11-755/18-797 Machine Learning for Signal Processing

Final Poster Presentations

Instructor: Bhiksha Raj

TAs: Zhiding Yu, Bing Liu

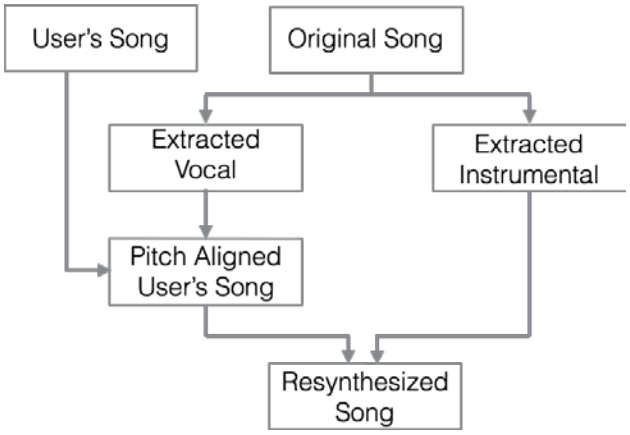
Fall 2015

1. SO YOU THINK YOU CAN SING?

Karishma Agrawal, Lu Lyu, Tomer Borenstein, Akshay Kulkarni

A classic Karaoke is an interactive entertainment tool where a person sings along instrumental version of songs. The choice of songs is limited to those that have an official instrumental version. A much cooler thing to do would be to let the user sing along any song that he or she likes, then extract the vocal out of the song and replace it with the user's recording. Also, we attempt to modify the pitch of the user singing to the original vocal, so anyone can be a 'perfect singer' in front of our program.

In this project we attempt to create a program that takes a song with instrumental and vocal components, separates the singer's vocals from the instrumental portion of the song, modifies the user's vocals to match the pitch and temporal nature of the original singer's vocals, and inserts the modified user's vocals back into the song to make it sound like the user was originally singing the song.



Non-Negative Factorization and Robust PCA is used for voice separation. The base frequency and onset detection techniques are used for aligning the pitch of the user and singer vocals. Finally the modified user vocals and the instrumental part are combined to produce the resynthesized song.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

2. SELF-PACED LEARNING IN MULTIMEDIA EVENT DETECTION WITH SOCIAL SIGNAL PROCESSING

Junwei Liang

With the increasing amount of user generated content (UGC) collection, multimedia video event detection has become an important research problem. This paper will present a method for automatically annotating consumer videos based on social signal in human communications for social events. We investigate adding social signal features such as facial emotions, crowd information and acoustic scene to multimedia event detection. We experiment with the social signal features and the basic state-of-the-art visual CNN feature and compare their performance using self-paced learning algorithm, a recently proposed learning regime inspired by the learning process of humans and animals that gradually incorporates easy to more complex samples into training. We also investigate the effects of different schemes and parameters of self-paced learning in such scenario, and compare our learning algorithm with simple batch training. Finally, we explore and compare fusion systems of different modalities and social signals. We test our system on the subset of data collection from TRECVID MEDTEST2013.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

3. IMPROVING INTONATION IN AUDIO BOOK SPEECH SYNTHESIS

Anuja Kelkar, Parag Goel, Shrimai Prabhumoye

Intonation (fundamental frequency, F0) is a key expressive component of speech that a speaker employs to convey his intent in delivering a sentence. It encodes a lot more information in the form of structure and type into an utterance than conveyed by the words. The prior work, Statistical Phrase/Accent Model (SPAM), presents an 'Accent group' based intonation model for statistical parametric speech synthesis.

We are using data from CXB audiobook, which contains 20 hours of US Female recordings. The recordings are already aligned to text, so there is no pre-processing required at our end for use in the algorithms. The data is labelled with syllables and phonemes.

The goal of this project is to develop interesting features to improve intonation of phrases in the audio books.

We have developed interesting features such as whether a word is an adjective or not, whether a word is a proper noun or not, whether a word is a subject in the sentence, whether a word is an object in a sentence, depth of a word in the parse tree, position of a word's parent in the dependency parse tree, etc. After adding these features to the prior SPAM model, we expect these features to provide more information regarding intonation of phrases in the text and to further improve the SPAM model results.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

5. YOUR KEYBOARD IS NOT YOUR FRIEND

Dustin Axman, Aleksander Bapst, Chen Liang

Protection of information from malicious intent is a topic of increasing concern in the security community as advances in machine learning continue to broaden the avenues of attack. One such attack which has generated interest in the last decade has been the potential for reconstruction of a user's pattern of keystrokes by analyzing the subtle differences between sounds made by individual keys in audio recordings of typing. Our project implements such attacks in an unsupervised manner.

Audio recordings of typed English text using laptop or webcam microphone recordings were used as our primary data source. Following segmentation of individual keystrokes, the gathered audio snippets were clustered using the k-means algorithm, and the sequence of labels were modeled as a hidden Markov model (HMM), which models the correlation of hidden states with emitted observations. The transition matrix and initial state probabilities were gathered using statistical analysis of a large body of English text, and used as input to the Baum-Welch algorithm for estimation of the emission matrix. The most likely hidden state sequence was predicted using the Viterbi algorithm and further refined using spelling correction.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

6. LEARNING SUCCESSFUL STRATEGY IN ADVERSARIAL GAMES

Chris Dunkers, Allard Dupuis, Mitch Kosowski, Aditya Sharma

One of the earliest problems in Artificial Intelligence has been to make computers become good at playing games. Most of the early work in this area was centered around traditional AI "search" algorithms like minimax search with alpha beta pruning. Recently, Google DeepMind came up with a more "machine learning" style solution when they made a computer learn the optimal strategy to win Atari games by playing the game multiple times. They developed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning. This was a big step towards using a human-like experience-based strategy to solve these problems

In this current work, we try to emulate a more "human learning" style network which is based on the fact that when humans first encounter a game they haven't played before, their learning begins by trying to learn the rules of the game which is followed by them trying to find a winning strategy. Then, over time and with more experience, learning of an optimal winning strategy takes place. In most cases this may take years to happen if it happens at all. In the current work, we propose a technique that first learns the rules of a game and then learns a strategy on top of this by experiencing the game itself. We have used Connect-4 as the game of our choice. This is an interesting technique because (1) it is similar to how humans learn to play games and (2) this method should be generalizable to different games including ones where search and other techniques perform poorly. Hence, with this approach it should be easy to extend the same system to learn to play and win a new game (with a grid-based structure) like Tic-Tac-Toe, Connect 5 etc.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

7. GESTURE PHASE SEGMENTATION

Paul Buchana, Omar Syed

In this project, we seek to aid human computer interaction by identifying nonverbal communication through the use of hand gesture analysis. Our dataset is comprised of individual frames from a recording of people telling a story from a book. We analyze the positions of the speaker's hands, neck, and spine to classify frames into the types of gestures. In our analysis, we present the performance of KNN and SVM classifiers in comparison to a RNN in which the temporal aspects of the data are considered. In particular, we want to maximize performance of classifying a certain gesture position called a stroke, as these gestures often have semantic meaning that can help understand human communication.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

8. LOOP QUERIER – SEARCHING THE RHYTHMIC PATTERN

Che-Yuan Liang, HangYang, Luchen Li

When there are thousands of beat loops with varies rhythmic patterns in your library, how to find out the beat in your mind? It is time consuming to listen through the entire loop beats to find out your favorite one. Instead, sending the query by tapping might be a feasible approach to spawn a list of candidates. However, the distance between different time series needs to be established before the searching or clustering.

The project use human tapping as training data to model the distance between rhythmic patterns. The distance function then enables us to finding the "similarity" of time series for sorting the candidates and further clustering or visualizing distance of the loops. The model is LSTM Recurrent Neural Network that takes 2 input time sequence with single hidden layer composed of 8 units that is self-recurrent, output the value ranging from 0 – 1.

The evaluation using mean reciprocal rank (MRR) achieves 0.5 on with this model at this time. The performance is compared with using correlation and the cost of dynamic time warping (DTW) as distance function. Also implemented an interface to search the loop, and plan to visualize the distance with the trained distance function.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

9. VISION-BASED MONTE CARLO LOCALIZATION FOR AUTONOMOUS VEHICLE

Wenda Xu

Localization is an essential functionality for the navigation of autonomous vehicles. For an autonomous vehicle to stay in a lane, centimeter level accuracy is required. However, commonly used solutions of coupling a high precision GPS with IMUs may only reach this accuracy in open sky environments. Moreover, this solution is too expensive for mass production. An alternative solution could be achieved by using other sensors such as camera and LIDAR to sense the surrounding environments. In this project, we use a camera to detect lane markings along the road. The relative position, heading, and curvature of the lane markings will be used to localize the vehicle to the road. We use a Particle Filter to estimate the vehicle state since it is able to manage multi-modal estimations. Data was collected by an autonomous vehicle, including video, odometry data from the car (e.g. wheel speed, steering wheel angle), and GPS data from a high precision GPS with cellular-transmitted RTK (Real Time Kinematic) correction. The GPS data is only used as ground truth. The performance of the proposed particle filter-based localization approach is validated by comparing the result to the high precision GPS data.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

11. BEATBOX-TO-DRUM CONVERSION

Jonathan Michelson, Devansh Zurale

MIDI drum tracking is prohibitively tedious. Musicians who desire specific realistic percussion in their recordings must funnel their creative ideas through the physical restrictions of a MIDI controller. Beatboxing, contrastingly -- where one imitates the different elements of a drum kit with his voice ¬-- is highly intuitive and widely used to convey or perform rhythmic ideas, suggesting its appropriateness as an input to a performance capture system. This potential is explored through creation of a preliminary translation system that employs signal processing and machine learning techniques to classify each unit of an input beatboxing sequence and to synthesize the corresponding drum pattern. The system focused on the speaker-dependent case and was trained on three drum elements (kick, snare, hihat) to scale the problem to a tractable size. Accurate results were obtained with this arrangement, warranting future development of a robust implementation.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

12. CITY-LOCATION ON FLICKR VIDEOS USING ONLY AUDIO

Guan-Lin Chao, Benjamin Elizalde, Ming Zeng

Can we identify the city that a web-video belongs to, based only on the audio?

City-identification on videos aims to determine the likelihood of a video belonging to a set of cities. In this work we present an audio-only content-based approach to identifying cities, so we don't use any other information such as the video's images, user-tags or geo-tags. Using the audio information provides insights into the extent to which city-scale geo-locations of videos are correlated with their acoustic features. Achieving success using only the audio would suggest the potential for further improvements when additional modalities are incorporated.

We present a novel approach that consists on representing the city's audio track with a collection of urban sounds using a Signal-Projection-on-Bases (Pseudo-inverse) technique. This method allows us to perform identification and provide semantic explanation of the acoustic content. We also explored different features and classifiers on the yearly MediaEval - Placing Task videos and the Urban Sounds datasets.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

13. FACIAL LANDMARKS BASED VIDEO FRONTALIZATION AND ITS APPLICATION IN FACE RECOGNITION

Niv Zehngut, Hill Ma

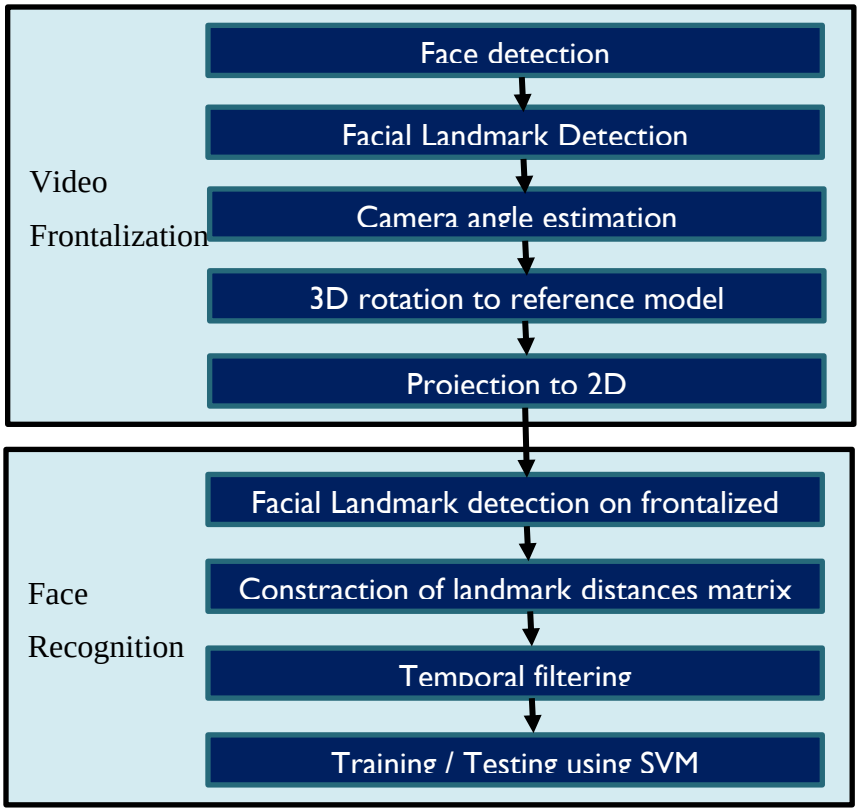
The usage of visual content in general, and video calls in particular, steadily grows wider with nowadays available communication bandwidth. However, many users still refrain from using video calls, even though they usually prefer visual interaction in-person on top of a regular phone call. Reasons for that phenomena includes sub-optimal camera angle and face distortion that are prevalent in videos taken by hand-held smartphones. Another disruption that users face during video calls is the constant need to keep their face in the middle of the frame. In this work, we present a real-time video frontalization method, which addresses both of these issues. We utilize recent advances in facial landmark detection as well as image-based face frontalization techniques, and generate smooth frontalized videos. We use this result to address the problem of face recognition in videos, a well sought-after task. Our solution uniquely relies on facial landmarks as features and incorporates temporal information, and hence can be used as additional classifier in various applications. On the gridcorpus database, we achieved a verification rate of 97% at false positive rate of 0.1%. Specifically, we employed an application-oriented testing scheme in which training is done based solely on 1-5 instances per subject, and testing is performed based on only 1 second videos.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):



Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

14. AUDIOSHOP

Raymond Xia, Mutian Fu, Lucas Crandall

The timbral characteristics of the singing voice is not well understood because of the complexity of its underlying temporal dynamics and spectral variations. Nevertheless, timbral modification is of interest to many applications. In this project, we attempt to allow users to modify the timbre of the original singer's voice to more closely match their own, while maintaining the quality of the rest of the music recording. Similar work has been done by exploiting physically meaningful features such as the spectral envelope and the voice onset/offset. Unlike those, our project makes a novel attempt to statistically learn the timbral quality of both the singer and the user, and morph them from one to the other. To learn and modify the timbral features of the voice, we combine methods such as latent component analysis and non-negative matrix factorization.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

15. PREDICTING AND CLASSIFYING RF SIGNAL STRENGTH IN AN ENVIRONMENT WITH OBSTACLES

Ervin Teng, Maxim Kovalev, Pedro Najera, Kai--Wen Liang

In order to plan and deploy multi-transmitter wireless networks, such as cellular networks, engineers must be able to predict how well the electromagnetic radio waves will propagate from the transmitter. RF signal propagation prediction is typically performed by a human expert using propagation models, derived from empirical field measurements, which predict RF path loss over a given distance. These models must be chosen and tuned to the specific environment in which the transmitter will be deployed; thus detailed information about the environment must be collected. After a prediction is performed, manual measurement of the resulting signal strength must be performed. This process of manual modeling and measurement is extremely time-consuming; network deployment is measured in weeks and months. In addition, it does not easily scale to the problem of three-dimensional network deployment in support of unmanned aerial vehicles.

This project seeks to automate the signal propagation modeling process. Using easy-to-collect point-of-view images of the environment from the transmitter, we seek to identify obstacles and thus radial paths from the transmitter where the propagation would be weaker. Using RF measurements collected by a drone during several field experiments, we associate signal strength loss with imagery taken from a transmitter-mounted camera, dividing a single panoramic images into segments with corresponding score, and computing features from each segment. Several techniques were investigated, such as Principal Component Analysis (PCA), ensemble boosting techniques, and neural networks, and we present the results along with the efficacy of each technique. The best result was obtained using RUSBoost with a specificity of ~78% for obstructed segments. We also demonstrate that results can be significantly improved by the collection of additional, vertically-offset images.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

Figure 1: Imagery scored by the RF signal loss due to obstacles.

Figure 2: Measurement device and sample of raw data.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

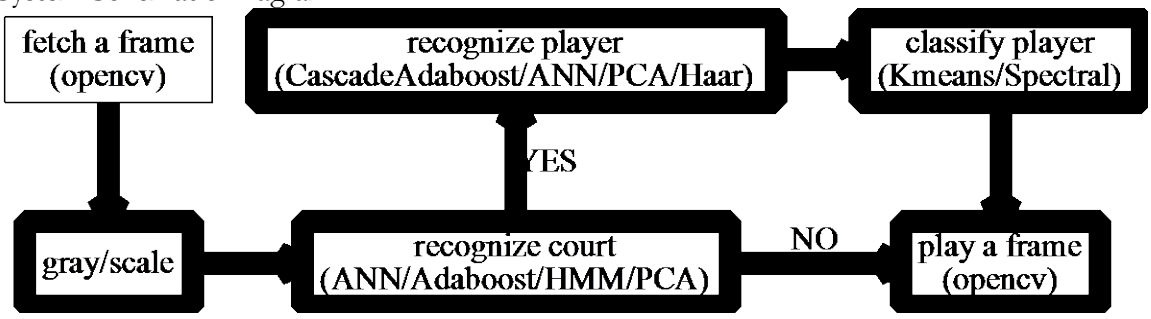
17. BASKETBALL PLAYER REAL-TIME DETECTION

Zhi Liu, Shushan Chen, Victor Shin-Deh Chao

Picture this: The Chicago Bulls is down 2 points to the Lakers. Five seconds left to the end of the game. Bulls' ball. Coach calls timeout. Group. Now, instead of taking out a whiteboard and plan out the last shot from scratch, Coach holds an Ipad which automatically pulls out 5 strategy plans that had highest chance against Lakers from history records. Coach picks one and displays it on the Ipad. End of timeout. Derek Rose hits the buzzer beater. Game.

This is why we aim to build a system that can automatically detect the position of NBA players on the court. The provided data of this system can be applied to further applications such as profiling strategy patterns. An automated system was implemented to perform player detection upon given video inputs. While the real time detector classifies teams and monitors the correct position of each player on the floor, it can provide an enormous amount of information to both professional and unprofessional basketball players. Not only can it make the work of coaches and analysts easier, common basketball lovers could also mimic and improve their game by learning a strategy or teamplay with a selection pool of the classic plays.

System Schematic Diagram:



Phase 1: System Activation Estimation: to predict whether the input contains a ball court.
Phase 2: Detection of Basketball Players.
Phase 3: Clustering Basketball Players to different teams.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

18. EVENT DETECTION IN SPATIO-TEMPORAL DATA

Kailiang Chen, Yilin Chen, Yi Wang

Detecting multiple event classes in videos has been a long-standing problem in computer vision. Standard approaches learn multi-class classifiers (or likelihood estimators) using many temporal segments of the events, while do not maintain classification consistency between adjacent or overlapping segments. As a result, during testing, the detectors tend to report different class types for adjacent or even overlapping test segments, producing temporally inconsistent class labels. Moreover, the detection processes is computationally costly, for one needs to evaluating multi-event detectors for all candidate segments to identify the class with the highest score. To address these challenges, we propose a novel alternative algorithm, distributed max-margin event detectors (DMMED), to detect multiple event classes. For DMMED, we adapted the feature extraction to Spark and are trying to improve the max-margin event detector to Petuum. Using DMMED, the detectors learn to sequentially discard the less confusing classes (instead of always reporting a class with highest score), and from the remaining classes, unique class identification is made for all partial segments. Moreover, DMMED sequentially reduces the number of classifier scores need to be computed. Experiments on 3D depth videos and RGB videos show that our approach achieves higher accuracy and efficiency than the standard Multi-class SVM-based detectors and other Max-Margin event detectors.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

19. ELECTRIC LOAD PREDICTION FOR AIRPORT BUILDINGS

Minkyung Kang

Given their large energy footprint and the availability of building energy management systems, airports are uniquely positioned to take advantage of demand response (DR) programs. One of the most essential components to utilize such DR opportunities is a baseline model, which is the estimate of what the load would have been in the absence of curtailment. Since the performance of peak-shaving DR programs is determined by the amount of curtailment during the peak period, it is important to have an accurate baseline to evaluate the DR strategies and estimate the DR savings.

This project aims to develop baseline models specifically intended for airport facilities. Specifically, I propose piece-wise linear regression models for predicting electricity demand using time-of-week, temperature, and flight schedule information. I hypothesize that flight schedule information would help explain a significant portion of the load after temperature and time-of-week information has been accounted for.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

20. UNSUPERVISED TEMPLATE LEARNING FOR BIRDSONG IDENTIFICATION SYSTEM

Shunsuke Aoki

Natural birdsongs are one of the most familiar natural sounds in the physical world, but it is difficult for general people to estimate a species without the help of an expert. For solving this issue, this project aims at designing bird species identification system with mobile computers. In our project, all of the acoustic data were collected via a smartphone and labeled manually as the ground-truth data by the ornithologists.

Our system used the features of the spectrograms as input features for the unsupervised template learning and also Random Forest Regression was applied. The classifier achieved 0.91 AUC in the evaluation part. The designed application would enable us to identify a bird from the songs in the actual forest and to entertain the natural environments in their daily lives.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):

21. REAL-TIME KEYWORD SPOTTING IN VIDEO GAMES

Nikolas Wolfe

Real-time keyword spotting is a challenging task, especially in the context of multiple simultaneous speakers. With children's speech, this task becomes even harder. Children, unlike adults, typically do not follow orderly turn-taking procedures and may interrupt each other at arbitrary times. If instructed to use only a restricted keyword vocabulary in order to play a voice-controlled video game, children may say any number of bogus out-of-vocabulary words. When having fun, they may shout, distort, elongate or fragment their words at random, they may sing or laugh, talk among themselves, and in general do whatever they please. With such noisy, non-standard speech, state-of-the-art keyword spotting systems often fail to perform well with even a simple two-word vocabulary task. This project details the implementation of a specialized real-time keyword spotting algorithm for a side-scrolling video game called Mole Madness developed at Disney Research Pittsburgh for the purposes of studying child turn-taking, engagement, and language use in an entertainment setting. We discuss the issues related to children's speech across several age groups and give quantitative performance assessments on a recorded dataset of children playing Mole Madness during a voluntary research study and data collection during the summer of 2015.

Score this Project

Originality: 1-5 (5 is highest):

Completeness: 1-5 (5 is highest):

Overall: 1-5 (5 is highest):