SPEAKER-INDEPENDENT DETECTION OF CHILD-DIRECTED SPEECH

Sebastian Schuster[†], Stephanie Pancoast[‡*], Milind Ganjoo[†], Michael C .Frank[°], Dan Jurafsky[◇†]*

[†] Department of Computer Science, Stanford University, Stanford, CA
[‡] Department of Electrical Engineering, Stanford University, Stanford, CA
[°] Department of Psychology, Stanford University, Stanford, CA
[◇] Department of Linguistics, Stanford University, Stanford, CA
{sebschu,pancoast,mganjoo,mcfrank,jurafsky}@stanford.edu

ABSTRACT

Identifying the distinct register that adults use when speaking to children is an important task for child development research. We present a fully automatic, speaker-independent system that detects child-directed speech. The two-stage system uses diarization-style voice activation techniques to extract speech segments followed by a supervised ν-SVM classifier trained on 1582 prosodic and log Mel energy features. The system significantly improves the state of the art, detecting child-directed speech with F1 of .66 (exact boundary) and .83 (within 1 second). A feature analysis confirms the importance of F0 features (especially 3rd quartile and range) as well as new features like the variance, kurtosis, and min of log Mel energy within a frequency band.

Index Terms— Speech Analysis, Child-directed Speech, Language Development, Prosody

1. INTRODUCTION

Significant evidence has shown that the language environment in early childhood is strongly correlated to that child's language performance both later in childhood and later in life [1]. The speech surrounding a child varies greatly, not only in content but also in prosody. Adults speak using a different register to children than they do to one another. This different speech register is referred to as *child-directed speech* (CDS), also know as Motherese. Huttenlocher et al. [1] found that the amount of CDS heard by a child is more positively correlated with developmental outcomes than are hereditary factors, and Weisleder and Fernald [2] found infants from low socioeconomic status families who experienced more child-directed speech in their day showed a larger vocabulary by the age of 24 months.

Because of the belief of the beneficial outcomes of child-directed speech, language development researchers are often interested in the quantity of CDS in data collected from a child's natural learning environment. The segmentation and labeling of the speech in a recording to determine the location and overall quantity of CDS is, however, very time-consuming. The goal of our work is to explore the features and classifiers that can best discriminate between child-directed and non-child-directed speech, as well as develop a speaker-independent system that could be used to identify the locations and labels of the CDS and non-CDS speech in a continuous audio recording.

Recent research has sought to address the automatic classification of CDS [3, 4, 5, 6]. Child-directed speech is typically characterized by elongated vowels, highly varying pitch contours [7, 8], and generally clearer speech [9]. Previous work uses binary classifiers [4, 5] or Hidden Markov Models [6] on pre-segmented audio, which is therefore not immediately applicable to a raw audio recording. Further, Mahdhaoui et al. [4] trained and tested on the same two speakers. Vosoughi et al. [3] perform their automatic detection on naturalistic recording but they also rely on the fact that only three speakers are present in the data, and can therefore identify the speakers and use speaker characteristics in some of their features. Results from both of these publications are not immediately applicable to researchers who may want to analyze a child's language environment without any prior knowledge of the speaker. In the case of Robinson-Mosher and Scassellati [5], the features and classifiers were speaker-independent, but the child-directed and non-child-directed utterances were read off a transcript or a book and are consequently not completely natural.

The novelty of our work is three-fold. Firstly, we use a larger set of features than has been applied in previous work. In addition we explore a wider range of classifiers than has been tried previously. Finally, and most prominently, we develop a system, that to the best our knowledge, is the first speaker-independent child-directed speech detector. The system including pre-trained models is available at `http://github.com/sebschu/cds-detector`.

*These authors contributed equally.

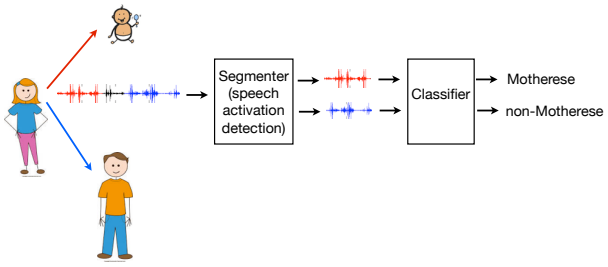

Fig. 1: Overview of system.

2. SYSTEM OVERVIEW

As illustrated in Figure 1, our speaker independent, end-to-end system takes an audio file (containing child-directed and non-child-directed speech), segments the speech, and then classifies each segment into child-directed versus non-child-directed. The two components are introduced in the following sections.

3. DATA

Our data consists of eleven audio files recorded in the Stanford Language and Cognition Lab [10]. All recordings were collected with a Shure Lavalier Wireless Microphone PG185 device at a sampling rate of 44.1 kHz. Each recording contains speech of a female caretaker, the female researcher (same for all recordings), and the infant. The age of the infants ranges from eight to sixteen months old.

After the recordings were collected, the annotations were performed in two passes. In the first pass, the annotator went through performing the segmentation and providing each segment with a label. In the second pass these segments were checked to ensure the labels were correct and consistent. In the end, the data consists of 74 minutes of usable speech segments, collected from approximately four and a half hours of recordings. Only segments that contain clean speech were labelled. Other sounds such as laughter, crying or other infant vocalizations were not labelled. Regions where speech and noise overlap significantly (e.g., the mother talking and the infant crying) were also not labelled.

4. CHILD-DIRECTED SPEECH SEGMENT CLASSIFICATION

4.1. Features

We experiment with two different feature sets. Our *baseline* feature set consists of the three acoustic features that were used by Vosoughi and Roy [3], namely fundamental frequency (F0) range, mean F0 and mean intensity. Our *extended* feature set consists of 1582 features that are statistical descriptors of F0, pitch, loudness, jitter, 8 log Mel frequency bands, Mel-frequency cepstral coefficients (MFCC), and 8 line spectral pair frequencies (LSP). We extract all features with openSMILE [11] using the configuration file from the INTERSPEECH 2010 paralinguistics challenge. We normalize all feature values in the training set such that each feature has zero mean and unit variance over all training examples. We apply the same transformations to the features of the test set.

	Training set	**Test set**
Positive examples	1402 (450)	216
Negative examples	450 (450)	124
Total examples	1852 (900)	340

Table 1: Training and test data set composition for the child-directed speech segment level classification experiments. The numbers in brackets correspond to the number of examples in the balanced training set.

4.2. Experiments

For our classification experiments, we split the data into a training and a held-out test set. We put all segments of one of the recordings in the test set to allow testing of how well the classifier performs on an unseen speaker. The segments of the remaining ten recordings are split randomly between the training and the test set to obtain approximately a 90:10 split between these two data sets. Table 1 shows the number of positive and negative examples in the two splits as well as the total for both the train and test set. Due to the experimental setup, our data contains more positive than negative examples. In order to balance the training data we randomly sample from all positives to obtain an equal number of examples from the two classes.

We train six different classifiers:

- a L_2-regularized SVM with a linear kernel (SVM-linear)
- a L_2-regularized SVM with a radial kernel (SVM-radial)
- a L_2-regularized ν-SVM with a radial kernel (ν-SVM)
- a L_1-regularized logistic regression classifier (Logistic Regression)
- a random forest classifier (Random Forest)
- a decision tree classifier (Decision Tree)
- a Gaussian Naive Bayes classifier (Naive Bayes)

For all classifiers we use the implementation provided by the scikit-learn[1] package. We optimize the hyperparameters for each of the classifiers using 10-fold cross validation.

[1] http://scikit-learn.org

Many classifiers achieve the best cross-validation results using the default parameters so we only modify the following parameters: We set the regularization strength C of the linear SVM to $C = 0.001$, of the logistic regression classifier to $C = 0.01$, set the number of trees for the random forest classifier to 50 and set the minimum number of samples per split for both tree classifiers to 3 and the maximum tree depth to 6. Also, we use entropy to measure the quality of a split for both tree classifiers.

4.3. Results and Discussion

We evaluate the performance of our classifiers using accuracy, precision, recall, and F1 at the segment level. Results for all classifiers using the *baseline* and the *extended* feature set are presented in Table 2.

The results indicate several trends. First of all, with the exception of the decision tree, we always get better performance in terms of accuracy and F1 using the *extended* feature set compared to using the three *baseline* features. Further, if we compare the performance of the different classifiers on the held-out test set, we can observe several differences depending on which feature set we use. Similar to Vosoughi and Roy [3], we also obtain the best results in terms of accuracy with a Naive Bayes classifier if we use the *baseline* features. However, if we use the *extended* feature set that contains many highly correlated features, Naive Bayes performs poorly. On the other hand, the SVM classifiers with radial kernels that are able to model feature interactions and make use of large feature sets give the best results on the test data. In terms of precision, the linear SVM trained with only the *baseline* features performs slightly better than any classifier using the *extended* features. However, in terms of recall, the classifiers using the *extended* feature set beat the classifiers using the *baseline* feature set by a large margin. The decision tree classifier is again an exception to this trend. For space reasons, we omit the results on the training data, which show a similar picture to the results on the test data. The only major difference is that the linear SVM classifier with the *baseline* features performs noticeably worse on the training set in terms of precision, which indicates that the good performance on the test set is more likely to be caused by properties of the test data rather than being a stable property of this classifier. The poor performance of the decision tree classifier can also be observed on the training data, which rules out the common problem of overfitting a decision tree.

In Figure 2 we illustrate how the performance varies (in terms of precision and recall) by speaker. The figure also shows the speaker that was excluded from the training data. We see that the unseen speaker's performance is almost on-par with the others. The accuracy for this speaker is 0.75 while precision and recall are 0.75 and 0.88 respectively. The plot indicates a high variance in performance between speakers. However, the number of test examples per speaker also

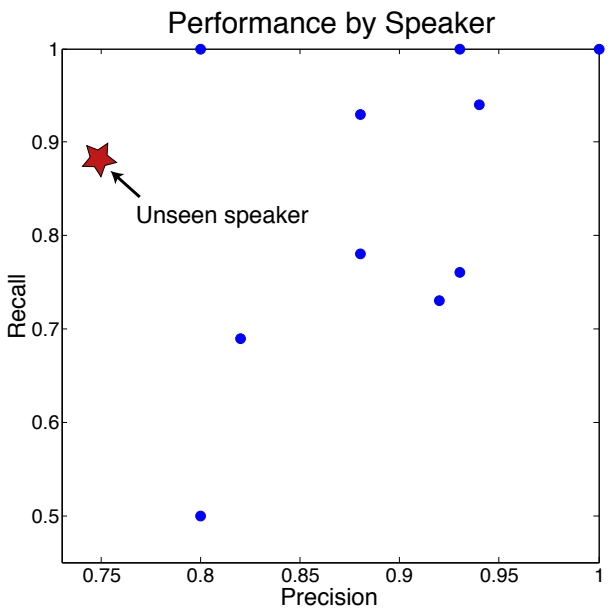

Fig. 2: Precision and recall for individual speakers, including one who was not included in the training set.

ranges from 2 to 32 for the speakers that are present in the training data and the speakers that deviate most from the average results (both positively and negatively) have fewer examples in the test set.

In order to determine the most discriminative features we perform an analysis of the feature weights of the logistic regression classifier. As it is very challenging to draw conclusions from MFCC features we remove them from the feature set for this analysis. This results in a drop of only 1% accuracy on the test set, so any conclusions from this analysis should also be valid for the larger feature set. The classifier assigns non-zero weights to 37 features. As observed in previous work, the most discriminative feature is a statistically stable descriptor of the fundamental frequency, namely the 3rd quartile of the F0 envelope.[2] This feature is highly positively correlated with child-directed speech. A high range in F0 is also positively correlated with CDS, confirming the results of previous work [3, 8, 7]. Most other features with high weights are descriptors of the log Mel energy within various frequency bands. We observe that both high variance and kurtosis (the "peakedness" of a curve) of the log Mel energy within a frequency band are positively correlated with CDS. This observation is consistent with previous findings that suggest that CDS has wider formant dispersion [9]. Further, very low energy minima in high-frequency bands negatively correlate with CDS.

[2]The envelope of the fundamental frequency is computed by replacing the F0 in unvoiced regions with the last non-zero value before the unvoiced region.

	Accuracy		Precision		Recall		F1 Score	
Classifier	Baseline	Extended	Baseline	Extended	Baseline	Extended	Baseline	Extended
SVM-linear	0.70	0.79	**0.88**	0.85	0.61	0.82	0.72	0.84
Log. regression	0.69	0.78	0.79	0.83	0.69	0.81	0.74	0.82
SVM-radial	0.71	0.80	0.81	0.83	0.71	0.86	0.76	0.84
ν-SVM	0.60	**0.82**	0.70	0.86	0.64	**0.86**	0.67	**0.86**
Random forest	0.71	0.78	0.78	0.81	0.77	0.84	0.77	0.83
Decision tree	0.70	0.64	0.81	0.74	0.70	0.68	0.75	0.71
Naive Bayes	0.71	0.72	0.81	0.78	0.72	0.79	0.76	0.78

Table 2: Results on the test set using different classifiers and feature sets (baseline and extended). All metrics are computed on the segment level. Best results for each metric are highlighted in **bold**.

5. SEGMENTATION

Up until this point we have only presented results on manually segmented audio files. In this section, we describe the two parts of our segmenter in detail and show how to automatically segment an audio file into short fragments of speech which can then be used as input for our classifier.

5.1. Voice Activation Detector

We perform four steps in the first pass to find the adult speech segments in a recording.

Voicing probability computation We use openSMILE [11] to compute the voicing probability for each frame of 10 ms. OpenSMILE uses the autocorrelation function which is computed by applying an inverse FFT to estimate voicing probabilities. We mark each frame as containing speech whose voice probability is above a threshold t_{voice}.

Smoothing The frame-by-frame voicing probability may vary drastically between adjacent frames. To smooth out these fluctuations in the probability of a speech signal, we applied a moving average filter of width w_{smooth} frames to obtain the smoothed voicing probability values for each frame.

Eliminating small gaps Very short segments of silence in between longer segments of speech are likely to be a result of recording errors and not actual silent portions. For this reason, we join all segments that are separated by a gap of less than g_{min} frames.

Eliminating small islands Likewise, very short segments of speech in between longer segments of silence are also likely to be a result of recording errors or signal noise. Therefore, for the final step we also eliminate all segments whose length is less than l_{min}.

	Training set	**Test set**
Positive examples	1155	38
Negative examples	348	38
Total examples	1503	76

Table 3: Training and test data set composition for noise elimination experiments. The positive class contains adult speech, either child-directed or non-child-directed, and the negative class contains everything else.

5.2. Noise Elimination

We noticed during preliminary experiments that some of the output from the voice activation detector included non-adult speech segments such as infant vocalizations or adult laugher. We refer to all these non-adult speech segments as "noise". We annotated such false positives in three of the videos, and built an extra classifier to post-process the first pass voice activation output. A positive label was provided to segments that contained either child-directed or non-child-directed adult speech, and a negative label was assigned to all other segments. Because this is a "filtering" step with an emphasis on minimizing false rejects, we used an imbalanced training set to bias the classifier to keep speech segments. The data set composition for the noise elimination experiments is presented in Table 3.

5.3. Experiments

For the segmentation experiments we work with the full, unsegmented recordings. Again, we split our data into a training and test set. The training set consists of the ten recordings that were part of the training set for the classifier experiments. The unseen speaker is used for testing.

We use grid search to optimize the four parameters of our first pass segmenter: $(t_{voice}, w_{smooth}, g_{min}, l_{min})$. The objective function that we are trying to maximize is the average F1-metric over all recordings in the training set. In

	Precision	Recall	F1	Accuracy
Oracle Training	0.83	0.73	0.78	0.88
Training	0.56	0.84	0.66	0.75
Oracle Test	0.67	0.73	0.70	0.89
Test	0.48	0.88	0.62	0.73
All CDS	0.59	0.77	0.66	0.83
Soft boundaries	0.83	0.83	0.83	0.84

Table 4: Results for the full pipeline. For the training data set we present the average of the individual results for each recording. In the gold standard for the *All CDS* metrics, segments in which child-directed speech and noise overlap and singing are also labelled as child-directed speech. The *Soft boundaries* metrics are computed using the same gold standard as for the *All CDS* metrics but we also mark 1-second chunks that are directly adjacent to CDS utterances as being correct if the classifier labelled them as containing CDS.

this experiment we make a prediction for every second of the recording. As gold annotations we use the timespans of the segments that were manually extracted for the previous experiment which correspond to all parts of the recording that contain clean speech. Our final parameters are a voicing probability threshold $t_{voice} = 0.3$, a smoothing window $w_{smooth} = 150$ frames, a minimum gap size $g_{min} = 10$ frames, and a minimum duration $l_{min} = 50$ frames.

We separately optimize the noise classifier at the segment level. We use the same feature set for the noise classifier as for the child-directed speech classifier to avoid extracting features twice. We find that, with this feature set, the radial basis function SVM performs the best, yielding a test accuracy of 0.92, recall of 0.92, and precision of 0.98. These results are significant enough to include the noise elimination step in our pipeline.

To evaluate the full pipeline, we compute the precision, recall, F1 and accuracy metrics. We label each 1-second chunk of the recording as containing CDS or not and then compare this to our gold annotations to compute the evaluation metrics. On the training set we take the average of the individual results for each recording. We compare our results to an oracle experiment in which we only test the two classifiers by using our gold annotations to segment the file.

5.4. Results and Discussion

The results of the full pipeline are shown in the upper part of Table 4. If we compare the results of the oracle experiments with the results from our fully automatic pipeline we can see that in both the training and the test recordings accuracy, F1, and, more significantly, precision go down. However, we actually see small improvements in terms of recall compared to the oracle experiments.

At a first glance, the precision of our pipeline seems to be very low. The numbers indicate that for every 1-second chunk that we correctly classify as containing child-directed speech, we also incorrectly classify another chunk as containing CDS. However, an extensive qualitative analysis shows that the performance is actually much better than the numbers indicate.

The main reason why precision is so low is that the manually annotated segments do not contain any pauses at all. So if the mother makes an utterance that includes a short pause, then this utterance is split into two segments in our gold annotations. The segmenter, on the other hand, typically extracts the entire utterance in such cases which hurts precision as the pause is also incorrectly marked as CDS. Modifying the segmentation parameters such that we generally obtain shorter segments with fewer pauses hurt recall significantly more than it helped precision, so we abandoned this approach.

Another reason for the low precision is the fact that the annotated segments only include clean speech segments. Segments that contain speech that overlaps with noises such as an infant crying while the mother is talking are not labelled. We observe several of these cases in our recordings and most of them are classified as containing CDS, a desirable outcome if one wants to measure the amount of child-directed speech that a child is exposed to.

Lastly, we observed in the test recording that the mother sings to the child several times. These parts of the recordings were not labelled as they technically do not contain either CDS or non-CDS speech. However, our classifier understandably marked all these regions as containing child-directed speech.

To show how these observations affect our results, we reannotate the test data. First, we also label all segments that contain either singing or unclean child-directed speech that overlaps with any form of noise as containing CDS (*All CDS*). Further, to show how short pauses in utterances influence the performance, we also compute all metrics with softer boundaries (*Soft boundaries*). The *Soft boundaries* are an extension of the *All CDS* with all 1-second chunks being considered correct if they are directly adjacent to a CDS segment in our gold standard, even if the classifier incorrectly labels them as containing CDS. That way segments that are slightly too long or segments that contain an utterance that is split into two parts in our gold annotations are still classified correctly. Results for these new metrics are presented in the lower part of Table 4. These results show that if we slightly modify the evaluation metrics such that they do not penalize any of the discussed observations, then we get significantly better results compared to evaluation on only the clean CDS segments.

Nevertheless, several errors still remain. The biggest challenge seems to be to separate noise and speech in case they partially overlap. If the mother is speaking to the child and, while she is speaking, the infant starts crying, both the utterance and the crying are typically extracted in one segment. We experimented with an automatic diarization system but

this system also failed at introducing the correct boundaries. Other common errors include the failure to detect very short segments of child-directed speech in some cases and the classification of laughter as CDS.

Based on this error analysis, we conclude that although there is still room for improvement, the detection pipeline is more useful in practical tasks than precision of clean CDS segments would indicate.

6. CONCLUSION AND FUTURE WORK

We presented an automatic speaker-independent child-directed speech detection pipeline that allows the automatic annotation of audio files. The addition of additional speech features gives significant improvements over a baseline system in detecting child-directed speech segments. Our pipeline also gives promising results and we expect this tool to prove highly useful for large-scale child development experiments that so far required a significant amount of manual work.

There are two areas of potential future work. Because of the lack of recordings containing male speakers our system was trained and tested only on female speakers. An obvious extension to our work would be to perform similar recordings with male speakers and to build additional models for male speakers. Further, as our qualitative analysis shows, more work can be done on improving the filtering of noise that overlaps with speech to further improve the pipeline.

7. ACKNOWLEDGEMENTS

We thank Andrew Maas for his valuable advice in numerous stages of this project. This material is based upon work supported by the National Science Foundation under Grant No. DGE-1147470.

8. REFERENCES

[1] J. Huttenlocher, W. Haight, A. Bryk, M. Seltzer, and T. Lyons, "Early vocabulary growth: Relation to language input and gender.," *Developmental Psychology*, vol. 27, no. 2, pp. 236, 1991.

[2] A. Weisleder and A. Fernald, "Talking to children matters early language experience strengthens processing and builds vocabulary," *Psychological Science*, vol. 24, no. 11, pp. 2143–2152, 2013.

[3] S. Vosoughi and D. Roy, "An automatic child-directed speech detector for the study of child language development," in *Interspeech*, 2012.

[4] A. Mahdhaoui, M Chetouani, C. Zong, R.S. Cassel, C. Saint-Georges, M. Laznik, S. Maestro, F. Apicella, F. Muratori, and D. Cohen, "Automatic motherese detection for face-to-face interaction analysis," in *Multimodal Signals: Cognitive and Algorithmic Issues*, vol. 5398, pp. 248–255. Springer Berlin Heidelberg, 2009.

[5] A. L. Robinson-Mosher and B. Scassellati, "Prosody recognition in male infant-directed speech.," in *IEEE International Conference on Intelligent Robots and Systems*, 2004.

[6] T. Inoue, R. Nakagawa, M. Kondou, and K. Shinohara, "Discrimination between mothers infant- and adult-directed speech using hidden Markov models.," *Neuroscience Research*, vol. 70, pp. 6270, 2011.

[7] D. L. Grieser and P. K. Kuhl, "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese.," *Developmental Psychology*, vol. 24, no. 1, pp. 14, 1988.

[8] S. Vosoughi and D. Roy, "A longitudinal study of prosodic exaggeration in child-directed speech," in *Speech Prosody*, 2012.

[9] N. B. Ratner, "Patterns of vowel modification in mother–child speech," *Journal of Child Language*, vol. 11, no. 03, pp. 557–578, 1984.

[10] M. .C. Frank, K. Simmons, D. Yurovsky, and G. Pusiol, "Developmental and postural changes in childrens visual access to faces," in *Proceedings of the 35th annual meeting of the Cognitive Science Society*, 2013, pp. 454–459.

[11] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838.