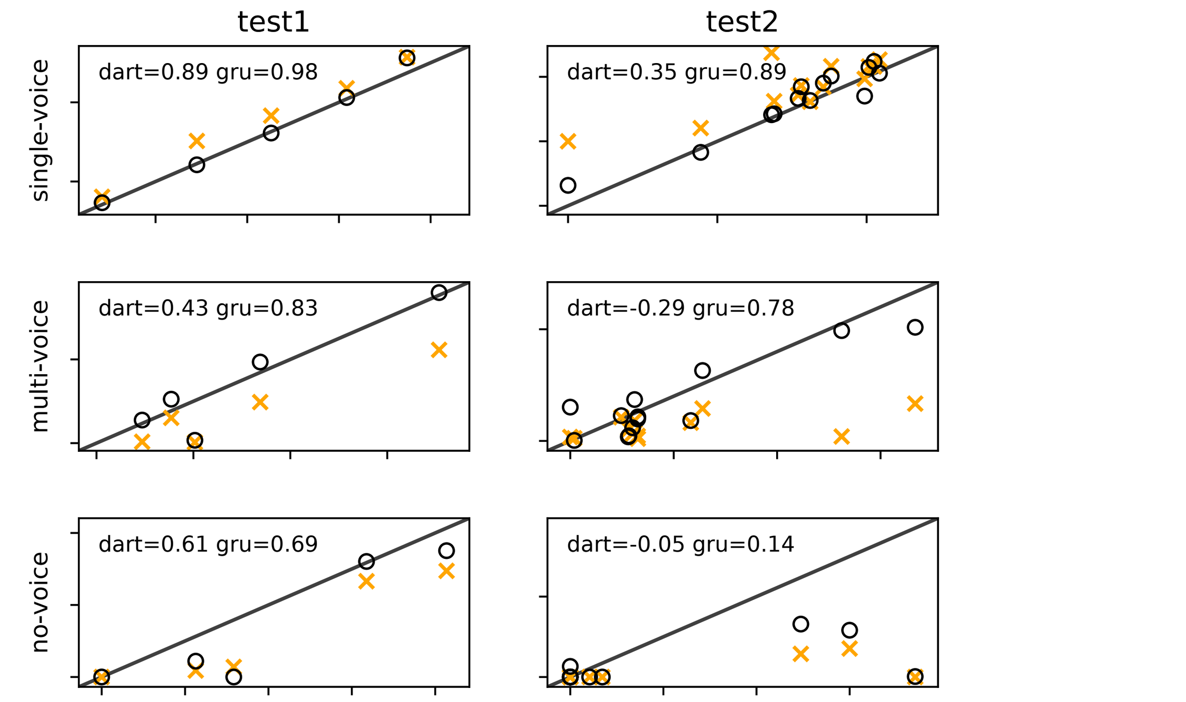

Figure: Correlation between the predicted amount of time (x-axis) spent on each activity for DART (orange x's) and GRU (black circles) relative to the ground truth (y-axis). R^2 listed in each subfigure.

- DART over-predicts single-voice while under-predicting multi-voice and no-voice.
- The average GRU coefficient of determination (R^2) across the test sets is 0.94 for single-voice and 0.81 for multi-voice.
- The GRU provides better estimates of time spent, especially for multi-voice.

Conclusions

- We propose deep and recurrent neural network approaches for identifying classroom activity and report improvements in frame error rate and F-measure.
- 32% to 45% relative reduction in frame error rate over previous state-of-the-art when generalizing to new class sessions from previously seen and unseen instructors, respectively.

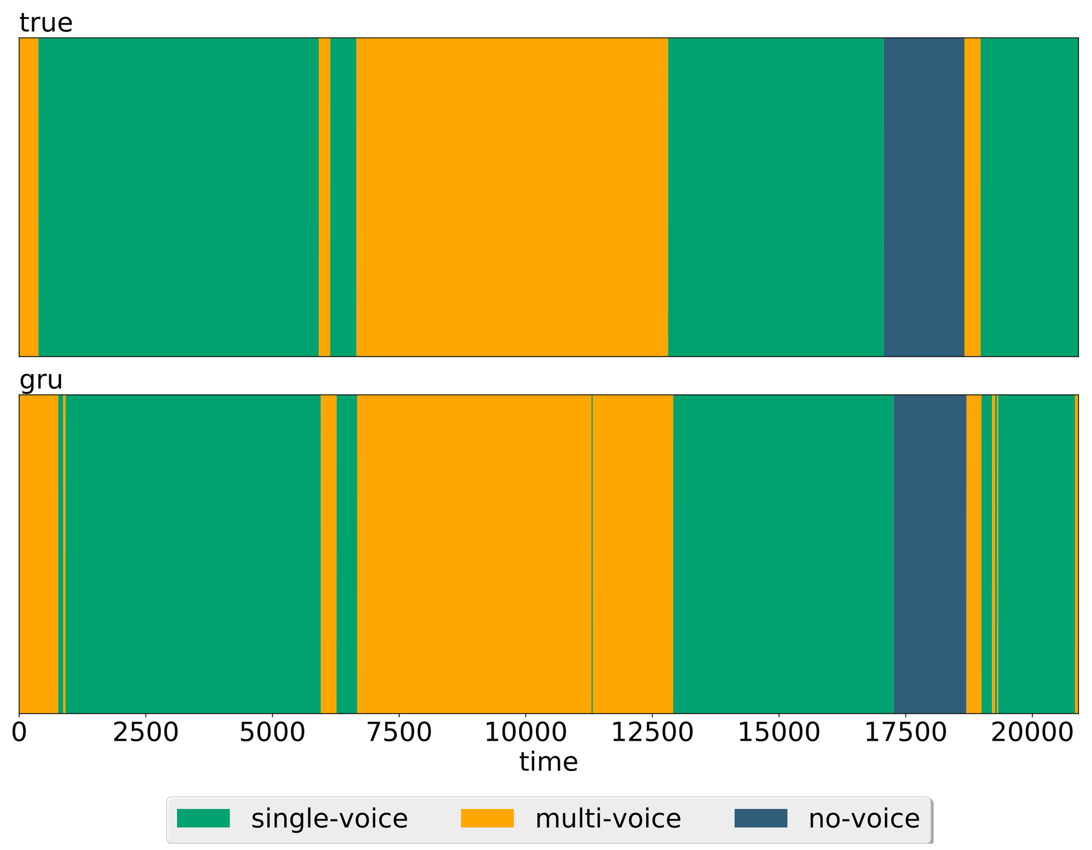

Figure: We show results from a class session in test1. The upper figure is ground truth and the lower figure is the GRU prediction. All detections less than 5s long were filtered out.