Obedient/Disobedient Emotion Recognition from Children' Speech*

Jiaqi Zheng
Aalto University
Espoo, Finland
jiaqi.zheng@aalto.fi

ABSTRACT

UPDATED—May 16, 2018. This paper explores the wrapper method of feature selection with several different pattern recognition classifiers. The goal of the exploration is to eventually train a model with relatively high accuracy in classifying the obedient and disobedient emotions. Here obedient means the children are willing to stop when restricted by online screen time, while disobedient implies the contrary. This study of attitudinal emotion recognition contributes to the potential future voice interaction scenario where a home smart device can supervise the children' online screen time. Moreover, in the current speech emotion recognition field, it remains unclear what specific acoustic features are more important for children compared with adults. This study went through data collection, feature extraction, wrapper method of feature selection, and finally compared the selected features with adults' speech. Therefore, it, on the one hand, gives the accuracy validation as a scientific proof; on the other hand, offers the insight about what acoustic features to focus on for children's voice. Regarding the results, Naive Bayes classifier achieved the highest accuracy of 95.1% for speaker-dependent recognition. Also among all kinds of features, the intensity is found to be much more important for children than adults, while Mel-frequency cepstral coefficients (MFCCs) and pitch (F0) matter more for adults.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Sound and Music Computing

Author Keywords

children's speech; emotion recognition; obedient/disobedient; online screen time; wrapper feature selection.

INTRODUCTION

The intelligent online screen time control has a significant industrial value with the smooth voice interaction. The market comes from the young adults because one of their primary concerns is the negative impact from the media on their children. According to the survey results [9], the most typical practical approach of getting rid of this negative influence is restricting online screen time. On the other hand, since the voice interaction has already gained attention recently, it is imaginable that the market size is considerable when combining the online screen time restriction with voice interaction.

There is not yet a robust and direct competitor, although similar products exist with part of such functionalities. One of the existing popular solutions of voice interaction is the intelligent client Alexa. However, it mostly focuses on general daily issues such as playing music and answering the weather. Although there are already some browser extensions for limiting online screen time, they only show textual notifications directly. Therefore, it is worthwhile to study the scientific proof of this restrictive interaction scenario. The study also contributes to finally helping the children to form a healthier Internet using habit.

As one of the concrete smart home device examples, the Finnish company F-Secure has the security router SENSE[1] . SENSE protects every connected device from online threats. In reality, F-Secure is already considering to expand the young adults' market by integrating voice supervision of online screen time with SENSE.

Therefore, with the ultimate goal of intelligent and constructive voice interaction in the online screen time restriction scenario, it is indeed first to interpret the obedient and disobedient emotions from children's responses. This paper only scopes and focuses on the emotion recognition part. Therefore, the research problem is how to classify the obedient and disobedient emotional states from children' speech signals. The overall objective is the relatively high accuracy for the recognition results. Sub-objectives include comprehensive extraction of acoustic features, efficient selection of the feature subset, and relatively excellent performance of the selected pattern recognition technique.

Related to the problem of classifying children's obedient and disobedient emotions from the acoustic features, there is already much research done in the field of affective computing that classifies prosody signals into basic and specific human emotions such as sadness, anger, happiness and fear. However, attitudinal emotions of obedient and disobedient are not that much studied. However, it is useful to distinguish them apart because it makes much sense for selecting a proper educative response by the intelligent agent.

In our scenario, obedient means the children are willing to obey the restrictive guidance. They might happily accept it directly, or agree to stop though feeling upset. Therefore, they may appear to be happy, peaceful or slightly down. Reversely, disobedient means the children are reluctant to obey the stop-watch-TV instruction. At that point, they might shout rudely, complain vehemently, or anxiously ask for a time extension.

*This student project was carried out in Spring 2018 in the Aalto University course Research Project in HCI (ELEC-E7861) under the supervision of Antti Oulasvirta.

[1] https://www.f-secure.com/en_US/web/home_us/sense

Therefore, the voice may sound stronger. The hypothesis is that all these features can be reflected in the acoustic signals so that feature-based classification is achievable. For instance, the voice intensity for disobedient should be stronger, and the loudness should be higher. More discussion about the differences in various speech features comes in the Feature Selection section.

The research process started from data collection, feature extraction, to the wrapper method of feature extraction where the performance of a specific classifier is used to evaluate the selected feature subset. The initially extracted baseline features cover almost all the possible aspects, including intensity, loudness, Mel-frequency cepstral coefficients (MFCCs), pitch (F0), probability of voicing, line spectral frequency (LSF), and Zero-Crossing Rate (ZCR). The baseline feature set is informative enough for the exploration of the optimal feature subset. Then the feature selection is conducted together with exploring and choosing the pattern recognition classifier, since the wrapper method searches all the possible combinations of the features, training each of them with the same algorithm, and returns the feature subset with the highest training accuracy. The following sessions of this paper are in the same sequence as stated above.

CHALLENGES AND LIMITATION

The first challenge is realism, literally called novelty and disruption effects when the children are not behaving the way close to their routine. For example, their tone might sound unnatural. However, the emotion recognition only makes sense if the training data is valid and realistic enough. To solve this, I tried to simulate the responses that the children might say as close as possible to the realism. This is also the goal of the background study that I performed as elaborated in the other session. Besides, I instructed them to feel relaxed at the beginning and kept observing whether they are behaving rightly. This cognition enabled me to give reminders at the proper time.

The second challenge is data collection control. One of the reasons why almost all of the previous work only used adults' speech is that it is easier for adults to follow the experimental instruction. Moreover, recruiting volunteers does not work because for children we need the permission from their parents.

The last challenge from the technical side is taking care of the feature selection and classifier selection at the same time since different classifiers suit different feature sets. The wrapper method of feature selection can solve this problem to some degree because its criteria of selecting the best feature set are right the performance of the chosen classifier. However, this method searches all the feature combination space to train the model respectively before coming to a result. Therefore, this has the threats of being highly time-consuming for some of the classifiers.

RELATED WORK

A considerable amount of previous affective computing research works on speech emotion recognition with different approaches and various classification goals. All of the works that I found focus on adults' speech signal but not children'. Besides, the average accuracy is roughly 70% - 85%. Lee [8] used Linear discriminant classification with Gaussian class-conditional probability distribution and k-NN methods to recognise the negative emotion with highest 80% accuracy. Roy [14] explored the approving and disapproving emotional classes with Fisher Linear Discrimination method, which achieved highest 88% accuracy for speaker-dependent recognition, and highest 65% for speaker-independent. Dellaert [3] proposed population hillclimbing recognition technique and completed highest 74.5% accuracy for four emotional categories. Kwon [7] compared various classifiers of SVM, linear discriminant analysis, and hidden Markov model and found that Gaussian SVM had the best performance of 96.3% for stressed/neutral style classification and 70.1% for four other emotional states. In conclusion, researchers have different approaches addressing different problems. However, these nevertheless give the general guidance that I should aim at an accuracy higher than 85%; I also discovered that KNN is a method worth noticing given the excellent performance achieved by several related works.

On the other hand, the significance of children's emotion status in educational scenarios is also recognised by researchers from different fields. One of the relevant works concentrates on the achievement emotion in the interactive education [11]. It argues that children's self-control over activities and outcomes and also their subjective values to the events have the most significant impact on achievement-related emotions. This cognition gives the implication that under our online screen time restriction scenario, children's positive feelings are possibly guidable. For example, through giving children the control of saving the show progress.

BACKGROUND STUDY AND USE CASE

The emotion recognition in this research applies explicitly to the online screen time restriction scenario, therefore, a proper design of the to-be-collected training sample sentences that children possibly response when hearing the restriction is especially important. Therefore, this session aims to figure this out by designing the interaction structure and specific use case under that situation. Additionally, this background study also attempted to validate the market demand and future opportunities.

The background study approach is collaborative design. Collaborative design means taking users in as design partners. This user research method is suitable for my case because first of all, the design goal is to identify the use case that is relatively close to users' daily life; therefore, their ability of design is relatively high. Secondly, collaborative design works synchronously, which enables the interaction between the researcher and the user. This corporation is extremely useful in my case since I am not a parent; this approach merges my technical and design related points of view with the user cultural background.

Concretely, the collaborative design was carried out with a 34-year-old PhD student at Aalto University, who is the mother of a 3-year-old boy. I hosted a design workshop that goes through "one diamond" design process, which means one divergence brainstorming plus one convergence result discussion. The

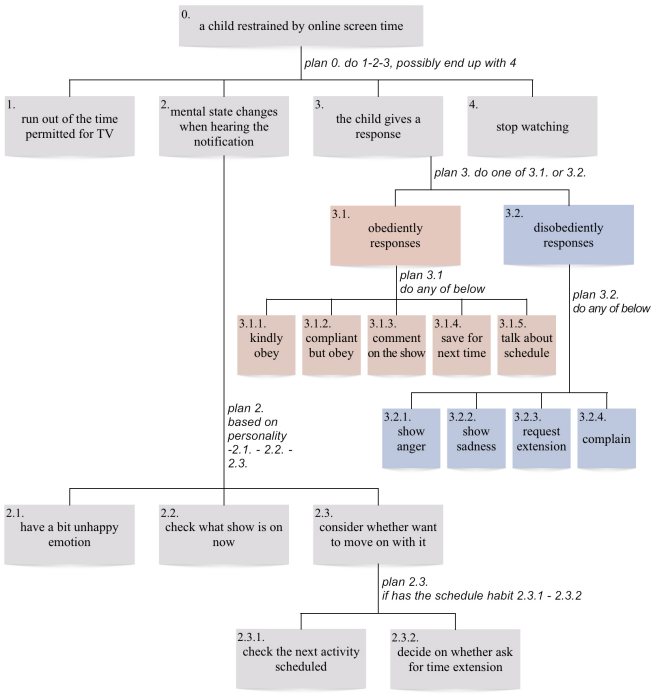

Figure 1. HTA: a Child Restrained by Online Screen Time

whole collaborative design process took around 30 min. I first started with a brief introduction and quick questions about whether she worries about Internet use of her son; how she currently plans to do to control his screen time; and whether she sees the value if child control is an attached functionality to a security router. These questions all aim to validate the market. The results of them were considerably positive: Parents do care a lot about child Internet control, and her current solution is just spending plenty of time accompanying the child by herself for better supervision. This phenomenon is indeed universe according to the research [9]. Recently, a few screen time apps have appeared; however, as far as I have found, none has been integrated smoothly with existing home devices.

After those fundamental questions above, the workshop started: both of us worked on brainstorming writing on a piece of paper about two well-sketched scenarios: What will parents say to set screen limit time; what might the child response when the machine notifies that the time is running out. We first worked independently for seven minutes, and then we exchanged the paper to build on top of peers' ideas for three more minutes. After that, we at the end discussed the ideas to see, what essential use cases that we agreed on are.

This workshop turned out to be productive. It led to some interesting findings that I have not previously considered. I presented the result conclusions of how the child might response when hearing restriction with the hierarchical task analysis (HTA) [1] as showed in the figure. HTA is helpful to demonstrate a process step by step, therefore is also more beneficial for building the insight into the reasons behind.

To summarise the key findings, importing the schedule by parents beforehand can potentially encourage the child to seek for healthier alternatives; however, when the child tries to negotiate time extension, it should still be parents who have the controlling power.

In conclusion, from the perspective of possible responding behaviour, as also analyzed in the figure, the possibilities are categorised as the following:

- Accept, although maybe complain.
- Positively inquire the sequential arrangement.
- Ask to save the on-going TV progress for next time.
- Show frustration and anger, negatively complain.
- Try to negotiate and seek for a time extension.

The laboratory experiment utilised these results by designing 30 obedient responses and 30 disobedient responses. Details are in the Data Collection session.

DATA COLLECTION: LAB EXPERIMENT

Since the plan is to conduct speaker-dependent emotion recognition, the goal of data collection is to obtain more valid audio data with either ab obedient or disobedient label. This data collection Controlled Laboratory Experiment was carried out in Kielo School[2] , which is an International primary school located in Helsinki. I used the method of Controlled Laboratory Experiment. Concretely, I invited three eight-year-old children to participate, two of whom are boys, the other is a girl. They individually went through the same procedure. I started with casual chatting until the participant looked comfortable with the environment. Then I introduced my research topic and gave the participant consent to both the child and the teacher. After that, I presented the scenario and asked the child to envision that they were right now watching the TV but got interrupted and stopped because of time limitation. I had prepared six cards with written sentences for them to read out one by one. 3 of the cards have obedient responses while the other 3 have disobedient ones. The three obedient cards have the same 30 sentences though in the different sequence, thus getting rid of the influence of context: on the first card, the sentences with similar meanings were put closer, while those on the other two cards were in totally random sequence. The case was the same for the disobedient side. Specific examples are shown in Figure 2 and Figure 3.

The children read the six cards following the order of obedient, disobedient, obedient, disobedient, obedient, and disobedient. Therefore, I finally collected 180 audio records with 90 obedient and 90 disobedient sentences for each child. So the total amount of the training samples obtained was 540.

The audio recording equipment was my MackBookPro with Quicktime Player, and the recorder was always on for every participant, thus making the environment more natural.

[2] http://www.kieloschool.fi

Figure 2. Experiment Instruction: the first obedient card

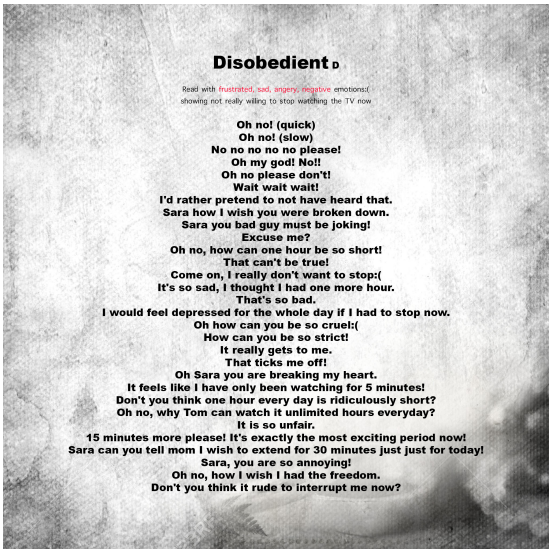

Figure 3. Experiment Instruction: the first disobedient card

DATA PREPARATION

Since some of the recorded audio is not valid training samples, data preparation process got filtered them out. For the original data, the three children each have 90 obedient speech and 90 disobedient ones, which add up to 540 in total. I left out the slow sentences which were because the child was not familiar enough with the phrases. Some other audio with turbulence or irrelevant people talking were also left out. Thus, as a result, there were finally 328 valid samples, 58 obedient 62 disobedient for the first boy, 33 obedient and 52 disobedient for the second girl, 61 obedient and 62 disobedient for the last boy.

ACOUSTIC FEATURE EXTRACTION

There are in principle two approaches for feature extraction. One is to extract a few target features; the other is to obtain the features as comprehensive as possible and then explore to find out the more crucial feature set. Although the first method is doable since almost all of the research agrees that pitch, energy-related features and MFCC coefficientsLe [8] are good indicators for emotions, I used the second approach because it is logically more informative and thus allows more possibilities.

Concretely, I used OpenSMILE toolkit [4] to perform the extraction, which is a prominent tool in the emotion recognition field with a comprehensive library. I used the baseline set of 988 acoustic features for each of the 328 audio files. The extracted features contain intensity, loudness, 12 kids of Mel-frequency cepstral coefficients (MFCCs), pitch (F0), probability of voicing, eight sorts of line spectral frequencies (LSF), zero-crossing rate and the delta regression coefficient of all those above. Furthermore, the max/min value, range, arithmetic mean, linear regression coefficients, linear and quadratic error, standard deviation, skewness, kurtosis, quartile, and three inter-quartile ranges are all computed. Below introduced the extracted feature categories in detail:

Intensity

The intensity means the strength of the voice. It is the sound power for each unit area. Obedient speech tends to have less intensity, while disobedient tends to be stronger in intensity.

Loudness

Loudness is closely related to the intensity. It has the similar relationship with the emotions that the disobedient ones are strong. However, the amplitude is the strength from the perspective of the ear's perception. Also, it is doubled with only a 10- fold increase in intensity according to the rule of thumb for loudness.

MFCCs

Mel-frequency cepstral coefficients (MFCCs) represents the short-term power spectrum, where "Mel" is the perceived pitch or frequency in a unit [8]. It is especially efficient in speaker-dependent content characterisation. Therefore, with different content and tone, the whole set of MFCCs vary.

F0

Pitch (F0), the fundamental frequency, is the lowest frequency of a periodic waveform. As one of the most basic acoustic

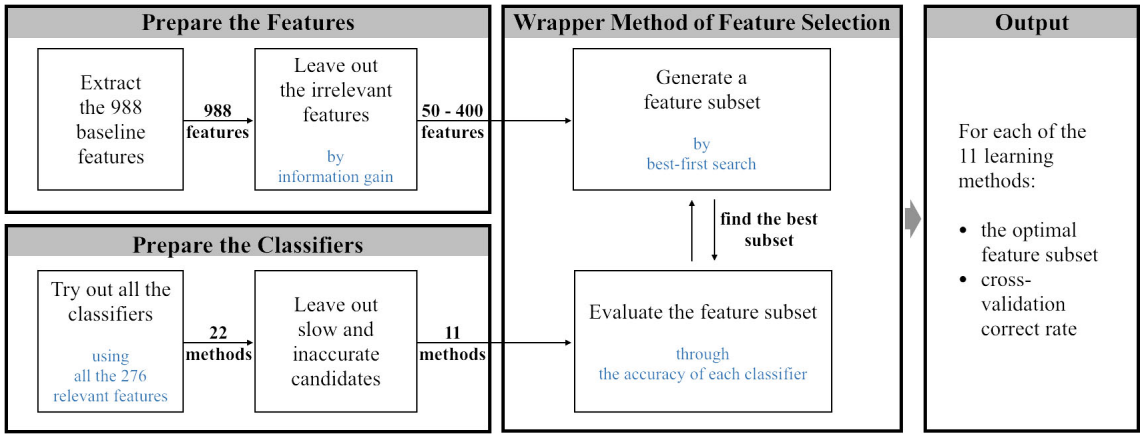

Figure 4. The methodology of the whole process regarding the data analysis. The core part is the Wrapper Method of Feature Selection in the middle. On the left side, the features and classifiers are input, while on the right it is the output.

features, almost all of the related works agree on its significance. For example, it has higher variation when people have stronger emotions, such as disobedient emotion.

Probability of Voicing

It is literately the probability distribution of voicing about whether it is present or the absent. It is therefore related to many features including speech rate and intensity. There it might has higher probability and denser distribution for the disobedient emotion.

LSF

Line spectral frequencies (LSF) represents transmission the linear prediction coefficients (LPC) [15], which analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz [10]. Therefore, similar to MFCCs, different tones with various emotions are presented differently with LSF.

Zero-Crossing Rate

The zero-crossing rate (ZCR) represents the rate of change in the speech signals. The stronger emotion of disobedient will probably have a higher changing rate. However, in the following comparison in the experiment, ZCR is not an efficient indicator.

WRAPPER METHOD OF FEATURE SELECTION

This session is the core part of data processing and modelling. The overall objective is to select the proper feature subset and then train the model with it to reach a high predictive accuracy. Such kind of Modeling and Simulation has the strength of repeatability. In other words, the trained model can enable the system to go through the same predictive process with other new speech inputs.

Feature selection is one of the essential steps in emotion recognition. It is a process of sorting out an optimal feature subset that enhances the efficiency in the model construction. It has notable advantages over training a model with all the baseline features, because it reduces the training time, improves the prediction performance and also facilitates a better understanding towards the data [5]. More frequently used feature selection methods are filters and wrappers. In this paper, I choose the wrapper methodology where the performance of a classifier is integrated into selecting the feature subset. This method is criticised to be time-consuming because of the amounts of computation compared with the filter method. The filter method directly selects the variable subset independently of any predictor. However, the wrapper method is remarkable in balancing the communication between the ideal feature set and the learning predictor, since the different learning method potentially has a distinctive feature subset with its most premium performance.

As the process overview shown in Figure 4, the wrapper method of feature selection needs feature preparation and classifier preparation before starting. This preparation is because both of the classifier and the features are the inputs for the wrapper method of feature selection:

Input: Prepare the Features

There are 988 baseline features while some of which are just irrelevant. Searching all the combinations without filtering them out is a waste of time. Therefore, I evaluated the worth of each feature by measuring the information gain concerning the classes. The gain was computed with InfoGainAttributeEval method of attribute selection in WEKA. As a result, out of the 988 baseline features, there are only 129 relevant features for Boy 1, 363 for Boy 2, 89 for Girl 1, and 276 for all.

Input: Prepare the Classifiers

I utilised Waikato Environment for Knowledge Analysis (WEKA) [6] as the data mining software to conduct the whole wrapper method of feature selection and the classification. There are in all 22 executive WEKA in-built classifiers for my case of data. However, some of them are time-consuming, it was therefore unreasonable to exhaust all the options.

To solve this problem, I trained all the 22 predictor candidates with the reduced 276 features. The absolute error rate is not essential at this point since the goal is to compare the

speed and the relative accuracy. After that, I left out the too-time-consuming ones, also abandoned the ones with an error rate higher than 30%. Finally, I got the 11 scoped range of learning methods. They are Bayesian Network, Naive Bayes, K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Sequential Minimal Optimization (SMO), Random Tree, Random Committee, Adaptive Boosting (AdaBoost), LogitBoost, Filtered Classifier, and Multiclass Classifier.

The predictor is now the sole variable in the wrapper method of feature selection because the prepared feature will not further change. The 11 classifiers are however used one at a time.

Processing: Feature Selection

With all the inputs settled, the feature selection process utilises the predictive performance of the classifier to score the subsets of features. In other words, the black-box generates a feature subset by best-first search, then evaluate it through the cross-validation error rate of the indicated classifier. This circle iterates until the optimal subset is discovered.

The best-first searching algorithm expands the candidate subset by the most promising nodes. And the cross-validation that I used is precisely the leave-one-out cross-validation (LOOCV) [13]. It trains the model with the left-out samples, and then the left-out sample for validation. This process rotates until all the data has been left out, then the average error is calculated.

Outputs

With an indicated classifier, the output is the optimal feature subset and the cross-validation accuracy. Since I have 11 classifiers as the different inputs, there are 11 sets of this kind of outputs for them respectively.

RESULTS

Accuracy

Table 1 demonstrates the results of the outputs of the highest accuracy for each of the 11 learning methods. As is also shown, to compare the performance of speaker-dependent recognition with speaker-independent, I did parallel recognition for Boy 1, Boy 2, Girl 1 and all data together. In summary, high accuracy is achieved. Naive Bayes obtained the highest accuracy of 95.1% for Boy 2. Following that, KNN reached the accuracy of 93.3% for Boy 1, and SGD completed 91.8% for Girl 1. The speaker-dependent recognition all accomplished at a correct rate higher than 90%, which is much beyond my initial goal of 85%. Although the speaker-independent recognition reasonably has a relatively lower accuracy of 89%, it is still higher than most of the previous work. Therefore, we can conclude that the wrapper method of feature selection for deciding on the feature subset and the model is an efficient approach in this children-specific obedient and disobedient acoustic emotion recognition. Also, it validates the initial hypothesis that high accuracy is reachable.

Figure 5 is only the graph presentation of the same results data. However, it intuitively gives another clear insight that the performance for different children varies much from the classifier to classifier. Each child has a distinguishing best suit of learning method.

	Correct Rate %			
Classifier	*B1 (120)*	*B2 (123)*	*G1 (85)*	*All (328)*
Bayesian Network	90.8	91.1	89.4	71.0
Naive Bayes	92.5	95.1	89.4	89.0
KNN	93.3	94.3	87.1	80.8
SGD	85.8	87.0	91.8	80.5
SMO	87.5	90.2	83.5	80.2
Random Tree	80.8	86.2	88.22	70.1
Random Committee	83.3	90.2	85.9	70.1
AdaBoost	78.3	91.1	87.1	81.4
LogitBoost	82.5	91.9	88.2	81.1
Filtered Classifier	83.3	91.1	88.2	82.6
Multiclass Classifier	86.7	91.9	89.4	80.5

Table 1. Classification results (correct rate) for various predictors using their optimal feature subsets respectively. It shows that high accuracy is reachable and each child has their different optimal classifiers.

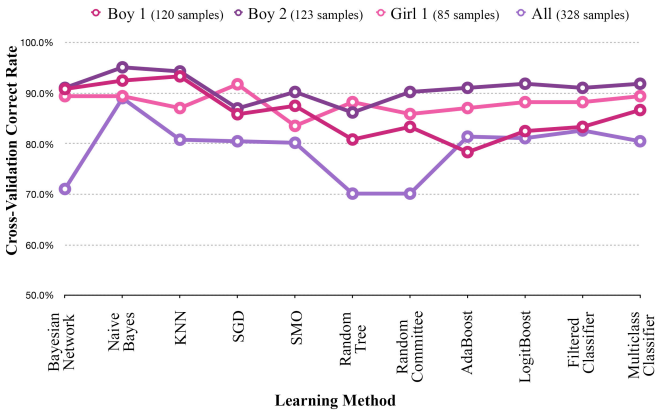

Figure 5. Classification results (correct rate) for various predictors using their optimal feature subsets respectively. It shows that speaker-dependent recognition has generally higher correct rates than the speaker-independent and that the performance varies from the classifier to classifier.

	Correct Rate %				
Class	*B1 (KNN)*	*B2 (NB)*	*G1 (SGD)*	*All (NB)*	*Average*
Obedient	91.0	93.7	90.6	86.2	90.4
Disobedient	95.8	96.7	92.5	91.7	94.2
Overall	93.3	95.1	91.8	89.0	92.3

Table 2. The performance of the optimal predictors for the obedient/disobedient classes. It shows that the disobedient samples are more accurately recognized than the obedient ones.

	Number of Related Features			
Selected Features	*B1 (KNN)*	*B2 (NB)*	*G1 (SGD)*	*All (NB)*
MFCC	8	4	2	10
LSF	4	3	0	7
Loudness	3	2	0	2
Intensity	1	0	3	0
Pitch (F0)	1	0	0	0
Voice Probability	0	1	0	3
ZCR	0	0	0	1

Table 3. Results of feature selection about the subsets of the models with the highest accuracy respectively. It shows that different children have different significance for feature subsets.

However, when looking in depth at each of the optimal classifications, I discovered that the obedient instances always have a lower correct rate than the disobedient samples. As demonstrated in Table 2, this difference exists for the children both separately and together. On average, the disobedient examples have 3.8% higher accuracy. The reason might be that some of the obedient attitudes have a slight frustration at the same time, which is emotionally similar to the disobedient ones. Therefore, a more accurate approach specifically for the obedient emotions has the potential to be explored in the future.

Selected Feature

Regarding the other outputs from the perspective of feature subsets, the results are presented separately in Table 3 and Figure 6 and Figure 7.

Table 3 only demonstrates the details of the feature subset of the learning method that gained the highest accuracy. It shows that different children have distinct selected feature set. For example, the best feature subset for Boy 1 is to train the KNN model with many MFCCs related features; the best option for Boy 2 is to build the Naive Bayes learning model with relatively fewer MFCCs related features.

Figure 6 is a more informative presentation of the importance of each feature category for different model respectively. It shows that in general, MFCC and ISF are much more significant than others. However, the case varies according to the learning methods. For example, Bayesian Network, Naive Bayes, KNN and SMO values MFCC and ISF more.

Similar to Figure 6, Figure 7 adds up the total occurrences of feature categories from the 11 models. However, it focuses on the aspect of individual differences. So we can conclude that the more significant features vary for individuals. For instance, the MFCC and ISF for Boy 1 are much more important than the other two children; while intensity does not appear to be a good indicator for him. On the other hand, the general importance can also be seen, especially that zero-crossing rate is almost irrelevant in speaker-dependent recognition when compared with other feature categories.

COMPARE CHILDREN'S AND ADULTS' FEATURE SETS

As an extra experiment, I also have the interest to investigate the difference in the more critical feature subset in emotion

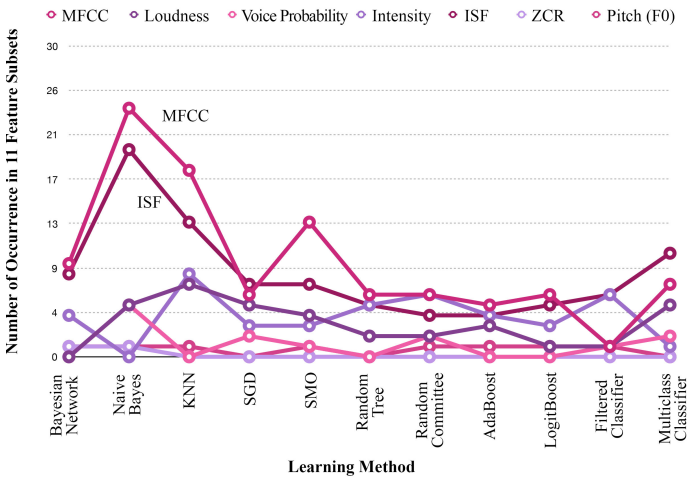

Figure 6. The total number of the occurrences regarding the feature categories for each kind of learning method. It demonstrates that MFCC and ISF are more important than other features categories, especially when using Bayesian Network, Naive Bayes KNN, or SMO as the learning method.

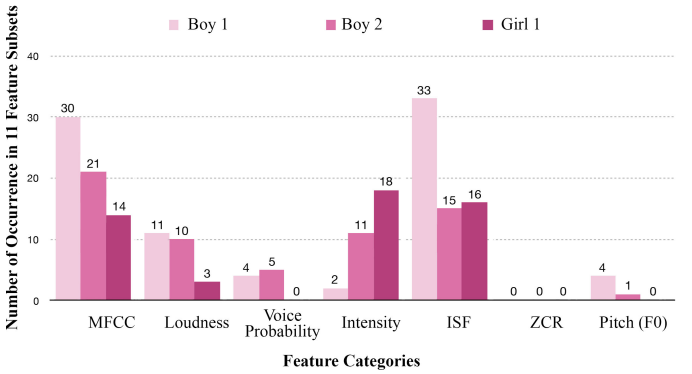

Figure 7. The total number of the occurrences regarding each feature categories that shows individual differences. It displays how different features matter differently for each child.

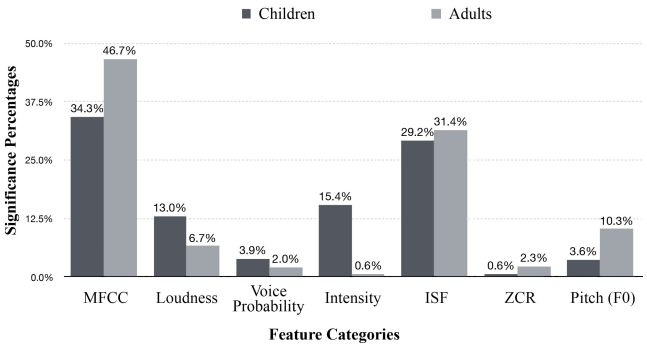

Figure 8. Significance Percentage for the Selected Features. It demonstrates most of the differences in MFCC, Intensity, and Pitch (F0).

recognition for children compared with adults. Therefore, I utilised the open audio resources of Berlin Emotional Speech Database [2] and went through the same process of how I did with the children's speech audio, the goal of which was to obtain the comparable feature subsets to explore the differences.

The database has initially 535 labelled data in all, with anger, boredom, disgust, anxiety, happiness, sadness and neutrality. To make it similar to my case where the classification targets are obedient and disobedient, I re-tagged happiness as the obedient emotion; and anger, boredom, disgust, and anxiety are re-tagged as disobedient. I did not consider the neutrality for the moment. However, there are significantly more disobedient samples than obedient, so I randomly picked and just left 100 obedient training samples, thus closer to the obedient amount of 71.

After leaving out irrelevant features and found the feature subsets for each of the same 11 learning methods, I calculated the Significance Percentage for each of the features selected out and added them up in each kind of acoustic feature categories. I defined the Significance as below, thus considering the correct rate at the same time:

Each feature category such as MFCC, Loudness and Intensity has a Significance coefficient that adds up from the 11 models

$$Significance = \sum_{i=1}^{11} \frac{number\ of\ occurrences}{subset\ size} \times accuracy$$

After attaining the Significance for each feature categories, I calculated the Significance Percentage where the denominator is the sum of the Significance for all categories. Thus I obtained the histograms as shown in Figure 4. There are interesting findings:

- MFCC related features are the most important for both sides; however, it is 12.4% more significant for adults.
- Pitch related features are 6.7% more important for adults.
- Intensity related features appear to be almost irrelevant for adults. However, it is 15.4% high significant for children.

CONCLUSIONS

This paper validates that we can technically recognise the obedient and disobedient emotions from children's acoustic features with high accuracy. It means that the intelligent voice interaction is potentially achievable. Besides, we can also conclude that the wrapper method of feature selection has the worth to be utilised and that personalised the model for each child (speaker-dependent) can reach even lower error rate.

However, the accuracy of Girl 1 and generally for the obedient emotions have the potential to be improved. I did not achieve as high cross-validation performance for Girl 1 partially because the girl was not fluent in reading the sentences. Thus the final valid training samples are only 88, which was relatively too few. Improvement could be achieved if more valid training samples can be collected. As for the recognition accuracy for the obedient emotion, one improving approach is to integrate the textual information together with the acoustic features. The Multi-Model Fusion approach [12] can merge the features from those two modalities.

Besides, through the additional experiment to compare the children and adults, we can imply that when conducting the voice recognition for children, it makes sense to focus more on voice intensity than when dealing with adults' speech signals.

FUTURE OPPORTUNITIES

We can see a considerable amount of future opportunities on top of this children-specific emotion recognition, although much future work is nevertheless required to empower an ideally educative online screen time restrictive voice interaction.

Emotional Guiding Responses

The successful recognition of obedient and disobedient emotions serves the ambition of educatively responding. In principle, the response should encourage the deferential attitude while educating the naughty children. To make the reaction approachable and more accessible for children, the AI model has to integrate pedagogy and psychology. The ideal case would be that the smart agent voice is capable of leading the situation, where the children are willing to stop watching the TV or surfing the Internet. For example, they can be attracted by some other healthier activities. One of the already technically doable ways to increase the accessibility is to polish the tone and characteristics of the agent's voice. For instance, it can be similar to the tone pattern of their parents; it also needs to be natural when switching between strict and gentle attitudes.

Concerning the difficulties, from the perspective of emotion recognition, it is generally technically feasible. However, there is still the challenge of interpreting the hidden emotions. Hidden emotions are those not that obvious to see even for real human beings. For example, the child may say something which is not his/her real feelings.

Most of the challenges are in building the responding strategy. Interpreting this strategy into sentences is relatively more achievable than that. This educative strategy is complicated because the realistic training samples are difficult to gain. A more efficient way is cooperating with educationalists who are

extraordinary in positive educating to design all the training samples, such as what sentences in responding a specific set of emotions. If a sufficient amount of high-quality training samples are obtained, this interaction model is technically doable. Furthermore, developers can write the algorithm of learning the individual personality, which is beneficial in implementing a more accurate education strategy.

Ethical Education

Since the children will sometimes have rude and improper expressions about their disobedient emotion, there we can integrate together with the ethical detection and education. For example, when a specific inappropriate word is detected, the educative strategy should tailor to moral education purpose. In that case, the smart agent might start a conversation related to manners rather than only the online screen time.

More Interaction Possibilities

Since this paper focuses on emotion recognition, more future opportunities other than online screen time scenario indeed exist. For example, the smart agent can even be a mental child guard. Concretely, it can detect the emotional changes and try to greet or talk to delight the child. If the child is still upset, or anything weird observed, the agent can further inform the parents. Notifying parents is a promisingly useful functionality since they do care about control power and supervision. In this regard, an extra feature might be tracking and analysing the child's emotion status, then gives the report and suggestions to parents.

REFERENCES

1. John Annett. 2003. Hierarchical task analysis. *Handbook of cognitive task design* 2 (2003), 17–35.
2. Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology*.
3. Frank Dellaert, Thomas Polzin, and Alex Waibel. 1996. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, Vol. 3. IEEE, 1970–1973.
4. Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.
5. Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.
7. Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.
8. Xuan Hung Le, Georges Quénot, and Eric Castelli. 2004. Recognizing emotions for the audio-visual document indexing. In *Computers and Communications, 2004. Proceedings. ISCC 2004. Ninth International Symposium on*, Vol. 2. IEEE, 580–584.
9. Sonia Livingstone and Ellen J Helsper. 2008. Parental mediation of children's internet use. *Journal of broadcasting & electronic media* 52, 4 (2008), 581–599.
10. Alan V McCree and Thomas P Barnwell. 1995. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and audio Processing* 3, 4 (1995), 242–250.
11. Reinhard Pekrun, Anne C Frenzel, Thomas Goetz, and Raymond P Perry. 2007. The control-value theory of achievement emotions: An integrative approach to emotions in education. In *Emotion in education*. Elsevier, 13–36.
12. Filip Povolny, Pavel MatËĞejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. 2016. Multimodal emotion recognition for AVEC 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 75–82.
13. Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. In *Encyclopedia of database systems*. Springer, 532–538.
14. Deb Roy and Alex Pentland. 1996. Automatic spoken affect classification and analysis. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. IEEE, 363–367.
15. Md Sahidullah, Sandipan Chakroborty, and Goutam Saha. 2010. On the use of perceptual Line Spectral pairs Frequencies and higher-order residual moments for Speaker Identification. *International Journal of Biometrics* 2, 4 (2010), 358–378.